

GARCH PRESS

GARCH

FINANCE, INSTITUTIONS AND
ECONOMIC DYNAMICS

VOLUME 1 / NUMBER 1 / 2026

INAUGURAL ISSUE

30 October 2026

Compute derivatives and technological change

Financial intermediation and institutional reform

Climate exposure, disclosure and volatility methods

A semiannual research publication

GARCH Quant LLC / GARCH Press

Volume 1, Number 1 | 2026

GARCH

Finance, institutions and economic dynamics

PUBLICATION INFORMATION

Title	GARCH
Issue	Volume 1, Number 1 (2026)
Frequency	Semiannual
Medium and language	Online publication; English
Publisher	GARCH Quant LLC
Publishing imprint	GARCH Press
Editor / contact	Fuli Yang dr@garch.press
Submissions	submissions@garch.press
Journal home	https://garch.press/journal/
ISSN	Not yet assigned
Publication date	30 October 2026

EDITORIAL AND PUBLICATION INFORMATION

This inaugural issue contains seven research articles selected by founder-editor Fuli Yang. The issue is editor-selected and self-published; it has not undergone independent external peer review.

Copyright remains with the respective authors. Articles are freely available to read and download. Reuse is subject to the rights stated in each article and the applicable original-source terms.

The original manuscript dates and working-paper version statements are retained within the articles. This journal edition has its own volume, issue and continuous pagination. Corrections and later versions will be dated and linked to the original record.

Contents

Volume 1, Number 1 | Inaugural issue

001 The Half-Life of a Futures Contract: Technological Obsolescence and Vintage Risk in Compute Derivatives

Fuli Yang

Pages 1-16

017 When Scale Outruns Fees: The Scale-Fee-Density Paradox in China's Trust Industry

Fuli Yang

Pages 17-29

030 From Disclosure to Resilience: Climate-Risk Transmission Frictions in China's Corporate Bond Market

Fuli Yang, Shiyu Lin

Pages 30-46

047 Trust as Resolution Infrastructure: Value Preservation, Creditor Coordination, and Bankruptcy Service Trusts in China

Fuli Yang

Pages 47-65

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

Fuli Yang

Pages 66-87

088 Unknown Is Not Zero: Heat Stress, Data-Center Power-Demand Exposure, and Utility Volatility

Fuli Yang

Pages 88-103

104 Factor Covariance in Yield-Curve Risk Forecasts: Evidence from U.S. Treasury Yields

Fuli Yang, Shiyu Lin

Pages 104-112

The Half-Life of a Futures Contract: Technological Obsolescence and Vintage Risk in Compute Derivatives

Fuli Yang

GARCH Institute, Harvard Square, Cambridge, MA 02138, USA

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

Working Paper v3.1 | 27 August 2026

Abstract

A futures contract has a legal maturity, but it can lose economic relevance before that date. This paper develops the concept of contract half-life for derivatives written on rapidly evolving technological vintages. Contract half-life is the time required for a hardware-specific contract's economic relevance - measured through commercial exposure, trading liquidity, and price-discovery importance - to fall to one-half of its reference level. The distinction matters for compute derivatives because the announced H100 and B200 futures are tied to named GPU generations whose commercial importance can migrate faster than their rental prices disappear. A simple vintage-migration model separates rental-price depreciation from benchmark-relevance depreciation and shows how innovation intensity, maturity, and liquidity network effects jointly determine economic expiry. The model yields four testable propositions: activity should migrate to a new vintage faster than the old rental index declines; distant maturities of an old-vintage contract should lose viability first; faster innovation should shorten contract half-life; and rolling or performance-adjusted benchmarks should extend durability while increasing conversion-model and governance risk. The paper proposes observable half-life measures based on open interest, volume, information shares, and commercial benchmark adoption, and explains why a still-published settlement index is insufficient evidence that a contract remains economically alive. Because these propositions require post-launch data, the paper also uses three completed benchmark transitions - the shift of a gasoline contract to an RBOB basis, the ultra-low-sulfur respecification of heating oil, and the inclusion of WTI Midland in Dated Brent - plus unsuccessful DRAM-futures proposals as identification templates. These analogues distinguish benchmark replacement, internal respecification, constituent revision, and launch failure; they are not estimates or formal bounds on compute-contract half-life.

Keywords: compute futures; contract half-life; technological obsolescence; vintage risk; benchmark design; liquidity migration

JEL Classification: G13; G14; L86; O33

1. Introduction

The calendar printed in a futures contract is not always the clock that matters economically. A contract can remain legally valid, retain a calculable settlement index, and continue to appear on an exchange's product list after the commercial risk it was designed to transfer has moved elsewhere. This possibility is familiar in a loose sense: exchanges regularly delist unsuccessful products, delivery grades are revised, and trading migrates between contracts. It becomes a first-order design problem when the underlier is a named technological vintage whose relevance can change during the maturity structure of the derivative itself.

Compute futures make that problem unusually visible. CME Group has announced cash-settled H100 and B200 rental-index futures for planned listing on 5 October 2026, subject to regulatory review. Both contracts are defined around specific NVIDIA GPU generations and settle to daily rental benchmarks constructed by Silicon Data (CME Group, 2026a, 2026b). Separately, Intercontinental Exchange (ICE) and Ornn announced on 19 May 2026 a suite of U.S.-dollar-denominated, cash-settled GPU compute futures based on the transaction-based Ornn Compute Price Index, subject to regulatory approval (ICE, 2026). The Commodity Futures Trading Commission (CFTC) has meanwhile requested public comment on the listing of compute derivatives, including the standardization of compute, idiosyncratic contract risks, effects on the cash market, and the possible role of perpetual products (CFTC, 2026). These product plans are public institutional events, but their trading outcomes remain prospective. Yet the economic unit being standardized can age much faster than grain, crude oil, or a broad equity index.

A new GPU generation does not make an older one physically vanish. H100 capacity can remain productive, rentable, and observable after B200 or later systems attract frontier workloads. Older generations may continue serving inference, fine-tuning, and cost-sensitive applications. What changes is the share of commercial risk for which the older vintage is the natural reference. That share can migrate before the old rental price reaches a low level and before an index provider loses the ability to publish a daily observation. A contract may therefore have a price but cease to be the market's focal hedge.

This paper names and formalizes that separation. Contract half-life is defined as the time required for the economic relevance of a hardware-specific derivative to fall to one-half of its reference level. Economic relevance is not the price of the underlying chip or rental service. It is a composite of three roles: the amount of commercial exposure naturally mapped to the contract, the share of trading liquidity concentrated in it, and its contribution to price discovery for the broader market. Legal expiry closes the contract. Economic expiry occurs when these roles have migrated sufficiently that the contract no longer coordinates the main risk-bearing activity it was intended to serve.

The central argument extends the benchmark-before-market framework in Yang (2026), SSRN Working Paper No. 7324139. That earlier paper examines how a settlement benchmark can help constitute a compute market before the underlying service is fully fungible and identifies a trilemma among tradability, exposure representativeness, and technological durability. The present paper takes one mechanism from that framework - technological durability - and develops it as an independent theory of contract life. It does not repeat the broader review, does not claim that announced contracts will attract realized liquidity, and does not use synthetic forward curves or simulated hedges as evidence of actual market success.

The contribution is fivefold. First, the paper distinguishes rental-price depreciation from benchmark-relevance depreciation. Second, it gives contract half-life a measurable definition rather than using obsolescence as a qualitative warning. Third, it introduces a simple two-vintage model in which innovation intensity, maturity, and endogenous liquidity migration determine economic life. Fourth, it uses three completed benchmark transitions and a never-alive control as identification templates that separate relevance decay, administrative redefinition, and launch failure without treating historical episodes as estimates of compute-contract half-life. Fifth, it compares fixed-vintage, rolling-vintage, performance-adjusted, and token-expenditure designs, showing that durability can be purchased only by accepting transition rules, workload assumptions, and benchmark-governance discretion.

This is a prospective theoretical and institutional analysis. The announced compute futures had not begun trading at the paper's cutoff date. The propositions are therefore framed for later falsification using actual volume, open interest, quote depth, information shares, and commercial adoption. A failure to attract trading is not repaired by simulated evidence; it is itself evidence about whether a hardware-specific contract can survive long enough to become a market institution.

2. Institutional Setting and Relation to Existing Work

2.1 From a rental index to a vintage-specific futures contract

The announced CME products convert a heterogeneous service into a monthly cash-settled claim. Silicon Data publishes separate rental indices for several GPU models, including A100, H100, H200, B200, and MI300X. Raw observations are converted to a per-hour, per-GPU rate, filtered, adjusted for observable contract characteristics, aggregated at the provider level, and then combined across providers (Silicon Data, 2025, 2026). The resulting number is more comparable than a raw cloud rate card, but it remains explicitly attached to a hardware model.

That attachment offers an early-market advantage. Hardware names are salient, observable, and already used in procurement. A contract labeled H100 is easier to explain than a latent-performance unit whose conversion depends on workload, precision, interconnect, software stack, and utilization. Salience can concentrate liquidity. At the same time, the label embeds a vintage clock. NVIDIA reports that DGX B200 can deliver substantially higher training and inference performance than DGX H100 for specified workloads, while later Blackwell Ultra and Rubin systems shift the frontier again (NVIDIA, 2024, 2025, 2026). Exact performance ratios are workload-dependent, but the direction of embodied technical change is not.

The settlement index can therefore remain statistically healthy while the contract's economic constituency changes. An old vintage may have many price observations because it is widely deployed, even as new investment, frontier workload demand, and marginal hedging needs move toward a successor. Index observability and contract relevance are separate state variables.

2.2 Three literatures that do not yet meet

The first relevant literature studies futures contract innovation, viability, and survival. Early work emphasizes commercial hedging demand, homogeneous exposure, sufficient speculative participation, and contract terms that avoid manipulation or delivery problems (Silber, 1981; Carlton, 1984; Black, 1986). Later evidence shows that most new derivatives begin with low activity and that network effects can make liquidity persistent once a focal contract emerges (Cavanaugh & Penick, 2018). This literature ordinarily treats the underlying commercial need as evolving more slowly than the product-launch process. Compute reverses that presumption: the product may be trying to establish a network while its named underlier is already being displaced. A related strand studies contracts that fail outright rather than age: Johnston and McConnell (1989) analyze the rise and fall of an established financial futures contract once the match between its terms and hedgers' exposure deteriorated. The boundary of the present contribution should therefore be stated positively. The existing success-and-failure literature studies contract failure - design flaws, insufficient hedging demand, or manipulability that prevent a market from forming or cause it to wither (Johnston & McConnell, 1989; Cavanaugh & Penick, 2018). This paper studies something else: the underlier being redefined by technology, so that the contract design succeeds and the settlement index remains healthy and computable while the economic exposure being hedged migrates to a successor vintage. In the first case the cause of failure is internal to the contract; in the second it is external to it. A further adjacent literature studies competition between trading venues and the migration of liquidity across exchanges rather than across vintages (Cantillon & Yin, 2011), a distinction that matters for identification and is developed in Section 8.2.

The second literature studies vintage capital and investment-specific technical change. New capital embodies superior technology, so an installed asset can lose economic value even when it remains physically productive (Solow, 1960; Greenwood, Hercowitz, & Krusell, 1997). Replacement models distinguish physical wear from obsolescence caused by the arrival of a better vintage (Cooley, Greenwood, & Yorukoglu, 1997). The diffusion of general-purpose technologies likewise emphasizes coexistence and displacement across

vintages (Jovanovic & Rousseau, 2005). Empirical depreciation research similarly separates deterioration from revaluation due to changing relative prices and technology (Hulten & Wykoff, 1981; Fraumeni, 1997). These models concern machines and investment. They do not ask what happens to a derivative whose settlement definition is itself a machine vintage.

The third literature concerns compute pricing and financialization. Bandi and Su (2026) analyze GPU rental indices, non-storability, term rentals, and synthetic forward curves. Fernández-Fuertes and Grégoire (2026) emphasize asymmetric basis risk across flexible buyers and hardware-locked sellers. Xing (2026) proposes token-based contracts, while Yang (2026) focuses on benchmark constitution and the standardization trilemma. These studies recognize technological change, but none makes the economic lifetime of a vintage-specific futures contract the central object.

Contract half-life joins these literatures. It treats market liquidity as an institution built through coordination, commercial exposure as a vintage-dependent state variable, and technology as a force that can move the object of hedging before legal maturity. The concept is not a synonym for delisting probability. Delisting is an exchange decision and often arrives late. Half-life is a threshold in the decline of economic relevance and can be measured while the contract remains listed.

3. What Exactly Has a Half-Life?

3.1 Price life, physical life, and contract life

A GPU has several clocks. Its physical service life ends when failure, maintenance cost, or technical incompatibility makes operation uneconomic. Its rental-price life describes the decline of the hourly rate earned by the hardware. Its frontier life ends when it is no longer the preferred technology for leading workloads. A derivative adds a fourth clock: the time during which the vintage remains a focal unit for transferring price risk.

These clocks need not coincide. Strong demand for cheaper inference can support the rental rate of an older GPU long after frontier training migrates. Conversely, a new architecture announcement can redirect long-horizon hedging and investment immediately even if the old spot rental index barely moves. A physical asset can therefore remain useful while the derivative written on it loses price-discovery leadership.

Let v denote a hardware vintage. Throughout the paper the event-time origin t_0 is the successor vintage's introduction date, and $u = t - t_0 \geq 0$ is event time, so that $u = 0$ at the successor's introduction. If a contract's normalization date differs from t_0 , it is written $u_{ref} = t_{ref} - t_0$. Define $C_v(u)$ as the normalized share of economically relevant commercial exposure naturally mapped to vintage v ; $L_v(u)$ as its normalized liquidity importance; and $I_v(u)$ as its normalized informational importance. Each component equals one at u_{ref} . A parsimonious relevance index is the weighted geometric mean

$$R_v(u) = C_v(u)^\alpha L_v(u)^\beta I_v(u)^\gamma, \quad \alpha + \beta + \gamma = 1 \quad (1)$$

where the weights reflect the research question. Equal weights provide a transparent baseline; hedger-centered work may place more weight on C , while exchange-centered work may emphasize L . In logarithms, the geometric index aggregates component decay rates as a weighted arithmetic average. A component approaching zero ultimately drives R_v toward zero, but the first-passage half-life need not equal the half-life of the fastest-decaying component. This distinction matters for measurement because C - the commercial component - is precisely the one for which the available proxies are weakest. Single-component half-lives and a two-component (L , I) version should therefore be reported as main results, as set out in Section 8.1. An arithmetic aggregation can be reported as an additional robustness measure.

The contract half-life is then

$$\tau_v(1/2; u_{\text{ref}}) = \inf \{ \tau \geq 0 : R_v(u_{\text{ref}} + \tau) \leq 0.5 R_v(u_{\text{ref}}) \} \quad (2)$$

This definition is an observable threshold, not a claim that relevance follows radioactive decay or that one constant governs the full path. The baseline sets $u_{\text{ref}} = 0$ when normalization occurs at the successor's introduction; equation (2) also permits a later joint-active normalization window. Relevance can recover after a supply shock, software improvement, export restriction, or renewed demand for a cost-effective vintage. Researchers should therefore report first-passage half-life, sustained half-life (remaining below one-half for a prespecified period), and any recrossing.

The threshold also differs from a rental-price half-life, τ_p , defined by the time for the vintage's rental index to fall by one-half. If trading and commercial exposure migrate while the old rental rate remains supported by residual users, τ_v can be far shorter than τ_p . This is the distinctive empirical claim.

3.2 Legal expiry versus economic expiry

Legal maturity is contractual and deterministic. Economic maturity is endogenous. A December H100 future can settle exactly as specified even if, by December, most forward-looking procurement risk is quoted in B200 or a newer generation. The old contract still performs settlement; it no longer performs coordination for the main market.

Economic expiry is not identical to zero volume. Thin trading can remain informative for a specialized constituency, and block activity may not appear as continuous order-book depth. Nor is high volume sufficient for economic life: temporary market-making incentives or speculative activity can support turnover without broad commercial adoption. The relevance index therefore separates trading, information, and commercial exposure instead of using one metric as a universal success test.

A further distinction is between a delivery-month contract and a product family. Individual futures always approach expiry. Contract half-life concerns the continuing economic relevance of a vintage-specific family across listed maturities, measured on constant-maturity or aggregated terms. It asks whether the H100 complex remains a focal market after controlling for routine rolling between calendar months.

4. A Simple Vintage-Migration Model

4.1 Commercial relevance after an innovation shock

Consider an incumbent vintage o and a successor n introduced at calendar date t_0 , so that event time is $u = t - t_0$ with $u = 0$ at the successor's introduction, exactly as in Section 3.1. Let λ denote effective innovation intensity: a reduced-form combination of performance improvement, expected availability, software readiness, financing capacity, and the probability that users can substitute toward the new vintage. Two features of λ should be fixed at the outset. It is a reduced-form scale parameter with units of inverse time that summarizes the realized intensity of substitution; it is not itself an arrival rate. The Poisson hazard λH introduced in Section 4.4 is the arrival rate of frontier-displacing events, and in a steady state $\lambda \approx \lambda H$ multiplied by the average substitution magnitude per event. Empirically the two are identified from different data - λ from realized adoption and exposure shares, λH from dated event counts - so they can be estimated and tested separately. The incumbent's natural commercial exposure share follows

$$C_o(u) = \exp(-\delta_c \lambda u) \quad (3)$$

where $\delta_c > 0$ captures workload substitutability. Hardware-locked sellers, regulated workloads, or installed software stacks reduce δ_c ; flexible buyers and frontier applications increase it. The old rental price need not fall at the same rate. Let

$$P_o(u) / P_o(0) = \exp(-\delta_p \lambda u) + q(u), \quad q(0) = 0 \quad (4)$$

where $q(u)$ captures residual scarcity, cost-sensitive demand, supply retirement, and common compute-market shocks. Even with a low δ_p , commercial exposure can migrate rapidly if the successor becomes the marginal unit for investment and frontier risk. Thus price persistence does not imply benchmark persistence.

The distinction is particularly important for service prices. Unlike the resale value of a chip, a rental rate reflects both embodied capability and the equilibrium of heterogeneous workloads. An older GPU can become more intensively used as it becomes cheaper. That response can slow price depreciation while accelerating the migration of high-value risk to the successor.

4.2 Liquidity migration and network effects

Liquidity is endogenous because traders prefer the contract other traders already use. Let $L_o(u)$ denote the incumbent's normalized cross-vintage liquidity importance after removing mechanical calendar-roll effects. A reduced-form transition written in cumulative-hazard form is

$$L_o(u) = \exp\{-\int_0^u [\delta_L \lambda + v A_n(s)] ds\} \quad (5)$$

where δL maps commercial migration into order flow, $A_n(s)$ in $[0,1]$ measures the successor contract's availability and adoption, and $v \geq 0$ captures the network amplification created by depth, tighter spreads, clearing familiarity, and benchmark reference in bilateral contracts. Before the successor is listed or operationally usable, A_n can be near zero and incumbent liquidity can persist despite declining commercial fit. Once the successor crosses a coordination threshold, migration can be discontinuous. The cumulative-hazard form ensures that $\delta L \lambda + v A_n(u)$ is the instantaneous decay rate even while adoption changes over time.

This structure allows both delay and acceleration. A strong incumbent network can initially protect the old contract. The same network logic then works against it after liquidity coordinates on the new vintage. The observed half-life is therefore not a mechanical chip-release interval; it depends on exchange timing, market-maker support, contract fungibility across clearing accounts, and whether the successor benchmark is accepted by commercial users.

The adoption path $A_n(u)$ is not exogenous; it is shaped by the exchange's choice of listing date a for the successor contract. Listing early (small a) facilitates spread trading and position transfer and lowers migration frictions, but it divides the order book while commercial exposure to the successor is still thin. Listing late (large a) protects the incumbent's network but makes the eventual migration more likely to display a liquidity gap. Let $Q_o(u; a)$ and $Q_n(u; a)$ denote usable liquidity levels rather than shares, $C_n(u)$ the commercial relevance of the successor, $\phi > 0$ the cost of simultaneous order-book fragmentation, and $g > 0$ the cost of leaving successor exposure without a usable contract. A minimal design problem is

$$\max_a W(a) = \int_0^T [Q_o(u; a) + Q_n(u; a) - \phi Q_o(u; a) Q_n(u; a) - g C_n(u) (1 - A_n(u; a))] e^{-ru} du \quad (5a)$$

The objective no longer collapses to the sum of two shares. An earlier listing reduces the uncovered-successor term but increases the interval over which both order books coexist and incur fragmentation costs. A later listing does the reverse. Under regularity conditions, the optimal listing date a^* moves earlier when innovation raises C_n more quickly or when the cost g of a liquidity gap rises, and later when the fragmentation cost ϕ rises. The formulation is deliberately reduced-form, but it makes the trade-off in Proposition 4 economically identified rather than merely descriptive.

4.3 Information migration

Informational importance $I_o(u)$ measures the old contract's contribution to the common efficient price of compute. It can be estimated from lead-lag relations, forecast-encompassing tests, Hasbrouck information shares, or Gonzalo-Granger component shares once synchronized observations are sufficient (Garbade & Silber, 1983; Gonzalo & Granger, 1995; Hasbrouck, 1995). In the model, let

$$I_o(u) = \exp\{-\int_0^u [\delta_l \lambda + \psi L_n(s)] ds\} \quad (6)$$

where ψ captures the feedback from successor liquidity to information production. Price discovery may migrate faster than volume when informed traders use the new contract while residual hedgers remain in the old. Alternatively, incumbent liquidity can make the old contract informative about common compute scarcity even after commercial demand begins to migrate. The path is empirical, but the model identifies what should be measured.

Combining equations (3), (5), and (6), take logarithms of equation (1):

$$\ln R_v(u) = \alpha \ln C_v(u) + \beta \ln L_v(u) + \gamma \ln I_v(u) \quad (7a)$$

Differentiating with respect to event time gives the instantaneous decay rate of relevance as the weighted arithmetic average of the three component decay rates:

$$\kappa_v(u) = \alpha \delta_c \lambda + \beta [\delta_l \lambda + v A_n(u)] + \gamma [\delta_l \lambda + \psi L_n(u)] \quad (7)$$

Because the component decay rates need not be constant while successor adoption and liquidity are still rising, the half-life is defined by an integrated-decay condition:

$$\int_0^{\tau_v(1/2)} \kappa_v(s) ds = \ln 2 \quad (8)$$

where $\kappa_v(s)$ is the instantaneous decay rate in equation (7). When $A_n(u)$ and $L_n(u)$ have converged to constants - or when their averages over the half-life interval are used - the familiar closed-form special case obtains:

$$\tau_v(1/2) \approx \ln(2) / \bar{\kappa}_v \quad (8')$$

Equation (8') is an approximation rather than an identity. Its error depends on the shape and speed of the successor-adoption and liquidity paths; it can be evaluated only after those paths or a stated numerical parameterization are supplied. The exact integrated condition in equation (8) should be used whenever adoption is still changing materially at the horizon of interest.

The expression makes the comparative statics transparent. Faster embodied innovation, easier workload substitution, a viable successor contract, and stronger liquidity-information feedback all shorten half-life. Contract half-life is shorter than rental-price half-life whenever $\kappa_v > \delta_p \lambda$ after accounting for residual-demand shocks.

4.4 Maturity and prospective obsolescence

A hardware-specific future with remaining maturity m transfers exposure from event time u to $u + m$. If frontier-displacing innovations arrive with Poisson hazard λ_H , distinct from the effective innovation intensity λ in Section 4.1, the probability that the named vintage remains the frontier reference through maturity is $\exp(-\lambda_H m)$. Prospective exposure and installed exposure are affected through different channels. Let ω_F and ω_S be nonnegative weights summing to one, and let $C_{i,o}^{\text{stock}}(u)$ denote the share of participant i 's installed exposure mapped to the incumbent vintage. Define the hedgeability index

$$J_{i,o}(u,m) = \omega_F \exp(-\lambda_H m) + \omega_S C_{i,o}^{\text{stock}}(u) |\text{Corr}[\Delta S_i(u,m), \Delta F_o(u,m)]| \quad (9)$$

The first term is a flow channel: it measures the probability that new frontier-oriented exposure still maps to the named vintage at maturity. The second is a stock channel: it measures the hedgeability of participant i 's installed exposure. The absolute correlation makes the stock term nonnegative. As maturity lengthens, frontier-reference survival and cross-hedge correlation may both fall, producing a maturity inversion in which far-dated old-vintage contracts become economically empty before nearby contracts. The effect is not conventional time-to-expiry decay; it is prospective technological mismatch.

The additive decomposition avoids multiplying two measures that may reflect the same substitution process. If prospective-flow exposure cannot be observed, researchers may report the stock-channel contribution $\omega S_{Ci,0}^{stock(u)} | \text{Corr}[\cdot] |$ alone. Because the omitted flow term is nonnegative, this contribution is a lower bound on the full index defined in equation (9), conditional on the stated weights; the unweighted correlation by itself is a diagnostic rather than a bound.

5. Four Propositions

Proposition 1 - Relevance migrates before price disappears. *Following an identifiable innovation shock and the availability of a successor contract, the incumbent vintage's shares of open interest and price discovery will decline faster than its rental-price index.*

The null is not that the old rental price is constant. Let P denote the incumbent's normalized rental index and let X be either its open-interest share or its Hasbrouck information share. For each co-primary outcome define $DX = (X_{60} - X_0) - (P_{60} - P_0)$ over the 60 trading days following the dated innovation event. Test $H_0: DX \geq 0$ against $H_1: DX < 0$ using a one-sided time-block bootstrap over daily observations within the event window, with Holm correction across the two outcomes and a familywise 5 per cent level. Proposition 1 is supported for an outcome only when the upper endpoint of its adjusted one-sided confidence interval is below zero. An interval containing zero is inconclusive, not a falsification; a statistically precise estimate in the opposite direction supports the competing claim that benchmark relevance is at least as persistent as rental price. Ordinary calendar rolls and common compute-market shocks should be removed before the statistic is constructed. The outcomes, sign convention, window, and inference rule are stated here before trading data become available.

Proposition 2 - Distant old-vintage maturities empty first. *For an incumbent hardware-specific product, liquidity and quoted depth will retreat from longer maturities before they retreat from nearby maturities as the probability of technological displacement rises.*

A clean test compares eligible constant-maturity segments within the same vintage and controls for the exchange's listing schedule. The proposition is weakened if long-dated commercial contracts create natural hedging demand specifically in the old vintage; such evidence would identify hardware lock-in rather than invalidate the framework. A power qualification also applies. In the first six to twelve months after launch, deferred-maturity liquidity is determined largely by the listing schedule and market-maker incentives, so that period has little power. Tests should use only near-versus-deferred maturity pairs that have both been continuously listed and quote-eligible for at least twelve months. Studies should report the minimum detectable deferred-versus-nearby share gap at 80 per cent power under the observed dependence and volatility of the share measures, so that an uninformative null is not reported as support for the alternative.

Proposition 3 - Innovation intensity shortens contract half-life. *Contract half-life will be shorter when performance improvements arrive faster, workloads are more substitutable across generations, and successor capacity becomes commercially available sooner.*

Innovation should be measured by dated events and realized adoption, not by product announcements alone. Shipment delays, software immaturity, power constraints, or export controls can lower effective λ even when advertised performance rises sharply. The proposition concerns the effective innovation intensity λ of Section 4.1, not the frontier-displacement hazard λ_H of Section 4.4. The two imply separable tests in a panel containing multiple vintage transitions: half-life should shorten with adoption-based λ , while, conditional on λ , the viability of distant maturities should shorten with event-count-based λ_H . A single H100-to-B200 transition can illustrate both mechanisms but cannot identify them independently without additional cross-vintage or cross-market variation.

Proposition 4 - Durability transfers risk into governance. *Rolling-vintage or performance-adjusted benchmarks will exhibit longer contract half-lives than fixed-vintage benchmarks, but their hedge errors will become more sensitive to conversion rules, workload weights, rebalancing dates, and benchmark-governance decisions.*

This is not a claim that abstraction is inferior. It identifies a design trade-off. A durable unit avoids repeated product death only by making equivalence across vintages a modeled and revisable judgment.

6. Historical Analogues and Identification Templates

The propositions above require post-launch data that do not yet exist. Historical episodes in other markets can nevertheless clarify what a future measurement exercise must observe. The evidence below consists of three completed benchmark transitions and one never-alive control. They illustrate fixed-label replacement, internal respecification, constituent revision, and launch failure; they do not correspond one-for-one to all four design families in Section 7. Table 1 summarizes the identification role of each episode.

6.1 Specification obsolescence and contract-family replacement (fixed vintage)

When a delivery specification is retired, a fixed-label contract may have to migrate to a successor specification. CME announced that the underlying basis of its miNY unleaded gasoline futures would change from conventional unleaded gasoline to RBOB beginning with the July 2006 contract month (CME Group, 2006). The contemporaneous rule change documents a successor specification and a dated transition point. It does not, by itself, establish how quickly open interest migrated across the broader gasoline complex; that claim requires archived contract-level volume and open-interest records. The episode therefore supplies an identification template for Proposition 1 - track activity by label before and after a successor specification becomes active - rather than a numerical estimate of half-life.

6.2 Respecification within an existing contract (rolling vintage)

The alternative is to redefine the underlier without changing the contract's identity. The NYMEX heating oil contract did precisely that: from the May 2013 contract month the deliverable grade changed from a maximum sulfur content of 2,000 ppm to ultra-low-sulfur diesel at 15 ppm (U.S. Energy Information Administration, 2013). Contract identity, and with it the possibility of preserving the existing liquidity network, survived; the underlying commodity was redefined. This is an institutional analogue of a rolling-vintage design. It illustrates the risk transfer in Proposition 4: avoiding a new contract label concentrates transition basis around the respecification date for hedgers whose physical exposure does not change simultaneously.

6.3 Changing constituents to extend benchmark life (performance-adjusted and governance-based)

A benchmark can also be kept relevant by widening what it includes. S&P Global Commodity Insights announced in 2022 that WTI Midland cargoes would be reflected in Dated Brent and Cash BFOE from the June 2023 loading month, and completed the inclusion in 2023 (S&P Global Commodity Insights, 2022, 2023). The stated purpose was to reinforce the robustness of the benchmark by expanding the deliverable supply reflected in the assessment. Whether correlation with any particular commercial exposure improved is an empirical question, not an assumption made here. The institutional lesson is that comparability across time, and the judgment of when a transition is warranted, become matters for consultation and governance. This is the closest analogue to a performance-adjusted compute unit whose conversion rule is revisable, and hence to Proposition 4.

6.4 The never-alive control: launch failure versus economic expiry

A fourth episode supplies a never-alive control. The Pacific Stock Exchange announced a proposal for DRAM-chip futures in 1989, while a 2003 industry report described renewed DRAM-futures ventures and noted that earlier attempts had not succeeded (Horovitz, 1989; Souza, 2003). These sources document repeated proposals, not a continuously traded contract family. The defensible conclusion is therefore that no durable exchange-traded DRAM market was established in those episodes; it is not that a previously focal contract experienced a measurable half-life. Proposed, approved, listed, quoted, and actively traded states must be coded separately. A half-life should be computed only after a prespecified joint-active threshold is met.

These episodes are analogues, not estimates, calibrations, or formal bounds. Without harmonized dates and activity records, the paper cannot infer that energy-market transition horizons are upper bounds for GPU contracts or that effective GPU innovation intensity λ is numerically larger. Once the relevant series are assembled, historical durations can serve as priors or scale checks. The LIBOR-to-SOFR transition supplies a complementary governance lesson: administrative replacement shifts attention to the fallback provisions of legacy contracts (Alternative Reference Rates Committee, 2022), which connects directly to the discussion of perpetual compute products in Section 9. Taken together, the cases motivate three observable series for a post-launch study: activity share by label, maturity structure around a dated redefinition, and an explicit state classification that prevents never-alive proposals from being treated as rapidly expiring contracts.

Episode	Transition mechanism	Closest design issue	What it identifies
Gasoline contract moved to an RBOB basis (2006)	Underlying specification changed at a dated contract month	Fixed-label replacement analogue	Template for measuring activity migration by label; no half-life estimate is made here
NYMEX heating oil respecified to ULSD (2013)	Underlier redefined inside an existing contract identity	Rolling/respecification analogue	Transition basis around a dated redefinition
WTI Midland included in Dated Brent (2023)	Constituent set widened through benchmark governance	Governance-based constituent revision	Durability-comparability trade-off and governance discretion
DRAM-futures proposals (1989 and 2003)	Proposals did not establish a durable exchange-traded market	Never-alive control; no family correspondence	Separation of launch failure from economic expiry

Table 1. Historical analogues and their identification role.

7. Contract Design Choices

The model suggests four main design families. They differ in where they place technological risk rather than whether they eliminate it.

Design	Durability	Main benefit	Risk transferred elsewhere
Fixed hardware vintage	Low to medium	Salient label; simple settlement; concentrated early liquidity	Rapid economic expiry; repeated successor launches; vintage basis
Rolling front-vintage index	Medium to high	Maintains exposure to commercially leading hardware	Roll discontinuity; rule gaming; forced migration; transition basis

Design	Durability	Main benefit	Risk transferred elsewhere
Performance-adjusted basket	Potentially high	Comparability across hardware generations	Workload-model risk; software dependence; conversion discretion
Token-expenditure index	Potentially high	Closer to application-layer cost	Provider pricing, model efficiency, quality, and margin enter the underlier

Table 2. Technological durability and the location of residual risk.

Fixed-vintage contracts minimize methodology discretion. Their weakness is the need to relaunch liquidity around each generation: short maturity limits, early successor listings, and spread products ease migration, yet every added vintage fragments the order book.

A rolling-vintage index preserves a recognizable rule but creates a discrete transition. The roll date becomes economically important: hedgers can face a gap between installed hardware and the new constituent, and participants may try to influence eligibility.

The comparison becomes sharper when the two designs are written as minimum-variance hedge-error variances over a fixed evaluation horizon H . Let $\rho_0(u; \lambda) = \rho_0 \exp(-\chi\lambda u)$, with $\chi > 0$, so that the correlation between the hedger's exposure and the incumbent future decays with effective innovation intensity. Using the optimal hedge ratio, the fixed-vintage error variance is

$$\text{Var}[e_{\text{fixed}}(H)] = \sigma_s^2 [1 - \rho_0^2 \exp(-2\chi\lambda H)] \quad (10)$$

Under a rolling design, suppose the exposure-futures correlation is maintained near an average level $\bar{\rho}$ over H . If N_h rolls occur within the horizon, each contributes a zero-mean transition error with variance σ_t^2 , and transition errors are mutually independent and uncorrelated with the continuous hedge error, then

$$\text{Var}[e_{\text{roll}}(H)] = \sigma_s^2(1 - \bar{\rho}^2) + N_h\sigma_t^2 \quad (11)$$

Equating equations (10) and (11) defines the critical intensity implicitly. Under the stated functional form, $\lambda^* = -(2\chi H)^{-1} \ln\{[\bar{\rho}^2 - N_h\sigma_t^2/\sigma_s^2]/\rho_0^2\}$, provided $0 < \bar{\rho}^2 - N_h\sigma_t^2/\sigma_s^2 < \rho_0^2$. The rolling design has lower error variance for $\lambda > \lambda^*$, while the fixed-vintage design dominates for $\lambda < \lambda^*$. The expression also makes the governance trade-off explicit: λ^* increases with transition-basis variance and with the number of rolls. If the independence assumptions fail, covariance terms must be added and λ^* should be estimated from the full error covariance matrix rather than from equation (11).

Performance adjustment converts hardware into a common effective-compute unit. Bandi and Su (2026) show the limit with H100-equivalent throughput: prices per normalized unit still differ across generations, because peak operations per second omit memory, interconnect, software, and utilization. The conversion rule is a hedonic model whose revision changes the exposure being hedged.

Token expenditure moves further downstream. It may be durable across chips, but it bundles hardware cost with model architecture, inference efficiency, and provider margins, so it is not a neutral substitute for GPU-hour risk.

8. Measuring Contract Half-Life After Launch

8.1 Observable components

The framework is designed for post-launch measurement with actual trading records. The commercial component C can be proxied by the share of disclosed rental contracts, provider quotations, capacity listings, procurement references, or financing documents mapped to each vintage. No single source is sufficient; the

measure should preserve provider, region, commitment term, and service-tier provenance. Coverage is a real constraint rather than a detail: procurement disclosures, lease agreements, and financing documents are difficult to obtain systematically and are rarely comparable across providers. A tiered strategy is therefore appropriate. Tier 1 uses the full three-component index $R = C^\alpha L^\beta I^\gamma$. Tier 2 drops the commercial component and reports $\tilde{R} = L^{\beta'} I^{\gamma'}$ with $\beta' + \gamma' = 1$, to be used whenever C is unavailable; because commercial exposure usually migrates before order flow does, omitting C biases the estimated half-life upward (too long), whereas under hardware lock-in with early liquidity migration the bias runs the other way, so the expected direction must be stated for the sample at hand. Tier 3 reports single-component half-lives and the disagreement among them. Main results should be reported at Tier 2, with Tier 1 as an extension where the data permit.

The liquidity component L should combine open-interest share, volume share, quoted spread, depth, zero-trade frequency, and maturity concentration. Open interest is more informative about risk transfer than one-day launch volume, while spread and depth distinguish a few large blocks from a continuously usable market. Routine expiration rolls must be removed by aggregating a vintage family or using constant-maturity series.

The information component I should begin with event-aligned lead-lag and forecast comparisons. Formal information-share methods require sufficiently frequent synchronized nonzero changes; they should not be forced onto sparse early data. The unit of analysis is relative contribution across vintages, not whether one contract alone predicts its own settlement index.

A transparent baseline can normalize each component to one over the first stable 20-trading-day window after both incumbent and successor products are active. Equal-weight geometric relevance then produces a first-passage and sustained half-life. Alternative weights, arithmetic aggregation, and single-component half-lives should be reported to show whether the result is driven by liquidity, commercial adoption, or information.

8.2 Event time and identification

Calendar time is an imprecise innovation clock. Relevant events include architecture announcements, general availability, major cloud-provider deployment, benchmark inclusion, software support, export-control changes, and the listing of a successor future. Each event changes a different element of λ or A_n . A credible study should preclassify events and avoid selecting only those followed by visible price movements.

The strongest early design is a within-market comparison of old and new vintage shares around dated events. The outcome is relative migration, not a causal estimate from two treated contracts. The announced CME/Silicon Data and ICE/Ornn families already create a prospective competing-benchmark setting (ICE, 2026). If both products begin trading, their relative liquidity, benchmark composition, and participant clienteles can be compared descriptively, but exchange choice and benchmark composition are jointly determined. Until the ICE/Ornn products actually begin trading, the comparison remains institutional rather than empirical.

Identification must also separate cross-venue from cross-vintage migration. An observed decline in an old vintage's share can reflect competition between venues rather than succession between vintages; the network-effect mechanism is the same in both cases, but the policy implication is not (Cantillon & Yin, 2011). The design implication is to compare vintage shares within a single exchange, holding the venue fixed, and to treat cross-venue share changes as a separate control rather than as evidence of obsolescence. The migration of Bund futures from LIFFE to DTB illustrates the identification logic for a cross-venue liquidity shift. It does not, without a separately estimated mapping, numerically calibrate the network-amplification parameter v .

No obsolescence conclusion should be drawn from a secular decline in H100 rent alone. Proposition 1 requires relative movement toward a successor in open interest or information importance. Missing data must

remain distinct from true zero volume. Provider quotes can support proxy cross-hedging analysis but cannot establish participant-level hedge effectiveness. These constraints preserve the distinction between a testable theory and a retrospective narrative.

Object	Primary measure	Timing	Failure criterion
Commercial relevance	Vintage share of mapped procurement, quotes, capacity, and contract references	Monthly or event time	No relative migration after successor availability
Liquidity relevance	Open-interest and volume share; spreads; depth; zero-trade frequency	Daily/intraday	Old-vintage activity is as persistent as its rental price
Information relevance	Lead-lag, forecast contribution, information/component share	Intraday/daily after data threshold	No shift in price discovery toward successor
Contract half-life	First passage and sustained passage of R_v below 0.5	From joint active window	Estimate is weight-sensitive or reflects missing observations

Table 3. Empirical implementation of contract half-life.

9. Implications for Exchanges, Hedgers, and Regulators

For exchanges, launch is not the end of design: a vintage ladder requires a migration protocol. Successor contracts should be listed early enough to support spreading and position transfer, but not so early that liquidity fragments across thin labels - the trade-off in equation (5a). Exchanges should also publish open interest by vintage and maturity.

For hedgers, the core risk is benchmark lock-in. A seller financing an H100 fleet may need a fixed-vintage hedge; a flexible buyer may prefer rolling exposure. Hedge quality cannot be inferred from a shared GPU label: hedge ratios must be estimated on actual cash flows (Ederington, 1979; Lien & Tse, 2002).

For benchmark administrators, methodology continuity is not economic continuity. An unchanged index preserves auditability while relevance decays; revised constituents preserve relevance while breaking comparability, as Section 6.3 shows. Governance should specify transition triggers, parallel publication, and back-history treatment (IOSCO, 2013).

For regulators, contract half-life is a customer-protection and market-integrity issue. A perpetual maturity does not solve vintage obsolescence: it removes calendar expiry while making benchmark substitution more consequential, as the LIBOR fallback experience in Section 6 shows. A perpetual rolling contract survives only through discretionary transitions.

The broader lesson is sharpest for semiconductor process nodes. A contract written on a named node inherits a testable criterion: when the volume-production ramp of the next node is shorter than half the contract's listed maturity horizon, the fixed-node contract's half-life will be shorter than its nominal term. Battery chemistries and telecommunications capacity raise the same issue.

10. Conclusion

Compute derivatives expose a clock that conventional futures analysis often leaves implicit. A contract may reach its legal settlement date and still have expired economically much earlier. The reason is not that the old underlier becomes worthless. It is that commercial exposure, liquidity, and price discovery migrate to a successor vintage while the old price remains observable.

Contract half-life provides a precise way to study that migration. It is the time for a vintage-specific derivative's economic relevance to fall by half, measured across commercial, trading, and informational functions. A simple model shows why faster innovation, easier substitution, longer maturity, and a viable successor contract shorten half-life. Fixed-vintage designs concentrate liquidity but age quickly; rolling and performance-adjusted designs last longer by transferring risk into transition rules and benchmark governance.

The propositions await real trading data. That is a feature rather than a defect of the research design. The concept is stated before the first announced CME compute contracts begin trading, so later evidence can show whether H100 and B200 activity actually migrates as predicted. Even a failure to launch will be informative: a benchmark can remain publishable without a futures contract becoming economically alive. The central question is therefore not only whether compute futures trade, but whether their definition of compute survives long enough to matter.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interests: The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability: No empirical dataset is analyzed in this conceptual paper. The post-launch measures proposed here are intended for future implementation using actual exchange, benchmark, and commercial-exposure data.

Ethics statement: This research did not involve human participants, personal data, animals, or interventions requiring ethical approval.

AI-use disclosure: Generative AI tools were used for language and document-production assistance. The author is responsible for the argument, source selection, verification, and final text.

Version and date: Working Paper v3.1, 27 August 2026.

References

- Alternative Reference Rates Committee. (2022). LIBOR legacy playbook. Federal Reserve Bank of New York. <https://www.newyorkfed.org/arrc/publications/overall-transition-materials>
- Bandi, F. M., & Su, Y. (2026). (Early) AI compute asset pricing. arXiv:2607.12156. <https://arxiv.org/abs/2607.12156>
- Black, D. G. (1986). Success and failure of futures contracts: Theory and empirical evidence. Monograph Series in Finance and Economics, Monograph No. 1986-1. Salomon Brothers Center for the Study of Financial Institutions, New York University.
- Cantillon, E., & Yin, P.-L. (2011). Competition between exchanges: A research agenda. *International Journal of Industrial Organization*, 29(3), 329-336.
- Carlton, D. W. (1984). Futures markets: Their purpose, their history, their growth, their successes and failures. *Journal of Futures Markets*, 4(3), 237-271. <https://doi.org/10.1002/fut.3990040302>
- Cavanaugh, G., & Penick, M. (2018). The lifecycle of exchange-traded derivatives. *Journal of Commodity Markets*, 10, 47-68. <https://doi.org/10.1016/j.jcomm.2018.05.007>
- CME Group. (2006). NYMEX to change underlying contract for NYMEX miNY unleaded gasoline futures. Press release, 19 May 2006.

- https://www.cmegroup.com/media-room/press-releases/2006/5/19/nymex_to_change_underlyingcontractfornymexminytmunleadedgasoline.html
- CME Group. (2026a). CME Group and Silicon Data to launch compute futures on October 5 to unlock new way to hedge AI risks. Press release, 11 August 2026. https://www.cmegroup.com/media-room/press-releases/2026/8/11/cme_group_and_silicondatatolaunchcomputeuturesonoctober5tounloc.html
- CME Group. (2026b). Initial listing of two (2) compute futures contracts. Special Executive Report SER-9785, 11 August 2026. <https://www.cmegroup.com/notices/ser/2026/08/ser-9785.html>
- Commodity Futures Trading Commission. (2026). Request for comment on the listing of compute derivatives contracts. 91 Federal Register 54259, 21 August 2026. <https://www.federalregister.gov/documents/2026/08/21/2026-17163/request-for-comment-on-the-listing-of-compute-derivatives-contracts>
- Cooley, T. F., Greenwood, J., & Yorukoglu, M. (1997). The replacement problem. *Journal of Monetary Economics*, 40(3), 457-499.
- Ederington, L. H. (1979). The hedging performance of the new futures markets. *Journal of Finance*, 34(1), 157-170. <https://doi.org/10.1111/j.1540-6261.1979.tb02077.x>
- Fernández-Fuertes, R., & Grégoire, V. (2026). Compute futures: Who is left exposed? SSRN Working Paper No. 7224223. <https://doi.org/10.2139/ssrn.7224223>
- Fraumeni, B. M. (1997). The measurement of depreciation in the U.S. national income and product accounts. *Survey of Current Business*, 77(7), 7-23.
- Garbade, K. D., & Silber, W. L. (1983). Price movements and price discovery in futures and cash markets. *Review of Economics and Statistics*, 65(2), 289-297.
- Gonzalo, J., & Granger, C. W. J. (1995). Estimation of common long-memory components in cointegrated systems. *Journal of Business & Economic Statistics*, 13(1), 27-35. <https://doi.org/10.1080/07350015.1995.10524576>
- Greenwood, J., Hercowitz, Z., & Krusell, P. (1997). Long-run implications of investment-specific technological change. *American Economic Review*, 87(3), 342-362.
- Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *Journal of Finance*, 50(4), 1175-1199. <https://doi.org/10.1111/j.1540-6261.1995.tb04054.x>
- Horovitz, B. (1989). Pacific Exchange may list futures in computer chips. *Los Angeles Times*, 30 May 1989.
- Hulten, C. R., & Wykoff, F. C. (1981). The measurement of economic depreciation. In C. R. Hulten (Ed.), *Depreciation, Inflation, and the Taxation of Income from Capital* (pp. 81-125). Urban Institute Press.
- Intercontinental Exchange (ICE). (2026). ICE and Ornn to launch GPU compute futures contracts. Press release, 19 May 2026. <https://ir.theice.com/press/news-details/2026/ICE-and-Ornn-to-Launch-GPU-Compute-Futures-Contracts/default.aspx>
- International Organization of Securities Commissions (IOSCO). (2013). Principles for financial benchmarks. Final Report FR07/13. <https://www.iosco.org/library/pubdocs/pdf/IOSCOPD415.pdf>
- Johnston, E. T., & McConnell, J. J. (1989). Requiem for a market: An analysis of the rise and fall of a financial futures contract. *Review of Financial Studies*, 2(1), 1-23. <https://doi.org/10.1093/rfs/2.1.1>
- Jovanovic, B., & Rousseau, P. L. (2005). General purpose technologies. In P. Aghion & S. N. Durlauf (Eds.), *Handbook of Economic Growth*, Vol. 1B (pp. 1181-1224). Elsevier.
- Lien, D., & Tse, Y. K. (2002). Some recent developments in futures hedging. *Journal of Economic Surveys*, 16(3), 357-396. <https://doi.org/10.1111/1467-6419.00172>

- NVIDIA. (2024). NVIDIA Blackwell platform arrives to power a new era of computing. News release, 18 March 2024. <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>
- NVIDIA. (2025). Inside NVIDIA Blackwell Ultra: The chip powering the AI factory era. NVIDIA Technical Blog, 22 August 2025. <https://developer.nvidia.com/blog/inside-nvidia-blackwell-ultra-the-chip-powering-the-ai-factory-era/>
- NVIDIA. (2026). NVIDIA HGX platform. Product information, accessed 27 August 2026. <https://www.nvidia.com/en-us/data-center/hgx/>
- Silber, W. L. (1981). Innovation, competition, and new contract design in futures markets. *Journal of Futures Markets*, 1(2), 123-155. <https://doi.org/10.1002/fut.3990010205>
- Silicon Data. (2025). Silicon Data announces major revision to A100 & H100 rental indices, launches first-ever B200 index, and introduces new hyperscaler rental benchmarks. 5 December 2025. <https://www.silicondata.com/news-room/silicon-data-announces-major-revision-to-a100-h100-rental-indices-launches-first-ever-b200-index-and-introduces-new-hyperscaler-rental-benchmarks>
- Silicon Data. (2026). GPU rental price indices: Daily benchmarks across the AI compute market. Accessed 27 August 2026. <https://www.silicondata.com/products/silicon-index>
- Solow, R. M. (1960). Investment and technical progress. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical Methods in the Social Sciences*, 1959 (pp. 89-104). Stanford University Press.
- Souza, C. (2003). DRAM futures trading stages a comeback. *EBN/EE Times*, 28 July 2003.
- S&P Global Commodity Insights. (2022). Platts to include WTI Midland in Dated Brent-related assessments from June 2023. Subscriber note, 8 June 2022. <https://www.spglobal.com/energy/en/news-research/latest-news/crude-oil/060822-platts-to-include-wti-midland-in-dated-brent-related-assessments-from-june-2023>
- S&P Global Commodity Insights. (2023). Platts Dated Brent now reflects WTI Midland crude oil, completes inclusion in Brent complex. Press release, 2 May 2023. <https://press.spglobal.com/2023-05-02-Platts-Dated-Brent-Now-Reflects-WTI-Midland-Crude-Oil%2C-Completes-Inclusion-in-Brent-Complex>
- U.S. Energy Information Administration. (2013). Heating oil futures contract now uses ultra-low sulfur diesel fuel. *Today in Energy*, 10 May 2013. <https://www.eia.gov/todayinenergy/detail.php?id=11211>
- Xing, Y. (2026). AI token futures market: Commoditization of compute and derivatives contract design. arXiv:2603.21690. <https://arxiv.org/abs/2603.21690>
- Yang, F. (2026). Futures before fungibility: Compute derivatives and the financialization of an unstandardized market. SSRN Working Paper No. 7324139. <https://doi.org/10.2139/ssrn.7324139>

When Scale Outruns Fees

The Scale–Fee-Density Paradox in China’s Trust Industry

Source manuscript dated 19 August 2026

Fuli Yang

GARCH Institute, Cambridge, MA, USA

Email: dr@garch.ac | ORCID: <https://orcid.org/0009-0005-4055-7561>

August 19, 2026

Abstract

China’s trust industry has returned to asset growth while its fee-generating capacity has not expanded at the same pace. This study documents that divergence using a balanced panel of nine trust institutions with scope-consistent public disclosures from 2021 to 2025 (45 firm-years). On a contemporaneous year-end basis, these institutions held RMB4.87 trillion of trust assets in 2021 (23.7% of the industry total) and RMB8.08 trillion in 2024 (27.3%). The primary measure is fee density: annual trust-business or net fee income divided by year-end trust assets under management (AUM), expressed in basis points. Across institutions, median cumulative AUM growth was 117.1%, whereas median fee-income growth was –32.8% and median fee density fell by 62.2%. Fee density declined at all nine institutions, yielding an exact two-sided sign-test p-value of 0.0039. A two-way fixed-effects log specification estimates an elasticity of fee income with respect to AUM of 0.623 (homoskedastic standard error 0.185), below the proportional benchmark of one. Point estimates remain below one across alternative windows, first differences, exclusion of explicitly caveated observations, and leave-one-institution-out checks, although the estimate excluding one influential institution is not distinguishable from proportional scaling. The evidence therefore supports a consistent descriptive pattern of scale–fee dilution, with magnitudes that vary across sample perturbations. It does not identify a causal mechanism and does not equate fee density with accounting profitability. The findings suggest that AUM growth is an incomplete performance signal during a transition toward asset-service and standardized asset-management activities.

Keywords: trust companies; assets under management; fee density; sustainable financial development; institutional disclosure; China

JEL classifications: G21; G23; L25

1. Introduction

Scale is usually treated as a sign of competitive strength in financial intermediation. A larger asset base can spread fixed technology, compliance, distribution, and professional costs; deepen client relationships; and increase bargaining power. Yet scale can also grow faster than monetization. New assets may enter low-fee mandates, standardized products, administrative service arrangements, or activities in which pricing is compressed. In that setting, assets under management rise while fee income per unit of assets falls. This paper calls that divergence the scale–fee-density paradox, while using the term cautiously: the empirical object is fee-generating capacity, not accounting profit.

The sustainability of financial development depends not only on fiduciary scale but also on whether institutions can support that scale with durable revenue capacity, transparent reporting, and adequate service infrastructure.

The issue is especially salient for Chinese trust institutions. Following the 2018 asset-management reforms and the subsequent contraction of channel and financing activities, the sector has been pushed toward standardized asset management, asset services, and charitable trusts. The 2023 business-classification rules formalized three broad categories—asset-service trusts, asset-management trusts, and charitable trusts—and divided them into 25 business types. This reclassification sharpened the distinction between balance-sheet-like financing, active investment management, and fiduciary or administrative service. It also made aggregate AUM harder to interpret as a homogeneous revenue base: identical amounts of assets can carry very different fee schedules, service intensities, and performance-remuneration potential.

Industry-level figures illustrate the tension. Official association data show that trust assets resumed expansion before operating outcomes fully recovered. But aggregate data cannot establish whether the same institutions expanded assets while experiencing thinner fee yields, and cross-sectional comparisons are vulnerable to persistent differences in business mix and disclosure scope. A further measurement break limits recent annual comparisons: the association's comparable full-year trust-business-income series ends in 2022, and its legacy functional classification of trust assets ends in 2023. A comparable annual industry fee-density series is therefore unavailable after 2022, making institution-level disclosures necessary to extend the analysis through 2025. The central question is narrow: within institutions that report reasonably comparable annual series, did fee or trust-business income scale proportionally with AUM during 2021–2025?

The paper contributes in three ways. First, it constructs a disclosure-based balanced panel with explicit scope normalization. The sample is intentionally small: nine institutions and 45 firm-years. It favors comparability over breadth and does not claim to represent the industry's population, although the institutions accounted for 23.7% of industry trust assets at end-2021 and 27.3% at end-2024 on a contemporaneous year-end basis. Second, it combines a directly interpretable fee-density measure with a log-log two-way fixed-effects elasticity. The ratio describes monetization per unit of AUM; the elasticity checks whether changes in the numerator are sub-proportional without relying only on a common-denominator ratio. Third, it emphasizes robustness and inferential restraint. The analysis reports endpoint sign tests, alternative time windows, first differences, observation exclusions, and leave-one-institution-out diagnostics, while declining to present conventional clustered inference with only nine clusters as decisive.

The result is clear in direction. From 2021 to 2025, eight of nine institutions expanded AUM, but all nine recorded lower fee density. Median institution-level AUM growth exceeded 100%, median fee income fell by roughly one-third, and median fee density declined by more than three-fifths. The baseline fixed-effects elasticity is 0.623: a 1% within-institution increase in AUM is associated with only a 0.62% increase in fee income, after common year effects. Every leave-one-out point estimate remains below one, although the estimate excluding one influential institution approaches proportional scaling and is not distinguishable from one. The broader pattern survives plausible sample perturbations, but the small, selected panel and nonuniform income labels preclude causal or population-wide claims.

The interpretation is operational rather than mechanical. The estimates are consistent with a shift toward lower-fee asset services, price competition, declining performance remuneration, or legacy-business runoff. They cannot distinguish among those channels. What they do establish is that asset growth alone overstated the development of the sampled institutions' fee engines during the observation window. For managers, analysts, and regulators, the implication is to pair AUM with fee yield, recurring revenue quality, business mix, and the cost of serving assets.

2. Institutional Background and Conceptual Framework

2.1 Trust institutions and the changing meaning of scale

Chinese trust institutions are licensed nonbank financial institutions that act as trustees and receive remuneration for managing trust affairs. Trust assets are legally separate from proprietary assets, so AUM is not equivalent to the institution's balance sheet. Economically, however, AUM indicates the volume of fiduciary relationships and assets administered or managed. Its revenue relevance depends on mandate type, pricing, duration, and whether remuneration is fixed, performance-linked, transaction-based, or embedded in a broader service arrangement.

The regulatory transition altered this mapping. The 2018 asset-management framework emphasized look-through regulation, reduced implicit guarantees, and constrained multi-layer channel structures. The 2023 classification notice then defined asset-service trusts, asset-management trusts, and charitable trusts as distinct categories. Asset-service trusts may generate stable but relatively thin administrative fees; asset-management trusts can generate management and performance fees but face market competition and cyclical returns; charitable trusts need not maximize commercial fee output. Consequently, an increase in aggregate AUM can reflect a change in mix rather than a proportionate expansion of monetizable, high-fee mandates.

2.2 Scale, capacity, and fee dilution

The asset-management literature documents decreasing returns to scale alongside the cost economies that a larger asset base can provide. In Berk and Green's (2004) equilibrium model, capital flows toward skill until decreasing returns dissipate abnormal performance. Chen et al. (2004) and Yan (2008) find that larger mutual funds can suffer weaker performance, particularly when liquidity and organizational complexity bind. Pástor, Stambaugh, and Taylor (2015) likewise document decreasing returns at the industry level and more nuanced fund-level effects. These studies concern investment performance, not trust-company fee income, but they establish a general point: financial-intermediation capacity need not scale linearly with assets.

For a trust institution, let F_{it} denote annual scope-consistent trust-business or net fee income and A_{it} denote year-end trust AUM. If fee income scaled proportionally, $F_{it} = kA_{it}$ and the fee yield k would remain constant. More generally, suppose $F_{it} = cA_{it}^\beta$. Then $\beta = 1$ implies proportional scaling, $\beta < 1$ implies declining fee density as scale rises, and $\beta > 1$ implies increasing fee density. The empirical analysis treats $\beta < 1$ as the defining signature of scale-fee dilution.

Four mechanisms could generate $\beta < 1$. First, composition may shift toward low-fee service or passive mandates. Second, larger institutional mandates may carry volume discounts. Third, performance remuneration may fall even when administered assets rise. Fourth, legacy high-margin financing activities may run off faster than new businesses monetize. Fixed compliance, technology, and professional costs could still produce cost-side economies, so declining fee density does not necessarily imply declining profit margins. Direct cost and profit data would be required to make that stronger claim.

3. Data and Measurement

3.1 Source hierarchy and sample construction

The dataset is assembled from annual reports of trust institutions, listed-parent annual reports, official company disclosures, and regulatory or exchange filings. A small number of annual values are drawn from authoritative reproductions of official reports when the original document is not machine-accessible. The source hierarchy gives priority to audited parent-company disclosures and trust-asset statements, followed by

official summaries and clearly identified secondary reproductions. Each observation is tied to a source note and a scope decision.

An institution enters the strict sample only if it has five consecutive annual observations from 2021 through 2025 for both year-end trust AUM and a scope-consistent measure of fee or trust-business income. The income series must remain economically comparable within an institution. Consolidated commission income is not used when it includes brokerage, asset-management-subsidiary, or other non-trust activities. When an annual report separately identifies base trust fees, performance remuneration, or other trust remuneration, those components are combined to construct trust-business income. When reports disclose only a consistently defined parent net fee measure, that series is retained as labeled.

These rules produce a balanced panel of nine institutions: BOCOM Trust, CITIC Trust, Jiangsu Trust, Yingda Trust, Hwabao Trust, Shanghai Trust, Shaanxi International Trust, Huaneng Guicheng Trust, and Huaxin Trust. The panel is selected on disclosure completeness and comparability. It is not random, and disclosure quality may correlate with ownership, listing status, size, or governance. The evidence should therefore be read as a documented pattern within the observed institutions, not an estimate of an industry-population parameter.

Although small in count, the panel is economically material. Combined year-end trust assets of the nine institutions were RMB4.869 trillion at end-2021, or 23.7% of the industry total of RMB20.55 trillion (China Trustee Association, 2022). At end-2024, the corresponding figures were RMB8.076 trillion and 27.3% of the industry total of RMB29.56 trillion (China Trustee Association, 2025). These are contemporaneous year-end ratios. The sample institutions held RMB10.546 trillion at end-2025, but the latest published industry total—RMB32.43 trillion—refers to June 2025; the date-mismatched figures are therefore not used to calculate a 2025 coverage ratio. The sample boundary follows the stated five-year completeness and scope-consistency rules. It does not imply that the nine institutions approximate the industry population or constitute a random sample.

3.2 Variables

AUM is the institution's year-end trust-asset balance, expressed in RMB billions. Fee income is annual trust-business income or net fee and commission income, expressed in RMB billions, using the narrowest comparable parent scope available. The primary descriptive outcome is fee density:

$$Fee\ density_{it} = 10,000 \times Fee\ income_{it} / AUM_{it} .$$

The multiplier expresses annual income as basis points of year-end AUM. A fee density of 20 basis points means that recorded annual trust-business or fee income equals 0.20% of the year-end asset base. This is a revenue-yield measure. It is not a contractual fee rate, because the numerator may include performance remuneration and the denominator is an end-of-period stock rather than average AUM. It is also not a profit margin, because operating costs, investment income, impairments, taxes, and proprietary-business results are excluded.

Ratio measures can generate misleading correlations when numerator and denominator are mechanically related or jointly measured with error (Kronmal, 1993). For that reason, the paper does not rely only on fee density. It also estimates the elasticity of fee income with respect to AUM in logs. Endpoint changes, year medians, and exact directional tests provide additional transparent summaries.

Table 1. *Institution-level endpoint changes, 2021–2025*

Institution	AUM growth	Fee-income growth	2021 density (bp)	2025 density (bp)	Density change
BOCOM Trust	23.8%	-48.2%	22.15	9.26	-58.2%
CITIC Trust	287.6%	-58.8%	59.06	6.27	-89.4%
Jiangsu Trust	185.0%	7.8%	21.99	8.32	-62.2%
Yingda Trust	87.9%	33.3%	31.76	22.54	-29.0%
Hwabao Trust	5.3%	-48.2%	32.46	15.96	-50.8%
Shanghai Trust	132.8%	-32.8%	21.63	6.24	-71.1%
Shaanxi International Trust	213.3%	15.0%	60.99	22.38	-63.3%
Huaneng Guicheng Trust	-17.6%	-80.1%	61.20	14.75	-75.9%
Huaxin Trust	117.1%	34.7%	54.10	33.57	-38.0%

Notes: AUM is year-end trust assets. Fee income is a within-institution scope-consistent trust-business or net fee measure. Density change is the percentage change in fee density, that is, in fee income divided by year-end AUM. Percentage changes are computed from unrounded underlying values.

3.3 Descriptive patterns

Table 1 shows that the paradox is not an artifact of a single extreme institution. Eight institutions expanded AUM, and one contracted by 17.6%. Fee income increased at only four institutions and fell at five. Yet fee density declined everywhere. The largest declines occur where rapid AUM expansion coincides with sharply lower fee income, but even institutions with positive fee growth experienced dilution because AUM grew faster.

Across the nine institutions, median cumulative AUM growth was 117.1%, median fee-income growth was -32.8%, and median fee-density change was -62.2%. The finding is directional as well as quantitative: all nine endpoint changes in density are negative. Under a null in which positive and negative changes are equally likely and independent across institutions, the exact two-sided sign-test probability of nine declines is 0.00390625. This diagnostic does not solve sample selection or dependence, but it shows that the common direction is unlikely under a symmetric sign null.

3.4 Industry context and the disclosure gap

Aggregate association data provide a limited benchmark for the years in which comparable series exist; they do not validate the panel's representativeness. Table 2 assembles the available annual industry series. Industry-wide trust-business income divided by year-end trust assets—an industry analogue of fee density—declined from 42.2 basis points in 2020 and 42.3 basis points in 2021 to 34.2 basis points in 2022, the last year for which the association published a full-year trust-business-income figure (China Trustee Association, 2021, 2022, 2023). The legacy functional composition of trust assets also shifted before that series ended: from end-2021 to end-2023, the financing-trust share fell from 17.43% to 14.5%, while the investment-trust share rose from 41.38% to 48.34% (China Trustee Association, 2022, 2024). Newer association releases discuss the

three-category framework, but those measures are not directly continuous with the legacy functional classification.

Two implications follow. First, the early-period aggregate fee yield moved in the same direction as the panel's fee-density measure, which provides contextual consistency but does not establish that the 2021–2025 panel result is representative of the industry. Second, no comparable full-year industry fee-density observation is available after 2022, and the legacy functional classification ends in 2023. Institution-level disclosures are therefore required to extend a scope-consistent annual fee-yield analysis through 2025; that extension is the specific value of the panel.

Table 2. *Industry aggregates from China Trustee Association releases*

Year-end	Trust assets (RMB trn)	Trust-business income (RMB bn)	Fee density (bp)	Fin. share (%)	Inv. share (%)	Svc./adm. share (%)
2020	20.49	86.45	42.2	23.71	31.46	44.84
2021	20.55	86.87	42.3	17.43	41.38	41.20
2022	21.14	72.27	34.2	14.55	43.92	41.53
2023	23.92	n.a.	n.a.	14.5	48.34	37.12
2024	29.56	n.a.	n.a.	n.a.	n.a.	n.a.

Notes: All figures are from China Trustee Association releases. Trust assets are industry year-end balances. Trust-business income is the association's annual industry figure; the comparable full-year series ends in 2022. The association subsequently reported cumulative trust-business income of RMB38.738 billion through 2023Q3, but no comparable 2023 full-year value. Industry fee density is trust-business income divided by year-end trust assets, in basis points. Shares refer to the legacy functional classification of trust assets into financing, investment, and service/administrative categories, which ends with the 2023 year-end release. The 2020 financing and investment shares are computed from reported amounts. At the time of writing, the latest official 2025 industry AUM observation was RMB32.43 trillion at June 2025 (China Trustee Association, 2025); it is not combined with end-2025 sample AUM to calculate coverage.

Table 3. *Cross-sectional medians by year*

Year	Median AUM (RMB bn)	Median fee income (RMB bn)	Median fee density (bp)
2021	590.4	1.326	32.46
2022	495.2	1.383	31.28
2023	531.8	1.100	21.57
2024	635.7	1.327	20.87
2025	730.6	0.981	14.75

Notes: Medians are calculated variable by variable across the same nine institutions in each year, so the median fee density is not the ratio of the median fee income to the median AUM. Because institution ranks change, the median of institution-level cumulative growth rates (117.1%) is not equal to the percentage change between the annual cross-sectional medians shown here. The balanced composition avoids entry and exit effects but does not make the sample representative of the industry.

4. Empirical Design

4.1 Endpoint and median diagnostics

The first design is deliberately nonparametric. It compares 2021 and 2025 AUM, fee income, and fee density for each institution; reports medians across institutions; and applies an exact sign test to the direction of density changes. This design is transparent and robust to the magnitude of outliers, but it ignores intermediate-year dynamics and common shocks.

4.2 Two-way fixed-effects elasticity

The main panel specification is:

$$\ln(F_{it}) = \alpha_i + \gamma_t + \beta \ln(A_{it}) + \varepsilon_{it},$$

where α_i are institution fixed effects and γ_t are year fixed effects. Institution effects absorb time-invariant differences in business model, ownership, reporting practice, and average pricing. Year effects absorb common macroeconomic, market, and regulatory conditions. The coefficient β is estimated from within-institution deviations in AUM and fee income after removing common year movements. The benchmark is $H_0: \beta = 1$, not $H_0: \beta = 0$. A coefficient below one means that fee income changes sub-proportionally with AUM.

The regression is descriptive. AUM is endogenous to strategy, client demand, business mix, performance, and risk capacity. Fee income and AUM can also be jointly affected by classification changes or market conditions. Fixed effects remove persistent level differences and common annual shocks, but they do not supply an instrument, treatment, or quasi-experimental source of variation. Accordingly, β is referred to as a within-panel elasticity association rather than a causal return-to-scale parameter.

4.3 Inference and robustness

The baseline calculations report homoskedastic standard errors as a diagnostic because conventional cluster-robust asymptotics are unreliable with only nine institution clusters. Cameron, Gelbach, and Miller (2008) show that standard clustered tests can over-reject with few clusters, and MacKinnon, Nielsen, and Webb (2023) emphasize cluster leverage and influence diagnostics. Rather than attaching strong significance claims to a fragile variance estimator, this paper focuses on coefficient stability, sign consistency, and leave-one-institution-out influence.

Five perturbations are pre-specified. First, the sample is truncated to 2021–2024, removing all 2025 observations. Second, it is shifted to 2022–2025, checking dependence on the 2021 endpoint. Third, three 2025 observations with explicit rounding or scope caveats are excluded. Fourth, the equation is estimated in first differences with year effects, so permanent institution heterogeneity is removed algebraically. Finally, each institution is omitted in turn.

5. Results

5.1 Scale expansion without proportional fee growth

The cross-sectional medians in Table 3 trace the divergence over time. Median AUM rose from RMB590.4 billion in 2021 to RMB730.6 billion in 2025, although the median path was uneven over time, falling to RMB495.2 billion in 2022 before recovering. Median fee density, by contrast, declined from 32.46 basis points to 14.75 basis points. The median annual fee-income level fell from RMB1.326 billion to RMB0.981 billion. These annual cross-sectional medians should be distinguished from the median of

institution-specific cumulative growth rates because institution ranks change across years. Even so, the endpoint medians show that the pattern is not driven solely by a denominator expansion: the numerator also weakened.

The institution-level results add important heterogeneity. CITIC Trust recorded the largest cumulative AUM expansion in the sample and one of the sharpest density declines. Jiangsu Trust and Shaanxi International Trust increased fee income over the window, but not nearly as quickly as AUM. Yingda Trust and Huaxin Trust also posted positive fee growth, yet density still declined. Hwabao Trust's AUM was nearly flat while fee income fell, showing that dilution can arise from numerator contraction even without major scale growth. Huaneng Guicheng Trust contracted in both AUM and fee income, with the latter falling faster.

5.2 Fixed-effects elasticity

The baseline two-way fixed-effects estimate of β is 0.623. Thus, after institution and year effects, a 1% increase in AUM is associated with a 0.62% increase in fee income. The estimate lies 0.377 below the proportional benchmark of one, an economically meaningful gap. Using the reported homoskedastic standard error of 0.185, the t-statistic against one is -2.04 . Given the small number of institutions and the absence of a causal design, that statistic is best read as a scale diagnostic rather than a definitive hypothesis test.

The elasticity result complements the ratio evidence. Fee density necessarily falls when fee income grows more slowly than AUM, but the log specification asks whether the same sub-proportional relation appears after removing institution averages and common year effects. It does. The result therefore cannot be explained solely by permanently lower fee yields at larger institutions or by an industry-wide annual decline common to every institution.

Table 4. *Elasticity robustness checks*

Specification	N	Elasticity	Homoskedastic SE	t vs. 1
Baseline institution + year FE	45	0.623	0.185	-2.04
Window 2021–2024	36	0.678	0.236	-1.37
Window 2022–2025	36	0.507	0.209	-2.36
Exclude three explicit 2025 scope/rounding caveats	42	0.653	0.194	-1.79
First differences + year FE	36	0.541	0.170	-2.70

Notes: Dependent variable is log fee income; the scale variable is log year-end AUM. Level specifications include institution and year fixed effects. The first-difference specification includes year effects. Standard errors are homoskedastic diagnostics, not small-cluster-robust inference.

6. Robustness and Interpretation

6.1 Time windows, measurement caveats, and first differences

Table 4 shows that sub-proportional scaling is stable across alternative designs. The 2021–2024 window produces an elasticity of 0.678; the 2022–2025 window produces 0.507. Excluding three 2025 observations with explicit scope or rounding caveats yields 0.653, close to the baseline. The first-difference estimate is 0.541. Across these specifications, the elasticity ranges from 0.507 to 0.678. Every point estimate lies below one, although in the 2021–2024 window ($t = -1.37$) and the caveat-exclusion sample ($t = -1.79$) the difference from one is not statistically significant at conventional levels.

The window comparison matters because 2021 captures a different phase of the sector's regulatory transition, while 2025 incorporates rapid growth in newer service and securities-investment activities. Both overlapping four-year windows remain below one, though the magnitudes differ (0.678 versus 0.507), which reduces concern that one endpoint mechanically generates the result. The caveat exclusion addresses a separate problem: annual reports use heterogeneous labels and sometimes rounded values. Removing the most explicit 2025 ambiguities does not reverse the finding.

6.2 Leave-one-institution-out influence

Leave-one-institution-out point estimates range from 0.321 to 0.888, and all nine lie below one. The highest estimate arises when CITIC Trust is omitted (0.888, $t = -0.68$ against one): its rapid AUM growth and fee contraction contribute meaningfully to the baseline result, and without it the remaining panel cannot be statistically distinguished from proportional scaling. Omitting Huaneng Guicheng Trust produces the lowest estimate, 0.321. The direction of every estimate is consistent with dilution, but the strength of the statistical evidence depends materially on one influential institution.

This exercise is especially useful in a small panel. With nine clusters, formal asymptotics are fragile, while influence can be inspected directly. The result should not be summarized as a precise universal elasticity of 0.623. A more defensible statement is that every reasonable sample perturbation examined yields subproportional fee scaling, with the estimated magnitude varying across institutions and windows.

6.3 What the evidence can and cannot explain

The most plausible interpretation is business-model dilution: new AUM carried a lower average revenue yield than legacy assets. The industry context in Table 2 is consistent with this interpretation but does not identify it: the financing-trust share fell while the investment-trust share rose between 2021 and 2023, and the industry-wide revenue yield declined from roughly 42 to 34 basis points before the comparable full-year income series ended. Yet the dataset does not observe mandate-level fee schedules or AUM by the 25 regulatory business types. A lower fee yield may reflect deliberate entry into scalable, capital-light services rather than deteriorating economics. If low-fee assets require little marginal cost and generate cross-selling or client-retention benefits, they can create value despite compressing fee density.

Performance remuneration is another possible channel. Where disclosed, it can fluctuate sharply with market outcomes and product realization cycles. AUM may grow continuously while performance fees are episodic. Price competition and institutional-client bargaining could also lower realized fee rates. Finally, runoff of legacy financing trusts can remove high-fee income faster than assets decline if replacement assets are standardized or administrative. Distinguishing these mechanisms requires mandate-level composition, average-AUM denominators, recurring versus performance income, and cost allocation.

7. Implications

7.1 Management implications

Trust institutions should not use AUM growth as a sufficient operating target. A more informative scorecard would decompose AUM by business category, client type, pricing model, duration, and service intensity; separate recurring base fees from performance remuneration and one-off compensation; and track both gross and contribution-margin yield. This would reveal whether new assets are entering strategically attractive low-cost mandates or merely diluting revenue without lowering unit cost.

The elasticity benchmark also provides a practical diagnostic. If an institution's multi-year fee-income elasticity remains below one, management should ask whether the gap reflects intentional mix, temporary

ramp-up, price concessions, or weak monetization. The answer changes the appropriate response. Intentional low-fee service growth may require automation and shared infrastructure; price compression may require differentiation; weak performance fees may require product and investment-capability changes; legacy runoff may require a transition plan rather than indiscriminate asset targets.

7.2 Investor and analyst implications

External analysis should pair trust AUM with fee density, recurring fee growth, and disclosure scope. Consolidated fee and commission income can include non-trust subsidiaries or activities, while parent-company totals can contain non-trust commissions. Comparisons based on unadjusted headline figures may therefore rank institutions by reporting perimeter rather than operating economics. A source-by-source normalization protocol is essential.

The evidence also cautions against interpreting a return to industry AUM growth as a return to prior earnings economics. A sector can expand in socially useful service trusts and standardized products while average fee yields decline. That outcome may be consistent with regulatory objectives and lower systemic risk, but it changes valuation, staffing, technology, and capital-allocation assumptions.

7.3 Regulatory and disclosure implications

The three-category framework improves conceptual clarity, but public quantitative disclosure remains uneven. More standardized reporting of year-end and average AUM by category, recurring base fees, performance remuneration, other trust remuneration, and parent versus consolidated scope would materially improve market discipline. A reconciliation from trust-business income to the income statement would reduce ambiguity without disclosing client-sensitive information. Aggregate disclosure has also become less continuous: the association's comparable full-year trust-business-income series ends in 2022 and the legacy functional classification ends in 2023, leaving no public full-year aggregate fee-yield series on a continuous basis thereafter. Restoring an annual income series under the new three-category definitions would allow the dilution documented here to be monitored at the industry level.

Regulators should also distinguish low fee density from weak fiduciary value. Asset services may be intentionally low-margin and high-volume. The relevant supervisory question is whether pricing covers the operational, technology, legal, and conduct risks of the service, not whether every business type matches historical financing-trust yields. Better cost and incident reporting would make that assessment possible.

For emerging financial systems, this reporting discipline supports both sustainable economic growth and stronger, more accountable financial institutions.

8. Limitations

The first limitation is selection. Nine institutions qualify because they disclose five years of comparable AUM and fee information. They may differ systematically from institutions with incomplete disclosure. Results cannot be extrapolated mechanically to the full sector. Industry-level corroboration is likewise bounded: the aggregate income series ends in 2022 and the functional classification in 2023. Second, annual-report terminology and accounting scope are not perfectly uniform. The construction prioritizes within-institution consistency, so the absolute level of fee density is more comparable over time within an institution than across institutions.

Third, the denominator is year-end AUM, whereas fee income accumulates over the year. Large intra-year flows can distort the yield. Average AUM would be preferable, but it is not disclosed on a consistent basis; where opening and closing trust-asset balances are reported, only a two-point average can be constructed. Fourth, fee density mixes base fees and, where scope requires, performance or other trust

remuneration. It is a realized revenue yield, not a posted schedule. Fifth, only five annual observations per institution are available, which limits dynamic modeling and makes serial dependence difficult to estimate.

Sixth, the fixed-effects elasticity is associational. Unobserved time-varying business mix, client composition, pricing, and risk can drive both AUM and fees. Seventh, the paper does not model costs. Scale–fee dilution can coexist with cost economies, so the results cannot establish declining operating margins, return on equity, or social welfare. Finally, small-cluster inference is intrinsically fragile. The analysis therefore emphasizes exact signs and coefficient stability rather than conventional significance thresholds.

9. Conclusion

Between 2021 and 2025, the nine trust institutions in this balanced disclosure sample expanded assets without proportionate growth in fee or trust-business income. The median institution-level change was +117.1% for AUM, –32.8% for fee income, and –62.2% for fee density. All nine institutions recorded lower fee density. The baseline within-panel elasticity of fee income to AUM is 0.623, and every alternative-window, caveat-exclusion, first-difference, and leave-one-out point estimate remains below the proportional benchmark of one, although one leave-one-out estimate is not distinguishable from one.

The scale–fee-density paradox is therefore real as a descriptive fee-yield pattern in this sample, but its label should not be overread. The analysis does not show that larger trust institutions are less profitable, that asset-service trusts destroy value, or that regulation caused the decline. It shows something narrower and operationally important: during a period of structural transition, AUM growth substantially outpaced the sampled institutions’ ability to convert assets into recorded fee income. Future research should connect mandate-level business mix, recurring and performance fees, average AUM, operating costs, and risk outcomes. Until then, scale should be treated as an input to economic performance, not a substitute for it.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests: The author declares no competing interests.

Ethics statement: The study uses publicly available institution-level information and involved no human participants, personal data, or animal subjects; ethics approval was not required.

Data availability: The constructed institution-year dataset is available from the author upon reasonable request. The underlying annual reports, regulatory or exchange filings, company disclosures, and association releases are publicly available from the sources cited in the manuscript.

Generative AI disclosure: OpenAI Codex (GPT-5) was used for language editing, formatting, and coding assistance; the author verified all content and takes full responsibility.

References

- Berk, J. B., & Green, R. C. (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112(6), 1269–1295. <https://doi.org/10.1086/424739>
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427. <https://doi.org/10.1162/rest.90.3.414>
- Chen, J., Hong, H., Huang, M., & Kubik, J. D. (2004). Does fund size erode mutual fund performance? The role of liquidity and organization. *American Economic Review*, 94(5), 1276–1302. <https://doi.org/10.1257/0002828043052277>

- China Banking and Insurance Regulatory Commission. (2023). Notice on standardizing the classification of trust-company trust business (CBIRC Regulation No. 1 of 2023).
<https://app.www.gov.cn/govdata/gov/202303/26/498271/article.html>
- China Banking Regulatory Commission. (2007). Measures for the administration of trust companies (CBRC Order No. 2 of 2007; first amended by the National Financial Regulatory Administration on 15 May 2025). <https://www.nfra.gov.cn/cn/view/pages/rulesDetail.html?docId=1209888>
- China Trustee Association. (2021). Analysis of trust company main business data at year-end 2020.
<http://www.xtxh.net/txth/statistics/46670.htm> (archived at
<https://web.archive.org/web/20240523083310/http://www.xtxh.net/txth/statistics/46670.htm>)
- China Trustee Association. (2022). Analysis of trust company main business data at year-end 2021.
<http://www.xtxh.net/txth/statistics/47592.htm> (archived at
<https://web.archive.org/web/20250212043150/http://www.xtxh.net/txth/statistics/47592.htm>)
- China Trustee Association. (2023). Analysis of trust company main business data at year-end 2022.
<http://www.xtxh.net/txth/statistics/48366.htm> (archived at
<https://web.archive.org/web/20241115035845/http://www.xtxh.net/txth/statistics/48366.htm>)
- China Trustee Association. (2024). Main business data of trust companies at year-end 2023.
<http://www.xtxh.net/txth/statistics/48366.htm>
- China Trustee Association. (2025). Review of trust-industry development in the first half of 2025.
<http://www.xtxh.net/txth/statistics/48366.htm>
- Kronmal, R. A. (1993). Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society: Series A*, 156(3), 379–392. <https://doi.org/10.2307/2983064>
- MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2023). Fast and reliable jackknife and bootstrap methods for cluster-robust inference. *Journal of Applied Econometrics*, 38(5), 671–694.
<https://doi.org/10.1002/jae.2969>
- National Financial Regulatory Administration. (2025). Measures for the administration of trust companies (NFRA Order No. 8 of 2025; issued 11 September 2025, effective 1 January 2026).
<https://www.nfra.gov.cn/cn/view/pages/rulesDetail.html?docId=1225642>
- Pástor, L., Stambaugh, R. F., & Taylor, L. A. (2015). Scale and skill in active management. *Journal of Financial Economics*, 116(1), 23–45. <https://doi.org/10.1016/j.jfineco.2014.11.008>
- People's Bank of China, China Banking and Insurance Regulatory Commission, China Securities Regulatory Commission, & State Administration of Foreign Exchange. (2018). Guiding opinions on regulating the asset-management business of financial institutions (PBC Document No. 106 of 2018).
<https://www.pbc.gov.cn/jinrongwendingju/146766/146770/c2e228ad641d4a85acdc9142da4f955d/index.html>
- Yan, X. (S.). (2008). Liquidity, investment style, and the relation between fund size and fund performance. *Journal of Financial and Quantitative Analysis*, 43(3), 741–767.
<https://doi.org/10.1017/S0022109000004270>

Appendix A. Leave-One-Institution-Out Diagnostics

Table A1. *Influence of individual institutions on the elasticity estimate*

Institution omitted	Elasticity	Homoskedastic SE	t vs. 1
BOCOM Trust	0.619	0.202	-1.88
CITIC Trust	0.888	0.165	-0.68
Jiangsu Trust	0.644	0.203	-1.76
Yingda Trust	0.634	0.180	-2.04
Hwabao Trust	0.650	0.208	-1.68
Shanghai Trust	0.635	0.202	-1.81
Shaanxi International Trust	0.572	0.218	-1.96
Huaneng Guicheng Trust	0.321	0.191	-3.55
Huaxin Trust	0.605	0.197	-2.01

Notes: Each row re-estimates the institution- and year-fixed-effects model after omitting one institution. All point estimates remain below one; the estimate excluding CITIC Trust is 0.888 ($t = -0.68$ against one) and is not distinguishable from proportional scaling. Standard errors are diagnostic.

Appendix B. Comparability Protocol

Item	Rule
Unit of analysis	Institution-year, 2021–2025.
AUM	Year-end trust assets, not consolidated managed assets or proprietary assets.
Income	Narrowest scope-consistent parent trust-business or net fee measure available.
Components	Base fees, performance remuneration, and other trust remuneration are combined only when the report identifies them as trust-business income.
Subsidiaries	Brokerage, asset-management-subsidary, and other non-trust commissions are excluded when separable.
Missing values	No AUM or fee-income value is imputed in the strict sample.
One-offs	Material one-off remuneration prevents entry into the strict sample unless it can be isolated consistently.
Rounding	Explicitly rounded observations are retained only with a caveat and tested through exclusion.
Interpretation	Fee density is a revenue yield; it is neither a contractual rate nor an accounting profit margin.

From Disclosure to Resilience: Climate-Risk Transmission Frictions in China's Corporate Bond Market

Fuli Yang* and Shiyu Lin

GARCH Institute, Harvard Square, Cambridge, MA 02138, USA

Fuli Yang: ORCID 0009-0005-4055-7561 | Shiyu Lin: ORCID 0009-0000-2471-3644

*Corresponding author: dr@garch.ac

Abstract

The paper examines why climate-related information does not consistently become a traceable signal in corporate bond prices, contracts, and secondary-market resilience. It defines transmission adequacy as the ability to connect a material exposure to the correct legal obligor, align it with the information date and financial horizon, document its treatment by credit intermediaries, and observe it in financing and stress-state indicators. A structured critical review of 49 verified academic and institutional sources, updated through 13 October 2026, organizes evidence around five links: disclosure, exposure measurement, pricing and contract design, credit intermediation, and market resilience. The synthesis identifies five corresponding frictions—legal-entity and asset-location mismatch, information-date and horizon misalignment, measurement non-comparability, absorption by ratings, guarantees and ownership, and limited price discovery in thinly traded bonds. For China, the article proposes Module 1, an issuer–emitter–bond map; Module 2, bond-relevant climate disclosure; Module 3, transparent intermediary treatment; Module 4, a climate-credit stress dashboard; and Module 5, pre-specified pilots and proportional scaling. These modules target traceability rather than an administratively imposed premium and are paired with burden safeguards, implementation metrics and failure interpretations.

Keywords: climate risk; corporate bonds; information disclosure; credit intermediation; market liquidity; China

JEL classifications: G12; G18; G32; Q54; Q58

1. Introduction

Climate risk can enter debt markets through expected cash flows, default probability, collateral values, refinancing capacity, risk appetite and market liquidity. Yet the empirical record does not deliver one stable climate premium (Giglio et al., 2021). Physical exposure is associated with municipal and corporate borrowing costs in several settings, but its magnitude depends on geography, maturity and adaptation capacity (Painter, 2020; Goldsmith-Pinkham et al., 2023; He et al., 2025). Carbon and transition measures are priced in some equity and bond samples, while other studies find that governance commitments or disclosure quality matter more than measured carbon exposure itself (Bolton & Kacperczyk, 2021, 2023; Lemma et al., 2021; Dong et al., 2026). Contract outcomes can move in a different direction from prices, and recent evidence shows that emissions can be associated with secondary spreads but not primary spreads (Ginglinger & Moreau, 2023; Goodell et al., 2025; Kim & Pouget, 2026).

A common response is to search for a better proxy or a more favorable specification. That approach mistakes an institutional question for a coefficient-selection problem. A risk can be economically material and still fail to appear as a uniform average premium when the legal obligor is not the measured entity, the disclosed information was unavailable at the decision date, the horizon of the exposure is longer than the bond, guarantees or implicit support absorb the signal, or trading is too sparse to reveal stress through transactions. Conversely, a significant association can reflect disclosure salience, regional conditions, issuer selection or liquidity rather than a structural default premium. The relevant policy question is therefore not simply whether climate risk is priced. It is whether the market's transmission chain is adequate for investors, intermediaries and supervisors to trace material climate information into credit decisions and stress monitoring.

The paper develops the concept of transmission adequacy. A climate-risk governance system is adequate when information can be traced from an operating exposure to the correct legal obligor, aligned to the information date and financial horizon, translated into a documented credit judgment, and observed in both financing terms and stress-state market indicators. This definition avoids two symmetric errors. It does not equate an insignificant mean coefficient with policy failure or economic irrelevance. It also does not treat any positive climate coefficient as proof of efficient risk pricing. Instead, it asks which link generated, transformed, absorbed or obscured the signal.

A concrete information-timing error illustrates the framework. China's 2025 national carbon-market expansion notice was dated 20 March but publicly released on 26 March. Treating bonds priced during the intervening days as already exposed violates the market's information set. Similar errors arise when a listed parent is substituted for a bond obligor or a registered office for operating assets. Transmission adequacy makes these lineage choices observable rather than burying them inside a risk score.

The contribution is threefold. First, the paper reorganizes a fragmented literature around a five-stage chain—disclosure, measurement, pricing and contract design, credit intermediation, and market resilience—rather than around isolated risk proxies. Second, it uses this chain to diagnose China-specific frictions, including listed-company versus bond-obligor mismatches, changing emissions-trading scope, concentrated bond ownership, and the distinction between timely disclosure and decision-useful disclosure. Third, it converts the diagnosis into a sequenced policy architecture with measurable implementation indicators, burden safeguards and failure interpretations. The objective is not to direct prices. It is to improve the informational and institutional conditions under which market participants can form defensible prices and contracts.

The remainder proceeds as follows. Section 2 explains the structured review method. Section 3 distinguishes risk measurement from risk transmission. Sections 4 and 5 synthesize the evidence and diagnose the five frictions. Section 6 develops the China-specific institutional assessment. Section 7 proposes a staged policy architecture and evaluation design. Section 8 presents a research agenda, and Section 9 concludes.

2. Review method and scope

The article uses a structured critical review rather than a meta-analysis. The empirical literature combines different risk constructs, markets, horizons, outcomes and identification strategies; pooling coefficients would give a false impression that each study estimates the same object. The review instead codes what information enters each study, how it is aligned to the issuer and decision date, which market margin is observed, and which institutional condition mediates the result.

Unlike broad climate-finance reviews such as Giglio et al. (2021), this article does not attempt to survey every climate-related financial market. It narrows the unit of synthesis to the operational links and frictions through which climate information can reach corporate bond prices, contracts, credit intermediation and stress-state resilience.

The core literature searches covered January 2000 through 4 September 2026; targeted literature and official-policy updates were completed on 13 October 2026. Core searches combined climate, physical risk, transition risk, carbon exposure, emissions trading and climate disclosure with corporate bonds, credit spreads, issuance, maturity, liquidity, volatility, tail risk, ratings, guarantees and China. Platform families were publisher sites and DOI/Crossref-style discovery, broad scholarly search, SSRN and NBER working-paper discovery, official Chinese regulatory and market-infrastructure sites, and official IFRS, BIS, European Commission and U.S. SEC sites. Backward checking began from major reviews and highly relevant bond studies. English- and Chinese-language items were eligible.

Academic studies were included when they examined corporate or municipal debt directly, or when a capital-market, disclosure, measurement or intermediary mechanism was portable to bond markets. Institutional items were included only when issued by the responsible authority or standard setter and material to disclosure, ETS scope, bond reporting, liquidity or supervisory design. Items were excluded when they were commentary without an identifiable method or rule, duplicates, unverifiable bibliographic records, purely technical climate-science papers without a finance channel, or studies whose only connection was a generic ESG label. Each included item was coded for market, period, risk dimension, transmission link, method, main finding, friction and policy relevance.

The initial evidence freeze on 12 August contained 26 items. A targeted update and full citation cross-check on 13 August added eleven items. A standards update on 4 September incorporated China's climate-specific disclosure standard issued in December 2025 and the ISSB's targeted IFRS S2 amendments issued in December 2025, producing 40 verified sources: 28 academic and 12 institutional. A targeted refresh on 13 October added three academic studies directly relevant to Chinese bond pricing and disclosure, alongside three authoritative institutional documents covering the Climate Standard application-guide consultation, the 2026 national carbon-market progress report and the 2026 ETS work notice. The coded evidence set therefore contains 49 sources (34 academic and 15 institutional). Three directly relevant studies by the authors are coded under the same inclusion rules: one published article on rating sorting and two SSRN working papers on measurement transport and transaction continuity. The working-paper status is identified in the references; their results are treated as provisional research evidence, not as peer-reviewed findings. The refresh was targeted rather than an exhaustive search of all research published after 4 September. The exploratory first-stage search did not preserve search-result hit counts, so the review does not reconstruct a retrospective PRISMA funnel. This is a structured, non-exhaustive critical review rather than a systematic review claiming complete retrieval. That boundary limits conclusions about literature prevalence but does not prevent transparent mechanism coding. Table 1 summarizes the coding dimensions, and Appendix A reports the review update trail and source-count changes.

The synthesis follows three rules. First, positive and null estimates are not counted as votes. Second, causal language is retained only when the design supports it; otherwise findings are described as associations. Third, official dates are coded according to the information set available to market participants. For example, the 2025 national carbon-market expansion is dated to its public release on 26 March rather than the document's internal signature date. The same discipline applies to annual-report announcement dates, ETS control years and compliance deadlines.

Verified means that the DOI or official URL resolves, the title and authors or issuing body match the authoritative record, the outlet and date match, and the coded finding has been checked against the abstract or official text. Recent 2024–2026 items received the same checks. Verification is bibliographic and interpretive quality control; it is not an endorsement of every empirical design or result.

Table 1. Evidence-coding framework

Review dimension	Coding question	Reason
Entity	Is exposure measured for the legal obligor, parent, subsidiary or operating asset?	Creditors hold a claim on a legal entity; group-level exposure is not automatically the obligor's exposure.
Time	Was the information public before pricing, rating or trading?	Using subsequently released information creates look-ahead bias.
Horizon	Does the exposure horizon match the bond and financial-planning horizon?	Long-horizon hazards need not affect short-dated debt in the same way.
Outcome	Is the study measuring price, access, size, maturity, covenant, liquidity or tail loss?	Climate risk can move different financing margins in different directions.
Intermediation	Do ratings, guarantees, ownership or underwriters transmit or absorb the signal?	Observed bond terms are reduced-form outcomes of multiple institutions.
Market state	Is the effect estimated in ordinary conditions or stress states?	Liquidity and credit interactions can strengthen during market stress.

3. From risk measurement to transmission adequacy

3.1 Three risk dimensions are not interchangeable

Physical, transition and perceived climate risk describe different objects. Physical metrics estimate exposure to hazards such as heat, precipitation, flooding or sea-level rise. Transition metrics capture policy, technology, carbon cost and demand shifts. Perception measures use disclosure or news to capture what managers and investors discuss. A headquarters weather index, an ETS coverage indicator and a climate-risk text count are therefore not alternative measurements of one latent variable. They can move independently and can enter credit decisions through different horizons and channels.

Granular establishment geocoding shows why geography must be tested rather than assumed: Addoum et al. (2020) combine daily temperature with establishment locations but find no significant average effect on establishment sales or productivity. The study therefore supports measurement at operating locations while also warning against presuming a nonzero average response. Municipal-bond research likewise shows that geographic exposure interacts with maturity and local capacity (Painter, 2020; Goldsmith-Pinkham et al., 2023). Text measures can add firm-level, time-varying information, but they require human and outcome validation and should distinguish physical, regulatory and opportunity exposure (Sautner et al., 2023). Direct legal coverage by an emissions-trading system is narrower than a high-carbon industry label, while a carbon-gap measure can capture broad forward transition pressure beyond current compliance obligations (Dong et al., 2026). More climate language can reveal risk, signal governance, or constitute imprecise ‘cheap talk’; these interpretations cannot be separated by word counts alone (Bingler et al., 2024).

3.2 Pricing is only one transmission margin

Corporate debt adjusts along multiple margins: whether firms issue, how much they raise, the maturity and covenant structure, the choice between bonds and loans, ratings and guarantees, and secondary-market liquidity. Ginglinger and Moreau (2023) connect physical risk to leverage through both borrower demand and lender supply. Lemma et al. (2021) find that climate commitment is associated with longer maturity even when carbon exposure does not have a robust debt-policy association. Goodell et al. (2025) report nonlinear maturity

responses that vary across firms and legal settings. These results caution against reading a stable price coefficient as necessary or sufficient evidence of risk transmission.

The same logic applies within bond markets. Coupon-minus-curve measures, secondary-market spreads and transaction returns are different reduced-form outcomes. Primary pricing reflects bookbuilding, underwriting competition, allocation and cross-subsidization as well as expected loss. Secondary spreads mix credit, liquidity and benchmark components. Kim and Pouget (2026) directly show why the distinction matters: carbon emissions are associated with secondary-market spreads in their setting but not primary spreads, with underwriting reputation conditioning the gap. Contract design can also absorb a risk through shorter maturity, lower volume or stronger protection before a large spread appears. A complete assessment therefore requires a vector of price, quantity, contract and resilience outcomes.

3.3 Transmission adequacy

Transmission adequacy is a process property. It does not require climate risk to increase every yield. Adaptation, credible transition plans, insurance, guarantees or policy support can reduce expected loss. Adequacy instead requires four forms of traceability: the measured exposure belongs to the obligor or is transparently attributed; the information predates the decision; the risk horizon and uncertainty are explicit; and intermediaries document how the information affects residual credit risk. A fifth requirement is observability during stress, when liquidity and refinancing constraints can amplify otherwise modest average effects.

Table 2. Five-stage climate-risk transmission framework

Link	Required function	Core friction	Observable diagnostic	Policy objective
Disclosure	Timely, comparable and financially connected information	Narrative volume is disconnected from exposure and financial effects	Timeliness; financial linkage; comparability	Make information decision-useful, not merely longer
Measurement	Map exposure to the legal obligor and material operating assets	Parent–subsidiary mismatch, registered-address proxies and broad sector labels	Entity, asset and scope coverage	Distinguish direct obligations from value-chain exposure
Pricing and contract	Reflect material risk across price and non-price terms	Selection, benchmark noise and contract substitution	Price, access and contract margins	Use consistent outcome and benchmark definitions
Credit intermediation	Translate information into documented credit judgment	Stale ratings, guarantees and implicit support can mask obligor exposure	Documented credit channel	Explain judgments without mechanical rating penalties
Resilience	Reveal liquidity and loss absorption in stress states	Thin trading and concentrated holders weaken transaction discovery	Liquidity and tail observability	Monitor state-contingent fragility and refinancing stress

4. What the evidence says—and does not say

4.1 Physical risk: exposure precision and financial capacity

Physical-risk evidence is strongest when geography and cash-flow channels are well aligned. Painter (2020) and Goldsmith-Pinkham et al. (2023) show that sea-level exposure can enter municipal yields, but the

effect varies with maturity and local conditions. He et al. (2025), using 30,087 Chinese bond-month observations, associate physical climate risk with wider corporate bond spreads through weaker internal cash flow and tighter external financing. These findings are economically coherent, but they do not validate every geographic proxy. A registered-address measure can combine hazard exposure with regional development, fiscal conditions and credit quality. The policy implication is not to discard physical metrics; it is to record asset coverage, mapping uncertainty and adaptation capacity.

4.2 Transition risk: legal scope, expectations and policy timing

Transition risk can enter markets before a direct carbon cash cost is material. Carbon pricing, technology, demand and future regulation alter expectations about cash flow and default risk. Global equity evidence documents carbon-transition premia (Bolton & Kacperczyk, 2021, 2023), while recent Chinese bond evidence associates carbon-gap and firm-level climate-policy-uncertainty measures with spreads, liquidity and non-price terms (Dong et al., 2026; Zhao et al., 2026). Yet broad transition exposure and direct ETS obligations remain distinct. The latter requires legal-entity matching, covered activity and event-time accuracy. Treating every power producer as fossil generation or using a document signature date before public release creates classification error precisely where the identification is supposed to be strongest. A recent China green-bond study further reports that greater climate-policy uncertainty is associated with lower corporate green-bond issuance costs, alongside stronger enforcement and investor attention (Lai et al., 2026). Because this result concerns labelled green-bond issuance, it should not be generalized to conventional corporate bonds or interpreted as a uniform climate-risk discount.

China's national carbon market illustrates why one policy date is insufficient. Public identification of a covered sector, formal release, the first control year, allowance allocation rules and compliance deadlines convey different information. The 2025 expansion notice for steel, cement and aluminium was publicly released on 26 March, although the document was dated 20 March (Ministry of Ecology and Environment, 2025). A credible bond-market analysis must preserve that difference and must distinguish formal inclusion from quantified stringency. Environmental-violation evidence further shows that concrete compliance events can affect corporate bond costs through refinancing risk (Ma et al., 2022). For governance, the same data lineage should be available to investors rather than reconstructed separately by each researcher or institution.

4.3 Disclosure: salience, comparability and financial connection

Disclosure can reduce uncertainty, reveal bad news, signal governance or amplify risk salience. Liu et al. (2026) find that climate information disclosure strengthens the association between climate risk and Chinese issuance spreads, while high-reputation underwriters mitigate the effect. In contrast, Xiang and Gong (2026) report that a TCFD-based disclosure measure is associated with lower Chinese corporate bond spreads, which they attribute to reduced information asymmetry and credit risk. Ke and Xu (2026), using Chinese listed-company bonds, find a positive association between climate-risk disclosure and spreads, with stronger effects in several investor- and issuer-level conditions. These results are not direct replications: the disclosure constructs, sample periods, bond universes and outcomes differ. Together they indicate that disclosure can make latent risk more salient or make issuer quality more legible; its price effect depends on what is disclosed, how the signal is interpreted and which financing margin is measured. Ben-Amar et al. (2024) show that analyst benefits are concentrated where disclosed physical risks are financially material. Bingler et al. (2024) demonstrate the opposite concern: climate narratives can be imprecise and disconnected from actions and emissions. Corporate green-bond evidence similarly indicates that labels are more credible when external certification and issuer behavior align (Flammer, 2021). These studies reject the simple assumption that more disclosure automatically lowers financing costs.

Chinese evidence using annual-report text also associates firm-level climate-risk exposure with wider corporate bond spreads, illustrating the importance of reproducible text definitions and issuer matching (Cang & Li, 2024). The IFRS S2 architecture offers a useful benchmark (IFRS Foundation, 2023). It separates physical risks, transition risks, opportunities, emissions, capital deployment, internal carbon prices and governance-linked remuneration, and it asks for current and anticipated effects on financial position, performance and cash flows. Subsequent ISSB educational material explains proportionality relief when quantitative effects are not separately identifiable or cannot be estimated without undue cost or effort (IFRS Foundation, 2025). The design lesson is to improve financial connectivity and comparability while retaining a transparent uncertainty channel—not to require spurious precision.

4.4 Liquidity and stress: why average pricing may be incomplete

Corporate bond liquidity is an independent transmission channel. Chen and Jiang (2021) find low liquidity in China but a relatively small estimated liquidity share of spreads in a pre-screened issuer market. Other Chinese evidence finds that multidimensional liquidity and credit-risk interactions are priced and strengthen during crisis conditions (Wu et al., 2022). The apparent tension is informative: low trading frequency, evaluated prices and institutional screening can weaken the relationship between transaction-based liquidity and average spread levels even while liquidity-credit amplification emerges under stress.

Climate-news innovations can also be priced as hedging states rather than as a uniform average spread premium (Huynh & Xia, 2021). Official market statistics reinforce the broader resilience diagnosis. For nonfinancial-enterprise debt-financing instruments in the interbank market at the end of November 2023—not the full universe of exchange and interbank credit bonds—the median number of holders per instrument was 12; 88% of instruments had no more than 20 holders. The top 50 investors held 50.2% of the volume and accounted for 63.8% of trading (People’s Bank of China, 2023). These aggregate statistics are not a ready-made instrument-level dashboard and do not map one-for-one to every proposed field. They establish the relevance of holder concentration and sparse trading, which a dashboard would combine with evaluated yields, valuation-provider dispersion, zero-trading days, turnover, refinancing and collateral indicators.

5. Five transmission frictions

5.1 Legal-entity and asset-location friction

The first friction arises because climate exposure follows physical assets and regulated activities, whereas creditors hold claims on legal entities. Listed-company data are often assigned to bonds issued by parents, subsidiaries or finance vehicles. Group affiliation is economically relevant but does not erase legal priority, guarantees or ring-fencing. A defensible system must preserve the legal obligor, map controlled operating assets and record how guarantees transfer risk. For physical hazards, it should report what share of material assets is geographically covered. For transition risk, it should distinguish direct regulated installations from value-chain and sector exposure. A recent China-bond measurement audit shows why parent or city proxies need an explicit bridge to the legal issuer and observed bond outcome (Yang, 2026a).

5.2 Information-date and horizon friction

A report can be accurate and still be unusable for a prior financing decision. Bond analysis requires the announcement date, reporting period and availability of the exact document used. Policy analysis additionally requires public-release dates, control years and compliance deadlines. Horizon alignment is equally important. A five-year note may not respond to a distant hazard in the same way as a 20-year obligation, and a near-term compliance cost differs from a long-run technology transition. Disclosures should therefore distinguish short-, medium- and long-term financial effects and state the assumptions linking scenarios to cash flows.

5.3 Measurement comparability and uncertainty friction

Climate indicators combine observed quantities, modeled hazards, text, ratings and policy classifications. Comparability fails when units, boundaries, normalization and missing-data treatment are opaque. A practical taxonomy should record metric owner, methodology version, legal scope, asset coverage, reporting date, estimation uncertainty and whether a value is measured, estimated or imputed. Multiple metrics can be appropriate—as recognized by the Basel Committee’s disclosure framework (Basel Committee on Banking Supervision, 2025)—but users need to know their strengths and limitations. Model disagreement should be retained as information rather than averaged into a falsely precise score.

5.4 Credit-intermediation friction

Ratings, guarantees, ownership and underwriting determine how an exposure reaches bond terms. A strong guarantor can absorb obligor-specific climate risk; state ownership can alter expected support; a rating can lag new information; and underwriters can validate or discount disclosures. Institutional investors themselves report that climate risks are financially relevant but often prefer engagement and risk management to immediate divestment (Krueger et al., 2020). These mechanisms are not nuisance controls. They are transmission institutions. Policy should require a documented explanation when climate information is financially material, including the horizon, mitigants and residual uncertainty. It should not prescribe a rating downgrade or a spread direction, which would replace credit judgment with administrative pricing. Related evidence from China documents issuer-grade compression and debt-grade sorting, indicating that rating categories can preserve public-signal differentiation even when issuer grades cluster near the top (Yang, 2026c).

5.5 Thin-trading and tail-observability friction

When holder bases are concentrated and trades are sparse, an average observed yield may change slowly even as the market becomes fragile. A recent matched-event study of China’s exchange-traded corporate bonds finds no common deterioration in transaction continuity after localized extreme-heat and extreme-rain events in its selected sample, while showing that zero-trading days, price age and resumption time capture different aspects of market function (Yang, 2026b). Climate-news, abnormal-weather and disaster evidence shows that salience and local experience can shift attention and portfolios, making unconditional averages incomplete (Alok et al., 2020; Choi et al., 2020; Engle et al., 2020). Risk can appear as valuation disagreement, lower turnover, a jump in evaluated yields, withdrawal of bids, shorter refinancing horizons or increased reliance on guarantees. Tail monitoring therefore needs a state-contingent design. Normal-period regressions should be complemented by event windows, local projections, quantile models or scenario exercises when adequate data exist. When coverage is inadequate, the correct conclusion is that the effect is not measurable with the current infrastructure—not that the risk is absent.

6. Institutional diagnosis for China

China-specific empirical evidence has expanded but remains thinner than the global evidence base. The diagnosis in this section therefore combines a limited set of China-focused bond studies with official disclosure, ETS and market-structure records. It should be read as an evidence-informed institutional assessment, not as a claim that every proposed friction has already been separately identified in Chinese bond data.

6.1 Disclosure coverage is advancing faster than cross-market linkage

China has substantially strengthened listed-company sustainability reporting. The Shanghai Stock Exchange issued trial sustainability-report guidelines in April 2024 and implementation guidance in January

2025 (Shanghai Stock Exchange, 2024, 2025). Debt-market periodic reporting is also highly timely: NAFMII reported 99.80% on-time annual-report disclosure and 99.65% on-time first-quarter disclosure among covered entities for the relevant 2025 cycle (National Association of Financial Market Institutional Investors, 2025). These are important compliance achievements. Yet timeliness and volume do not guarantee that climate metrics are linked to the bond's legal obligor, financial effects, guarantee structure or maturity horizon.

The national framework has now moved beyond a generic ESG architecture. China issued Sustainability Disclosure Standards for Business Enterprises No. 1—Climate (Trial) in December 2025, following the 2024 Basic Standard (Ministry of Finance et al., 2025). The new standard creates a common climate-reporting reference, but it does not by itself solve bond-level legal-entity linkage or establish universal instrument-level disclosure. The equity and debt disclosure systems still begin from different regulatory objects. A listed parent may issue some debt directly while subsidiaries or affiliated entities issue other instruments. A climate report at the listed-group level can be valuable, but a bond investor also needs the obligor's position, upstream support and guarantee. Cross-market identifiers and consolidation bridges are therefore as important as another narrative section.

6.2 ETS expansion raises the value of legal-scope and event-time infrastructure

The national ETS expanded from fossil-fuel power generation to steel, cement and aluminium management in 2025. The reform creates a natural demand for issuer–emitter–bond linkage, but simplistic classifications are hazardous. Non-fossil generators should not be treated as covered fossil-power entities solely because their industry code says power. Conglomerates can have covered and uncovered activities. Policy documents also distinguish public release, control year, allocation and compliance. A common, versioned scope table would reduce duplicated manual classification and make both investment and policy evaluation more credible.

6.3 Market concentration makes resilience data a first-order issue

The interbank market is large, but size does not imply continuous instrument-level trading. The holder statistics cited above show that many nonfinancial debt instruments are owned by a small number of institutions. This structure can support relationship-based monitoring and stable holdings, yet it can also delay public price discovery and concentrate liquidity pressure. A climate-credit monitoring framework should therefore combine market prices with valuation and ownership diagnostics rather than importing a transaction-only template from a more liquid market.

7. Policy architecture and sequencing

The policy architecture rests on a simple rule: build traceability before attempting to infer or influence prices. This is now an implementation problem rather than a purely prospective disclosure proposal: China has a national climate standard, while the bond market still lacks a consistently observable bridge from that standard to the obligor, guarantee structure, maturity and secondary-market state. Each phase has a burden safeguard, an implementation indicator, a market outcome and a failure interpretation. This converts policy support from a list of desirable measures into an evaluable sequence.

Table 3. Policy sequence and evaluation design

Phase	Action	Burden safeguard	Implementation indicator	Outcome to evaluate
0. Baseline	Publish definitions, entity maps, public event dates and current coverage gaps	Use existing identifiers and filings before requesting new data	Share of live bonds with validated obligor and issuer–emitter linkage	Reduction in unmatched and ambiguous observations

Phase	Action	Burden safeguard	Implementation indicator	Outcome to evaluate
1. Disclosure	Add a bond-relevant climate schedule to covered reports	Phased thresholds and qualitative relief when quantitative effects are not supportable	Timeliness, metric completeness, assurance and financial-effect coverage	Analyst disagreement, issuance dispersion and extraction cost
2. Intermediation	Require ratings, guarantors and underwriters to explain material climate inputs	Explain judgments; do not mandate rating changes or a pricing direction	Reports documenting horizon, mitigants and residual uncertainty	Rating outlook, guarantee reliance and underwriting allocation
3. Stress monitoring	Construct a liquidity and climate-credit dashboard	Aggregate or anonymize sensitive holder data	Coverage of trades, evaluated yields, zero-trade days, valuation dispersion and concentration	Jump risk, tail spread widening and refinancing stress
4. Pilot evaluation	Pilot in exposed sectors with transparent comparison groups	Pre-register outcomes and review dates	Completeness, pre-trend balance and treatment support	Price, contract, access and stress-state outcomes
5. Scale	Scale modules that improve traceability or resilience	Proportionality by issuer size and measurement capability	Cost per usable field, error and restatement rates, market use	Lower information friction without impaired financing access

Note: A failure to observe a price response is not by itself a failed policy outcome. Each phase is evaluated first on whether it improves traceability, comparability or stress observability.

7.1 Module 1: issuer–emitter–bond linkage

The foundational module is a versioned crosswalk connecting bond code, legal obligor, listed parent, guarantor, controlled operating entities and regulated emissions installations. The crosswalk should record effective dates and confidence levels and should preserve rejected matches. For physical risk, an asset-location coverage field should state the proportion of material operations represented. For transition risk, direct legal coverage, broader value-chain exposure and scenario sensitivity should be separate fields. This infrastructure supports investors, rating agencies, researchers and supervisors without prescribing any valuation.

7.2 Module 2: financially connected climate disclosure

A bond-relevant climate schedule should focus on information that can enter credit analysis: material assets and locations; direct and indirect emissions boundaries; current obligations; short-, medium- and long-term risks; capital expenditure and adaptation; insurance and contractual protection; and current or anticipated effects on cash flow, assets, liabilities and financing. China's Climate Standard now provides a domestic implementation reference alongside IFRS S2 (International Sustainability Standards Board, 2025). The two standards make financial effects, scenario assumptions, governance, metrics and targets more central, but they also preserve the practical problem of measurement uncertainty and scope. IFRS S2's 2025 targeted amendments further illustrate that comparability must be balanced with implementation relief. Domestic bond-market implementation should therefore translate the Climate Standard into legal-obligor, maturity and refinancing fields rather than simply reproduce a longer sustainability narrative.

7.3 Module 3: transparent credit intermediation

Rating, guarantee and underwriting institutions should disclose how material climate information enters their process, not disclose proprietary model weights. A concise explanation can identify the horizon

considered, the exposure or scenario, mitigants, the role of a guarantor and the residual uncertainty. For a guaranteed bond, the analysis should distinguish obligor exposure from guarantor capacity and correlation. For state-owned issuers, analysts should separate operating risk from assumptions about support. Supervisory review should test consistency, not require a predetermined climate penalty.

7.4 Module 4: climate-credit stress dashboard

The dashboard should combine daily or monthly evaluated yields, transaction yields, zero-trading indicators, turnover, valuation-provider dispersion, holder concentration, rating outlooks, guarantee dependence, refinancing calendars and climate-policy or hazard events. The People's Bank of China statistics above provide an aggregate motivation for holder and trading indicators, not an instrument-level substitute for the dashboard. Aggregate views can be published while granular holder data remain protected. The objective is to detect when an exposure that is quiet in average pricing becomes material through valuation disagreement, liquidity withdrawal or refinancing concentration. Scenario exercises can compare ordinary conditions with policy-stringency, hazard and forced-sale states.

7.5 Module 5: pre-specified pilots and proportional scaling

Policy pilots should freeze exposure definitions, information dates, outcomes and minimum sample-support thresholds before analysis. The registration record should be publicly time-stamped and administered by a debt-market self-regulatory body or an independent academic repository that is institutionally separate from the evaluation outcome. Price, contract, access and stress outcomes should all be reported; null and adverse results should not trigger post hoc redesign. Scaling decisions should consider the cost per usable field, error rates, restatements, market use and financing-access effects. If a field produces compliance volume but little comparability, it should be redesigned. If a pilot lacks adequate treated entities or market observations, the correct response is to improve data or change the evaluation horizon, not to broaden treatment after seeing results.

7.6 Implementation cost, infrastructure reuse and international lessons

The architecture is not costless, and this article does not claim a numeric budget without an audited implementation study. Its sequencing nevertheless separates cost types. Module 1 requires the largest upfront coordination effort because legal entities, installations, bonds and effective dates must be reconciled, but it should impose little recurrent issuer burden when existing identifiers and official filings are reused. Module 2 creates incremental reporting and assurance costs and therefore needs phased thresholds and proportionality. Modules 3 and 4 mainly require standardized explanations and secure linkage of data already held by market participants or infrastructures; sensitive holder information should remain protected and only aggregates should be published. Pilot budgets should report staff time, vendor expense, usable fields, match errors, restatements and extraction-cost savings before scale-up.

International experience cautions against treating maximal disclosure as synonymous with durable implementation. The EU's CSRD and ESRS architecture advances connected risk-and-impact reporting, while the stop-the-clock directive, the December 2025 political agreement on simplification, and July 2026 delegated acts illustrate burden, timing and value-chain constraints (European Commission, 2026). The U.S. SEC's 2024 climate rule connected material risks to strategy, operations and financial condition. The Commission stayed the rule in April 2024 and voted in March 2025 to end its defense in the consolidated litigation, underscoring legal durability and implementation uncertainty (U.S. Securities and Exchange Commission, 2024, 2025). IFRS S2 offers a more interoperable baseline with proportionality relief. For China, the lesson is to preserve comparability and financial connectivity while phasing scope, reusing infrastructure, documenting cost and embedding review points.

8. Research and evaluation agenda

The framework generates testable questions rather than a single prediction. First, does legal-entity matching change estimated climate coefficients relative to parent-level or sector-level assignment? Second, do report-date and public-release-date alignment reduce apparent anticipation? Third, does financially connected disclosure reduce analyst or valuation disagreement even when mean spreads do not fall? Fourth, do guarantees and ratings absorb climate signals in normal periods but transmit them through outlooks, covenants or refinancing terms under stress? Fifth, does a combined transaction-and-valuation dashboard predict liquidity withdrawal better than turnover alone?

Empirical work should impose feasibility gates. A secondary-market tail study, for example, needs sufficient issuers with usable daily histories, enough genuine exposure changes, high coverage and nondegenerate outcomes. If these conditions are not met, sophisticated tail models cannot repair the design. Researchers should publish coverage tables, legal-entity quality controls, event-date lineage and minimum detectable effects. Where causal identification is weak, descriptive stress mapping can still be useful if its limits are explicit.

Evaluation should also distinguish process success from market outcomes. An identifier map can succeed by reducing ambiguous matches even before yields change. A disclosure module can succeed by increasing financial-effect coverage and reducing extraction costs without lowering borrowing costs. An intermediation module can succeed by producing consistent, auditable credit narratives. A stress dashboard can succeed by detecting valuation disagreement and refinancing concentrations. Only after these process links are operating should policymakers interpret the direction and magnitude of market-price responses.

9. Conclusion

The climate-bond debate is often framed too narrowly: either markets price climate risk or they do not. The evidence supports a more useful question. Climate signals pass through a chain of disclosure, measurement, financing choices, credit intermediaries and market liquidity. Each link can transmit, transform, absorb or obscure the signal. Conflicting estimates are therefore not merely an econometric inconvenience; they can reveal where information and institutions are misaligned.

For China, the next step is not to administratively manufacture a climate premium. It is to make material climate risk traceable to the correct obligor and assets, align information with decision dates and horizons, connect disclosure to financial effects, require transparent intermediary judgment, and observe liquidity and refinancing stress in a concentrated market. A sequenced architecture—identifier infrastructure, bond-relevant implementation of the Climate Standard, credit-intermediation transparency, stress monitoring and pre-specified pilots—can improve resilience while preserving proportionality and market judgment.

This approach also changes how policy evidence is read. A weak average price response need not mean that governance is unnecessary; it may indicate missing linkage, measurement or observability. A significant coefficient need not mean that the market is efficient; it may reflect salience, selection or stress. Transmission adequacy provides a stricter and more constructive standard: material risks should be traceable, comparable, documented in credit decisions and visible when they threaten financing capacity or market resilience.

9.1. Policy and evidence update through 13 October 2026

The policy environment has developed since the SSRN version posted in September. On 30 September 2026, the Ministry of Finance published a consultation draft of the application guide for China's Climate Standard No. 1 and nine industry guides covering electricity, steel, coal, oil, fertilizer, aluminium, hydrogen, cement and automobiles. The consultation deadline is 26 October 2026 (Ministry of Finance, 2026). These materials

provide more operational guidance for connecting climate matters to business activities and financial effects, and make sector-specific exposure mapping more feasible. They remain consultation drafts: they do not themselves establish a new mandatory reporting perimeter or effective date. The article's recommendations should therefore be read as a design proposal consistent with emerging guidance, not as a description of binding implementation requirements.

The Ministry of Ecology and Environment published the Progress Report of China's National Carbon Market (2026) on 20 September, following its release at the 15 September China Carbon Market Conference. The report is a current official source for the development of both the national emissions trading system and the voluntary emissions-reduction market. The 2026 ETS work notice also specifies 2025 compliance and allocation steps for power, steel, cement and aluminium-smelting entities, including a 31 December 2026 deadline for surrender of 2025 allowances. These developments strengthen the case for dated, legal-entity and installation-level exposure measures. They do not retroactively change the paper's literature synthesis or establish bond-pricing effects. (Ministry of Ecology and Environment, 2026a, 2026b). For any future bond-level study, the relevant treatment date must be the public release or applicable compliance date, and a sector label should not be treated as proof that a specific issuer or facility is covered.

9.2. Literature update through 13 October 2026

The targeted literature refresh adds three external studies that sharpen the review's central point about heterogeneous transmission. The review also codes three directly relevant studies by the authors, using the same source and relevance checks as for other work. Xiang and Gong (2026) link TCFD-based climate disclosure to lower corporate bond spreads through reduced information asymmetry and credit risk; Ke and Xu (2026) report higher spreads associated with climate-risk disclosure in a Chinese listed-bond sample; and Lai et al. (2026) find lower issuance costs under greater climate-policy uncertainty for Chinese corporate green bonds. The latter is a labelled-bond result, not evidence about the full conventional bond market. The authors' studies add evidence on measurement transport (Yang, 2026a), transaction continuity (Yang, 2026b), and rating compression and sorting (Yang, 2026c). These studies reinforce a conditional account: transparency can improve the quality of issuer information while also increasing the salience of downside exposure, and policy uncertainty can affect labelled issuance through enforcement and investor attention. They do not support a single direction of climate pricing. Their distinct samples, measures and outcomes are why this article synthesizes mechanisms instead of treating estimates as votes.

Yang (2026c) is coded as a companion empirical study in the evidence base, not as evidence generated by this review. The two SSRN papers are identified as working papers in both the text and references. This distinction matters: the updated rules and the companion studies improve context and traceability, but this manuscript remains a structured critical review rather than an original empirical paper.

Appendix A. Review protocol and update trail

Appendix A records the review update trail and the limits of the retrieval protocol. The exploratory first-stage search did not preserve hit counts, so no retrospective PRISMA funnel is reconstructed.

Table A1. Evidence-base update record

Stage	Date	Coverage and action	Included evidence
Initial freeze	12 Aug 2026	Mechanism-led searches, official-source retrieval and backward checking; source identity and coded findings verified	26 items: 18 academic, 8 institutional

Stage	Date	Coverage and action	Included evidence
Revision update	13 Aug 2026	Targeted searches for major reviews, measurement, investor behavior, primary-secondary differences and international implementation	11 added: 9 academic, 2 institutional
Final coded set	13 Aug 2026	Duplicates and unverifiable records excluded; every item assigned review stage, date added and verification basis	37 items: 27 academic, 10 institutional
Final standards update	4 Sep 2026	China Climate Standard No. 1 and ISSB IFRS S2 amendments added; text and policy implications rechecked	2 institutional items added; final coded set: 40 items (28 academic, 12 institutional)
Policy refresh	13 Oct 2026	Targeted verification of China's draft Climate Standard application guide and sector guides, 2026 national carbon-market progress report, and ETS work notice	3 institutional documents added; updated coded set: 43 items (28 academic, 15 institutional)
Literature and policy refresh	13 Oct 2026	Targeted update: three bond studies, three official updates and three related author studies coded	49 sources: 34 academic, 15 institutional

Note: The review is structured and non-exhaustive. It does not claim PRISMA-complete retrieval or estimate literature-wide effect prevalence.

Declarations

Author contributions (CRediT): Fuli Yang: Conceptualization, Methodology, Writing—original draft, and Project administration. Shiyu Lin: Investigation, Validation, and Writing—review and editing. Both authors reviewed and approved the final manuscript and accept responsibility for the work.

Funding: No external funding was received for this article.

Conflict of interest and institutional position: Both authors are affiliated with GARCH Institute. GARCH Institute is an independent research institution, and this article uses only public academic and institutional sources; it does not use proprietary client, transaction or internal institutional data. The authors declare no financial or personal relationships that could have influenced the analysis.

Ethics approval: Not applicable. This structured review uses publicly available academic and institutional sources and involves no human participants, personal data, interventions or identifiable private information.

Data availability: This review generated no new dataset. The source list, screening criteria and review update trail are provided in the article.

Generative AI disclosure: OpenAI ChatGPT (GPT-5 and GPT-6) was used to assist with literature organization, drafting, language revision, consistency checks and document formatting. The authors selected and verified sources, determined the argument and policy recommendations, reviewed and revised all generated material, and assume full responsibility for the article. The October 2026 update added a targeted check of Chinese climate-disclosure implementation guidance and national carbon-market materials.

References

- Addoum, J. M., Ng, D. T., & Ortiz-Bobea, A. (2020). Temperature shocks and establishment sales. *Review of Financial Studies*, 33(3), 1331–1366. <https://doi.org/10.1093/rfs/hhz126>
- Alok, S., Kumar, N., & Wermers, R. (2020). Do fund managers misestimate climatic disaster risk? *Review of Financial Studies*, 33(3), 1146–1183. <https://doi.org/10.1093/rfs/hhz143>
- Basel Committee on Banking Supervision. (2025). A framework for the voluntary disclosure of climate-related financial risks. Bank for International Settlements. <https://www.bis.org/bcbs/publ/d597.htm>
- Ben-Amar, W., Herrera, D. C., & Martinez, I. (2024). Do climate risk disclosures matter to financial analysts? *Journal of Business Finance & Accounting*, 51(7–8), 2153–2180. <https://doi.org/10.1111/jbfa.12778>
- Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2024). How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164, 107191. <https://doi.org/10.1016/j.jbankfin.2024.107191>
- Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics*, 142(2), 517–549. <https://doi.org/10.1016/j.jfineco.2021.05.008>
- Bolton, P., & Kacperczyk, M. (2023). Global pricing of carbon-transition risk. *Journal of Finance*, 78(6), 3677–3754. <https://doi.org/10.1111/jofi.13272>
- Cang, H., & Li, C. (2024). Corporate climate risk and bond credit spreads. *Finance Research Letters*, 67, 105741. <https://doi.org/10.1016/j.frl.2024.105741>
- Chen, Y., & Jiang, L. (2021). Liquidity risk and corporate bond yield spread: Evidence from China. *International Review of Finance*, 21(4), 1117–1151. <https://doi.org/10.1111/irfi.12322>
- Choi, D., Gao, Z., & Jiang, W. (2020). Attention to global warming. *Review of Financial Studies*, 33(3), 1112–1145. <https://doi.org/10.1093/rfs/hhz086>
- Dong, Y., Liu, Y., Wang, T., & Zhang, L. (2026). Climate transition risk and corporate bond pricing. *Emerging Markets Review*, 73, 101493. <https://doi.org/10.1016/j.ememar.2026.101493>
- Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020). Hedging climate change news. *Review of Financial Studies*, 33(3), 1184–1216. <https://doi.org/10.1093/rfs/hhz072>
- European Commission. (2026). Corporate sustainability reporting. Updated 3 July 2026. https://finance.ec.europa.eu/financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en
- Flammer, C. (2021). Corporate green bonds. *Journal of Financial Economics*, 142(2), 499–516. <https://doi.org/10.1016/j.jfineco.2021.01.010>
- Giglio, S., Kelly, B., & Stroebel, J. (2021). Climate finance. *Annual Review of Financial Economics*, 13, 15–36. <https://doi.org/10.1146/annurev-financial-102620-103311>
- Ginglinger, E., & Moreau, Q. (2023). Climate risk and capital structure. *Management Science*, 69(12), 7492–7516. <https://doi.org/10.1287/mnsc.2023.4952>
- Goldsmith-Pinkham, P., Gustafson, M. T., Lewis, R. C., & Schwert, M. (2023). Sea-level rise exposure and municipal bond yields. *Review of Financial Studies*, 36(11), 4588–4635. <https://doi.org/10.1093/rfs/hhad041>
- Goodell, J. W., Palma, A., Paltrinieri, A., & Piserà, S. (2025). Firm-level climate change risk and corporate debt maturity. *Journal of International Money and Finance*, 152, 103275. <https://doi.org/10.1016/j.jimonfin.2025.103275>

- He, F., Ren, X., Wang, Y., & Lei, X. (2025). Climate risk and corporate bond credit spreads. *Journal of International Money and Finance*, 154, 103297. <https://doi.org/10.1016/j.jimonfin.2025.103297>
- Huynh, T. D., & Xia, Y. (2021). Climate change news risk and corporate bond returns. *Journal of Financial and Quantitative Analysis*, 56(6), 1985–2009. <https://doi.org/10.1017/S0022109020000757>
- IFRS Foundation. (2023). ISSB issues inaugural global sustainability disclosure standards. <https://www.ifrs.org/news-and-events/news/2023/06/issb-issues-ifrs-s1-ifrs-s2.html>
- IFRS Foundation. (2025). Disclosing information about anticipated financial effects of sustainability-related risks and opportunities. <https://www.ifrs.org/news-and-events/news/2025/08/disclosing-information-anticipated-financial-effects/>
- International Sustainability Standards Board. (2025). Amendments to greenhouse gas emissions disclosures—Amendments to IFRS S2. <https://www.ifrs.org/projects/completed-projects/2025/amendments-to-disclosure-of-greenhouse-gas-emissions-s2/>
- Kim, D., & Pouget, S. (2026). Do carbon emissions affect the cost of capital? Primary versus secondary corporate bond markets. *Journal of Corporate Finance*, 97, 102932. <https://doi.org/10.1016/j.jcorpfin.2025.102932>
- Ke, H., & Xu, X. (2026). Climate-risk information disclosure and corporate bond credit spreads [in Chinese]. *Finance & Economy*, 2026(2), 1–13, 26. <https://doi.org/10.19622/j.cnki.cn36-1005/f.2026.02.001>
- Krueger, P., Sautner, Z., & Starks, L. T. (2020). The importance of climate risks for institutional investors. *Review of Financial Studies*, 33(3), 1067–1111. <https://doi.org/10.1093/rfs/hhz137>
- Lai, F., Tong, Q., Xu, D., & Chen, L. (2026). The paradox of climate policy uncertainty and financing costs: The policy-driven effect on China's green bond market. *Journal of Emerging Market Finance*. Advance online publication. <https://doi.org/10.1177/09726527261485521>
- Lemma, T. T., Lulseged, A., & Tavakolifar, M. (2021). Corporate commitment to climate change action, carbon risk exposure, and a firm's debt financing policy. *Business Strategy and the Environment*, 30(8), 3919–3936. <https://doi.org/10.1002/bse.2849>
- Liu, D., Sang, Z., Wan, H., & Wang, B. (2026). Does climate information disclosure affect climate risk pricing in bond issuances? *Australian Accounting Review*. <https://doi.org/10.1111/auar.70028>
- Ma, R., Ji, Q., Zhai, P., & Yang, R. (2022). Environmental violations, refinancing risk, and the corporate bond cost in China. *Journal of International Financial Management & Accounting*, 33(3), 480–504. <https://doi.org/10.1111/jifm.12154>
- Ministry of Ecology and Environment of the People's Republic of China. (2025). Notice on incorporating the steel, cement, and aluminium smelting industries into national carbon emissions trading market management. Publicly released 26 March 2025. https://www.mee.gov.cn/xxgk2018/xxgk/xxgk03/202503/t20250326_1104736.html
- Ministry of Ecology and Environment of the People's Republic of China. (2026a). Notice on work arrangements for the 2026 national carbon emissions trading market [in Chinese]. https://www.mee.gov.cn/xxgk2018/xxgk/xxgk06/202602/t20260209_1143900.html
- Ministry of Ecology and Environment of the People's Republic of China. (2026b). Progress report of China's national carbon market (2026). https://www.mee.gov.cn/ywgz/xdqhbh/wsqtz/202609/t20260920_1166890.shtml
- Ministry of Finance of the People's Republic of China, Ministry of Ecology and Environment, Ministry of Foreign Affairs, National Development and Reform Commission, Ministry of Industry and Information Technology, Ministry of Commerce, People's Bank of China, State-owned Assets Supervision and

- Administration Commission, National Financial Regulatory Administration, & China Securities Regulatory Commission. (2025). Sustainability Disclosure Standards for Business Enterprises No. 1—Climate (Trial) (Cai Kuai [2025] No. 34).
https://kjs.mof.gov.cn/zhengcefabu/202512/t20251225_3980202.htm
- Ministry of Finance of the People's Republic of China. (2026). Consultation on the application guide for China Sustainability Disclosure Standard No. 1—Climate (Trial) and nine industry application guides [in Chinese].
https://kjs.mof.gov.cn/gongzuotongzhi/202609/t20260929_3998329.htm
- National Association of Financial Market Institutional Investors. (2025). Periodic-report disclosure compliance for nonfinancial enterprise debt-financing instruments [in Chinese].
https://www.nafmii.org.cn/zlgl/pjywb/pjyw/tb/202505/t20250526_322265.html
- Painter, M. (2020). An inconvenient cost: The effects of climate change on municipal bonds. *Journal of Financial Economics*, 135(2), 468–482. <https://doi.org/10.1016/j.jfineco.2019.06.006>
- People's Bank of China. (2023). Financial market operation in November 2023 [in Chinese].
<https://www.pbc.gov.cn/goutongjiaoliu/113456/113469/2025092212553627845/index.html>
- Sautner, Z., van Lent, L., Vilkov, G., & Zhang, R. (2023). Firm-level climate change exposure. *Journal of Finance*, 78(3), 1449–1498. <https://doi.org/10.1111/jofi.13219>
- Shanghai Stock Exchange. (2024). Guidelines for listed-company sustainability reports (trial) [in Chinese].
https://www.sse.com.cn/aboutus/mediacenter/hotandd/c/c_20240412_10753197.shtml
- Shanghai Stock Exchange. (2025). Sustainability report preparation guidance [in Chinese].
https://star.sse.com.cn/aboutus/mediacenter/hotandd/c/c_20250117_10770282.shtml
- U.S. Securities and Exchange Commission. (2024). The enhancement and standardization of climate-related disclosures for investors. Adopted 6 March 2024; stayed 4 April 2024. <https://www.sec.gov/rules-regulations/2024/03/s7-10-22>
- U.S. Securities and Exchange Commission. (2025). SEC votes to end defense of climate disclosure rules. 27 March 2025. <https://www.sec.gov/newsroom/press-releases/2025-58>
- Wu, J., Cao, J., Liu, D., & Li, S. (2022). Liquidity, credit risk, and their interaction on the spreads in China's corporate bond market. *Discrete Dynamics in Nature and Society*, 2022, 2996704.
<https://doi.org/10.1155/2022/2996704>
- Xiang, Y., & Gong, J. (2026). The value of climate disclosure: Evidence from corporate bond credit spreads. *Finance Research Letters*, 103, 110156. <https://doi.org/10.1016/j.frl.2026.110156>
- Yang, F. (2026a). From city proxies to legal entities: How measurement choices change climate-risk inference in Chinese corporate bonds [Working paper]. SSRN. <https://ssrn.com/abstract=7301319>
- Yang, F. (2026b). Trading through the shock: Physical climate events and transaction continuity in China's corporate bond market [Working paper]. SSRN. <https://ssrn.com/abstract=7299058>
- Yang, F. (2026c). Where public climate signals remain visible: Issuer compression and debt-rating sorting in China. *Journal of Financial Reporting and Accounting*. <https://doi.org/10.1108/JFRA-08-2026-0832>
- Zhao, Y., Nagayasu, J., Tian, X., & Geng, Y. (2026). Corporate bond pricing under climate policy uncertainty: Evidence from firm-level exposure. *Structural Change and Economic Dynamics*, 79, 459–481.
<https://doi.org/10.1016/j.strueco.2026.06.003>

Trust as Resolution Infrastructure

Value Preservation, Creditor Coordination, and Bankruptcy Service Trusts in China

Fuli Yang

GARCH Institute

Cambridge, MA, USA

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

Source manuscript v1.0 dated 18 August 2026

Author independence statement: This paper expresses the author's independent academic views and does not represent any institution or transaction participant.

JEL Classification: G23, G28, G33, K22

Keywords: bankruptcy service trust; corporate restructuring; creditor recovery; fire-sale discount; value preservation

Abstract

Bankruptcy service trusts have become an increasingly common component of large corporate restructurings in China. Among the 144 listed companies whose reorganization plans were approved by courts between 2007 and 2025, the share of plans repaying creditors with trust beneficial interests rose from 4 percent over 2019–2022 to 36 percent in 2024 and 40 percent in 2025. Yet the economic contribution of these trusts is easily misstated: restructuring plans may treat trust units as debt satisfaction while underlying assets remain unmonetized, whereas evaluations based only on immediate cash ignore the value of avoiding forced sales. This paper first develops a two-period model of deferred disposition in which the fire-sale discount, expected asset recoverability, and governance-dependent agency costs jointly determine when a service trust preserves value; four hypotheses follow as comparative statics, including a net-effect boundary for asset complexity. It then combines a census-based sampling frame compiled from public plan records and professional annual reports, a source-graded audit of 20 cases, a focal positive mechanism analysis of the Founder Group restructuring, and a structured comparison of the 2024 listed-company cohort. Of the 20 audited cases, 12 have a verified plan, seven a verified establishment, and only two provide case-level monetary evidence meeting the audit threshold: Bohai Steel reports cumulative cash distributions, while HNA Group's trust units are reported at a Level 3 fair-value carrying amount in affiliated companies' audited financial statements. Because realized trust distributions remain sparse, the paper specifies a package-level market-implied recovery measure and an exploratory roadmap of three prospective research designs built on the 2023 trust-business classification reform, the 2024 Supreme People's Court and CSRC Minutes, and within-case creditor-class variation, stating explicitly the conditions under which each would support causal inference. The evidence to date supports an institutional conclusion: where viable operations can be separated from difficult but potentially recoverable assets, a service trust preserves contingent creditor rights and allows the core restructuring to close without a forced sale. Whether it raises discounted recovery is the empirical hypothesis this design makes testable and auditable.

1. Introduction

Corporate restructuring faces a basic timing problem. Courts and creditors need a plan that can be approved and executed within a bounded period, but distressed assets often require more time to verify, disentangle, operate, or sell. A rapid disposition may close the case while destroying value. A prolonged proceeding may preserve optionality but impose administrative costs, uncertainty, and continued creditor conflict. Bankruptcy service trusts have emerged in China as an institutional response to this mismatch between the legal clock and the economic clock.

The mechanism is straightforward in form but ambiguous in economic effect. A debtor or restructuring platform places designated assets or equity interests into a trust. Creditors receive trust units, and the trustee and beneficiary governance bodies administer the assets beyond the formal restructuring period. This arrangement can separate viable operations from historically encumbered assets, prevent fragmented enforcement, and retain the upside from later recovery. It can also postpone loss recognition: a trust unit may legally satisfy a claim even when its eventual cash value is uncertain or negligible. The same structure can therefore represent either resolution infrastructure or deferred disappointment.

The phenomenon is no longer marginal. Regulators formally recognized risk-disposal service trusts, including enterprise bankruptcy service trusts, in the trust-business classification reform effective June 2023 (CBIRC, 2023). More than thirty trust companies now conduct this business, with an outstanding scale reported above RMB 1.5 trillion (Sina Finance, 2025). Among listed companies, 144 court-approved reorganization plans have accumulated since the Enterprise Bankruptcy Law took effect in 2007; only two of the 49 plans approved during 2019–2022 used trust beneficial interests to repay creditors, but adoption appeared in five of 15 plans in 2023 — a year straddling the June 2023 reform — four of 11 plans in 2024, and six of 15 in 2025 (AllBright Law Offices, 2025, 2026). Adoption is accelerating precisely as the Supreme People’s Court and the China Securities Regulatory Commission tighten standards for plan feasibility and execution (SPC & CSRC, 2024; CSRC, 2025).

This paper asks: when do bankruptcy service trusts preserve recoverable value, rather than merely delay the recognition of losses? The question matters for three reasons. First, the risk-disposal service-trust market is already large enough to affect the architecture of major restructurings. Second, the conventional plan-level recovery rate is not an adequate performance measure when it combines cash, equity, retained debt, and trust units at face or assessed values. Third, China remains short of micro-level evidence on bankruptcy recovery and post-plan execution, a gap emphasized by Hotchkiss et al. (2023).

The paper makes four contributions. First, it develops a simple two-period model of deferred disposition that nests the two rival interpretations — resolution infrastructure and deferred disappointment — as regions of the same parameter space, and derives four hypotheses as comparative statics. Second, it defines a case-level evidence chain, anchored in a census of court-approved listed-company reorganizations, that prevents proposed trusts, established trusts, and distributing trusts from being treated as equivalent observations. Third, it develops the Founder Group restructuring as a focal positive mechanism case and compares its logic with recent listed-company restructurings. Founder is not used to prove that trust distributions were high. It is used to show how a trust can allow a complex restructuring to close while preserving residual asset claims that would otherwise be abandoned or sold under pressure. Fourth, it specifies a measurement and research-design strategy — package-level market-implied recovery from defaulted-bond and debt-repayment share prices, combined with an exploratory roadmap of institutional variation from the 2023 classification reform, the 2024 Minutes, and within-case creditor-class assignment — that converts the framework from a research agenda into a testable design whose identifying assumptions are stated, and flagged where they are not yet satisfied.

The paper advances a bounded positive interpretation. Service trusts cannot repair an asset that lacks economic value, and delay can conceal weak execution. Their plausible contribution is greatest when rapid sale would

create a material discount, the assets retain credible operating or recovery value, and governance rules make post-plan administration accountable. The model makes this boundary explicit, and the empirical design makes it testable.

2. Literature and analytical foundations

2.1 Liquidation value and forced-sale discounts

Shleifer and Vishny (1992) show that liquidation value depends on who can buy an asset when its owner is distressed. Natural buyers are often firms in the same industry, yet those buyers may also be financially constrained during a sector downturn. Pulvino (1998, 1999) identifies distressed-sale discounts in commercial aircraft transactions, while Maksimovic and Phillips (1998), Strömberg (2000), and Bernstein, Colonnelli, and Iverson (2019b) show that bankruptcy reallocates assets across owners and uses rather than merely converts them into cash. The service-trust rationale is therefore strongest when a thin market, buyer constraints, title defects, or incomplete asset verification prevent competitive disposition at the plan date.

Delay nevertheless has two sides. Andrade and Kaplan (1998) document the economic costs associated with severe distress, and Bris, Welch, and Zhu (2006) show that procedure, duration, and cost vary materially across bankruptcy cases. Benmelech and Bergman (2011) further demonstrate that bankruptcy can depress collateral values outside the debtor itself. A trust creates an option to wait, but the option is valuable only when the expected improvement in net realizable value exceeds financing, administration, deterioration, and agency costs. The relevant benchmark is discounted net distribution from disciplined disposition, not face-value trust units versus immediate cash.

2.2 Creditor coordination and governance

Distress also creates collective-action and control problems. Creditors differ in priority, information, liquidity needs, and litigation strategy. Ayotte and Morrison (2009) find pervasive creditor control in Chapter 11, while Ivashina, Iverson, and Smith (2016) show that claim ownership and trading reshape creditor concentration. Iverson (2018) demonstrates that institutional congestion changes restructuring outcomes. A service trust can reduce fragmentation by transforming bilateral enforcement claims into beneficial interests in a common vehicle, but it cannot remove distributive conflict; it relocates that conflict into the vehicle's governance rules.

Post-reorganization evidence reinforces the need for accountability. Hotchkiss (1995) links governance turnover to post-bankruptcy performance, and Hotchkiss and Mooradian (1997) show that specialized distressed investors can contribute monitoring and control. Bankruptcy can also create local spillovers (Bernstein, Colonnelli, Giroud, & Iverson, 2019a). Beneficiaries therefore require verifiable asset schedules, valuation bases, trustee and manager duties, voting and removal rights, periodic cash-flow reporting, distribution rules, and a credible end date. Without these controls, the same structure that centralizes recovery may become an opaque warehouse for difficult assets.

2.3 Bankruptcy resolution in China

China's formal bankruptcy system has expanded alongside corporate debt and bond defaults, but implementation remains uneven. Jiang (2014) and E. Lee (2011) identify judicial discretion, government involvement, and limited institutional familiarity as constraints on reorganization practice. Using case-level evidence, Li and Ponticelli (2022) find that specialized bankruptcy courts shorten case duration and affect insolvency outcomes; Hotchkiss et al. (2023) synthesize the continuing enforcement frictions and scarcity of detailed recovery data. Comparative evidence also shows that court capacity and the design of reorganization and liquidation rules affect finance and investment (Ponticelli & Alencar, 2016; Rodano, Serrano-Velarde, &

Tarantino, 2016). Service trusts should thus be viewed as complements to adjudication: a post-plan organizational form for assets that cannot be fully disposed of inside the judicial timetable, not a substitute for court quality or legal enforcement.

2.4 Chinese trust law: property segregation and fiduciary accountability

The legal case for service trusts rests on asset partitioning as well as contractual flexibility. Organizational-law scholarship treats the separation of an asset pool from the personal creditors of its controllers as a distinct legal technology (Hansmann & Kraakman, 2000). Articles 15–18 of China’s Trust Law distinguish trust property from the settlor’s and trustee’s other property and restrict enforcement against the trust fund, while Articles 25–29 impose duties of loyalty, prudence, and effective management and require the trustee to manage trust property separately from its own property and from other trusts. Yet the transplantation of the trust into a civil-law setting produced continuing uncertainty over title, beneficiary rights, and the relationship between contract and property (Tan, 2001; Qu, 2003; R. Lee, 2009; Foster, 2010).

Accordingly, “trust property independence” should not be treated as an automatic label. Valid transfer, registration, separate management, third-party effects, and statutory exceptions must be established transaction by transaction (Ho, 2003; Ho, Lee, & Jin, 2013; Lyu, 2015; Jing, 2024). The agency-cost account of trust law also makes clear that segregation without fiduciary accountability is incomplete (Sitkoff, 2004). For bankruptcy service trusts, creditor value depends on loyalty, prudence, reporting, beneficiary remedies, and continuing oversight after plan completion (J. Lee, 2021; Tian, 2021; Dong, 2025). Governance and disclosure are therefore components of the value-preservation mechanism, not procedural decoration.

3. Institutional setting

3.1 Regulatory recognition and market scale

China’s trust-business classification reform, effective June 1, 2023, formally recognizes risk-disposal service trusts as one of five categories of asset service trusts, and distinguishes market-oriented restructuring trusts from enterprise bankruptcy service trusts established for debtors undergoing court-supervised reorganization, composition, or liquidation (CBIRC, 2023). In a bankruptcy service trust, the trust company serves as trustee for assets designated under a court-approved plan, while creditors become beneficiaries through allocated trust units. The trust may hold equity in an asset platform, residual claims, receivables, real estate, litigation rights, or interests in operating subsidiaries. By early 2025, more than thirty trust companies conducted risk-disposal service-trust business with a reported outstanding scale above RMB 1.5 trillion (Sina Finance, 2025).

Two regulatory developments frame the sample period. The classification reform is a supply-side event: it legitimized and standardized the product, lowering the institutional cost of including a trust in a plan. The Minutes of the work symposium on the adjudication of listed-company bankruptcy reorganization cases, issued jointly by the Supreme People’s Court and the CSRC on December 31, 2024 (Fa [2024] No. 309), together with the CSRC’s companion Guideline No. 11, are a demand-side event: they tighten requirements for plan feasibility, disclosure, and execution standards, including explicit completion criteria and extended supervision when plans are not fully executed (SPC & CSRC, 2024; CSRC, 2025). Both events inform the research designs in Section 5.3.

3.2 Adoption among listed companies

Court-approved reorganization plans of listed companies provide a census-based sampling frame compiled from public plan records and professional annual reports; the author’s independent case-by-case re-verification of the full list is part of the census extension described in Section 5. From the Enterprise Bankruptcy Law’s

entry into force in 2007 through the end of 2025, 144 listed companies had reorganization plans approved by courts, including six in 2019, 13 in 2020, 19 in 2021, 11 in 2022, 15 in 2023, 11 in 2024, and 15 in 2025 (AllBright Law Offices, 2026). Trust beneficial interests as a repayment instrument were rare before the June 2023 classification reform — two of the 49 plans approved during 2019–2022 — but appeared in five of 15 plans in 2023, a year that straddles the reform’s June 1 effective date, four of 11 in 2024, and six of 15 in 2025 (AllBright Law Offices, 2025, 2026). Relative to the 2019–2022 pre-reform baseline of roughly 4 percent, adoption among approved plans has risen by an order of magnitude, which both motivates the research question and supplies the variation the research designs examine. Because 2023 mixes pre- and post-reform approvals, all reform-based comparisons assign cases by each plan’s approval date rather than by calendar year.

3.3 Service-trust forms

Table 1 summarizes the main structural forms observed in practice.

Table 1. Service-trust forms, functions, and principal risks

Form	Typical asset position	Primary economic function	Principal risk
Sale-style	Non-core or difficult assets left after viable assets are sold	Separate residual assets and preserve supplementary recovery	Residual pool may have little realizable value
Going-concern	Operating businesses retained within the trust structure	Maintain operations and allow later strategic disposition	Operating losses and governance complexity
Asset-holding	Specific properties, equity stakes, receivables, or claims	Avoid forced-sale discount and complete title or recovery work	Delay, carrying costs, and weak liquidity
Hybrid	Cash, equity, retained debt, and trust units combined	Balance creditor liquidity and debtor feasibility	Nominal recovery may obscure economic recovery

The forms share one feature: the trust separates the formal completion of the restructuring plan from the economic completion of asset recovery. That separation is the source of both its value and its measurement problem.

4. Model, evaluation framework, and hypotheses

4.1 A two-period model of deferred disposition

A parsimonious two-period comparison makes the boundary between resolution infrastructure and deferred disappointment explicit. The model is in the spirit of the investment-timing literature (McDonald & Siegel, 1986; Dixit & Pindyck, 1994) but is deliberately not a full real-option model: deferral leads to disposition with certainty one period later, and no post-deferral stopping decision is modeled. This keeps the mechanism transparent; an optimal-stopping extension in which the trustee observes the realized state before choosing between sale and continued holding is left for the econometric companion. At the plan date, a reorganizing debtor holds a residual asset pool with fundamental value A and retained operations with going-concern value V . Immediate disposition of the pool realizes $(1 - \delta)A$, where $\delta \in [0, 1]$ is a fire-sale discount that increases with buyer financial constraints, market thinness, unresolved title defects, and incomplete verification (Shleifer & Vishny, 1992; Pulvino, 1998).

A service trust defers disposition by one period. The pool then realizes $\tilde{\theta}A$, where $\tilde{\theta}$ has mean $\hat{\theta} \leq 1$ reflecting the probability that title cure, litigation recovery, and orderly marketing restore value, net of deterioration.

Deferral costs $\kappa(g)A$, comprising financing, administration, and agency costs, where g indexes governance and disclosure quality with $\kappa'(g) < 0$: stronger beneficiary voting, reporting, and termination rules reduce the agency component (Sitkoff, 2004). Creditors discount delayed, illiquid recoveries at rate r . In addition, immediate disposition is not instantaneous: it must be executed inside the reorganized perimeter over a disposition window of marketing, auction, and title-transfer procedures. During that window the unresolved pool imposes a one-time contamination cost cV on retained operations — blocked transfers, continued enforcement, and impaired investor entry — so c measures the transaction loss that forced in-perimeter disposition inflicts on going-concern value. The trust avoids this loss by moving the pool outside the perimeter at the plan date, before disposition work begins. Retaining the pool within the reorganized entity constitutes a third alternative whose ranking depends jointly on the duration of contamination, expected recovery, and governance costs; it is therefore left to an extension of the model, and the analysis proceeds on the binary comparison.

The two complete payoffs can now be stated. Under immediate disposition, the estate realizes the fire-sale proceeds while retained operations bear the contamination cost:

$$W_I = (1 - c)V + (1 - \delta)A.$$

Under the trust, retained operations are immediately freed of contamination, and the pool is realized one period later net of deferral costs:

$$W_T = V + \frac{[\dot{\theta} - \kappa(g)]A}{1 + r}.$$

The trust weakly dominates immediate disposition, $W_T \geq W_I$, if and only if

$$\frac{\dot{\theta} - \kappa(g)}{1 + r} + \frac{cV}{A} \geq 1 - \delta. (1)$$

The term cV/A thus has a concrete timing interpretation: it is the contamination cost avoided at the plan date, expressed per yuan of residual assets, and it enters undiscounted because the separation benefit accrues immediately while the recovery payoff arrives one period later.

A final clarification concerns whose value W measures. W_I and W_T are the total value available for distribution under the plan, not creditor recovery directly. If creditors capture a fraction $\lambda \in (0, 1]$ of going-concern value through plan instruments such as shares and retained debt, while pool proceeds accrue to them in full, the creditor-value ranking replaces cV with $c\lambda V$ in condition (1) and is otherwise unchanged. The model therefore speaks to creditor recovery under the maintained assumption that λ is invariant to the disposition mode; where a plan changes λ itself — for example, through dilution of creditor equity — the mapping from preserved enterprise value to creditor recovery is an empirical question, which is what the measurement rules of Section 4.3 are designed to address.

Condition (1) partitions the parameter space into the two rival interpretations. When the fire-sale discount δ is large, expected recoverability $\dot{\theta}$ is high, governance keeps κ low, and separation protects substantial operating value cV , the trust is resolution infrastructure. When δ is small or $\dot{\theta}$ is low — the pool is simply worthless — deferral only postpones recognition, and weak governance (high κ) makes the region of deferred disappointment larger. Because $\tilde{\theta}$ is realized only through disposition and distribution, condition (1) also implies that nominal trust units carry no information about which region a case occupies; only realized, dated cash flows or market prices do.

4.2 Five-layer evaluation framework

Consistent with the model, a service trust should be evaluated on five layers rather than a single plan recovery rate (Table 2).

Table 2. Five-layer evaluation framework

Layer	Measure	Expected service-trust effect
Immediate liquidity	Cash paid at or shortly after approval	Lower or neutral because recovery is deferred
Realized recovery	Discounted cash distributions and verifiable disposition proceeds	Positive only if value preservation exceeds delay costs
Continuity	Plan completion, survival, listing status, employment, operating recovery	Positive when separation protects viable operations
Coordination	Execution time, plan acceptance, litigation, centralized asset control	Positive in fragmented and complex creditor structures
Accountability	Asset, valuation, governance, reporting and distribution disclosure	Determines whether delay remains disciplined

4.3 Measuring discounted realized recovery

At a common observation date T , package-level discounted realized recovery is $DRR(T;r) = [C_0 + \sum_{0 \leq t \leq T} D_t / (1+r)^t + MT / (1+r)^T] / Q$. C_0 is immediate cash; D_t is cash actually distributed by the trust; MT is the verifiable cash-equivalent value of disposed property, freely tradable equity, or retained debt at T ; and Q is the full confirmed claim principal exchanged for the recovery package. If the analysis isolates the trust sleeve, both numerator and denominator must be restricted consistently to the claim portion assigned to trust interests. Unmonetized or legally unresolved assets remain economically related to Q but are excluded from the numerator. Their appraised value is disclosed separately with valuation date, basis, encumbrances, and haircut scenarios; it is not labeled realized recovery.

The discount rate r represents the creditor's opportunity cost for a delayed, illiquid, and risky recovery. Because no single rate is institutionally privileged, the study pre-specifies a scenario grid rather than presenting one rate as an estimate. The 3% case is a low-risk time-value benchmark; 8% is the intermediate reference case for an illiquid corporate-credit claim; and 15% is a severe distress and illiquidity hurdle. These are transparent sensitivity anchors, not estimates of a debtor's weighted average cost of capital or a contemporaneous distressed-debt yield. Case-specific analysis will additionally report a duration-matched sovereign yield plus a pre-specified credit and illiquidity spread. Undiscounted cash is shown alongside all DRR estimates, and rates are not selected after observing outcomes. The observation horizon T is fixed across compared cases. If C_0 , D_t , Q , or timing is unavailable, DRR is reported as not estimable rather than imputed from trust-unit face value. Zero distributions and missing disclosure are coded separately.

4.4 Hypotheses

The four hypotheses follow from the comparative statics of condition (1).

Hypothesis 1 (fire-sale discount, from $\partial/\partial\delta$). When forced-sale discounts are material and governance is adequate, service trusts are expected to lower or leave unchanged immediate liquidity but improve discounted cumulative recovery relative to immediate disposition. Evidence status: model-derived; not yet tested in the present sample.

Hypothesis 2 (complexity as a net effect, from the joint dependence of δ , θ , and κ on asset conditions). Asset heterogeneity, unresolved title or litigation, and temporarily constrained natural buyers raise the fire-sale discount δ , which favors the trust; but the same conditions may simultaneously lower expected recoverability θ and raise administration and litigation costs κ , which oppose it. The trust's relative advantage therefore increases only when the discount effect dominates the recoverability and cost effects: complexity creates a net-effect boundary, not an unconditional benefit. Evidence status: mechanism-consistent case evidence; the sign of the net effect is an empirical question rather than a model implication.

Hypothesis 3 (separation, from the contamination term cV). Separating residual assets from retained operations is expected to be associated with plan completion and operating continuity. Evidence status: descriptive association only; investor capital, debt-to-equity conversion, court coordination, and accounting effects are co-interventions.

Hypothesis 4 (governance, from $\kappa(g)$). Because $\kappa'(g) < 0$, stronger beneficiary governance and post-establishment disclosure enlarge the region in which condition (1) holds and increase the likelihood that preserved option value is converted into realized creditor recovery. Governance is not a necessary condition in the model: a sufficiently large δ or cV/A can satisfy condition (1) even under weak governance. Evidence status: model-derived and testable; current evidence is insufficient for outcome validation.

5. Evidence and method

5.1 Sampling frame and case construction

The sampling frame combines two populations. For listed companies, the census of 144 court-approved reorganization plans (Section 3.2) identifies every plan and every trust adoption, eliminating selection on discoverability. For unlisted enterprise groups, candidates are located through the National Enterprise Bankruptcy Reorganization Case Information Network, court-approved plans, trustee recruitment and establishment notices, company annual reports, and trustee disclosures. Professional and legal commentary is used to locate documents, not as the sole source of monetary outcomes.

The present audit covers 20 candidate restructurings from 2018–2026, with an audit cut-off of August 18, 2026 for all case-level statements. The 20-case audit is a pilot evidence audit rather than a subsample of the census: it combines the trust-adopting listed-company plans identified in the census with major unlisted group restructurings located through the public sources above. The census defines the sampling frame for the next-stage listed-company analysis; the pilot audit establishes the evidence-grading protocol that the census extension will apply to all 144 plans. A case enters the performance sample only after three stages are distinguished: (1) the approved plan confirms a service-trust arrangement; (2) establishment or effective registration is verified; and (3) at least one post-establishment observation is available, such as an asset disposition, distribution, annual-report valuation, extension, or termination event. Proposed trusts and ongoing trustee recruitments remain in a prospective sample.

Table 3 reports the evidence funnel at the audit cut-off. Stages are cumulative; the grade is a separate evidence-quality label and should not be read as a lifecycle stage.

Table 3. Evidence funnel at the audit cut-off (August 18, 2026)

Evidence gate	Cases	Share of candidates	Interpretation
Candidate universe	20	100%	Located from public court, exchange, company, trustee, and professional sources

Evidence gate	Cases	Share of candidates	Interpretation
Approved plan verified	12	60%	Authoritative plan confirms a service-trust arrangement
Establishment verified	7	35%	Effective establishment or equivalent registration is documented
Post-establishment observation	7	35%	At least one operation, disposition, valuation, extension, or termination observation
Observable monetary evidence	2	10%	A recovery-relevant realized distribution or disposition-proceeds amount (A1), or a fair-value carrying amount reported in audited financial statements (A2), is observable

Table 4 provides the grade-to-stage crosswalk. P, E, O, and M denote verified plan, verified establishment, post-establishment observation, and observable monetary evidence. “NV” means not verified at the audit cut-off. Grade D is an exclusion override: a case that is only proposed, recruiting a trustee, explicitly not established, or otherwise ineligible remains D even if a plan document is available.

Table 4. Grade-to-stage crosswalk

Grade	Stage pattern P/E/O/M	Reproducible inclusion rule	Cases
A1	1 / 1 / 1 / 1	All three lifecycle gates verified and a cumulative or otherwise recovery-relevant realized distribution or disposition-proceeds amount sufficient to evaluate recovery performance is observable	1
A2	1 / 1 / 1 / 1	All three lifecycle gates verified and a carrying or fair-value amount for trust units reported in audited financial statements is the principal monetary evidence, without a disclosed cumulative recovery total; an isolated annual cash receipt does not by itself constitute cumulative recovery evidence	1
B1	1 / 1 / 1 / 0	Trust established and post-establishment evidence available, but no qualifying observable monetary evidence	5
B2	1 / NV / NV / 0	Plan architecture verified; establishment or later execution remains unverified; not subject to the D override	4

Grade	Stage pattern P/E/O/M	Reproducible inclusion rule	Cases
C	NV / NV / NV / 0	Credible case-level or secondary evidence, but the authoritative plan has not yet been verified	7
D	Override	Proposed, recruiting, explicitly not established, or otherwise ineligible for performance analysis	2

The crosswalk reconciles every total. A (2) plus B1 (5) gives the seven established and post-establishment cases. Adding B2 (4) and the one D case with a verified plan gives 12 plan-verified cases. Only the two A cases contain observable monetary evidence, they can be named, and their evidence differs in kind. The A1 case is the Bohai Steel Group restructuring: professional public sources report that the property trust established in January 2020 (CCB Trust “Caidie No. 1”) completed distributions of collected management, operating, and disposition proceeds totaling RMB 812,002,121.16 by December 31, 2021 (Yanglee Trust Net, n.d.). The underlying trustee report has not yet been publicly traced, so the amount is graded as professionally reported cumulative cash distributions rather than an independently audited figure; tracing the trustee or administrator source document is a task of the census extension. The A2 case is the HNA Group consolidated restructuring of 321 companies: units of the HNA bankruptcy-reorganization special service trust are reported at Level 3 fair-value carrying amounts — measurements based on unobservable inputs — in the audited financial statements of affiliated listed companies; for example, Gongxiao Daji reports a year-end 2024 carrying value of RMB 8.958 billion for its trust units, classified as fair value through profit or loss, while noting that no valuation report for the trust was obtained for 2024 (Gongxiao Daji Group, 2025). The same report records a trust distribution of RMB 9.4879 million received in 2024; under the A1 rule this isolated annual cash receipt corroborates that the trust is operating but does not by itself constitute cumulative recovery evidence, so the case remains A2. A Level 3 fair-value carrying amount reported in audited financial statements is a valuation, not a realized recovery, and the two cases are not pooled in any performance statement. The remaining coding decisions follow the written protocol described in the Declarations and will become independently reproducible when the source index is released with the case-level dataset. The published headline grade remains A/B/C/D = 2/9/7/2, with A = A1 + A2 and B = B1 + B2. The current performance core is therefore A + B1 = seven cases; B2 and C support architecture or disclosure analysis only, and D is excluded.

5.2 Market-implied recovery

Sparse trust distributions do not make recovery unobservable, because markets price the instruments creditors receive. Three price sources complement the DRR rules of Section 4.3. First, many debtors in the sample have defaulted exchange- or interbank-market bonds that continue to trade; the bond price immediately after plan approval is a market-implied valuation of the entire recovery package (Jankowitsch, Nagler, & Subrahmanyam, 2014). Subtracting the value of the package’s cash and equity components does not, however, isolate the trust-unit sleeve: the residual also reflects seniority and collateral differences, enforcement delay and litigation risk, liquidity discounts, taxes and trustee fees, and plan-execution risk. Absent an explicit decomposition and strong assumptions, the measure is therefore reported as package-level market-implied recovery, not as a trust-specific valuation. Second, plans that repay claims with shares specify an explicit debt-repayment conversion price; the post-lockup market price of those shares yields a realized value for the equity component, narrowing — but not eliminating — the residual attributable to trust interests and the other unpriced factors. Third, transfer prices in the claims-trading market, where distressed-asset managers purchase claims after plan approval, provide a transaction-based valuation (Ivashina, Iverson, & Smith, 2016), although

such transfers are frequently unobserved. Each source is dated, verifiable, and independent of trust-unit face values. Coverage is itself an empirical question: defaulted bonds do not always trade continuously and claims-trading prices are often private. The census extension will therefore begin with a coverage audit table reporting, for each of the 144 companies, whether defaulted-bond prices, debt-repayment share prices, and claims-trading prices are observable, so that symmetry between trust and non-trust plans is demonstrated rather than assumed.

5.3 Comparative designs and an exploratory identification roadmap

Adoption is endogenous. Large, opaque, and heavily indebted groups are more likely to use service trusts. A raw comparison with simpler restructurings would therefore load case severity onto the trust indicator. Three sources of institutional variation are candidates for addressing this problem. Together they constitute an exploratory identification roadmap, subject to the coverage and statistical-power audits described below, rather than a committed causal design; none is treated as established identification at this stage.

First, the June 2023 classification reform (Section 3.1) is a supply-side event, but it is a nationwide policy with no untreated group. Exposure can be measured as the pre-reform presence of trust companies active in risk-disposal business in the debtor's province and court district, yielding a continuous-treatment-intensity design rather than a canonical difference-in-differences (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021; Callaway, Goodman-Bacon, & Sant'Anna, 2024). The exposure measure is itself endogenous: it may proxy for bankruptcy-court capacity, regional financial development, the local weight of state-owned and large enterprises, and the maturity of the special-assets market. The design is therefore a heterogeneous-exposure design whose validity must be established, requiring at minimum an exposure variable frozen at a pre-reform date, evidence of parallel pre-reform adoption trends, controls for bankruptcy-court specialization, regional financial development, debt scale, and ownership type, and placebo tests using pseudo-reform years and alternative exposure measures. The nationwide reform itself is not treated as an exogenous instrument. Two further constraints keep this design exploratory. Because trust companies operate nationally, local pre-reform presence may capture relationship networks or regional deal flow rather than a binding supply constraint, so the exposure audit will report both province-level and business-relationship-based measures. And with only two trust adoptions among the 49 pre-reform plans, pre-trend evidence and statistical power are necessarily limited; the design's feasibility will be reported before any estimate is. Consistent with the approval-date assignment rule in Section 3.2, the pre-reform baseline is the 49 plans approved during 2019–2022 with two trust adoptions, not the 2019–2023 window, because 2023 straddles the reform's effective date.

Second, the December 2024 Minutes and the companion CSRC guideline tightened plan-execution and disclosure standards for listed companies (SPC & CSRC, 2024; CSRC, 2025). Because the Minutes were issued on December 31, 2024, they cannot have shaped the design of the 11 plans approved during 2024, and the post-treatment window currently contains essentially a single annual cohort (2025). Comparing trust-plan disclosure quality and execution behavior around this demand-side event is therefore an interrupted institutional comparison, or post-reform cohort analysis, rather than a quasi-experiment; it bears on Hypothesis 4 descriptively and will gain identifying content only as later cohorts accumulate and as the rules' transition provisions, acceptance dates, and scope boundaries generate sharper treatment and exemption groups.

Third, within-case creditor-class variation holds case-level severity fixed but does not by itself isolate the trust component. Within a single plan, small claims typically receive full cash while claims above a threshold receive mixed packages of shares and trust units; the classes therefore differ by construction in legal treatment and package composition, and case fixed effects cannot separate small-claim protection policy, claim size, claim type and collateral status, and the bundling of shares with trust units. The comparison is accordingly reported as a within-case descriptive valuation comparison, not as identification of the valuation of trust interests. Where a plan specifies a strict monetary threshold, a local comparison of claims just below and just

above the threshold can approach identification, provided the threshold is non-manipulable, claims on either side are otherwise comparable, creditor-level market or realized recovery is observable, and the treatment change at the threshold is not confounded by multiple simultaneous instruments.

Where cross-case comparison is still required, the design combines within-case liquidation benchmarks, matched non-trust restructurings, and common observation horizons; entropy balancing and doubly robust adjustment will be used when the expanded sample supports them. Until then, results are described as structured comparative evidence, not causal treatment effects.

5.4 Present evidentiary limits

The evidence presently supports a mixed-method paper. Plan architecture and continuity outcomes are observable across several cases, whereas actual trust distributions are sparse. This asymmetry is itself substantively important: the trust may help a restructuring become legally executable while remaining economically unresolved for beneficiaries.

6. Founder Group: a focal positive mechanism case

Evidence and independence protocol. All case facts in Sections 6.1–6.3 are drawn solely from cited public court, company, investor, trustee, legal, or news materials. Statements labeled “Public-record fact” report those materials; statements labeled “Analytical inference” are the author’s interpretation. No confidential institutional, client, beneficiary, trustee, or transaction information is used. The analysis represents the author’s personal and independent academic views only and is unrelated to, and does not represent, the views of the author’s former employer, Ping An Trust, or any other project participant.

6.1 Why the case is informative

Public-record fact. Founder Group and four affiliated companies entered substantive consolidated restructuring. Public materials report approximately RMB 62.2 billion in assets, RMB 146.9 billion in liabilities, and RMB 130.9 billion in court-confirmed claims, including about RMB 116.0 billion in ordinary claims; the group included information technology, healthcare, finance, urban development, and other businesses, including interests connected to six listed companies (Yicai, 2021; JunZeJun Law Offices, 2025).

Analytical inference. The case is informative because the restructuring problem was not only aggregate insolvency. It also involved a timing and perimeter problem: potentially viable operating platforms had to be transferred or rehabilitated while disputed, weak, off-book, or title-imperfect assets could not be fully resolved within the court timetable. This inference is based on the disclosed asset classification and plan architecture, not on internal experience.

6.2 Separation of operating value and residual uncertainty

Public-record fact. The restructuring used a sale-style architecture. Retained assets were used to form New Founder and business platforms. Public legal materials identify Zhuhai Huafa, Ping An (implemented through Ping An Life), and Shenzhen Tefa/related entities as investors and report a cash investment range of approximately RMB 53.725–73.3 billion, depending on the acquired equity percentage (JunZeJun Law Offices, 2025; CICC, n.d.). Residual assets — including some title-defective, litigation-pending, or off-book assets — remained in the old restructuring entities. Public reports state that Ping An Trust participated in the design of a third-party property trust for clearing, confirming, managing, and disposing of those assets, with net proceeds available for supplementary creditor distributions (Yicai, 2021; JunZeJun Law Offices, 2025).

Analytical inference. The trust's defensible positive contribution was architectural: it prevented residual uncertainty from blocking the transfer and rehabilitation of retained operations while preserving creditors' contingent upside. This does not imply that every residual asset had positive value. Some were publicly reported with an appraised value of zero (JunZeJun Law Offices, 2025). The trust is therefore interpreted as a bridge between legal completion and economic completion, not as evidence that the residual pool was worth its nominal claim allocation. In terms of condition (1), the disclosed facts — thin markets for disputed and title-imperfect assets, residual uncertainty obstructing investor entry into the retained platforms, and a governed deferral structure — are qualitatively consistent with a high fire-sale discount δ and material contamination costs cV , but neither parameter is estimated here. On that qualitative reading, the case sits in the region where the model predicts the trust is resolution infrastructure.

6.3 Time, option value, and governance

Public-record fact. The disclosed trust term was 36 months, with a further 36-month extension possible if the governance bodies did not adopt a termination decision. Creditors not paid in full received beneficial interests in the residual pool, and later net proceeds were designated for supplementary distribution (Yicai, 2021; JunZeJun Law Offices, 2025).

Analytical inference. The term converted an immediate-disposition problem into a governed recovery option. Its value depends on subsequent title resolution, litigation recovery, asset sales, operating cash flow, expenses, and distributions. Founder is accordingly a positive mechanism case, not a completed recovery-rate success case. It shows how a highly complex restructuring may isolate historical problems and preserve supplementary recovery rights; this paper will not attribute a high realized return to the trust until public distribution evidence is available. The main comparative conclusions will also be reported after excluding Founder.

7. Comparative evidence from recent listed companies

The 2024 listed-company cohort illustrates how service trusts operate inside hybrid recovery packages. The cases differ in cash thresholds, stock allocations, trust-unit design, and residual asset quality, but share a common objective: complete the restructuring and separate difficult assets without requiring an immediate sale. Case facts are drawn from the court-approved reorganization plans and subsequent exchange filings (Helytai, 2024; Orient Landscape, 2024; Aonong Biology, 2024; Flower King, 2024; Zhongli Group, 2024).

Table 5. Institutional contribution and financial outcome in the 2024 cohort

Case	Institutional contribution (0–2)	Company-level financial outcome (0–2)	Attribution boundary
Helytai	2 — verified residual-asset trust within a cash/equity/trust package	1 — net assets returned positive and delisting-risk warning was removed	Company outcome reflects the entire plan; trust cash distributions remain unobserved
Orient Landscape	2 — non-retained assets separated; claims above RMB 100,000 combine shares and trust interests	1 — risk warnings were later removed	Plan reports a 26.82% liquidation benchmark, but trust interests are not realized cash
Aonong Biology	2 — divested subsidiary equity placed behind trust units for the largest claims	1 — year-end net assets returned positive	Plan expressly states that trust-unit quantity does not equal final cash distribution

Case	Institutional contribution (0–2)	Company-level financial outcome (0–2)	Attribution boundary
Flower King	1 — hybrid trust architecture verified at plan level; establishment chain incomplete	1 — plan completed and restructuring-related delisting warning removed	Other risk warnings remained; continuity is not full recovery
Zhongli Group	1 — non-retained assets assigned to a service-trust structure at plan level	0 — post-plan outcome not yet verified	Insufficient evidence for a positive performance claim

Scores separate two concepts. Institutional contribution is coded 0 = no verified trust role, 1 = verified plan architecture, and 2 = verified separation or governance mechanism. Company-level financial outcome is coded 0 = unavailable or adverse, 1 = favorable company-level outcome within a multi-instrument plan, and 2 = trust-specific realized cash or disposition proceeds. No listed-company case currently receives a 2 on the financial-outcome dimension. The table therefore does not treat removal of a delisting-risk warning as a trust outcome. In the econometric companion design of Section 5, these ordinal codes are replaced by continuous outcomes — market-implied package values, dated distributions, and execution durations — so that the scoring serves only as a descriptive bridge.

The interim finding is institutional. Service trusts appear to expand the feasible set of restructuring packages by separating difficult assets and centralizing residual creditor rights. Favorable balance-sheet or listing outcomes occur alongside investor capital, debt-to-equity conversion, court coordination, and accounting effects and cannot be attributed solely to the trust. The stronger recovery claim remains subject to actual distributions and common-horizon DRR estimates.

8. Discussion and policy implications

8.1 What service trusts positively accomplish

Service trusts create value chiefly by changing organization and timing. They can isolate assets that are difficult to sell, aggregate creditors' residual rights, allow viable operations to emerge without historical contamination, and preserve an orderly route to later recovery. These functions are particularly valuable for enterprise groups with numerous subsidiaries, mixed asset quality, extensive guarantees, and uncertain title chains.

8.2 What they do not accomplish

The trust form does not eliminate underlying economic losses. Assigning one trust unit for one yuan of adjusted debt is a legal allocation rule, not a valuation result. Nor does the completion of the restructuring plan prove that the trust has completed its work. Regulators, courts, and researchers should avoid describing nominal trust interests as realized recovery.

8.3 A disclosure architecture

A minimum disclosure regime should include: a dated schedule of trust assets; valuation methods and assumptions; title, pledge, freeze, and litigation status; trustee and asset-manager responsibilities; annual operating and disposition reports; cash receipts and expenses; beneficiary distributions; extensions and their rationale; and a final comparison of plan-implied and realized recovery. These disclosures would make the “time for value” claim auditable and reduce the risk that service trusts become opaque repositories for losses. They would also allow market prices to discipline the structure: with a dated asset schedule and cash waterfall,

the market-implied valuations of Section 5.2 can be reconciled against reported net asset values, making divergence itself an early-warning signal.

Table 6. Minimum disclosure blocks and their evaluation use

Disclosure block	Minimum fields	Evaluation use
Asset perimeter	Asset identifier, ownership chain, valuation date/method, encumbrances, litigation, related-party status	Tests whether valuable and difficult assets were actually transferred
Governance	Trustee and manager duties, beneficiary voting, committee authority, conflicts, fee rules, removal and replacement	Tests accountability and agency-cost controls
Cash waterfall	Opening cash, receipts by asset, operating costs, fees, taxes, reserves, distributions by beneficiary class	Separates gross realization from net creditor recovery
Time and status	Establishment date, milestones, unsold assets, extensions, reason for delay, termination trigger	Measures whether “time for value” remains disciplined
Outcome reconciliation	Claim denominator, plan-implied value by instrument, actual cash dates, equity value, DRR sensitivity	Makes nominal-versus-realized recovery auditable

9. Conclusion

Bankruptcy service trusts should be judged neither by immediate cash alone nor by nominal trust units. Their economic role is to reconcile the compressed legal timetable of restructuring with the longer period needed to verify, operate, and dispose of complex assets. The model in Section 4 makes the boundary precise: the trust preserves value when the avoided fire-sale discount and protected operating value exceed governance-dependent deferral costs. The present evidence supports a positive institutional interpretation: Founder and several listed-company plans show how a trust can separate difficult assets, centralize residual creditor rights, and prevent unresolved property from obstructing rehabilitation.

The evidence does not yet establish a positive recovery effect. Only seven of 20 candidate cases have a verified post-establishment observation, and only two provide case-level monetary evidence meeting the audit threshold — one as professionally reported cumulative cash distributions, the other as a Level 3 fair-value carrying amount reported in audited financial statements. Founder remains a mechanism case, and listed-company balance-sheet or listing outcomes arise from the whole restructuring package. The next empirical step is specified rather than merely promised: extend the audit to the full census of 144 court-approved listed-company plans, collect package-level market-implied recovery from defaulted-bond, share, and claims-trading prices, and evaluate whether the exploratory designs in Section 5.3 — the 2023 classification reform exposure design, the post-Minutes cohort comparison, and within-case creditor-class variation — can estimate the effect of trust adoption under the DRR rules, coverage and power audits, and validity conditions defined here. If governance and disclosure convert preserved optionality into transparent distributions, service trusts can mature into credible resolution infrastructure; if not, they risk becoming vehicles for deferred loss recognition.

Declarations

Author independence and conflict of interest. This paper expresses the author's personal and independent academic views only. It is unrelated to, and does not represent, the views of the author's former employer, Ping An Trust Co., Ltd., or any other institution or transaction participant. The author was previously affiliated with Ping An Trust, which participated in the Founder Group restructuring discussed as a focal case. The study relies solely on publicly available materials and does not use confidential client, beneficiary, trustee, or transaction information. Public-record facts and analytical inferences are separately labeled in the focal-case section. Three safeguards address the potential conflict: all coding decisions follow a written protocol, and a public source index will be released with the case-level dataset so that the coding can be independently reproduced; comparative results will be reported with Founder excluded in the subsequent full-sample analysis; and independent replication of the case coding is expressly invited.

Funding. No external funding was received for this study.

Data availability. The case-level dataset is being constructed from publicly available court, exchange, company, and trustee disclosures. A source-indexed version will accompany a later manuscript version to the extent permitted by source access and redistribution rules.

Author contribution. Fuli Yang: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Writing – original draft, Writing – review and editing.

AI assistance for this version. AI tools assisted with editorial organization, language editing and document formatting. The author remains responsible for the manuscript.

References

Academic literature: bankruptcy economics and institutions

Andrade, G., & Kaplan, S. N. (1998). How costly is financial (not economic) distress? Evidence from highly leveraged transactions that became distressed. *Journal of Finance*, 53(5), 1443–1493.

<https://doi.org/10.1111/0022-1082.00062>

Ayotte, K. M., & Morrison, E. R. (2009). Creditor control and conflict in Chapter 11. *Journal of Legal Analysis*, 1(2), 511–551. <https://doi.org/10.1093/jla/1.2.511>

Benmelech, E., & Bergman, N. K. (2011). Bankruptcy and the collateral channel. *Journal of Finance*, 66(2), 337–378. <https://doi.org/10.1111/j.1540-6261.2010.01636.x>

Bernstein, S., Colonnelli, E., Giroud, X., & Iverson, B. (2019a). Bankruptcy spillovers. *Journal of Financial Economics*, 133(3), 608–633. <https://doi.org/10.1016/j.jfineco.2018.09.010>

Bernstein, S., Colonnelli, E., & Iverson, B. (2019b). Asset allocation in bankruptcy. *Journal of Finance*, 74(1), 5–53. <https://doi.org/10.1111/jofi.12740>

Bris, A., Welch, I., & Zhu, N. (2006). The costs of bankruptcy: Chapter 7 liquidation versus Chapter 11 reorganization. *Journal of Finance*, 61(3), 1253–1303. <https://doi.org/10.1111/j.1540-6261.2006.00872.x>

Hotchkiss, E. S. (1995). Postbankruptcy performance and management turnover. *Journal of Finance*, 50(1), 3–21. <https://doi.org/10.1111/j.1540-6261.1995.tb05165.x>

Hotchkiss, E. S., & Mooradian, R. M. (1997). Vulture investors and the market for control of distressed firms. *Journal of Financial Economics*, 43(3), 401–432. [https://doi.org/10.1016/S0304-405X\(96\)00900-2](https://doi.org/10.1016/S0304-405X(96)00900-2)

Ivashina, V., Iverson, B., & Smith, D. C. (2016). The ownership and trading of debt claims in Chapter 11 restructurings. *Journal of Financial Economics*, 119(2), 316–335. <https://doi.org/10.1016/j.jfineco.2015.09.002>

Iverson, B. (2018). Get in line: Chapter 11 restructuring in crowded bankruptcy courts. *Management Science*, 64(11), 5370–5394. <https://doi.org/10.1287/mnsc.2017.2808>

Maksimovic, V., & Phillips, G. (1998). Asset efficiency and reallocation decisions of bankrupt firms. *Journal of Finance*, 53(5), 1495–1532. <https://doi.org/10.1111/0022-1082.00063>

Ponticelli, J., & Alencar, L. S. (2016). Court enforcement, bank loans, and firm investment: Evidence from a bankruptcy reform in Brazil. *Quarterly Journal of Economics*, 131(3), 1365–1413. <https://doi.org/10.1093/qje/qjw015>

Pulvino, T. C. (1998). Do asset fire sales exist? An empirical investigation of commercial aircraft transactions. *Journal of Finance*, 53(3), 939–978. <https://doi.org/10.1111/0022-1082.00040>

Pulvino, T. C. (1999). Effects of bankruptcy court protection on asset sales. *Journal of Financial Economics*, 52(2), 151–186. [https://doi.org/10.1016/S0304-405X\(99\)00007-0](https://doi.org/10.1016/S0304-405X(99)00007-0)

Rodano, G., Serrano-Velarde, N., & Tarantino, E. (2016). Bankruptcy law and bank financing. *Journal of Financial Economics*, 120(2), 363–382. <https://doi.org/10.1016/j.jfineco.2016.01.016>

Shleifer, A., & Vishny, R. W. (1992). Liquidation values and debt capacity: A market equilibrium approach. *Journal of Finance*, 47(4), 1343–1366. <https://doi.org/10.1111/j.1540-6261.1992.tb04661.x>

Strömberg, P. (2000). Conflicts of interest and market illiquidity in bankruptcy auctions: Theory and tests. *Journal of Finance*, 55(6), 2641–2692. <https://doi.org/10.1111/0022-1082.00302>

Academic literature: investment timing, recovery pricing, and econometric methods

Callaway, B., Goodman-Bacon, A., & Sant’Anna, P. H. C. (2024). Difference-in-differences with a continuous treatment (NBER Working Paper No. 32117). National Bureau of Economic Research. <https://www.nber.org/papers/w32117>

Callaway, B., & Sant’Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230. <https://doi.org/10.1016/j.jeconom.2020.12.001>

Dixit, A. K., & Pindyck, R. S. (1994). *Investment under uncertainty*. Princeton University Press.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2), 254–277. <https://doi.org/10.1016/j.jeconom.2021.03.014>

Jankowitsch, R., Nagler, F., & Subrahmanyam, M. G. (2014). The determinants of recovery rates in the US corporate bond market. *Journal of Financial Economics*, 114(1), 155–177. <https://doi.org/10.1016/j.jfineco.2014.06.001>

McDonald, R., & Siegel, D. (1986). The value of waiting to invest. *Quarterly Journal of Economics*, 101(4), 707–727. <https://doi.org/10.2307/1884175>

Academic literature: China bankruptcy resolution

Hotchkiss, E., John, K., Li, B., Ponticelli, J., & Wang, W. (2023). Default and bankruptcy resolution in China. *Annual Review of Financial Economics*, 15, 369–385. <https://doi.org/10.1146/annurev-financial-110921-014557>

Jiang, Y. (2014). The curious case of inactive bankruptcy practice in China: A comparative study of U.S. and Chinese bankruptcy law. *Northwestern Journal of International Law & Business*, 34(3), 559–582.

Lee, E. (2011). The reorganization process under China’s corporate bankruptcy system. *International Lawyer*, 45(4), 939–974.

Li, B., & Ponticelli, J. (2022). Going bankrupt in China. *Review of Finance*, 26(3), 449–486. <https://doi.org/10.1093/rof/rfab023>

Academic literature: trust law and fiduciary governance

- Dong, X. (2025). On the role of bankruptcy administrators in enterprise bankruptcy service trusts [in Chinese]. *Advances in Social Sciences*, 14(3), 28–36. <https://doi.org/10.12677/ass.2025.143186>
- Foster, F. H. (2010). American trust law in a Chinese mirror. *Minnesota Law Review*, 94(3), 602–651. <https://doi.org/10.24926/265535.1389>
- Hansmann, H., & Kraakman, R. (2000). The essential role of organizational law. *Yale Law Journal*, 110(3), 387–440.
- Ho, L. (2003). *Trust law in China*. Sweet & Maxwell Asia.
- Ho, L., Lee, R., & Jin, J. (2013). Trust law in China: A critical evaluation of its conceptual foundation. In L. Ho & R. Lee (Eds.), *Trust law in Asian civil law jurisdictions* (pp. 80–98). Cambridge University Press. <https://doi.org/10.1017/CBO9781139149945.009>
- Jing, H. (2024). Third-party and bankruptcy effects under Chinese trust law: Comparisons with English trust law. *Asian Journal of Comparative Law*, 19(2), 216–232. <https://doi.org/10.1017/asjcl.2024.8>
- Lee, J. (2021). The irreducible core of trustee duties in East Asian trusts. *Trusts & Trustees*, 27(4), 302–310. <https://doi.org/10.1093/tandt/ttab013>
- Lee, R. (2009). Conceptualizing the Chinese trust. *International & Comparative Law Quarterly*, 58(3), 655–669. <https://doi.org/10.1017/S0020589309001183>
- Lyu, K. (2015). Re-clarifying China's Trust Law: Characteristics and new conceptual basis. *Loyola of Los Angeles International and Comparative Law Review*, 36, 447–495.
- Qu, C. Z. (2003). The doctrinal basis of the trust principles in China's Trust Law. *Real Property, Probate and Trust Journal*, 38(2), 345–376.
- Sitkoff, R. H. (2004). An agency costs theory of trust law. *Cornell Law Review*, 89(3), 621–684.
- Tan, Z. (2001). The Chinese law of trusts: A compromise between two legal systems. *Bond Law Review*, 13(1), 224–249. <https://doi.org/10.53300/001c.5371>
- Tian, S. (2021). A comparative analysis of the fiduciary duties of loyalty of trustees in English common law and Chinese law. *Trusts & Trustees*, 27(10), 974–988. <https://doi.org/10.1093/tandt/ttab088>
- Primary legal, regulatory, and case materials
- People's Republic of China. (2001). *Trust Law of the People's Republic of China* (especially arts. 15–18 and 25–29). <https://www.nfra.gov.cn/cbircweb/download/downloadPdf?docId=879933&itemId=927&zcfg=0>
- People's Republic of China. (2006). *Enterprise Bankruptcy Law of the People's Republic of China*.
- China Banking and Insurance Regulatory Commission. (2023). Notice on regulating the classification of trust business of trust companies [in Chinese; effective June 1, 2023]. https://jrj.beijing.gov.cn/jrgzdt/202303/t20230328_2946033.html
- Supreme People's Court & China Securities Regulatory Commission. (2024). Minutes of the work symposium on properly adjudicating bankruptcy reorganization cases of listed companies (Fa [2024] No. 309) [in Chinese]. <https://www.csrc.gov.cn/csrc/c100028/c7529727/content.shtml>
- China Securities Regulatory Commission. (2025). Guideline No. 11 for the supervision of listed companies: Matters relating to bankruptcy reorganization of listed companies [in Chinese]. <https://www.csrc.gov.cn/csrc/c100028/c7544550/content.shtml>

National Financial Regulatory Administration. (2025). Measures for the administration of trust companies [in Chinese]. <https://www.nfra.gov.cn/cn/view/pages/ItemDetail.html?docId=1225658&generaltype=0&itemId=915>

National Enterprise Bankruptcy Reorganization Case Information Network. Case plans, rulings, and execution notices. <https://pccz.court.gov.cn/>

JunZeJun Law Offices. (2025). Risk-disposal service trusts and the Founder Group case [in Chinese]. <https://www.junzejun.com/Publications/1018014d7d56d2-5.html>

China International Capital Corporation. (n.d.). Founder Group restructuring transaction overview [in Chinese]. https://www.cicc.com/news/details311_68903.html

Yicai. (2021). Public report on the Founder Group restructuring service-trust structure [in Chinese]. <https://www.yicai.com/news/101133366.html>

Helytai Technology Co., Ltd. (2024). Reorganization plan [in Chinese]. <https://static.cninfo.com.cn/finalpage/2024-12-24/1222116030.PDF>

Beijing Orient Landscape & Environment Co., Ltd. (2024). Reorganization plan [in Chinese]. <https://static.cninfo.com.cn/finalpage/2024-12-24/1222112814.PDF>

Fujian Aonong Biology Science and Technology Group Co., Ltd. (2024). Reorganization plan [in Chinese]. <https://static.cninfo.com.cn/finalpage/2024-12-11/1221982063.PDF>

Flower King Eco-Engineering Co., Ltd. (2024). Reorganization plan [in Chinese]. https://www.sse.com.cn/disclosure/listedinfo/announcement/c/new/2024-11-16/603007_20241116_V4MD.pdf

Jiangsu Zhongli Group Co., Ltd. (2024). Announcement on court approval of the reorganization plan [in Chinese]. <https://file.finance.qq.com/finance/hs/pdf/2024/12/12/1221999786.PDF>

Gongxiao Daji Group Co., Ltd. (2025). 2024 annual report [in Chinese]. <https://static.cninfo.com.cn/finalpage/2025-04-29/1223396530.pdf>

Market and professional sources

AllBright Law Offices. (2025). Annual report on the bankruptcy reorganization of Chinese listed companies, 2024 [in Chinese]. <https://www.allbrightlaw.com/CN/10475/1d7e4d40b81d7fba.aspx>

AllBright Law Offices. (2026). Annual report on the bankruptcy reorganization of Chinese listed companies, 2025 [in Chinese]. <https://www.allbrightlaw.com/CN/10475/e747fb460b48b6a8.aspx>

Sina Finance. (2025). Risk-disposal service trusts exceed RMB 1.5 trillion as trust companies accelerate entry [in Chinese]. <https://finance.sina.com.cn/jjxw/2025-02-28/doc-inemziti7023487.shtml>

Yanglee Trust Net. (n.d.). A brief discussion of risk-disposal service trusts [in Chinese]. <https://www.yanglee.com/Information/Details.aspx?i=133813>

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.

Unknown Is Not Zero: Heat Stress, Data-Center Power-Demand Exposure, and Utility Volatility

Fuli Yang

GARCH Institute, Harvard Square, Cambridge, MA 02138, USA

SSRN version 1.2 — September 2026

Abstract

This paper asks whether service-territory heat stress and disclosed data-center power-demand exposure jointly amplify the volatility of U.S. listed electric utilities. Its principal finding is methodological but empirically consequential: pre-disclosure silence is not a verified zero. The study constructs an auditable point-in-time chain from dated operating or binding customer-load megawatts to the regulated utility, the applicable EIA Form 861 peak-load denominator, a representative NOAA station, daily parent-company returns, and lag-available financial controls. Ten issuers satisfy a broad data-center four-quarter evidence rule over 2021–2025, although the measure is not uniformly AI-specific and one issuer is predominantly exposed to cryptocurrency hosting. Reclassifying 8,685 legacy zero-coded issuer-days as unknown leaves 3,855 positive-disclosure issuer-days and 212 complete portfolio observations across nine variance-recursion segments. The measurement path is non-monotone: the uncentered interaction is 0.143 in the legacy zero-coded panel and 0.060 when OTTR is excluded; after the common-positive restriction it reaches -2.000 at the coefficient bound, while centering yields an interior estimate of -0.118 . Before centering, $\text{Corr}(C, C \times A) = 0.979$ and the interaction VIF is 29.4; centering lowers the VIF to 1.09. In the final Student-t GARCH-X specification, AIC and BIC select the no-interaction model and a 499-replication conditional parametric bootstrap does not reject it ($LR = 1.638$, $p = 0.324$). The current design therefore fails to detect heat-by-load amplification; it does not establish that the population effect is zero. H3 remains untested, and the pre-specified Markov-switching extension is not estimated. Positively, the audit identifies how exposure coding and coordinate construction can manufacture apparent interaction evidence, supplies a reusable regulatory-data architecture, and defines an advance rule under which a longer panel can retest the same hypotheses without redefining the original evidence.

Keywords: heat stress; data centers; artificial intelligence infrastructure; electricity demand; conditional volatility; GARCH-X; measurement audit

JEL classifications: C32; C58; G12; G14; Q54

1. Introduction

Climate risk and artificial intelligence are often treated as separate sources of financial change. Their physical infrastructure intersects in the electricity system: data centers require grid connections and substantial power capacity, while extreme heat raises cooling demand and tightens operating conditions. For listed utilities, the empirical question is whether disclosed incremental data-center load changes how local heat stress is translated into return volatility.

The scale of the physical connection has changed quickly. Berkeley Lab estimates that U.S. data centers consumed about 176 TWh in 2023, approximately 4.4% of national electricity consumption, and projects a wide 2028 range of 325–580 TWh, or 6.7–12.0% of U.S. electricity use (Shehabi et al., 2024). The width of that range is itself economically relevant. Utilities must plan generation, transmission, distribution, and financing before realized demand is known, while duplicated interconnection requests and uncertain completion rates can inflate pipeline measures. For financial identification, announced demand is therefore not a sufficient exposure measure: contractual status, legal-entity matching, denominator scope, and public timing all matter.

The central question is whether the volatility response to this joint pressure is super-additive. If climate stress and AI-related load merely contribute independent risks, a conventional additive variance specification is sufficient. If, however, contracted data-center load makes a utility more sensitive to heat-related grid stress—or if heat stress makes the market update the value of incremental AI load more abruptly—then the interaction should enter conditional variance directly. A second possibility is that the interaction operates through state dynamics: dual stress may raise the probability of entering a high-volatility regime or lengthen the time spent there. These are distinct economic claims and require distinct estimands.

This distinction matters because much of the existing empirical evidence is compatible with several competing interpretations. A positive association between climate news and volatility may reflect a common energy-price shock. A positive association between AI exposure and volatility may be little more than a technology, growth, or capital-expenditure factor. A post-2022 interaction between two trending indices may capture a period effect rather than a mechanism. The empirical design therefore uses utility-level, physically interpretable quantities. Climate exposure is tied to the service territory. AI exposure is tied to dated, operational or contractually binding customer-load megawatts. The denominator is the matched utility system peak load from EIA Form 861. Both components are lagged and standardized using expanding information sets.

The design is narrower than a general study of AI adoption. It does not infer AI use from a technology index, corporate text, or market classification. It studies operating or binding data-center electricity demand because that channel can be stated in physical units, matched to a utility system, and dated to the public information set. The audit also shows why the label matters: one qualifying load sequence, Otter Tail, predominantly reflects cryptocurrency hosting. The empirical estimand is therefore data-center power-demand exposure, with AI infrastructure as a motivating application rather than a verified treatment for every issuer.

The paper makes three contributions. First, it turns scattered load disclosures into a reproducible point-in-time measurement chain linking the parent security, operating utility, load object, EIA denominator vintage, weather territory, and public-information date. Second, it documents a concrete identification result: treating pre-disclosure silence as zero and multiplying uncentered regressors can make a climate-by-load interaction appear better supported than it is. Table 4A consolidates the full measurement-to-estimate path and distinguishes the OTTR label sensitivity from the common-positive and centering decisions. Third, the paper implements an advance rule that converts a failed first-stage interaction test into a disciplined stopping decision. Regime switching is reserved for a future sample that adds independent exposure variation, rather than used to rescue a weak single-state result.

The final broad data-center register contains ten listed parents: AEE, AEP, AES, CMS, D, EXC, IDA, MDU, OTTR, and PPL. The audit preserves the original filings, entity mappings, denominator vintages, weather records, market data, financial-control provenance, and corrections. The daily public-information panel contains 11,541 complete issuer-days. Requiring positive disclosed exposure for all ten parents produces a common window from 13 February through 31 December 2025, with 212 portfolio observations across nine variance-recursion segments.

The M1 estimates do not provide detectable support for the proposed amplification mechanism, but they do deliver an informative result about the design. In the legacy zero-coded panel, the uncentered interaction estimate is 0.143; excluding OTTR, whose tagged load is predominantly cryptocurrency hosting, reduces it to 0.060. These are label and coding sensitivities, not a sequential common-positive estimate. Reclassifying pre-disclosure zeros as unknown reduces the analysis to the common-positive window, where the uncentered interaction reaches -2.000 at its imposed bound. Centering the main effects produces an interior estimate of -0.118 , both information criteria select the no-interaction model, and a conditional parametric bootstrap yields $p=0.324$. Because M1 fails the advance condition, H3 remains untested and the paper does not estimate the Markov-switching extensions; this preserves a clean benchmark that can be updated when additional public disclosure history becomes available.

Section 2 develops the economic mechanisms. Section 3 documents the institutional setting and audited measurement. Section 4 states the directional hypotheses. Section 5 presents the staged estimation and stopping rule. Section 6 explains the identification limits. Section 7 reports data readiness and M1 results. Section 8 concludes.

2. Related Literature and Economic Mechanisms

2.1 Climate risk and financial volatility

Climate risk can affect asset prices through expected cash flows, financing costs, insurance and adaptation expenditures, policy exposure, and discount-rate variation. Evidence on carbon risk shows that investors distinguish firms by emissions and carbon intensity, while broader climate-finance research emphasizes the difficulty of converting long-horizon hazards into dated market innovations (Bolton and Kacperczyk, 2021; Giglio, Kelly, and Stroebel, 2021). A related measurement-layer problem appears when an issuer-level label compresses variation that remains visible at the underlying security or operating-entity level. In Chinese corporate bonds, public climate signals remain detectable in security-level pricing even when issuer credit ratings compress them (Yang, 2026b). That result motivates the present decision not to assign a parent-company climate or AI label mechanically: exposure is instead mapped to the regulated operating utility and its service territory. Physical risk is especially relevant for utilities because temperature and severe weather affect demand, operating costs, reliability, and investment requirements within geographically defined service territories.

Granular temperature evidence strengthens the case for location-specific measurement. Addoum, Ng, and Ortiz-Bobea (2023) link establishment-level temperatures to quarterly earnings and find economically meaningful temperature sensitivity across a broad set of industries. Acharya et al. (2022) exploit regional heat exposure across municipal bonds, corporate bonds, and equities, showing why headquarters-level or national averages can conceal the relevant asset location. These studies motivate the service-territory

mapping used here, but they do not test whether a newly expanding class of electricity demand changes the volatility transmission of heat stress.

For volatility measurement, the relevant insight is that climate information need not have a constant effect. A heat anomaly can be economically minor when capacity margins are ample and material when the grid is tight. The same event can also be priced differently across firms because service territories, generation mixes, and regulatory recovery mechanisms differ. This motivates a conditional-variance model with issuer-specific exposure rather than a national climate-news index alone.

2.2 AI adoption, correlated decisions, and market instability

AI can reduce idiosyncratic forecast error while increasing commonality in beliefs and actions. The distinction between forecasting performance and market stability is central: a model can be individually useful yet systemically destabilizing if many agents rely on similar inputs, architectures, or recommendations. Existing work discusses algorithmic herding, performative feedback, and cognitive dependency, but direct identification is difficult because adoption is often latent and because measures of AI exposure overlap with technology-sector composition (Yang, 2026a).

Recent theory makes the nonlinearity explicit. Meng and Chen (2026) model performative prediction, correlated algorithmic signals, and cognitive dependency as mutually reinforcing channels, implying a convex relationship between adoption and systemic coupling. Laboratory-style evidence using language-model agents is more ambiguous: Hansen and Lee (2025) find that AI agents can rely more heavily on private information than human subjects in some settings, yet can also herd when profit-maximizing incentives make such behavior optimal. Taken together, these studies provide mechanism language rather than a ready-made field measure. The present paper therefore does not infer trading-system adoption; it studies a separate, observable infrastructure channel through which AI expansion can alter the cash-flow sensitivity of utilities.

Electricity demand supplies a more concrete adoption margin. A binding large-load agreement reflects infrastructure commitment and expected consumption rather than generic AI enthusiasm. It can alter the utility investment path, load forecast, financing need, and regulatory bargain before the facility is fully operational. At the same time, contracted megawatts are not equivalent to realized consumption. The empirical measure therefore distinguishes operational load, binding approved load, and requested pipeline capacity, and treats disclosure dates as the beginning of the public information set.

2.3 Conditional volatility and regimes

ARCH and GARCH models formalize volatility clustering through the conditional second moment (Engle, 1982; Bollerslev, 1986). The GJR specification allows negative innovations to have a different effect from positive innovations of the same magnitude (Glosten, Jagannathan, and Runkle, 1993). This asymmetry is essential here because an apparent climate-by-AI interaction could otherwise be a relabeled leverage effect. Mixed-frequency extensions show how slow-moving macroeconomic variables can enter longer-run volatility components (Engle, Ghysels, and Sohn, 2013), while Markov-switching models allow parameters or transition probabilities to vary across latent states (Hamilton, 1989; Gray, 1996).

The paper uses these tools sequentially. A single-state GJR-GARCH-X is the first estimable benchmark. A two-state model is considered only if the data support the interaction and the simpler model

passes diagnostics. The regime framework then asks whether dual stress changes volatility within a state, changes the probability of leaving the low state, or changes the probability of remaining in the high state. This sequencing protects the economic interaction from being manufactured by an unstable latent-state fit.

2.4 The unresolved intersection

The three literatures leave a specific empirical intersection unresolved. Climate-finance studies generally treat electricity demand and grid capacity as part of the background through which weather affects firms. AI-finance studies generally measure adoption through text, patents, occupational exposure, or technology-sector affiliation. Volatility studies can accommodate nonlinear states but cannot identify an economic interaction when the underlying exposures are not separately measured. Consequently, finding that both climate and AI variables predict volatility would not establish that they reinforce each other.

The paper closes this gap by defining an estimand before selecting a nonlinear model. The numerator of AI exposure is a dated physical commitment; the denominator is the capacity scale of the utility serving that load; climate stress is local to the operating territory; and the interaction is constructed only after each component passes an independent audit. The Markov state is then used to decompose the interaction, not to rescue it. This measurement-before-regime ordering is the paper's central methodological commitment.

3. Institutional Setting, Sample, and Measurement

3.1 Sample construction

The study begins with U.S. listed electric-utility parents from 1 January 2021 through 31 December 2025. The original thirteen-name tariff register remains part of the provenance record. The empirical portfolio is instead the ten-parent disclosure-positive register—AEE, AEP, AES, CMS, D, EXC, IDA, MDU, OTTR, and PPL—formed by applying the frozen operating-or-binding customer-load rule. This is a retrospectively audited research sample, not an ex ante investable index.

Layer	Definition	Advance rule
Returns	Daily adjusted-close log returns; portfolio and issuer panels	At least 1,000 valid returns per included issuer and at least 90% common-date coverage
Climate	Lagged service-territory cooling-load stress	No revised or future information; representative station and territory mapping audited
Data-center load	Operating or binding customer-load MW divided by matched system peak MW	At least 10 issuers with four consecutive quarterly observations
Joint support	Audited C and A with verified zero-versus-unknown status	At least 50 complete observations in every central 2×2 cell
Model	Single-state GJR-GARCH-X before any regime model	Stable optimum and acceptable residual diagnostics

3.2 Returns and market controls

The dependent variable is 100 times the daily log total return. The principal series is an equal-weighted portfolio of the ten eligible parents, rebalanced at the start of each quarter; weights drift with total returns within the quarter. Market controls are the SPY return, XLU return, VIX log return, and

natural-gas return. Lag-available firm controls are portfolio-weighted noncurrent debt/assets, cash/assets, quarterly capital expenditure/assets, and regulatory assets/assets. AES regulatory assets remain an annual disclosure carried as the latest public observation with its source period and age preserved.

3.3 Physical climate stress

The primary climate variable is heat or cooling-load stress derived from daily maximum and minimum temperature at a representative service-area station in GHCN-Daily (National Centers for Environmental Information, 2026). Daily average temperature is converted into cooling degree days relative to 65°F. The anomaly is computed against a 1991–2020 calendar-day baseline and scaled by the historical monthly standard deviation. Heat is primary because it is the physical channel that directly overlaps with incremental data-center cooling demand and therefore gives the climate-by-load interaction a single interpretable mechanism. Cold snaps, storms, disasters, and outage megawatts can affect load and reliability through different signs, durations, and operational bottlenecks; combining them into one index would mix mechanisms and add timing noise. They remain pre-specified complementary physical-risk tests. For AES Ohio, the representative station is Dayton International Airport (GHCN-Daily USW00093815).

3.4 Data-center power-demand exposure

For utility i in quarter q , data-center power-demand exposure is $A_{i,q} = MW_{i,q} / \text{Peak}_{i,q}$. The numerator includes operational or contractually binding customer-load megawatts that can be assigned to the matched utility; requested or speculative pipeline megawatts and generator-side power-purchase agreements remain separate. The denominator is the matched utility system peak from the most recent Form EIA-861 vintage publicly available at the information cutoff (U.S. Energy Information Administration, 2026). Exposure begins after the disclosure date. This variable measures disclosed data-center commitment relative to system scale; it is not realized electricity use or a direct measure of firm-level AI adoption.

Evidence family	Numerator object	Treatment
A: utility customer load	Operational or binding data-center customer-load MW	Primary exposure; divide by matched utility system peak load
A-pipeline	Requested, queued, or nonbinding load MW	Separate robustness measure; never combined with committed MW
B: generator PPA	Generation output or PPA MW linked to AI/data-center counterparty	Separate event/robustness family; no utility-load denominator
Tariff/regulatory event	Large-load tariff, docket, or order without customer MW	Institutional evidence only; not relabeled as AI adoption

3.5 The interaction and no-lookahead timing

The first data build standardized climate and load variables using expanding information sets and coded exposure as zero before the first qualifying disclosure. The audit shows that this zero means “no qualifying disclosure observed,” not verified absence. The primary M1 sample therefore treats those dates as unknown and retains only positive disclosed exposure. For numerical identification, C and A are demeaned in the fixed 212-day portfolio sample and the interaction $(C - \bar{C})(A - \bar{A})$ is divided by its sample standard deviation. Main effects remain in the variance equation. This centering convention was fixed

before the final profile and bootstrap calculations and is reported because GARCH-X persistence restrictions can otherwise depend on arbitrary covariate origins.

3.6 AES Ohio as a measurement example

AES Ohio illustrates the distinction between a source-qualified observation and a complete exposure series. The listed security is AES Corporation. Its regulated Ohio operating utility is The Dayton Power and Light Company, doing business as AES Ohio, and the corresponding EIA Form 861 utility identifier is 4922. The 31 October 2024 SEC-filed earnings release states that agreements for an additional 900 MW brought 2024 year-to-date data-center load growth at AES Ohio to 2.1 GW (AES Corporation, 2024). Because the disclosure identifies the utility, customer class, contractual status, quantity, and publication date, 2.1 GW enters family A from that date forward.

The observation is nevertheless handled conservatively. It is not assigned to quarters preceding the filing. A later 310 MW statewide retail-supply item is recorded separately because it describes a different object and cannot be added mechanically to the utility load total. The EIA denominator uses only Dayton Power & Light's reported system peak, not consolidated AES generation or non-Ohio operations. Daily climate exposure is represented by Dayton International Airport station USW00093815. This chain—listed parent, legal utility, load object, peak-load denominator, weather territory, and public date—is required for every load-side observation.

3.7 Network capacity and regulatory confounds

Data-center siting is not random. Developers may select utilities with spare reserve margin, favorable generation mix, uncongested transmission, faster interconnection processing, or a regulatory regime that permits timely cost recovery. Those same characteristics can affect heat-event resilience, outage exposure, capital spending, and the speed at which financial information is incorporated. Thus, residualizing AI exposure on size, capital expenditure, leverage, and broad region does not remove grid-level confounding. A positive climate-by-load coefficient could otherwise reflect pre-existing network quality or regulatory maturity.

Available firm controls do not remove non-random data-center siting. Reserve margin, generation mix, transmission constraints, interconnection processing, and regulatory cost recovery may affect both load placement and heat resilience. Coverage of these network variables is not sufficient for a causal design in the current sample. Accordingly, the M1 estimates are conditional associations and the failure to reject the null is not evidence that the underlying physical mechanism is absent.

3.8 Exposure concentration and leave-one-out diagnostics

The predeclared concentration rule remains relevant to any future expansion: no broad cross-utility claim should rest on one unusually complete issuer. In the present version, leave-one-issuer-out regime diagnostics are not activated because M1 fails before M2. The OTTR exclusion is nevertheless reported because its audited load is predominantly cryptocurrency hosting rather than verified AI demand.

Excluding OTTR from the legacy full-history portfolio reduces the uncentered interaction estimate from 0.143 to 0.060. Restricting the same legacy specification to the 212 common-positive dates reverses

the sign and creates boundary problems. These comparisons motivated the unknown-aware classification and centered profile analysis; they are sensitivity diagnostics rather than estimates selected for their sign.

3.9 Audited data-foundation checkpoint

Ten issuers satisfy the broad data-center load-side four-quarter evidence rule: MDU, AEP, AES, AEE, EXC, D, IDA, PPL, CMS, and OTTR. Disclosure entry dates range from 9 January 2023 for OTTR to 12 February 2025 for Dominion. The source register records issuer filings and regulatory materials, including the AEP Ohio tariff filing, Ameren and Exelon investor materials, Dominion slides, and MDU annual reports (AEP Ohio, 2024; Ameren Corporation, 2024; Dominion Energy, 2025; Exelon Corporation, 2024; MDU Resources Group, 2024, 2025). G1–G3 therefore pass for returns, broad load evidence, and climate coverage. The legacy 2×2 grid has a minimum cell of 555 observations, but that count uses pre-disclosure zeros. Once those observations are classified as unknown, it cannot serve as a verified low-exposure control grid; confirmatory G4 support is therefore unresolved. Table 5 reports the nonduplicated gate summary.

The denominator and climate mappings use the same operating utility and point-in-time public vintages. For Exelon, the qualifying 2024Q4–2025Q3 window uses ComEd's 2023 EIA vintage peak of 22,467 MW; the 2024 vintage peak of 21,560 MW enters only in 2025Q4. AEP Ohio uses 7,914, 8,104, and 9,014 MW for the successive 2022–2024 vintages after their release dates. NOAA temperature records are converted from tenths of degrees Celsius before computing Fahrenheit temperatures, cooling degree days, normals, and heat stress; nonblank quality flags are excluded.

4. Hypotheses

H1 — variance amplification. The climate-by-load interaction increases conditional variance, with the largest effect expected in the high-volatility state.

H2a — **state entry**. Dual stress lowers the probability of remaining in the low-volatility state.

H2b — **state persistence**. Dual stress raises the probability of remaining in the high-volatility state and increases expected high-state duration.

H3 — **exposure gradient**. The marginal response to climate stress is larger for the High-A portfolio than for the Low-A portfolio.

H1 is tested at M1 and reported regardless of sign. H2a, H2b, and H3 remain directional design commitments but are not estimated because the advance rule stops the sequence before M2. No transition or heterogeneity result is implied.

4.1 Mechanism-specific observable implications

The cash-flow channel predicts that the interaction is strongest when heat stress coincides with high committed-load intensity and when cost recovery is delayed or uncertain. The information channel predicts larger effects around disclosures that revise contracted load or capital requirements, but not around generic AI news unrelated to the utility territory. The system-pressure channel predicts that the interaction is more visible in the upper tail of conditional variance and in low-to-high transition probabilities than in the conditional mean.

These implications also define informative failures. A climate effect with no High-A gradient supports a general weather-risk channel but not the proposed interaction. An AI-load main effect that survives during low climate stress may indicate financing or growth uncertainty rather than dual pressure. A transition effect without stable state separation is treated as model failure. A significant interaction driven only by future exposure leads or by pipeline requests is classified as timing or measurement contamination.

5. Empirical Framework

5.1 Single-state benchmark

The M1 benchmark is a standardized Student-t GARCH-X model. Returns follow $r_t = \mu + \phi r_{t-1} + b'X_t + \varepsilon_t$, with $\varepsilon_t = \sigma_t z_t$. After the boundary audit sets the empirically negligible GJR leverage term to zero, the reported variance recursion is $\sigma_t^2 = (\omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2) \exp[\lambda^c(C_{t-1} - \bar{C}) + \lambda^A(A_{t-1} - \bar{A}) + \lambda^{cA}Q_{t-1}]$, where Q is the centered product divided by its sample standard deviation. The exponential term preserves positivity; λ^{cA} is the log-variance scaling associated with a one-standard-deviation change in the centered interaction coordinate.

5.2 Markov-switching extension

The planned two-state model remains a contingent extension. If M1 had passed, low- and high-volatility states would have been ordered by conditional variance and transition probabilities would have included the same centered exposure coordinates. Because M1 provides no incremental interaction support, estimating this more heavily parameterized model would violate the advance rule.

No transition coefficient, marginal effect, state occupancy, or expected duration is reported. This is a substantive stopping decision: latent states are not used to manufacture an interaction that the simpler observed-state benchmark does not support.

5.3 Estimation, inference, and diagnostics

M1 is estimated by conditional maximum likelihood from dispersed starts. The 212 observations form nine segments because a new segment is declared whenever one or more candidate trading dates are excluded by the common-positive or public-information rule. At each boundary, the conditional-variance recursion is reset to the same variance of the first 60 OLS residuals; σ^2 is not propagated across the omitted dates. The mean equation retains the last observed portfolio return as r_{t-1} , so the reset applies only to the latent variance state. The bootstrap uses the identical fixed reset vector and re-estimates the initialization variance in every draw. The boundary audit compares nonnegative free-GJR and gamma-fixed-zero specifications under three segment-initialization multipliers. An additional fixed-zero warm start corrected a local optimum in the null model. The centered specification then fixes gamma at zero, uses 20 dispersed starts for the interaction model, and verifies the Student-t likelihood independently.

Inference uses a 499-replication parametric bootstrap under the centered no-interaction model, conditional on the observed regressor matrix. Each simulated sample recomputes the first-60-observation OLS residual variance used for segment initialization and is fitted from two null and three interaction starts. All 499 replications produce valid solutions; the observed LR statistic is 1.638 and the plus-one

bootstrap p-value is 0.324. A descriptive profile-likelihood interval is also reported, but it is not treated as a substitute for the bootstrap. The fixed-design bootstrap conditions on the observed regressors and segment pattern; it is not a causal test or a regime-occupancy power analysis.

Stage	Specification	Question	Advance condition
M0	Event-window realized-volatility contrasts	Are sign and timing consistent with the physical mechanism?	Descriptive validation only
M1	Single-state GJR-GARCH-X	Is there stable additive or interactive variance scaling?	Convergence and residual diagnostics
M2	Two-state model, constant transitions	Is amplification concentrated in a high-volatility state?	Separated, occupied, stable states
M3	Time-varying transitions	Does dual stress affect entry and persistence?	Effective transition variation and bootstrap feasibility
M4	High-A/Low-A and bounded robustness families	Is the response exposure-specific and definition-robust?	Report regardless of sign

5.4 Finite-sample feasibility and parameter budget

The realized parameter budget is 16 free parameters in the centered no-interaction model and 17 in the interaction model, including ten mean parameters, the Student-t tail parameter, variance dynamics, and exposure slopes. The apparent daily panel is much larger than the identifying time series: the unknown-aware portfolio contains 212 observations across nine segments, while exposure changes occur only at disclosure and denominator-vintage dates. This distinction is central to feasibility.

Rather than report post hoc “observed power,” which is mechanically tied to the realized test statistic, the paper reports a bootstrap-dispersion sensitivity calculation. Under the fitted null, the standard deviation of the 499 unrestricted interaction estimates is 0.101. A local-normal approximation therefore gives an 80%-power minimum detectable absolute coefficient of $(1.96+0.842) \times 0.101 = 0.284$ at a two-sided 5% level. Because λ^{CA} is a log-variance multiplier, $|0.284|$ corresponds to approximately a 33% variance change per one-standard-deviation movement in the centered product. This threshold is economically material in the selected no-interaction model: its fitted conditional variance rises by about 41% from the sample 25th percentile (0.192) to the 75th percentile (0.272), and by about 76% from the 5th percentile (0.183) to the 95th percentile (0.322). It also exceeds the 95th percentile of within-segment one-day absolute log-variance changes (0.079, approximately 8%). A 33% interaction multiplier is therefore slightly smaller than the model’s interquartile cross-date variance spread but much larger than an ordinary one-day variance movement. The current design is informative about large amplification and remains weak for moderate effects. This is a diagnostic approximation conditional on the observed design, not a guarantee of power against every nonlinear alternative.

The originally planned M3 Monte Carlo exercise is not run because the sequence stops at M1. Its state-occupancy criteria cannot validate a regime model after the prerequisite observed-state interaction has failed. Any future return to M2/M3 requires a longer positive-exposure history or additional independently audited issuers, followed by a newly frozen feasibility design.

6. Identification and Reporting Design

Physical load intensity reduces overlap with a traded technology factor, but it does not create exogenous treatment. The ten-parent register measures data-center commitments, and OTTR is predominantly tied to cryptocurrency hosting. Market, sector, energy, and firm controls are included, while network capacity and regulatory confounds remain incompletely observed. The estimates are therefore descriptive conditional-volatility evidence.

Common-trend and disclosure problems are addressed by retaining the legacy coding only as a sensitivity case and treating pre-disclosure observations as unknown in the primary M1 sample. This choice reduces the usable portfolio history to 212 dates. The paper reports that loss of support rather than converting non-disclosure into a low-exposure control group.

The original central-cell count is not sufficient for confirmatory interaction inference because the low-A cells partly rely on unverified zeros. Before recentering in the unknown-aware sample, the product is highly aligned with the heat main effect: $\text{Corr}(C, C \times A) = 0.979$ and the interaction VIF is 29.4. After centering, $\text{Corr}(C - \bar{C}, Q) = -0.127$, $\text{Corr}(A - \bar{A}, Q) = -0.249$, and the interaction VIF falls to 1.09. Centering therefore removes the mechanical coordinate collinearity and the artificial coefficient-bound result, but it does not create new exposure dates or independent economic variation. Table 4A summarizes this sequence and shows that the 0.060 estimate is an OTTR-exclusion sensitivity, not an intermediate common-positive estimate. The remaining limitation is power and effective exposure support, not conventional linear multicollinearity in the centered regressors.

Table 4A. Measurement decisions and the interaction estimate

Stage	Exposure/sample decision	N	λ_{CA}	ΔAIC	VIF	Diagnostic interpretation
Legacy panel	Pre-disclosure silence coded zero; all ten issuers	1,118	0.143	-2.736	1.93	Pilot AIC favors interaction; no calibrated inference
Label sensitivity	Legacy coding; OTTR excluded	1,119	0.060	+1.097	1.59	AIC reverses; crypto-hosting label matters
Unknown-aware	All ten issuers positive; uncentered $C \times A$	212	-2.000	-0.037	29.39	Coefficient bound; $\text{Corr}(C, C \times A) = 0.979$
Final centered M1	Unknown-aware; centered and scaled product	212	-0.118	+0.362	1.09	Interior; bootstrap $p = 0.324$

Note. The rows are diagnostic specifications, not a sequence of confirmatory hypothesis tests. The -2.000 estimate is at the imposed coefficient bound and is not interpreted economically.

Model adequacy is evaluated across initialization multipliers, leverage restrictions, starts, information criteria, profile likelihood, and the conditional bootstrap. The no-interaction model is preferred under AIC and BIC in every boundary/initialization configuration. These diagnostics support the stop decision, while their fixed-design and short-sample limitations prevent a structural or causal conclusion.

The preserved data can support future work on heat-sensitive utility volatility, data-center load disclosure, denominator vintages, and event timing. A future interaction test should extend the post-disclosure history and obtain independently varying exposures. Storm/outage measures, generator PPAs, GARCH-MIDAS, Wasserstein diagnostics, and Markov switching remain separate extensions rather than specifications searched after observing M1.

7. Data Readiness and M1 Results

7.1 Audited measurement sample

The data audit yields ten broad data-center exposure sequences and 11,541 issuer-days with market, climate, and latest-public financial controls. Of the 12,540 issuer-day records in the underlying panel, 8,685 pre-disclosure zero codes are reclassified as unknown and 3,855 observations retain positive disclosed commitments. The requirement that all ten portfolio constituents have positive disclosed exposure and complete controls leaves 212 portfolio observations across nine variance-recursion segments from 13 February to 31 December 2025. Table 5 records how this audit changes the advance decision; Table 7 preserves issuer-level descriptive heterogeneity without treating it as an H3 test.

Table 5. Data-readiness and advance status

Gate	Observed evidence	Status
G1 Returns	Ten-issuer portfolio returns available; 212 unknown-aware common-positive dates	PASS for M1
G2 Load evidence	Ten broad data-center sequences; OTTR mainly crypto hosting; not ten AI-specific measures	PASS only for broad DC load
G3 Heat stress	Matched NOAA service-territory series; corrected 0.1°C units	PASS
G4 Joint support	Legacy minimum cell 555 relies partly on unverified pre-disclosure zeros	UNRESOLVED
G5 M1 Regime advance	Centered bootstrap LR=1.638, p=0.324; AIC/BIC prefer null All prerequisites required	FAIL STOP

Note. G2 validates a broad data-center load measure only. It does not certify ten AI-specific adoption or AI-load measures.

7.2 Single-state interaction evidence

The uncentered legacy estimates are sensitive to sample definition and coefficient bounds. Under the primary unknown-aware sample, centering C and A and scaling their product to unit standard deviation yields an interior interaction estimate of -0.118 . The interaction lowers negative log likelihood by 0.819, but the additional parameter raises AIC from 323.171 to 323.533 and BIC from 376.876 to 380.595. Thus both criteria select the no-interaction specification in this sample. The bootstrap p-value of 0.324 fails to reject the no-interaction model; it does not establish that the population interaction is zero.

Table 6. Centered Student-t GARCH-X comparison

Specification	NLL	Parameter s	AIC	BIC	λ_{CA}	Bootstrap result
No interaction	145.585	16	323.171	376.876	—	Reference
Centered interaction	144.767	17	323.533	380.595	-0.118	LR=1.638; p=0.324

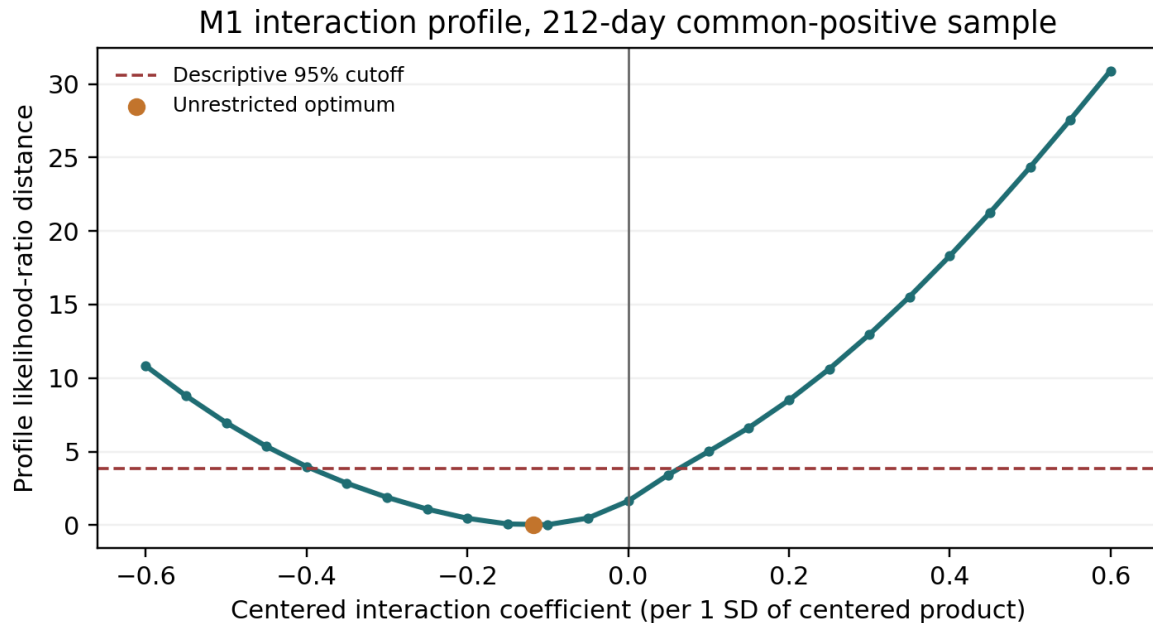


Figure 1. Profile likelihood for the centered interaction coefficient. The dashed line is the descriptive one-parameter 95% cutoff; the interpolated interval is $[-0.395, 0.064]$ and includes zero.

7.3 Interpretation and stopping rule

The estimates do not establish amplification, absorption, or a causal null. They show that the audited 2021–2025 sample lacks detectable incremental support for the proposed interaction after disclosure uncertainty and coordinate dependence are handled. This failure to reject is scientifically useful because it locates the present bottleneck: effective exposure variation, rather than conventional centered-regressor collinearity or optimizer failure. The approximate 80%-power threshold of $|0.284|$ indicates that economically smaller interactions remain unresolved. In accordance with the predeclared sequence, G5 is marked FAIL and M2/M3 are not estimated. Every source decision, denominator vintage, coding change, and optimization attempt is retained, allowing a longer post-disclosure panel to revisit the same hypothesis without redefining the evidence after seeing the result.

7.4 Issuer-level descriptive coverage

The portfolio test deliberately avoids treating issuer-days as independent replications of a common quarterly exposure. Table 7 nevertheless reports the cross-sectional scale hidden by aggregation. The wide range is economically plausible under the frozen definition because the numerator is disclosed operating or binding customer load and the denominator is the matched regulated utility's system peak—not the consolidated parent's size or current realized data-center consumption. At the low end, EXC uses 500 MW relative to Commonwealth Edison peaks of 22,467 MW through 2025Q3 and 21,560 MW in 2025Q4. At the high end, Dominion uses 14,000 MW relative to Virginia Electric & Power peaks of 17,957–18,409 MW; PPL uses 5,000 MW relative to 6,898–7,303 MW; and AES uses 2,100 MW relative to Dayton Power & Light peaks of 3,241–3,280 MW. MDU's range reflects a dated numerator change from 180 MW to 530 MW against a much smaller 652–675 MW utility peak. These ratios measure

prospective committed-load burden relative to system scale and can therefore be large without implying that the load was already operating at that share. The numerator/denominator pairs were rechecked against the point-in-time source register; no denominator remapping is introduced in this revision. The figures document heterogeneity in levels but do not estimate H3: disclosure windows differ, load objects are not uniformly AI-specific, and no issuer-level significance tests are used to revive a failed portfolio gate.

Table 7. Issuer-level positive-disclosure coverage and exposure intensity

Issuer	First positive date	Positive days	Mean A (%)	A range (%)	Mean r (%)
AEE	2024-11-08	286	3.39	3.37–3.47	0.82
AEP	2024-05-14	410	7.34	6.66–7.58	0.87
AES	2024-11-01	291	64.64	64.02–64.79	2.02
CMS	2024-11-01	291	3.06	3.06–3.08	0.84
D	2025-02-13	222	77.45	76.05–77.96	1.01
EXC	2024-11-11	285	2.25	2.23–2.32	0.90
IDA	2023-12-29	503	3.87	3.69–3.92	0.82
MDU	2024-02-23	466	52.23	27.56–81.28	1.15
OTTR	2023-01-10	747	10.45	10.12–11.04	1.22
PPL	2024-08-05	354	71.35	68.47–72.48	0.84

Note. A is committed data-center customer-load MW divided by the matched utility peak-load denominator. “Positive days” counts disclosed-positive issuer-days through 31 December 2025. Mean |r| is descriptive and is not adjusted for market or firm controls.

8. Conclusion

This paper asks whether disclosed data-center electricity demand changes how local heat stress is translated into utility-stock volatility. The immediate answer is deliberately limited: the present common-positive sample does not detect incremental amplification. The durable contribution is that the study makes the test well defined. It provides a point-in-time chain linking parent securities, operating utilities, customer-load evidence, system-peak denominators, weather territories, returns, and lag-available financial controls. It also demonstrates that broad data-center exposure is not uniformly AI-specific, that pre-disclosure silence is not a verified zero, and that thousands of issuer-days can collapse to only 212 portfolio observations across nine variance-recursion segments once common information is required.

The M1 evidence fails to detect a climate-by-load amplification effect; it is not evidence that the effect is absent. H3 remains untested. The positive empirical lesson is that measurement choices determine whether an apparent interaction exists. Reclassifying disclosure silence as unknown changes the admissible sample, while centering reduces the interaction VIF from 29.4 to 1.09 and removes an artificial coefficient-bound result. AIC and BIC then select the no-interaction specification, the bootstrap does not reject it, and the approximate 80%-power threshold is $|0.284|$ log-variance units. That threshold is slightly below the fitted null model’s 41% interquartile cross-date variance spread, showing that the design can detect only economically large amplification with reasonable power. Stopping before regime switching is consequently part of the contribution rather than an unfinished analysis. The frozen evidence register and estimation archive create a reusable baseline: when longer disclosure histories or more independently varying audited exposures become available, the same gate can be rerun and the nonlinear extension can proceed without changing the original definitions. The paper therefore converts an inconclusive interaction estimate into a transparent measurement result, a falsifiable update rule, and a durable empirical asset.

Declarations

Conflict of interest. The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics approval. Not applicable. The study uses publicly available market, regulatory, corporate-disclosure, energy-system, and weather data and involves no human participants.

Data availability. The frozen replication archive is `Climate_AI_GARCH_Reusable_Data_v0.21`, comprising core processed data and code, SEC evidence, NOAA/EIA source material, and SHA-256 manifests. M1 sensitivity and centered-profile supplements v0.22–v0.23 preserve every optimizer start, likelihood check, bootstrap draw, and reporting input. The SSRN v1.2 closeout package preserves the centered VIF calculation, segment map, minimum-detectable-effect and conditional-variance calibration, issuer-level numerator/denominator audit, citation crosswalk, and the exact upload manuscript. Public-source terms and redistributability remain those of the original providers.

Use of generative AI. Generative AI tools assisted language editing, source triage, code scaffolding, and bibliographic cross-checking. They did not generate or impute observations, decide which megawatt evidence qualified, recode failed observations to obtain a preferred sign, select specifications after viewing significance, or authorize advancement after a failed gate. The author manually reviewed source selection, variable definitions, estimation diagnostics, calculations, citations, and claims.

References

- Acharya, V. V., Johnson, T. C., Sundaresan, S. M., and Tomunen, T. (2022). Is physical climate risk priced? Evidence from regional variation in exposure to heat stress. NBER Working Paper No. 30445. <https://doi.org/10.3386/w30445>.
- Addoum, J. M., Ng, D. T., and Ortiz-Bobea, A. (2023). Temperature shocks and industry earnings news. *Journal of Financial Economics*, 150(1), 1–45. <https://doi.org/10.1016/j.jfineco.2023.07.002>.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1).
- Bolton, P., and Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics*, 142(2), 517–549. <https://doi.org/10.1016/j.jfineco.2021.05.008>.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007. <https://doi.org/10.2307/1912773>.
- 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.
- Giglio, S., Kelly, B., and Stroebe, J. (2021). Climate finance. *Annual Review of Financial Economics*, 13, 15–36. <https://doi.org/10.1146/annurev-financial-102620-103311>.

- 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>.
- Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1), 27–62. [https://doi.org/10.1016/0304-405X\(96\)00875-6](https://doi.org/10.1016/0304-405X(96)00875-6).
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384. <https://doi.org/10.2307/1912559>.
- Hansen, A. L., and Lee, S. J. (2025). Financial stability implications of generative AI: Taming the animal spirits. Finance and Economics Discussion Series 2025-090. Board of Governors of the Federal Reserve System. <https://doi.org/10.17016/FEDS.2025.090>.
- Meng, S., and Chen, X. (2026). Artificial intelligence and systemic risk: A unified model of performative prediction, algorithmic herding, and cognitive dependency in financial markets. arXiv:2604.03272. <https://doi.org/10.48550/arXiv.2604.03272>.
- Shehabi, A., Smith, S. J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M. A. B., Holecek, B., Koomey, J. G., Masanet, E., and Sartor, D. (2024). 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, LBNL-2001637. <https://doi.org/10.71468/P1WC7Q>.
- Yang, F. (2026a). Artificial intelligence and financial volatility: From model augmentation to market disruption—A systematic critical review. SSRN. <https://doi.org/10.2139/ssrn.7196800>.
- Yang, F. (2026b). Where public climate signals remain visible: Issuer compression and debt-rating sorting in China. *Journal of Financial Reporting and Accounting*, forthcoming.

Primary data and disclosure sources

- AES Corporation (2024). Third Quarter 2024 Earnings Release, filed 31 October 2024. U.S. Securities and Exchange Commission.
- U.S. Energy Information Administration (2026). Annual Electric Power Industry Report, Form EIA-861, 2001–present.
- National Centers for Environmental Information (2026). Global Historical Climatology Network Daily (GHCN-Daily).
- AEP Ohio (2024). AEP Ohio Files Plan to Secure Grid Resources for Data Centers, Protect Residential Customers, 13 May 2024; Public Utilities Commission of Ohio Case No. 24-0508-EL-ATA.
- Ameren Corporation (2024). Third Quarter 2024 Earnings Presentation, 7 November 2024.
- Dominion Energy, Inc. (2025). Fourth Quarter 2024 Earnings Call Slides, 12 February 2025.
- MDU Resources Group, Inc. (2024, 2025). Annual Reports on Form 10-K for fiscal years 2023 and 2024.
- Exelon Corporation (2024). Investor Meetings, November 2024, slide 16: Historical Data Center Load in Northern Illinois.

Factor Covariance in Yield-Curve Risk Forecasts: Evidence from U.S. Treasury Yields

Fuli Yang* Shiyu Lin†

GARCH Institute, Harvard Square, Cambridge, MA 02138, USA

3 September 2026

Abstract

Yield-curve risk is often assembled in two steps: estimate separate volatilities for Nelson–Siegel factors, then map those volatilities back to individual maturities. The convenient diagonal map silently treats level, slope, and curvature innovations as conditionally uncorrelated. We show that this is not a harmless simplification for U.S. Treasury yields. Negative level–slope comovement makes diagonal forecasts too large at the front end, with a much smaller effect at long maturities. In a strictly out-of-sample exercise from 2008 to 2026, rolling CCC- and DCC-GARCH forecasts that retain factor covariance reduce QLIKE loss relative to diagonal GARCH at horizons of one, five, and twenty trading days. At two years, DCC lowers mean QLIKE by 0.77, 0.82, and 0.72, respectively, and reduces mean-squared error by 35–60 percent. The advantage narrows sharply with maturity. Two-sided interval exceedances provide the same diagnosis: the diagonal model covers too often because its intervals are excessively wide, whereas covariance-aware forecasts are much closer to nominal coverage. The improvement is not universal—it weakens at longer maturities and reverses for some twenty-day forecasts during the 2022–2023 tightening—but its maturity pattern follows directly from the Nelson–Siegel loadings. Factor covariance is therefore part of the forecast object, not a dispensable refinement to marginal volatility.

Keywords: Nelson–Siegel; yield-curve risk; multivariate GARCH; covariance forecasting; Treasury yields.

JEL classification: C32; E43; G12; G17.

1 Introduction

Many yield-curve risk systems are modular. A curve is compressed into level, slope, and curvature factors; a volatility model is fitted to each factor; and the resulting variances are mapped back to two-, five-, or ten-year yields. Each step is familiar. The problem lies in the join between them. Three well-estimated marginal variances do not form a covariance matrix, and a maturity is exposed to a linear combination of all three factors. Setting the missing off-diagonal terms to zero changes the risk forecast itself.

This paper asks whether that change matters outside the estimation sample. The question is deliberately narrower than choosing a globally optimal term-structure model. We hold the Nelson–Siegel representation and the marginal GARCH models fixed, and vary only the treatment of factor covariance. The benchmark discards covariance. Constant conditional correlation (CCC) restores a rolling correlation matrix. Dynamic conditional correlation (DCC) also lets that matrix move with incoming shocks. A backward-looking rolling-sample-covariance estimator provides a model-light comparison. We evaluate all four forecasts against subsequent changes in the *observed* Treasury constant-maturity yields, rather than against fitted Nelson–Siegel yields.

The answer is economically large and statistically visible at the front end. In the descriptive 2003–2026 sample, the level–slope innovation correlation is -0.84 . Because the Nelson–Siegel level and slope loadings are both positive, the omitted cross term is negative in the full quadratic form. Removing it raises the mapped variance. Holding marginal factor variances fixed, the diagonal-to-full conditional-variance ratio has a median of 2.90 at two years, 1.76 at five years, 1.37 at ten years, and 1.11 at thirty years.

That accounting result survives a harder forecasting test. We estimate models on rolling five-year windows, refit every 63 observations, and generate daily forecasts from January 2008 through August 2026. At two years, DCC-GARCH lowers average QLIKE loss relative to diagonal GARCH by 0.765 at one day, 0.823 at five days, and 0.724 at twenty days. The corresponding mean-squared-error ratios are 0.646, 0.537, and 0.396. CCC also improves on the diagonal benchmark, though less than DCC. The gains become smaller with maturity, just as the loading algebra predicts.

*Corresponding author. Email: dr@garch.ac.

†ORCID: [0009-0000-2471-3644](https://orcid.org/0009-0000-2471-3644).

Interval coverage makes the practical meaning transparent. For a nominal 5 percent two-sided exceedance rate, the one-day diagonal forecast produces only 0.8 percent exceedances at two years. It is not prudently calibrated; it is much too wide. CCC and DCC produce 5.1 and 4.9 percent. At twenty days the corresponding rates are 0.9, 3.9, and 5.2 percent. Covariance-aware models do not dominate every cell—five- and ten-year one-day intervals still undercover, and long-horizon improvements are imprecise at ten years—but they remove the diagonal model’s most pronounced front-end distortion.

We make three contributions. First, we isolate an exact and auditable channel through which factor dependence becomes maturity-specific risk error. Existing work establishes the usefulness of Nelson–Siegel factors and time-varying second moments; we focus on the final mapping from factor covariance to tenor risk. Second, we show that the mapping error has out-of-sample consequences under identical marginal volatility forecasts. This distinguishes a genuine covariance result from a comparison in which one model also changes the factor variances. Third, we document the boundary of the result. The benefit is strongest from two to five years, modest at ten to thirty years, and not uniformly positive during the 2022–2023 tightening. The paper therefore argues for preserving covariance, not for treating DCC as a universal winner.

The design also rules out two broader interpretations. This is not a bond-return predictability study: no expected-return model enters. It is not a portfolio-timing exercise: there is no allocation rule whose performance can hide a persistent risk bias. The empirical object is the conditional variance of yield changes at a stated tenor and horizon. Keeping that object narrow makes the comparison falsifiable.

The rest of the paper proceeds as follows. Section 2 links the question to term-structure and volatility forecasting. Section 3 defines the covariance map, competing forecasts, and real-time protocol. Section 4 reports the anatomy of covariance omission. Section 5 presents out-of-sample losses, interval coverage, and episode-level results. Section 6 discusses robustness and limitations. Section 7 concludes.

2 Related literature and empirical prediction

The level, slope, and curvature interpretation of Treasury yield movements predates the Nelson–Siegel forecasting literature and appears clearly in the principal-component evidence of [Litterman and Scheinkman \(1991\)](#). [Nelson and Siegel \(1987\)](#) supplied a parsimonious parametric loading system, and [Diebold and Li \(2006\)](#) gave its factors a dynamic forecasting interpretation. Subsequent work allowed the loading decay and factor dynamics to vary over time ([Koopman et al., 2010](#)), while [Christensen et al. \(2011\)](#) derived an affine arbitrage-free counterpart. These contributions make a low-dimensional risk representation natural, but none implies that factor innovations are independent. Our fitted-factor design is intentionally less structural: it isolates the second-moment mapping rather than asking the covariance model to repair pricing restrictions in the curve specification.

Term-structure models with stochastic volatility explicitly recognize time variation in second moments. [Hautsch and Yang \(2012\)](#) embeds stochastic volatility in a Nelson–Siegel system, while [Hautsch and Ou \(2012\)](#) studies interest-rate risk under stochastic factor volatility. [Shin and Zhong \(2017\)](#) show that realized-volatility information can improve bond-yield density forecasts. A parallel risk-management literature evaluates interest-rate covariance forecasts by the portfolio and interval decisions they support ([Ferreira and Lopez, 2005](#)). Our question is more operational and more tightly controlled. Once a researcher or risk manager has three factor volatility forecasts in hand, can the covariances be dropped when the forecasts are projected back to maturities? This question arises even in systems that do not attempt a fully joint state-space estimation.

The multivariate-GARCH literature provides several ways to retain dependence; [Bauwens et al. \(2006\)](#) survey the central specifications and their trade-offs. CCC imposes a fixed conditional correlation matrix while allowing marginal variances to change ([Bollerslev, 1990](#)). DCC permits parsimonious correlation dynamics ([Engle, 2002](#)). Its standard two-step estimator also has known interpretation and consistency limitations, motivating corrected formulations ([Aielli, 2013](#)). We therefore use DCC as a parsimonious forecasting competitor, not as a structural law for factor dependence. Comparing it with CCC separates two propositions: whether nonzero correlation matters at all, and whether allowing it to evolve adds further value.

Forecast evaluation follows two principles. First, the realized variance proxy can be noisy, so the loss function should not mechanically reward a model for matching that noise. The broad volatility-forecast comparisons of [Hansen and Lunde \(2005\)](#) demonstrate why out-of-sample ranking can depend on the chosen loss and proxy. QLIKE is robust to conditionally unbiased volatility proxies and is therefore the primary loss ([Patton, 2011](#)). For multivariate volatility forecasts, [Laurent et al. \(2012, 2013\)](#) likewise emphasize formal out-of-sample comparison and loss-function consistency. Mean-squared error is reported because scale errors are economically interpretable. Equal-loss inference follows the forecast-comparison logic of [Diebold and Mariano \(1995\)](#), with a HAC adjustment for overlapping horizons. Second, a risk model should also be judged by the events it is designed to cover. We report unconditional and independence diagnostics for two-sided forecast intervals following [Kupiec \(1995\)](#) and [Christoffersen \(1998\)](#).

The loading structure yields a directional prediction before any forecast is run. With positive level and slope loadings and negative level–slope covariance, deleting the cross term inflates tenor variance. Since the slope loading decreases with maturity, the inflation should be largest at short maturities and decay along the curve. If that mechanism is relevant out of sample, the loss advantage of covariance-aware forecasts should display the same maturity gradient.

3 Data and empirical design

3.1 Treasury yields and Nelson–Siegel factors

The data are daily Federal Reserve H.15 constant-maturity Treasury yields obtained through FRED. The cross section contains 3- and 6-month yields and 1-, 2-, 3-, 5-, 7-, 10-, 20-, and 30-year yields. The descriptive sample runs from 2 January 2003 through 27 August 2026. The out-of-sample evaluation begins on 2 January 2008, leaving approximately five years for the first estimation window. Yield units are percentage points.

For maturity τ in years, the Nelson–Siegel curve is

$$y_t(\tau) = \beta_{0t} + \beta_{1t} \frac{1 - e^{-\lambda\tau}}{\lambda\tau} + \beta_{2t} \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right) = \mathbf{\Lambda}(\tau)' \boldsymbol{\beta}_t, \quad (1)$$

where $\lambda = 0.7308$ in the main specification. On every date with at least six observed maturities, we estimate $\boldsymbol{\beta}_t$ by cross-sectional ordinary least squares. The risk innovations are daily factor changes $\Delta\boldsymbol{\beta}_t$.

This fitted-factor approach is intentionally transparent. It avoids claiming that the three factors are observed without error, but it also keeps the covariance calculation reproducible from public data. The realized forecast target is not the fitted curve. For each reported tenor, we square the subsequent change in the corresponding raw CMT series. This choice asks the model to forecast an external observation rather than its own projection residual.

3.2 The covariance map

Let $\mathbf{H}_t = \text{Var}_t(\Delta\boldsymbol{\beta}_t)$ be the conditional factor covariance matrix. A one-day tenor-variance forecast is the quadratic form

$$V_t^{\text{full}}(\tau) = \mathbf{\Lambda}(\tau)' \mathbf{H}_t \mathbf{\Lambda}(\tau). \quad (2)$$

The diagonal forecast preserves the three conditional marginal variances while setting covariances to zero:

$$V_t^{\text{diag}}(\tau) = \sum_{i=0}^2 \Lambda_i(\tau)^2 H_{ii,t}. \quad (3)$$

Their difference is not an approximation residual of unknown origin. It is exactly

$$B_t(\tau) = V_t^{\text{diag}}(\tau) - V_t^{\text{full}}(\tau) = -2 \sum_{i < j} \Lambda_i(\tau) \Lambda_j(\tau) H_{ij,t}. \quad (4)$$

Equation (4) is useful because it separates magnitude from sign. A diagonal map is not inherently conservative. It is conservative in this sample only because the dominant omitted covariance is negative and multiplies positive loadings.

For $h \in \{1, 5, 20\}$ trading days, each GARCH model produces a path of factor covariance forecasts $\{\widehat{\mathbf{H}}_{t+k|t}\}_{k=1}^h$. We sum their tenor projections,

$$\widehat{V}_{t,h}(\tau) = \sum_{k=1}^h \mathbf{\Lambda}(\tau)' \widehat{\mathbf{H}}_{t+k|t} \mathbf{\Lambda}(\tau), \quad (5)$$

which treats forecasted daily innovations as serially uncorrelated. The realized proxy is $[y_{t+h}(\tau) - y_t(\tau)]^2$ from the raw CMT series.

3.3 Competing covariance forecasts

All GARCH specifications share zero-mean Student- t GARCH(1,1) marginal models for the three factor changes. The four forecasts differ only in how dependence is handled.

1. *Diagonal GARCH* sets the conditional correlation matrix to the identity.

2. *CCC-GARCH* combines the same marginal variances with the correlation matrix of standardized residuals in the current estimation window.
3. *DCC-GARCH* updates a correlation state $\mathbf{Q}_t$ according to

$$\mathbf{Q}_t = (1 - a - b)\bar{\mathbf{Q}} + a\mathbf{z}_{t-1}\mathbf{z}'_{t-1} + b\mathbf{Q}_{t-1}, \quad (6)$$

then standardizes $\mathbf{Q}_t$ into a correlation matrix. The parameters a and b are estimated by Gaussian correlation QML after fitting the marginal models.

4. *Rolling sample covariance* uses the sample covariance of factor changes over the previous 60 observations and scales its mapped daily variance by h .

The rolling-sample-covariance forecast is valuable precisely because it is not another elaborate parametric model. If it produces the same maturity ranking, the result is less likely to be an artifact of CCC or DCC likelihood choices. We reserve “realized volatility” for estimators constructed from higher-frequency observations; the present benchmark is simply a backward-looking daily sample covariance.

3.4 Real-time protocol and forecast evaluation

The first forecast uses the preceding 1,260 factor changes. Parameters are re-estimated every 63 observations on a rolling 1,260-observation window. Between refits, GARCH variances and DCC states are updated daily using only innovations already observed. No pre-2003 observations enter, and no full-sample correlation estimate enters the out-of-sample exercise. The protocol produces 4,667 one-day forecasts; slightly fewer observations are available at longer horizons.

The primary loss is

$$L_{t,h}^Q = \log \hat{V}_{t,h} + \frac{RV_{t,h}}{\hat{V}_{t,h}}, \quad (7)$$

where $RV_{t,h}$ is the squared raw-yield change. Lower QLIKE is better. We also report MSE and the ratio of each model’s MSE to diagonal GARCH. For loss differences, heteroskedasticity-and-autocorrelation-consistent standard errors use a Bartlett kernel with lag $\max(5, 2h)$ (Diebold and Mariano, 1995; Newey and West, 1987). This adjustment is essential because the five- and twenty-day forecast outcomes overlap.

For coverage, the models define symmetric Gaussian intervals with half-width $z_{0.975}\sqrt{\hat{V}_{t,h}}$. The distributional assumption is deliberately held fixed across models so that coverage differences reflect variance forecasts rather than different tail quantiles. Multi-day coverage tests use nonoverlapping observations. We report the observed exceedance frequency and p -values from unconditional-coverage and exceedance-independence tests.

4 The anatomy of covariance omission

The full-sample CCC correlations are

$$\text{Corr}(\Delta\beta_0, \Delta\beta_1) = -0.84, \quad \text{Corr}(\Delta\beta_0, \Delta\beta_2) = -0.02, \quad \text{Corr}(\Delta\beta_1, \Delta\beta_2) = -0.22.$$

The level-slope term is therefore the natural place to look for the maturity gradient.

Table 1: The measurement effect of omitting factor covariance

Tenor	Mean V^{full}	Mean V^{diag}	Med($V^{\text{diag}}/V^{\text{full}}$) [95% interval]	Rolling 60d median
2 years	0.00283	0.00752	2.897 [2.784, 3.006]	3.247
5 years	0.00320	0.00554	1.760 [1.736, 1.782]	1.785
10 years	0.00294	0.00407	1.375 [1.367, 1.381]	1.362
30 years	0.00313	0.00348	1.109 [1.107, 1.111]	1.101

Notes: 5,917 daily factor changes, 3 January 2003–27 August 2026. The CCC- t comparison holds marginal conditional variances fixed. Intervals use 5,000 circular moving-block resamples of length 60, conditional on the fitted full-sample model. The last column uses only the preceding 60 factor changes. Variance units are squared percentage points.

Table 1 shows that the diagonal map is not a small perturbation. At two years, it produces almost three times the variance of the full map on the median day. The ratio falls monotonically and is close to one by thirty years. The backward-looking estimator yields essentially the same ordering, with an even larger median ratio at two years.

The mechanism is also stable through time. Figure 1 plots the level–slope correlation estimated in each rolling out-of-sample window. It remains negative throughout and ranges from approximately -0.97 to -0.63 . This persistence explains why the diagonal error is not confined to a single policy episode.

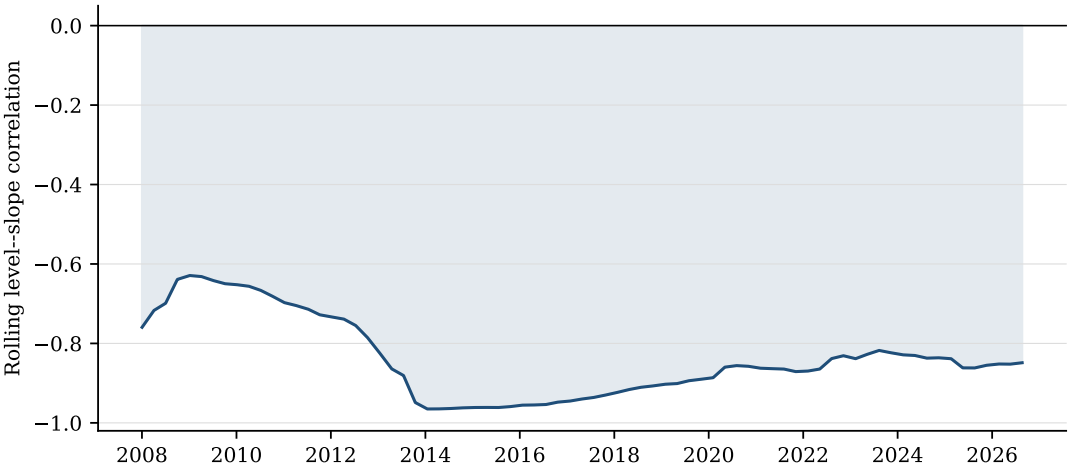


Figure 1: Rolling level–slope correlation. Each point uses standardized marginal-GARCH residuals from the preceding 1,260 observations at an out-of-sample refit date.

Decomposing Eq. (4) confirms that level–slope covariance accounts for about 79 percent of the mean two-year diagonal-minus-full difference. The level–curvature and slope–curvature terms are smaller. The absolute level–slope contribution declines from 0.00372 squared percentage points at two years to 0.00032 at thirty years. This is the algebraic source of the forecast prediction tested next.

5 Out-of-sample results

5.1 Forecast losses

Table 2 is the main result. Panel A reports each covariance-aware model’s mean QLIKE difference relative to diagonal GARCH; negative entries indicate improvement. Panel B reports MSE ratios, with values below one favoring the covariance-aware model.

Table 2: Out-of-sample variance-forecast performance, 2008–2026

h	Tenor	Panel A: QLIKE difference vs. diagonal			Panel B: MSE ratio vs. diagonal		
		CCC	DCC	Rolling 60d	CCC	DCC	Rolling 60d
1	2y	−0.528***	−0.765***	−0.743***	0.678	0.646	0.670
1	5y	−0.096***	−0.145***	−0.102***	0.897	0.872	0.886
1	10y	−0.021**	−0.035***	−0.018	0.973	0.960	0.967
1	30y	−0.014***	−0.017***	−0.010	0.984	0.974	0.988
5	2y	−0.554***	−0.823***	−0.879***	0.553	0.537	0.510
5	5y	−0.110***	−0.170***	−0.159***	0.784	0.742	0.720
5	10y	−0.031*	−0.052***	−0.052**	0.933	0.899	0.892
5	30y	−0.021***	−0.025***	−0.028***	0.969	0.952	0.963
20	2y	−0.447***	−0.724***	−0.806***	0.423	0.396	0.346
20	5y	−0.104*	−0.154**	−0.142*	0.747	0.682	0.630
20	10y	−0.024	−0.051	−0.051	0.941	0.890	0.867
20	30y	−0.023**	−0.028***	−0.039**	0.975	0.952	0.963

Notes: Forecasts use rolling 1,260-observation windows and quarterly parameter refits. Realizations are squared changes in raw CMT yields. Negative QLIKE differences and MSE ratios below one favor the column model. HAC inference uses lag $\max(5, 2h)$. *, **, and *** denote 10, 5, and 1 percent significance for equal mean QLIKE loss.

Three features matter. First, retaining covariance improves the two-year forecast in every horizon and loss function. The DCC MSE reduction grows from 35 percent at one day to 60 percent at twenty days. Second, DCC typically improves on CCC, especially at the front end. Correlation dynamics add information beyond a nonzero rolling correlation matrix. Third, the maturity gradient is steep. Improvements remain detectable at five years, are much smaller at ten years, and are economically modest at thirty years even when statistically significant.

The rolling-sample-covariance benchmark is competitive. It records the lowest two-year MSE at five and twenty days and the largest twenty-day QLIKE improvement. That result is useful rather than embarrassing: the paper’s claim concerns preserving covariance, not the superiority of a particular parametric estimator. Both the parametric and model-light routes reject the diagonal restriction.

Figure 2 plots the running mean DCC-minus-diagonal QLIKE difference for twenty-day forecasts. The two-year line stays well below zero across most of the evaluation. The long-maturity lines lie much closer to zero. The graph also reveals a deterioration around the recent tightening cycle, which motivates the episode analysis below.

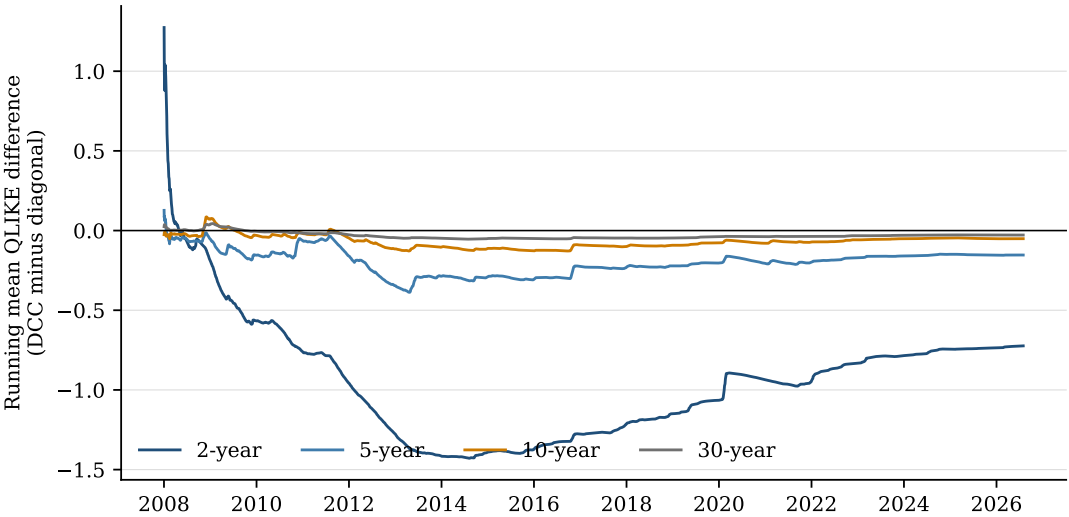


Figure 2: Running mean QLIKE difference for twenty-day forecasts. Negative values favor DCC-GARCH. Outcomes overlap; statistical inference is reported in Table 2 using HAC standard errors.

5.2 Forecast-interval coverage

Average losses do not show whether a model is systematically too wide or too narrow. Table 3 reports two-sided exceedance rates for nominal 95 percent intervals.

Table 3: Two-sided 95 percent forecast-interval exceedance rates

Horizon	Model	2 years	5 years	10 years	30 years
1 day	Diagonal GARCH	0.8***	2.0***	3.0***	3.4***
	CCC-GARCH	5.1	6.7***	6.2***	4.3**
	DCC-GARCH	4.9	6.9***	5.8***	4.2***
20 days	Diagonal GARCH	0.9***	1.7***	3.0	1.7***
	CCC-GARCH	3.9	7.3	6.4	3.0
	DCC-GARCH	5.2	6.9	7.3	3.0

Notes: Entries are percentages. The nominal exceedance rate is 5 percent. Twenty-day tests use every twentieth forecast to avoid overlapping coverage events. Stars denote rejection of correct unconditional coverage using the Kupiec test: ** 5 percent and *** 1 percent. The same Gaussian critical value is used for all models.

The diagonal model’s front-end intervals are far too wide. Its one-day two-year interval has a mean half-width of 0.165 percentage points, compared with 0.089 for DCC, and records fewer than one fifth of the intended exceedances. At twenty days the mean half-width is 0.745 percentage points under the diagonal model and 0.405 under DCC. DCC restores the two-year exceedance rate almost exactly to nominal at both horizons.

Coverage elsewhere is not perfect. At one day, CCC and DCC intervals are somewhat too narrow at five and ten years, while thirty-year intervals are slightly wide. At twenty days, sample sizes are much smaller after nonoverlapping selection and the tests have correspondingly low power. The important point is narrower: treating negative factor covariance as zero creates severe and predictable front-end overcoverage.

5.3 Episode dependence

The full-sample averages could conceal a single crisis payoff. We therefore report twenty-day DCC-minus-diagonal QLIKE differences for the global financial crisis, the 2020 pandemic, the 2022–2023 tightening, and all other dates.

Table 4: Twenty-day DCC QLIKE difference across episodes

Episode	2 years	5 years	10 years	30 years
Global financial crisis	−0.728	−0.145	0.034	0.026
Pandemic	−1.498	−0.857	−0.360	−0.089
2022–2023 tightening	0.330	0.103	0.092	0.032
Other dates	−0.820	−0.148	−0.057	−0.036

Notes: Entries are mean QLIKE for DCC-GARCH minus mean QLIKE for diagonal GARCH. Negative values favor DCC. The global financial crisis is September 2008–June 2009; the pandemic period is March–December 2020; tightening is January 2022–December 2023.

The covariance gain is not merely a crisis effect. At two years it remains large on dates outside the three named episodes. It is especially strong during the pandemic, when factor dependence moved quickly and the dynamic correlation model had room to adjust. During the 2022–2023 tightening, however, DCC loses to the diagonal model at every reported maturity for twenty-day QLIKE. This reversal does not overturn the full-sample result, but it warns against interpreting a historically negative level–slope covariance as a fixed structural hedge. Rapid policy repricing can alter both the persistence of correlation and the serial dependence ignored in Eq. (5).

6 Robustness, scope, and limitations

The original covariance measurement is stable to three checks. A purely backward-looking 60-day covariance produces the same sign and maturity ranking. Splitting the descriptive sample into 2003–2012, 2013–2019, and 2020–2026 yields two-year median diagonal/full ratios of 2.86, 3.17, and 2.53. Re-estimating factors and volatility models over $\lambda \in \{0.40, 0.60, 0.7308, 1.00\}$ gives two-year ratios between 2.69 and 3.29, while thirty-year ratios remain between 1.09 and 1.17. These exercises show that the front-end concentration is not created by one decay calibration.

The out-of-sample design is deliberately harder than the descriptive calculation. All correlation estimates are rolling; the raw Treasury yield, not the NS fitted value, supplies the realization; and the model is evaluated at three pre-specified horizons. The comparable performance of the rolling 60-day sample covariance also shows that the main finding is not dependent on a particular DCC optimizer.

Several limitations remain. First, Eq. (5) sums daily conditional variances and excludes autocovariances of factor changes. That approximation is common but can matter at the twenty-day horizon, particularly during rapid tightening. A joint model for factor means and covariance would address it, at the cost of changing more than the off-diagonal risk object. Second, the Student- t assumption enters the marginal GARCH estimates, but interval coverage is compared using a common Gaussian critical value. This choice isolates variance calibration; it is not a claim that Gaussian tails are adequate for Treasury changes. Third, CMT yields are interpolated benchmark series rather than tradable zero-coupon yields. Replication with zero-coupon curves and intraday realized measures would sharpen the economic interpretation.

Finally, a larger variance estimate is not automatically a safer risk estimate. The diagonal model appears conservative because it overstates front-end variance in this sample. Yet the cost is distorted relative risk across maturities and intervals that cover far more often than intended. A risk manager can choose an explicit capital buffer after obtaining a calibrated forecast. Embedding a maturity-dependent buffer accidentally by deleting covariance is harder to defend.

7 Conclusion

The central empirical fact in this paper is simple: Nelson–Siegel factor covariance survives the map back to maturity. For U.S. Treasury yields, strongly negative level–slope dependence subtracts substantial variance at the front end. A diagonal map deletes that offset, inflating two-year risk far more than thirty-year risk.

The out-of-sample evidence shows that this is more than descriptive accounting. Rolling CCC, DCC, and sample-covariance forecasts generally reduce QLIKE and MSE relative to diagonal GARCH, with the largest

gains at two years. DCC also converts severely overwide front-end forecast intervals into intervals close to nominal coverage. The advantage weakens along the curve and is not uniform across policy episodes, especially at twenty-day horizons during the recent tightening.

The practical conclusion is therefore precise. Separate factor volatility models can remain useful, but their outputs should not be mapped to tenor risk through an identity correlation matrix without testing the implied cross terms. Covariance is not an optional decoration on three marginal forecasts. It is part of the maturity-level risk forecast they are meant to produce.

Data and code availability

The Federal Reserve H.15 constant-maturity yield series are publicly available through FRED. Series identifiers, raw data, estimation code, forecast outputs, and figure code are included in the replication package accompanying this manuscript.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies

During preparation of this manuscript, the authors used OpenAI's ChatGPT for language editing and code assistance. The authors reviewed and edited all output and take full responsibility for the content.

A Implementation details

At each refit date, the three marginal models are estimated on the previous 1,260 observations after multiplying factor changes by 100 for numerical stability. Conditional variances are transformed back to squared percentage points before tenor mapping. CCC correlations use standardized residuals from the same rolling window. DCC parameters satisfy $a \geq 0$, $b \geq 0$, and $a + b < 0.999$. Across 75 refits, the median estimates are $a = 0.066$ and $b = 0.927$; the median rolling CCC level-slope correlation is -0.858 .

Within each 63-observation block, parameters are fixed but states are not. Before forecasting observation t , the marginal variance recursion uses the shock observed at $t - 1$, and the DCC recursion uses standardized residuals through $t - 1$. After recording all horizon forecasts for t , the realized factor change updates the states used at $t + 1$. This ordering prevents a same-day shock from entering its own forecast.

Sampling intervals in Table 1 are descriptive circular moving-block intervals conditional on the fitted full-sample model; they do not incorporate parameter-estimation uncertainty. By contrast, significance in Table 2 is based on the out-of-sample loss-difference series. The two forms of uncertainty answer different questions and are not combined.

References

- Aielli, G.P., 2013. Dynamic conditional correlation: On properties and estimation. *Journal of Business & Economic Statistics* 31, 282–299. <https://doi.org/10.1080/07350015.2013.771027>.
- Bauwens, L., Laurent, S., Rombouts, J.V.K., 2006. Multivariate GARCH models: A survey. *Journal of Applied Econometrics* 21, 79–109. <https://doi.org/10.1002/jae.842>.
- Bollerslev, T., 1990. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics* 72, 498–505. <https://doi.org/10.2307/2109358>.

- Christensen, J.H.E., Diebold, F.X., Rudebusch, G.D., 2011. The affine arbitrage-free class of Nelson–Siegel term structure models. *Journal of Econometrics* 164, 4–20. <https://doi.org/10.1016/j.jeconom.2011.02.011>.
- Christoffersen, P.F., 1998. Evaluating interval forecasts. *International Economic Review* 39, 841–862. <https://doi.org/10.2307/2527341>.
- Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. *Journal of Econometrics* 130, 337–364. <https://doi.org/10.1016/j.jeconom.2005.03.005>.
- Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13, 253–263. <https://doi.org/10.1080/07350015.1995.10524599>.
- Engle, R.F., 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics* 20, 339–350. <https://doi.org/10.1198/073500102288618487>.
- Ferreira, M.A., Lopez, J.A., 2005. Evaluating interest rate covariance models within a value-at-risk framework. *Journal of Financial Econometrics* 3, 126–168. <https://doi.org/10.1093/jjfinec/nbi005>.
- Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics* 20, 873–889. <https://doi.org/10.1002/jae.800>.
- Hautsch, N., Ou, Y., 2012. Analyzing interest rate risk: Stochastic volatility in the term structure of government bond yields. *Journal of Banking & Finance* 36, 2988–3007. <https://doi.org/10.1016/j.jbankfin.2012.06.020>.
- Hautsch, N., Yang, F., 2012. Bayesian inference in a stochastic volatility Nelson–Siegel model. *Computational Statistics & Data Analysis* 56, 3774–3792. <https://doi.org/10.1016/j.csda.2010.07.003>.
- Koopman, S.J., Mallee, M.I.P., van der Wel, M., 2010. Analyzing the term structure of interest rates using the dynamic Nelson–Siegel model with time-varying parameters. *Journal of Business & Economic Statistics* 28, 329–343. <https://doi.org/10.1198/jbes.2009.07295>.
- Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* 3, 73–84. <https://doi.org/10.3905/jod.1995.407942>.
- Laurent, S., Rombouts, J.V.K., Violante, F., 2012. On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics* 27, 934–955. <https://doi.org/10.1002/jae.1248>.
- Laurent, S., Rombouts, J.V.K., Violante, F., 2013. On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics* 173, 1–10. <https://doi.org/10.1016/j.jeconom.2012.08.004>.
- Litterman, R., Scheinkman, J., 1991. Common factors affecting bond returns. *Journal of Fixed Income* 1, 54–61. <https://doi.org/10.3905/jfi.1991.692347>.
- Nelson, C.R., Siegel, A.F., 1987. Parsimonious modeling of yield curves. *Journal of Business* 60, 473–489. <https://doi.org/10.1086/296409>.
- Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708. <https://doi.org/10.2307/1913610>.
- Patton, A.J., 2011. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160, 246–256. <https://doi.org/10.1016/j.jeconom.2010.03.034>.
- Shin, M., Zhong, M., 2017. Does realized volatility help bond yield density prediction? *International Journal of Forecasting* 33, 373–389. <https://doi.org/10.1016/j.ijforecast.2016.11.003>.