

Fragile When It Matters: Regimes, Spillovers, and the Structural Determinants of Gold ETF Tracking Errors

Ahmad Shauqi bin Haji Mohamad Zubir

shauqi@umt.edu.my

Universiti Malaysia Terengganu

Che Mohd Imran bin Che Taib

Universiti Malaysia Terengganu

Zalailah binti Salleh

Universiti Malaysia Terengganu

Nik Nor Amalina binti Nik Mohd Sukrri

Universiti Malaysia Terengganu

Akmalia binti Mohamad Ariff

Universiti Malaysia Terengganu

Siti Nadhirah binti Mohamad Fauzi

Universiti Malaysia Terengganu

Research Article

Keywords: gold ETFs, tracking error, Markov regime switching, TVP-VAR connectedness, structural breaks, cross-sectional dependence, authorised participants

Posted Date: August 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10401436/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Fragile When It Matters: Regimes, Spillovers, and the Structural Determinants of Gold ETF Tracking Errors

Ahmad Shauqi bin Haji Mohamad Zubir*
ORCID ID: 0000-0002-3492-8988

Che Mohd Imran bin Che Taib*
ORCID ID: 0000-0002-6013-0427

Zalailah binti Salleh*
ORCID ID: 0000-0001-7246-4160

Nik Nor Amalina binti Nik Mohd Sukrri*
ORCID ID: 0000-0002-1536-2755

Akmalia binti Mohamad Ariff*
ORCID ID: 0000-0001-6854-554X

Siti Nadhirah binti Mohamad Fauzi*
ORCID ID: 0000-0001-9185-6530

*Faculty of Business, Economics and Social Development, Universiti Malaysia Terengganu, 21030, Kuala Nerus, Terengganu, Malaysia.

Corresponding author:

Ahmad Shauqi bin Haji Mohamad Zubir
E-mail: shauqi@umt.edu.my
Telephone: +6096333794
Fax: +6096334222

Abstract

Gold exchange-traded funds (ETFs) are marketed as crisis insurance, yet little is known about how their tracking fidelity behaves in the very conditions that motivate holding them. This study characterises the dynamics and determinants of tracking errors for 71 physically backed gold ETFs across 10 national markets, combining four quantitative frameworks: Markov regime-switching models with time-varying transition probabilities, connectedness measures from a time-varying parameter vector autoregression, panel estimation under cross-sectional dependence, and multiple structural break tests, applied to daily data from 2019 to 2026 with regime histories reaching back to 2004. Tracking errors are governed by distinct calm and stress regimes whose volatilities differ by multiples, and equity market volatility drives both entry into and persistence within the stress regime. Nearly two-thirds of the variance in country-level tracking errors is attributable to cross-market spillovers, and the direction reverses the natural

prior: shocks flow from Switzerland and most Asia-Pacific markets to a United States that absorbs rather than exports them. Country fixed effects array markets on a structural hierarchy from 11.3 to 98.3 basis points that maps onto arbitrage infrastructure, and five synchronised structural breaks interact with this hierarchy multiplicatively, so the same shock that raises US tracking errors by roughly 13 basis points raises Australian or Korean errors by more than 50. The safe-haven function of gold ETFs is least reliable in the markets and episodes where investors need it most.

Keywords: gold ETFs; tracking error; Markov regime switching; TVP-VAR connectedness; structural breaks; cross-sectional dependence; authorised participants

JEL classification: C32, C33, C58, G11, G15, G23

1 Introduction

In March 2020, at the moment global investors were fleeing into gold, the instruments most of them now use to get there (Cheng et al., 2020) began to misbehave. Authorised participants faced funding constraints, physical delivery chains seized, and gold ETF prices decoupled from the benchmarks they exist to replicate. The fragility was not unforeseen: Pagano et al. (2019) had warned that voluntary, concentrated authorised participation could allow ETF prices to detach from the value of their underlying assets precisely in stressed markets. The episode exposed a question the tracking literature has never systematically answered: not how large gold ETF tracking errors are on average, but how they behave over time, how they move across borders, and what determines which markets suffer most when the arbitrage machinery is stressed.

The average-level question has a mature literature. Tracking deviations have been documented for US gold products and across countries, and their microstructure determinants are increasingly well identified: Bae and Kim (2020) establish causally that ETF illiquidity generates tracking errors through authorised participant arbitrage frictions, Kim et al. (2026) show that underlying market accessibility conditions arbitrage effectiveness, and Orlando (2025) documents that authorised participant concentration has created structural bottlenecks in the creation and redemption mechanism. What this body of work treats as fixed, however, is precisely what crises reveal to be variable. If arbitrage capacity is the binding constraint on tracking, then tracking errors should switch between identifiable regimes as stress conditions activate and relax that constraint, should transmit across markets through the correlated positioning of global arbitrageurs and correlated investor flows (Broman, 2016), and should respond to common shocks in proportion to each market's structural arbitrage capacity. Each of these predictions is dynamic, and none can be tested with the static toolkit.

This study tests all three on the most comprehensive multi-country gold ETF panel assembled for the purpose: 71 physically backed funds aggregated into 10 national markets, observed daily from April 2019 to February 2026, a window containing the COVID-19 pandemic, the Russia-Ukraine war, the SVB banking stress, a sustained gold rally, and the April 2025 tariff shock, supplemented by regime-switching estimation on fund-level histories reaching back to 2004. Four econometric frameworks target the four dynamic questions. Markov regime-switching models, extended to time-varying transition probabilities driven by the VIX and economic policy uncertainty, characterise the regime structure of tracking errors and its observable triggers. The

TVP-VAR connectedness framework of Antonakakis et al. (2020) maps how tracking deviations propagate across the 10 markets and identifies net transmitters and receivers. Panel estimation with Pesaran's (2006) common correlated effects mean group estimator quantifies the structural country hierarchy while accounting for the pervasive cross-sectional dependence that invalidates standard inference. Bai and Perron (2003) tests locate the permanent structural shifts around which the regime dynamics organise.

The paper makes four contributions. First, it provides the first regime-switching characterisation of gold ETF tracking errors, establishing that a two-state description with economically large volatility differentials fits essentially every fund tested, and that entry into and persistence within the high-error state are systematically driven by observable volatility stress. Second, it constructs the first cross-country connectedness network of tracking errors and documents a directional reversal of the natural hypothesis: tracking shocks flow from the high-error periphery to the low-error centre, with the United States acting as the system's shock absorber rather than its source. This network is defined on tracking-error innovations and is conceptually distinct from return-based ETF connectedness of the kind Li and Jiang (2025) document across cross-country ETF markets; we examine return-based connectedness for gold ETFs in separate work. Third, it estimates the structural hierarchy of tracking efficiency across markets under cross-sectional dependence, identifying a three-tier structure spanning nearly an order of magnitude, and traces it to arbitrage infrastructure, market access, and distance from the locus of gold price discovery. Fourth, it documents that the structural hierarchy and the crisis dynamics interact multiplicatively rather than additively, a compounding that conventional constant-parameter tracking risk assessments necessarily understate, and one with direct implications for regulators in markets seeking to develop gold ETF sectors.

Beyond these substantive findings, the study offers a computational template for system-level tracking analysis: country aggregation keeps the network estimable where a fund-level system would require 5,041 coefficients per lag, forgetting-factor Kalman filtering delivers time-varying parameters without the computational burden of MCMC, and the four frameworks, estimated independently on a common fund universe, validate one another through the cross-framework agreement documented in Section 4.5.

The remainder of the paper proceeds as follows. Section 2 reviews the literature and states the hypotheses. Section 3 describes the data, the aggregation strategy, and the four estimation frameworks. Section 4 presents the results in the order regime structure, connectedness, structural determinants, and breaks, closing with the synthesis across frameworks. Section 5 discusses the findings, and Section 6 concludes.

2 Literature Review and Hypotheses

2.1 Regime dependence in ETF pricing

Several strands of evidence suggest that ETF pricing efficiency is state-dependent rather than stable. Broman (2016) shows that the commonality in ETF misvaluation intensifies sharply during financial stress, with the own-style comovement loading rising from 0.958 before the global financial crisis to 1.096 during it, consistent with correlated non-fundamental demand overwhelming arbitrage capacity exactly when funding liquidity contracts. Chen and Xu (2023)

find that the destabilising effect of ETF ownership on underlying volatility is significantly magnified during the COVID-19 pandemic. Hattori and Yoshida (2023) document that the price effects of the Bank of Japan's ETF purchases shift across four distinct policy regimes and amplify in crisis, direct evidence that policy conditions constitute a regime-switching mechanism for ETF price dynamics. Han (2024) identifies inflection points in ETF connectedness around the Russia-Ukraine and Israel-Hamas conflicts. And the quantile evidence of Lee and Chen (2020a, 2020b, 2021) shows that the sensitivity of cross-border ETF returns to volatility, geopolitical risk, policy uncertainty, and sentiment switches sign across the return distribution, patterns invisible to mean-based estimation and symptomatic of underlying regime structure. Single-market evidence on gold ETFs specifically points in the same direction while disagreeing on the details. Evidence from single markets divides on direction: Indian gold ETFs lose sensitivity to gold in bear markets (Kaur & Singh, 2018), while Tehran-listed funds show the opposite, with higher bear-market betas and no residual explanatory power for the bull-bear split once volume, intraday volatility, and NAV deviations are controlled for (Kaviani et al., 2024). The disagreement is itself diagnostic: a bull and bear classification imposed from the direction of the gold price need not coincide with the latent states that actually govern replication, so regime structure has to be estimated rather than assumed. What none of these studies examines is the tracking error itself as the regime-switching object, despite the fact that the arbitrage mechanism that generates tracking errors (Bae & Kim, 2020) is precisely the capacity that stress episodes impair.

2.2 Cross-market transmission

The connectedness literature provides both the tools and the precedents for studying system-level tracking dynamics. Diebold and Yilmaz (2012, 2014) develop the variance decomposition framework and demonstrate country-level aggregation of entity-level data, and Demirer et al. (2018) show that country composites of firm-level series capture cross-border transmission robustly. Antonakakis et al. (2020) refine the framework with time-varying parameters. Applications consistently find that transmission intensity is crisis-specific rather than uniform; Papathanasiou et al. (2025), for instance, document that geopolitical conflict amplified cross-market volatility spillovers more than the pandemic did, a heterogeneity that recurs in the tracking-error network estimated below. Within the ETF literature, Shrestha et al. (2023) show that the locus of price discovery among structurally similar ETFs shifts across crisis and calm regimes, and Xu et al. (2020) establish the informational channel for cross-market propagation, finding that ETFs contribute roughly 73 percent of new information relative to their underlying indices and that informed ETF trading transmits to the underlying with a lag. If tracking deviations reflect the interaction of correlated demand shocks (Broman, 2016) with locally heterogeneous arbitrage capacity (Bae & Kim, 2020; Kim et al., 2026), they should propagate across borders through the creation and redemption activity that transmits commonality between funds and underlying markets (Marta & Riva, 2025), and the natural prior is that the largest and most liquid market originates the shocks that others import.

2.3 Structural determinants of cross-country heterogeneity

A third literature explains why tracking capacity differs across markets in the first place. Marszk and Lechman (2024) identify financial development, market infrastructure, and ICT capacity as robust drivers of ETF adoption across European markets. Kim et al. (2026) supply the sharpest causal evidence, exploiting minimum trading unit reforms in Japan, Singapore, and Vietnam to

show that reduced trading barriers raise institutional participation, strengthen arbitrage, and narrow ETF mispricing over a 12 to 18 month adjustment horizon. Orlando (2025) documents that authorised participant concentration and issuer-level risk management vary substantially across jurisdictions, Chen et al. (2017) show that home-market institutional quality moderates how global conditions transmit into ETF returns, and Dannhauser (2017) demonstrates that ETF activity itself reshapes the liquidity of underlying markets. Osterhoff and Kaserer (2016) link tracking precision to underlying market liquidity in the German setting. Evidence on the speed of arbitrage points the same way. Purohit and Malhotra (2015) find that deviations between price and net asset value in Indian ETFs persist for roughly three days, and up to five for some funds, where developed-market studies typically document same-day correction, direct evidence that the creation and redemption mechanism clears mispricing slowly where participation is thin. Collectively, this work predicts a persistent country hierarchy in tracking efficiency, but it has not been estimated on a common footing across markets, under common shocks, and with econometrics that respect the strong cross-sectional dependence such common shocks induce.

2.4 Hypotheses

The three literatures reviewed above generate predictions that are respectively dynamic, directional, and structural, and none is testable with a static toolkit. Three hypotheses organise the empirical work, each matched to one of the estimation frameworks described in Section 3, with the structural break analysis disciplining the first and third jointly.

H1 (Regime dependence). *Gold ETF tracking errors are governed by distinct low-error and high-error regimes, with transitions into and persistence within the high-error regime driven by observable uncertainty conditions.* The prediction follows from the arbitrage mechanism. If tracking errors are generated by frictions in authorised participant arbitrage (Bae & Kim, 2020), and if stress impairs funding liquidity and dealer capacity at the same moment that correlated non-fundamental demand intensifies (Broman, 2016), then the tracking process should exhibit not mild unconditional heteroskedasticity but a latent stress state that activates when volatility rises and persists while it stays elevated. Crisis evidence from adjacent ETF settings points the same way, with pricing distortions amplifying across policy regimes (Hattori & Yoshida, 2023) and during the pandemic (Chen & Xu, 2023). Observationally, regime activations should cluster around the global financial crisis, COVID-19, the Russia-Ukraine war, and the April 2025 tariff shock, with the VIX and economic policy uncertainty shifting both the entry and the persistence margins.

H2 (Tracking spillovers). *Tracking deviations transmit significantly across national gold ETF markets, and the United States, given its dominance in gold price discovery and ETF liquidity, is the dominant net transmitter of tracking shocks.* The existence component follows from common global drivers of arbitrage capacity, correlated arbitrageur positioning, and the creation and redemption activity through which fund trading reaches underlying markets (Marta & Riva, 2025). The directional component is a test between rival logics rather than a foregone conclusion. Price-discovery reasoning implies that the centre transmits: the US hosts price formation and leads the system informationally (Xu et al., 2020), so its shocks should be the ones others import. Arbitrage-capacity reasoning implies the opposite geography, because tracking innovations originate where replication is noisy and are absorbed where arbitrage is deep, so the periphery should transmit and the centre should receive. H2 states the price-discovery prior; the estimated network decides between them.

H3 (Structural heterogeneity). *Cross-country differences in tracking performance are large, persistent, and systematically related to structural characteristics, including arbitrage infrastructure, market accessibility, and institutional depth, rather than reducible to time-varying macro conditions.* The prediction assembles the causal evidence that market access conditions arbitrage effectiveness (Kim et al., 2026), that authorised participant concentration varies across jurisdictions (Orlando, 2025), and that underlying market liquidity governs tracking precision (Osterhoff & Kaserer, 2016). A corollary sharpens what structure means under stress: if structural capacity scales the response to common shocks rather than merely shifting the level of tracking errors, break amplitudes should be proportional to a market's structural position, so that stress and structure interact multiplicatively rather than additively. Support for H3 therefore requires decisive country effects under inference robust to cross-sectional dependence, an interpretable mapping from the fixed effect hierarchy to the identified channels, and break amplitudes that widen down the hierarchy.

3 Data and Methodology

3.1 Sample, aggregation, and variables

The fund universe originates from 81 gold exchange-traded products in LSEG Datastream, screened to physically backed products with meaningful benchmark linkage. The analyses in this paper use the 71 physical gold ETFs that belong to the 10 national markets with structurally coherent gold ETF sectors: the United States (7 funds), Canada (8), the United Kingdom (5), Germany (2), Switzerland (8), Australia (4), China (13), India (19), Japan (3), and South Korea (2). Funds from seven smaller markets (Turkey, Saudi Arabia, Hong Kong, South Africa, Taiwan, Thailand, and New Zealand) are reported in the companion cross-sectional study but excluded here, because they lack the within-market fund counts and the institutional coherence required for meaningful country-level aggregation: the constituent markets span fundamentally different currency regimes, gold market structures, and tracking profiles, and averaging across them would create a composite that represents no coherent economic unit. The exclusion of heterogeneous residual groupings follows standard practice in the connectedness literature (Diebold & Yilmaz, 2014; Demirer et al., 2018).

The fund-level tracking error is $e_{i,t} = r_{i,t} - r_{B,t}$, where $r_{i,t}$ is the daily log return of fund i and $r_{B,t}$ is the return on the COMEX gold benchmark aligned to the fund's denomination currency (Gold_USD, Gold_EUR, Gold_GBP, and other local-currency conversions based on contemporaneous exchange rates), with a narrow alignment lag of zero to one trading day to mitigate non-synchronous closes across time zones. Currency matching ensures that the residual reflects genuine tracking deviation rather than mechanical translation. For the system-level analyses, the country-level tracking error is the equal-weighted cross-sectional average of available fund-level absolute tracking errors,

$$|TE|_{c,t} = \frac{1}{N_{c,t}} \sum_{i \in c} |e_{i,t}|, \quad (1)$$

expressed in basis points, with time-varying composition so that each fund contributes for the full duration of its available history. Country-level aggregation keeps the connectedness system tractable (a 71-variable VAR would require estimating 5,041 coefficients per lag at each filtering

step against roughly 1,771 observations) and follows the precedent of aggregating entity-level series into country composites (Diebold & Yilmaz, 2014; Demirer et al., 2018; Han, 2024; Shrestha et al., 2023). The common daily panel spans April 2019 to February 2026, its start dictated by the youngest market group (South Korea).

The regime-switching analysis operates at the fund level and applies the most stringent screening, retaining the 24 ETFs whose benchmark correlation is high enough for a two-state decomposition of the tracking error to be econometrically meaningful; the retained funds span the United States (7), Canada (8), the United Kingdom (5), Switzerland (3), and Germany (1), with correlations between 0.701 and 0.986 and fund histories reaching back to November 2004. The geographic concentration is an outcome of the screening rather than a design choice: markets with weak benchmark linkage generate residuals that do not represent tracking error in the first place. The time-varying transition probability extension is estimated at the weekly frequency for computational stability, fitting successfully for 22 funds. The complete fund listings for each analytical sample in the study are provided in the Online Appendix (Tables A1 to A4).

Macro-financial covariates comprise the CBOE Volatility Index (VIX), the US Economic Policy Uncertainty index (EPU; Baker et al., 2016), the US Dollar Index (DXY), the CBOE Gold Volatility Index (GVZ; Luo et al., 2016), the Geopolitical Risk index (GPR; Caldara & Iacoviello, 2022), annualised gold return volatility, and the absolute daily gold return. All regressors are standardised to mean zero and unit variance so that coefficients are directly comparable in basis points per standard deviation.

3.2 Markov regime switching with time-varying transitions

The fund-level tracking error follows a two-regime Gaussian Markov switching process,

$$e_{i,t} = \mu_{S_t} + \phi_{S_t} e_{i,t-1} + \sigma_{S_t} \varepsilon_t, \quad \varepsilon_t \sim N(0,1), \quad (2)$$

where the latent state $S_t \in \{1,2\}$ distinguishes a low-error regime (small μ , σ) from a high-error regime, and evolves as a first-order Markov chain with transition matrix

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, \quad p_{ij} = \Pr(S_t = j \mid S_{t-1} = i). \quad (3)$$

Estimation follows Hamilton (1989) with smoothed state probabilities recovered by the Kim (1994) filter. The time-varying transition probability (TVTP) extension replaces the constant p_{ij} with logistic functions of lagged VIX and EPU, so that observable uncertainty conditions shift both the probability of entering the high-error regime and the probability of remaining in it. Effects are summarised as the change in transition probability, in percentage points, associated with a one standard deviation increase in each driver.

3.3 TVP-VAR connectedness

System dynamics are modelled with a time-varying parameter VAR of order one for the country-level tracking error vector $\mathbf{e}_t$,

$$\mathbf{e}_t = \Phi_{1,t} \mathbf{e}_{t-1} + \mathbf{u}_t, \quad \mathbf{u}_t \sim N(0, \Sigma_t), \quad (4)$$

estimated with the Koop and Korobilis (2014) Kalman filter approach using forgetting factors $\kappa_1 = \kappa_2 = 0.99$, which accommodates smooth variation in coefficients and error covariances without the computational burden of MCMC (Antonakakis et al., 2020). The generalised forecast error variance decomposition of Pesaran and Shin (1998) at horizon $H = 10$ days gives the share of market i 's forecast error variance attributable to shocks in market j ,

$$\theta_{ij,t}(H) = \frac{\sigma_{jj,t}^{-1} \sum_{h=0}^{H-1} (\mathbf{e}_i^\top A_{h,t} \Sigma_t \mathbf{e}_j)^2}{\sum_{h=0}^{H-1} \mathbf{e}_i^\top A_{h,t} \Sigma_t A_{h,t}^\top \mathbf{e}_i}, \quad (5)$$

normalised row-wise to $\tilde{\theta}_{ij,t}$. The total connectedness index and the net directional position of market i are

$$TCI_t(H) = \frac{1}{N} \sum_{i \neq j} \tilde{\theta}_{ij,t}(H) \times 100, \quad NET_{i,t} = TO_{i,t} - FROM_{i,t}, \quad (6)$$

where $TO_i = \sum_{j \neq i} \tilde{\theta}_{ji}$ and $FROM_i = \sum_{j \neq i} \tilde{\theta}_{ij}$. The estimation sample comprises 1,771 daily observations from 10 April 2019 to 18 February 2026, with connectedness measures reported from 21 January 2020 after a 200-observation initialisation, yielding 1,570 daily observations spanning COVID-19, the Russia-Ukraine war, the SVB stress, and the 2025 tariff shock.

3.4 Panel estimation under cross-sectional dependence

Structural heterogeneity is tested through three nested panel specifications for the country-level absolute tracking error. The fixed effects benchmark is

$$|TE|_{c,t} = \alpha + a_c + \Gamma_{c,t}^\top \beta + \varepsilon_{c,t}, \quad (7)$$

where a_c are country fixed effects and $\Gamma_{c,t}$ stacks the seven standardised macro regressors. The dynamic specification adds the lagged dependent variable,

$$|TE|_{c,t} = \psi |TE|_{c,t-1} + \Gamma_{c,t}^\top \beta + a_c + \varepsilon_{c,t}, \quad (8)$$

in the spirit of Arellano and Bond (1991). Because common global shocks induce cross-sectional dependence that invalidates standard clustered inference, dependence is tested with the Pesaran (2021) CD statistic, and the preferred specification is the common correlated effects mean group (CCE-MG) estimator of Pesaran (2006),

$$|TE|_{c,t} = \alpha + a_c + \Gamma_{c,t}^\top \beta_c + \delta_{1c} \overline{|TE|}_t + \delta_{2c}^\top \overline{\Gamma}_t + \varepsilon_{c,t}, \quad (9)$$

which augments each country's regression with cross-sectional averages as proxies for unobserved common factors and allows fully heterogeneous slopes (Chudik & Pesaran, 2015). The panel is balanced with $N = 10$ and $T = 1,721$, giving 17,210 observations after lag construction, a moderate- N , large- T configuration well suited to both the CD test and CCE-MG asymptotics.

3.5 Bai-Perron structural breaks

Permanent shifts in the level of tracking errors are located with the Bai and Perron (2003) multiple structural break methodology applied to the 20-day moving average of each country's

$|TE|$ series and to the cross-country mean, the smoothing chosen to filter bid-ask bounce and non-synchronous trading noise while preserving lower-frequency structural dynamics. The sup-F statistic for m breaks against none is

$$\text{sup}F_T(m; q) = \frac{(RSS_0 - RSS_m)/m}{RSS_m/(T - (m + 1)q - p)}, \quad (10)$$

with the number of breaks selected by the Bayesian information criterion and confirmed by sequential sup-F tests of m against $m - 1$ breaks, using trimming parameter $\varepsilon = 0.10$ (minimum segment length of roughly 175 observations).

4 Results

4.1 Tracking error regimes and their drivers

Table 1 summarises the daily constant-transition regime-switching estimates at the market level; fund-level estimates appear in the Online Appendix (Table A5). The two-state structure is present in every market, and its economic magnitude is the first key result: median high-regime tracking error volatility is a multiple of low-regime volatility, roughly 3.5 to 4 times in the US (38.09 against 9.99 basis points) and around 3 to 3.6 times in Canada, Switzerland, the UK, and Germany. Median high-regime occupancy ranges from 17.0 percent of days (US) to 23.6 percent (Germany), with typical high-regime spells lasting around five trading days. Tracking error is not a process with mild unconditional heteroskedasticity; it is a process with a latent stress state that is visited regularly, briefly, and violently.

Table 1. Market-level medians from the daily Markov regime-switching estimation

Market	N funds	Median corr.	Median TE SD (bps)	Median σ low (bps)	Median σ high (bps)	Median high-regime share	Median high-regime duration (days)
Canada	8	0.804	64.62	29.55	106.59	0.204	4.5
Switzerland	3	0.786	65.71	37.50	112.91	0.221	5.9
UK	5	0.803	63.78	35.77	114.77	0.208	5.1
Germany	1	0.883	50.88	26.95	88.04	0.236	5.7
US	7	0.982	19.37	9.99	38.09	0.170	4.6

Notes. Two-regime Gaussian Markov switching model, Eq. (2) to (3), estimated on daily fund-level tracking errors against currency-matched COMEX benchmarks; entries are within-market medians across funds. Fund histories span November 2004 to February 2026 where available.

The US occupies the same position here as in level terms: both regimes are compressed (its high-regime volatility of 38 basis points is comparable to other markets' low-regime volatility), but the proportional regime differential is just as large. Deep arbitrage infrastructure lowers the level of tracking errors in both states; it does not exempt a market from having states.

The TVTP extension tests the driver component of H1 directly. Table 2 reports the cross-fund evidence. The VIX channel is strong and consistent: a positive VIX shock raises the probability of entering the high-error regime for 83 percent of funds and raises the probability of remaining in it for 87 percent, with median effects of +1.75 and +4.66 percentage points per standard deviation respectively. Volatility stress therefore acts on both margins, activating the stress state

and prolonging it, with the persistence margin the stronger of the two. The EPU channel is present but heterogeneous, positive for regime entry in 57 percent of funds and for persistence in only 39 percent, suggesting that policy uncertainty operates through market-specific or event-specific pathways rather than as a uniformly signed global driver, a heterogeneity consistent with the country-varying uncertainty transmission documented by Lee and Chen (2021).

Table 2. Time-varying transition probability evidence on regime drivers (weekly specification, 22 funds)

Margin	Share of funds with positive VIX effect	Median effect of +1 SD VIX (pp)	Share of funds with positive EPU effect	Median effect of +1 SD EPU (pp)
Entry: P(low to high)	0.83	+1.75	0.57	+0.37
Persistence: P(high to high)	0.87	+4.66	0.39	-0.47

Notes. TVTP-MRS with transition probabilities specified as logistic functions of lagged VIX and EPU, estimated at the weekly frequency.

Table 3 confronts the crisis-trigger component of H1 with the event record. Relative to a 2017 to 2019 baseline in which the average cross-sectional high-regime probability is 0.138, the global financial crisis and COVID-19 windows raise it to 0.661 and 0.653, uplifts exceeding 50 percentage points, and the April 2025 tariff window raises it to 0.347, an uplift of 20.9 points. The Russia-Ukraine window produces a much smaller uplift of 3.2 points in this weekly specification, which should be read as heterogeneous transmission intensity across shocks rather than as evidence against regime dependence: as the structural break results below show, the 2022 shock registered strongly in European markets and only with a long lag in Asia, a dispersion that a global cross-sectional average dilutes.

Table 3. Cross-sectional high-regime probability by event window

Event window	Average cross-sectional P(high)	Uplift versus baseline (pp)
Baseline (2017-2019)	0.138	0.0
Global financial crisis (2008-2009)	0.661	+52.3
COVID-19 (2020)	0.653	+51.5
Russia-Ukraine (2022)	0.170	+3.2
Liberation Day tariffs (2025)	0.347	+20.9

Notes. Weekly TVTP-MRS specification. Entries are averages of smoothed high-regime probabilities across funds within each window.

Figure 1 plots the cross-sectional average of smoothed high-regime probabilities over time, a tracking-stress index whose peaks align with the crisis catalogue and whose synchronisation across funds indicates that the regime process is anchored in common macro-financial stress rather than fund-level idiosyncrasies.

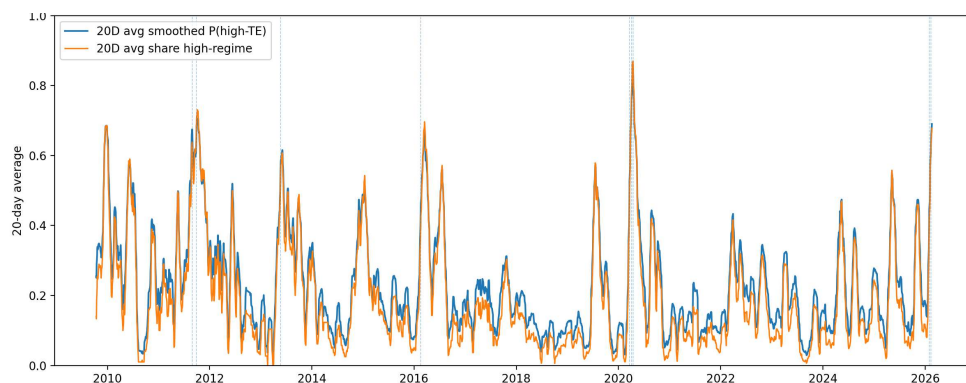


Fig. 1. Smoothed cross-sectional average of high-regime probabilities from the daily MRS model

Taken together, the regime evidence supports H1 in its structured form: the latent two-state description is pervasive, the VIX is a robust driver of both regime entry and regime persistence, and the great global shocks of the sample register as sharp, synchronised elevations of stress-state probability.

4.2 The tracking error network

Table 4 presents the full-sample average connectedness matrix for the 10 markets. The total connectedness index averages 64.57 percent: just under two-thirds of the 10-day-ahead forecast error variance in a country's tracking errors is attributable to shocks originating elsewhere in the system. Tracking error is a system-level phenomenon, not a collection of idiosyncratic national processes, and the first component of H2, the existence of significant spillovers, is confirmed emphatically (a one-sample test of the TCI against zero yields $t = 766.0$).

Table 4. Full-sample average connectedness matrix (percent, $H = 10$ days)

	US	CA	UK	DE	CH	AU	CN	IN	JP	KR	FROM
US	66.1	6.5	4.0	3.7	4.3	3.4	3.1	3.0	2.8	3.1	33.9
CA	5.9	48.6	5.1	6.0	7.7	4.7	5.7	4.9	5.5	5.8	51.3
UK	2.1	4.5	37.2	13.0	21.4	4.7	4.4	5.0	3.6	4.2	62.9
DE	2.2	4.7	13.5	39.7	17.4	4.9	4.5	5.2	3.7	4.2	60.3
CH	1.8	5.1	18.2	13.9	30.1	6.1	6.5	7.1	5.4	5.9	70.0
AU	0.9	3.3	3.6	3.5	6.3	24.8	15.5	13.3	12.5	16.4	75.3
CN	0.9	3.7	3.3	3.7	6.4	15.4	24.8	13.0	12.3	16.4	75.1
IN	1.1	3.5	3.9	4.2	7.3	13.8	14.0	27.5	10.2	14.5	72.5
JP	0.8	3.6	2.9	2.7	5.6	13.7	14.3	11.0	30.8	14.5	69.1
KR	1.0	3.8	3.3	3.2	5.9	16.0	16.2	13.4	12.7	24.5	75.5
TO	16.7	38.7	57.8	53.9	82.3	82.7	84.2	75.9	68.7	85.0	TCI = 64.57
NET	-17.1	-12.8	-5.2	-6.4	+12.4	+7.5	+9.1	+3.5	-0.4	+9.4	

Notes. Entry (i, j) is the share of market i 's 10-day-ahead tracking error forecast variance attributable to shocks from market j , averaged over 21 January 2020 to 18 February 2026. TO and FROM are column and row sums off the diagonal; NET = TO minus FROM.

The matrix has a pronounced block structure. North America is a relatively self-contained pair with high own-variance shares (66.1 percent for the US, 48.6 for Canada) and thin links to everything else. Europe is a tightly integrated triangle in which the Switzerland-UK link is the single strongest bilateral connection in the network (21.4 and 18.2 percent in the two directions), the footprint of the deeply integrated Zurich and London gold trading venues with overlapping hours and shared custodial networks. Asia-Pacific is a dense, symmetric cluster with bilateral flows of 10 to 16 percent among Australia, China, India, Japan, and South Korea, markets that share trading hours, correlated exposure to dollar-Asia currency dynamics, and correlated physical gold demand.

The second component of H2 concerns direction, and here the data reject the hypothesis and reverse it. The US is a persistent net receiver of tracking shocks (NET of -17.1 percent, negative on 98.4 percent of sample days), as are Canada (-12.8 percent, a net transmitter on only 3.4 percent of days), Germany (-6.4 percent), and the UK (-5.2 percent), while Japan sits essentially at neutrality (-0.4 percent). The dominant transmitters are Switzerland (+12.4, positive on 98.3 percent of days), South Korea (+9.4), China (+9.1, positive on 93.4 percent of days), Australia (+7.5), and India (+3.5). Tracking shocks flow from the high-tracking-error periphery to the low-tracking-error centre.

The reversal has a coherent economic reading rather than being an anomaly to be explained away. US funds track almost perfectly, so they generate small tracking innovations with little variance to export; their forecast error variance is instead partially accounted for by the large, volatile innovations arriving from markets where arbitrage capacity is thin. This is the supply-side arbitrage channel of Bae and Kim (2020) operating in the cross-section: markets with weaker arbitrage infrastructure originate larger tracking shocks, and those shocks propagate through correlated currency movements, global gold price dynamics, and the correlated positioning of cross-market arbitrageurs, with the correlated-demand mechanism of Broman (2016) amplifying transmission during stress. The deep US authorised participant ecosystem then functions as the system's shock absorber, attenuating rather than amplifying what arrives. Switzerland's strong transmitter role is consistent with its position as the world's dominant gold refining hub, whose tracking innovations carry supply-chain information into both the European and Asian networks.

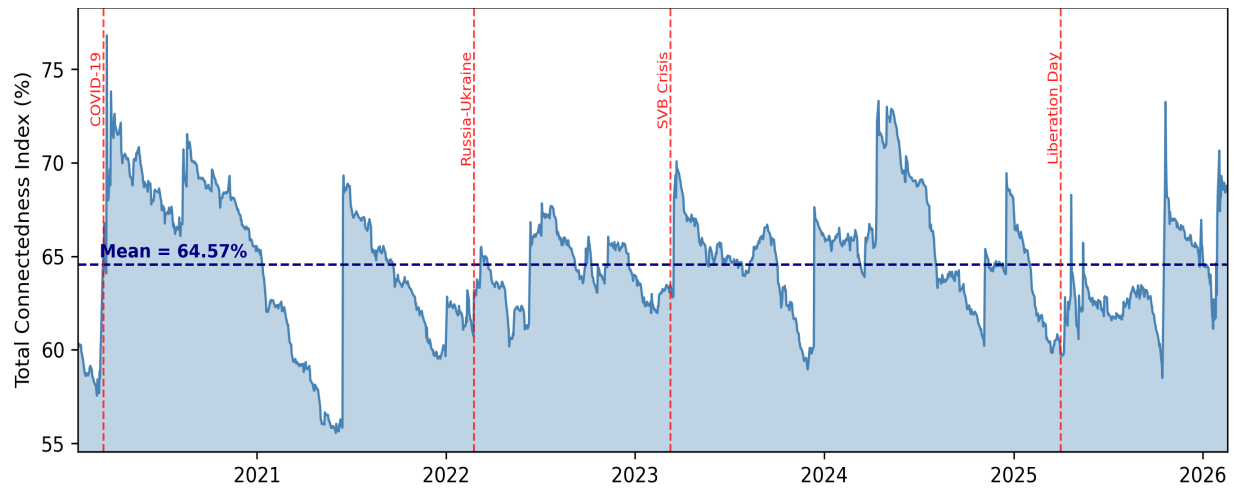
Table 5 summarises the sub-period dynamics. The COVID year produced both the highest sub-period connectedness, a mean TCI of 67.27 percent against the full-sample 64.57, and the sharpest directional stress: the US net receiver position deepened to -32.3 percent while Canada, measured on bullion funds alone, deepened to -15.6 percent, the network signature of the March to April 2020 disruption of creation, redemption, and physical delivery. The Russia-Ukraine window shows connectedness close to the full-sample mean with the lowest volatility of any window ($SD = 1.71$), a stable transmission regime during a regionally concentrated shock. The 2023 to 2025 period is moderately elevated (mean TCI of 64.64 percent) and contains the second highest reading of the sample, the April 2024 peak of 73.3 percent during the sustained gold rally. Connectedness during COVID is significantly higher than in the post-COVID recovery ($t = 15.29$). The findings are robust to excluding South Korea and extending the sample to 2014: the TCI from the 9-market specification correlates at 0.986 with the main series over the overlapping window, and the transmitter and receiver orderings are qualitatively unchanged.

Table 5. Sub-period connectedness summary

Period	N days	Mean TCI	Max TCI	SD TCI	NET US	NET CN	NET CA
COVID (2020)	246	67.27	76.80	3.73	-32.3	+7.8	-15.6
Post-COVID (2021-2022)	516	63.11	69.31	3.02	-15.7	+9.9	-12.7
Russia-Ukraine (from 24 Feb 2022)	220	64.52	67.82	1.71	-11.4	+9.8	-10.7
Recent (2023-2025)	774	64.64	73.30	2.83	-14.4	+8.8	-12.7

Notes. Full-sample mean TCI = 64.57 percent. NET entries are period means of daily net directional connectedness.

Figure 2 plots the dynamic TCI with the crisis events marked, showing the COVID surge to the sample maximum of 76.8 percent in March 2020, the decline to the sample minimum of 55.6 percent in mid 2021, and a sustained 2024 elevation that peaks at 73.3 percent in April of that year. The April 2025 tariff announcement itself produces no spike; the index sits just below 60 percent on the day and climbs over the following weeks, consistent with tariff uncertainty repricing tracking transmission gradually rather than instantaneously. Figure 3 renders the network as a directed graph, with arrows flowing from the green transmitter nodes, Switzerland and the Asia-Pacific markets, toward the red receiver nodes of North America and northern Europe.

**Fig. 2.** Dynamic Total Connectedness Index (TCI)

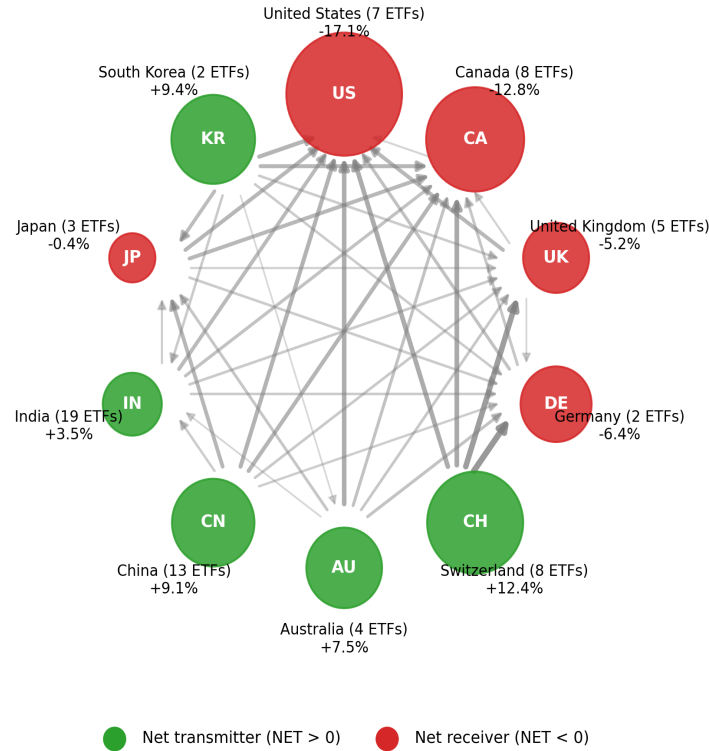


Fig. 3. Net Pairwise Connectedness Network

4.3 The structural hierarchy

Table 6 reports the panel estimates and diagnostics. The fixed effects model explains 36.3 percent of the variation in country-level tracking errors from seven macro regressors and country intercepts alone, and the F-test on the country effects is decisive ($F = 512.13$, $p < 0.001$): time-invariant national structure is a first-order determinant of tracking performance, the core of H3. The dynamic specification adds meaningful persistence ($\psi = 0.237$), indicating that roughly a quarter of any day's tracking deviation carries into the next day, the footprint of arbitrage that corrects mispricing over days rather than instantaneously. The Pesaran CD statistic of 72.31, with an average pairwise residual correlation of 0.260, decisively rejects cross-sectional independence, which both invalidates naive clustered inference and motivates the CCE-MG estimator; augmenting each country's regression with cross-sectional averages raises mean explanatory power to 55.7 percent, and the common factor loading is precisely estimated (1.000, $SE = 0.226$). A substantial share of tracking error variation is driven by global conditions that country-by-country regressions cannot see.

Table 6. Panel models of country-level absolute tracking errors

Panel A. Model comparison and coefficients (basis points per standard deviation)

Variable	FE, Eq. (7)	Dynamic, Eq. (8)	CCE-MG, Eq. (9)
VIX	-3.071** (1.316)	-1.446 (1.316)	heterogeneous
DXY	5.040*** (0.595)	3.330*** (0.589)	heterogeneous

Variable	FE, Eq. (7)	Dynamic, Eq. (8)	CCE-MG, Eq. (9)
GVZ	16.739*** (2.176)	11.282*** (2.177)	heterogeneous
EPU	-2.781*** (0.626)	-2.081*** (0.624)	heterogeneous
GPR	3.429*** (0.641)	2.142*** (0.623)	heterogeneous
Gold volatility	2.467** (1.142)	0.501 (1.132)	heterogeneous
Absolute gold return	21.391*** (0.967)	21.224*** (1.000)	heterogeneous
Lagged dependent (ψ)		0.237*** (0.016)	
Common factor loading			1.000*** (0.226)
R ²	0.3627	0.4024	0.5572 (mean)

Panel B. Diagnostics

Test	Statistic	p-value	Implication
Pesaran CD (cross-sectional dependence)	72.313	< 0.001	Dependence pervasive; CCE required
F-test, country fixed effects	512.131	< 0.001	Structural heterogeneity decisive
Mean Durbin-Watson	1.541		Positive serial correlation
Average pairwise residual correlation	0.260		Unobserved common factors

Notes. $N = 10$ markets, $T = 1,721$ days (April 2019 to February 2026), 17,210 observations. Heteroskedasticity robust (HC1) standard errors in parentheses for FE and dynamic models; Pesaran (2006) mean group standard errors for CCE-MG. Asterisks denote significance at the 10 (*), 5 (**), and 1 (***) percent levels. Under CCE-MG the mean group averages of the macro coefficients are individually insignificant because country-specific slopes take offsetting signs; the country-level distributions appear in the Online Appendix (Table A11).

Two macro drivers dominate. A one standard deviation increase in the absolute gold return raises country tracking errors by roughly 21.4 basis points, essentially undiminished by the inclusion of the lag, so large price moves generate immediate same-day deviations. Gold implied volatility (GVZ) adds 11 to 17 basis points, attenuating but remaining strong in the dynamic model, so it operates through both contemporaneous impact and persistence. Both effects are exactly what the arbitrage friction mechanism predicts: large moves and elevated volatility widen dealer spreads and raise inventory risk, impairing the creation and redemption process. Dollar strength (DXY) adds 3 to 5 basis points, the mechanical FX conversion friction borne by non-US funds, and geopolitical risk adds 2 to 4 basis points, consistent with disruption to physical gold logistics. The negative EPU coefficient of roughly -2 to -3 basis points looks counterintuitive until read through the safe-haven channel: policy uncertainty attracts gold inflows and trading volume, which deepens ETF liquidity and compresses tracking errors, the mirror image of the geopolitical channel that disrupts the physical side. Equity volatility (VIX) carries a negative coefficient that reaches 5 percent significance in the static model but loses it once the lagged dependent variable enters, so any equity-fear effect operates through persistence rather than contemporaneous impact; once gold-specific volatility is controlled for, tracking errors respond to gold market stress rather than to equity market fear per se.

The country fixed effects, reported in Table 7, quantify the structural hierarchy after stripping out macro conditions. Three tiers emerge. The US stands alone at 11.4 basis points. Canada (38.6), the UK (44.8), and Germany (49.3) complete the efficient tier below 50 basis points, with Switzerland (55.7) alone in a moderate band; the Canadian position, earned on a compact roster of bullion share classes half of which are currency hedged, shows that near efficient tracking does not require US scale. The five Asia-Pacific markets then cluster remarkably tightly between 90.3 and 98.3 basis points, a common ceiling of institutional friction spanning capital controls

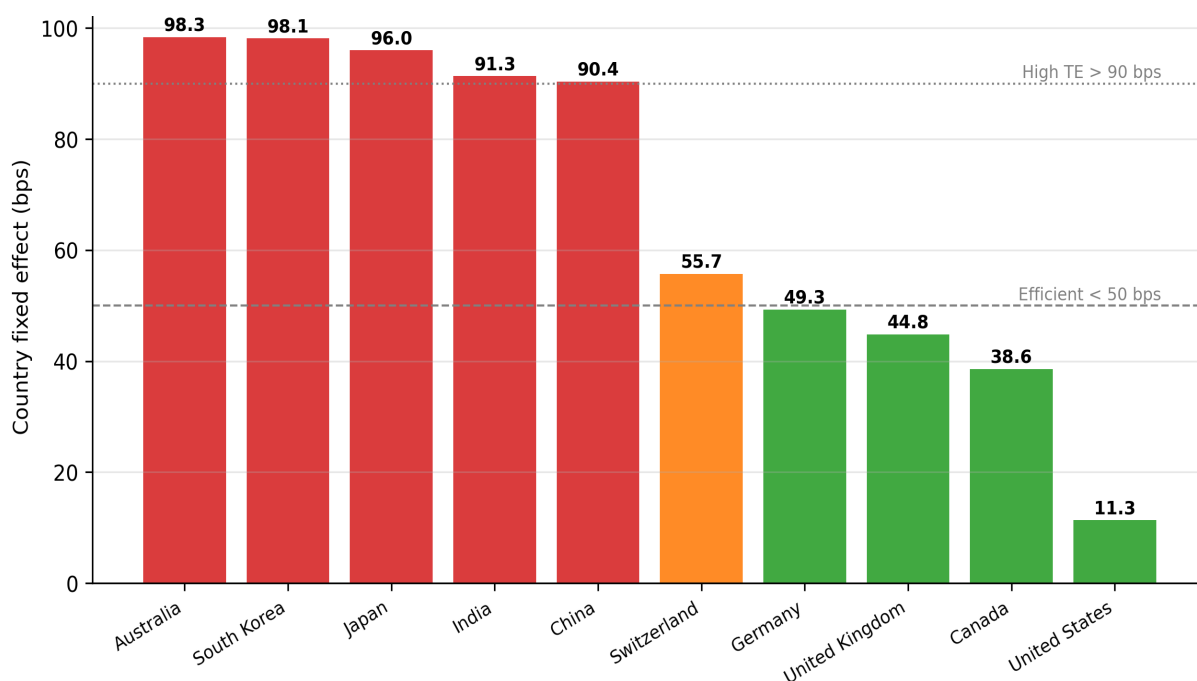
(China, India), restricted authorised participant participation and limited in-kind creation and redemption, and, tellingly, sheer distance from the locus of gold price discovery: Australia, a developed economy with a sophisticated financial system, records the highest fixed effect in the sample at 98.3 basis points, evidence that time-zone separation from COMEX price formation matters independently of financial development. The roughly 87 basis point gap between the US and Australia is the economic price of structural market differences in maintaining gold price fidelity.

Table 7. Country fixed effects and tracking performance classification

Market	N funds	Mean TE (bps)	Median (bps)	SD (bps)	Fixed effect (bps)	Rank	Classification
US	7	11.4	7.1	13.1	11.3	1	Efficient
Canada	8	38.6	31.1	28.3	38.6	2	Efficient
UK	5	44.8	33.3	41.8	44.8	3	Efficient
Germany	2	49.3	37.3	42.5	49.3	4	Efficient
Switzerland	8	55.7	44.3	42.5	55.7	5	Moderate
China	13	90.3	68.3	82.6	90.4	6	High TE
India	19	91.3	69.2	80.2	91.3	7	High TE
Japan	3	96.0	70.5	89.2	96.0	8	High TE
South Korea	2	98.1	76.1	87.8	98.1	9	High TE
Australia	4	98.3	75.8	85.4	98.3	10	High TE

Notes. Classification thresholds: efficient below 50 bps, moderate 50 to 90 bps, high above 90 bps. Fixed effects from Eq. (7). Medians below means throughout indicate positive skewness, most pronounced in the high-TE tier.

The heterogeneous CCE-MG slopes, reported in full in the Online Appendix (Table A11), carry the mechanism further. The cleanest divide runs through the common factor loadings and the model fit. In the five Asia-Pacific markets the loading on the cross-sectional mean ranges from 1.43 to 1.84 and the country-level R^2 from 0.70 to 0.82, so tracking errors there are largely an imported phenomenon, explained by the same global conditions that move every other market. In the efficient tier the loadings fall between 0.05 and 0.52 and the R^2 never reaches 0.50, with the US at the extreme ($\delta = 0.05$, $R^2 = 0.29$): its tracking process is governed almost entirely by idiosyncratic microstructural dynamics, the competitive arbitrage and market making that compress deviations independently of global conditions. Sensitivity to realised gold volatility leans the same way, negative in Canada (2.40 basis points per standard deviation lower), muted in the US, and strongly positive in Australia (4.58 higher), although the pattern is less uniform across the periphery. The implied volatility slopes invert that ordering, positive for the US and negative for Australia and China, which is not a contradiction but a consequence of the specification: country slopes in Eq. (9) are estimated conditional on the common factor, and because the common component itself rises with gold implied volatility, markets that load heavily on the factor have little GVZ response left to attribute, while the US, nearly orthogonal to the common factor, displays its response directly. The economically robust statement is therefore the division of labour between common and idiosyncratic drivers, and it aligns market by market with the connectedness results: the high-loading markets are the network's transmitters and the idiosyncratic markets its receivers, a cross-framework consistency that strengthens the structural interpretation of H3. Figure 4 displays the fixed effect hierarchy as a ranked bar chart with the tier thresholds marked.



Green = Efficient (<50 bps), Orange = Moderate (50 to 90 bps), Red = High TE (>90 bps).

Fig. 4. Country Fixed Effects in |TE| Panel Model

4.4 Structural breaks and the multiplicative interaction

The Bai-Perron analysis locates the permanent shifts around which these dynamics organise. The BIC selects five breaks for all 11 series (the 10 markets plus the cross-country mean), and every sequential sup-F statistic rejects at the 1 percent level, so the six-regime characterisation is not a statistical artefact of any single market. Table 8 reports the break dates and segment statistics for the cross-country mean; the full break diagnostics and country-level segment statistics appear in Online Appendix Tables A12 to A24.

Table 8. Structural breaks and regime statistics, cross-country mean absolute tracking error

Panel A. Estimated break dates

Break	Date	Nearest event (within 60 days)	Sequential sup-F
1	2020-02-28	COVID-19 declaration (11 Mar 2020)	175.86***
2	2020-11-03	Post-COVID normalisation, vaccine announcements	54.01***
3	2023-06-29	Late monetary tightening cycle	128.79***
4	2024-04-01	Onset of sustained gold rally	54.54***
5	2025-04-17	Liberation Day tariffs (2 Apr 2025)	109.03***

Panel B. Segment statistics (basis points)

Regime	Window	N days	Mean	Median	SD	Skewness
1. Pre-COVID baseline	2019-05-08 to 2020-02-28	211	57.0	52.8	16.6	0.48

Regime	Window	N days	Mean	Median	SD	Skewness
2. COVID crisis	2020-03-02 to 2020-11-03	176	85.1	75.5	30.3	1.02
3. Normalisation	2020-11-04 to 2023-06-29	684	63.6	60.6	13.4	1.15
4. Trough	2023-06-30 to 2024-04-01	194	48.6	45.7	11.5	0.32
5. Gold rally	2024-04-02 to 2025-04-17	271	66.9	63.6	14.9	0.81
6. Liberation Day	2025-04-21 to 2026-02-18	216	84.9	68.6	35.4	0.99

*Notes. Bai-Perron tests on the 20-day moving average of the cross-country mean absolute tracking error, trimming $\varepsilon = 0.10$, breaks selected by BIC and confirmed by sequential sup-F tests. *** denotes significance at the 1 percent level.*

The regime trajectory is strongly non-monotonic. From a pre-COVID baseline of 57.0 basis points, the mean tracking error jumps 49 percent to 85.1 in the COVID regime, mean-reverts only partially through the tightening cycle, reaches its sample trough of 48.6 in the calm of 2023 to early 2024, re-elevates through the gold rally, and returns to COVID-era levels of 84.9 basis points, with the highest volatility of the sample, after the April 2025 tariff shock, a 74.7 percent increase over the trough. Crisis regimes are also right-skewed (skewness near 1 in regimes 2 and 6), so stress manifests as occasional extreme deviation episodes layered on an elevated base rather than as uniform deterioration.

Break timing across markets reveals the geography of shock transmission. The COVID and Liberation Day breaks are nearly simultaneous across nine of the ten markets, falling within an eight trading day window and within twenty one days of the announcement respectively, the signature of truly global shocks. The intermediate breaks disperse: the UK, Germany, and Switzerland detect breaks near the February 2022 invasion of Ukraine, while the US and the Asia-Pacific markets break in mid 2023, consistent with the geopolitical shock reaching the wider system through the delayed monetary tightening channel rather than directly. Two idiosyncratic patterns validate the market decomposition. Germany's fourth break falls in December 2022, at the peak of the European energy crisis, a regime shift absent from the otherwise similar UK series. Canada follows a chronology of its own: measured on bullion funds alone, its series contains no sharp COVID break, peaks across the Russia-Ukraine tightening cycle, and descends into 2025 while every other market climbs, the footprint of a roster in which half the funds are currency hedged share classes whose tracking responds to hedging costs and rate differentials rather than to the global crisis calendar.

The most consequential result of the break analysis is the interaction between regime shifts and the structural hierarchy of Section 4.3. Break timing is common; break amplitude is not. Moving from the trough regime to the Liberation Day regime, the US shifts from 7.2 to 19.9 basis points, an increase of roughly 13 basis points, while Australia shifts from 68.1 to 122.2 (+79 percent), Japan from 69.2 to 126.6 (+83 percent), and South Korea from 68.9 to 128.3 basis points (+86 percent), increases of 54 to 59 basis points. Across the full sample span, the US moves between regime means of 7.2 and 19.9 basis points, so even its worst regime remains far below the best regime of any Asia-Pacific market. Structural market quality and crisis exposure therefore combine multiplicatively: weak arbitrage infrastructure raises the level of tracking errors in every state and amplifies the response to every shock. Figure 5 shows the cross-country mean series with the estimated breaks and regime means, and Figure 6 condenses the interaction into a single heatmap of market-by-regime mean tracking errors, in which developed-market rows shift shade only mildly across regime columns while peripheral-market rows darken dramatically in the crisis columns.

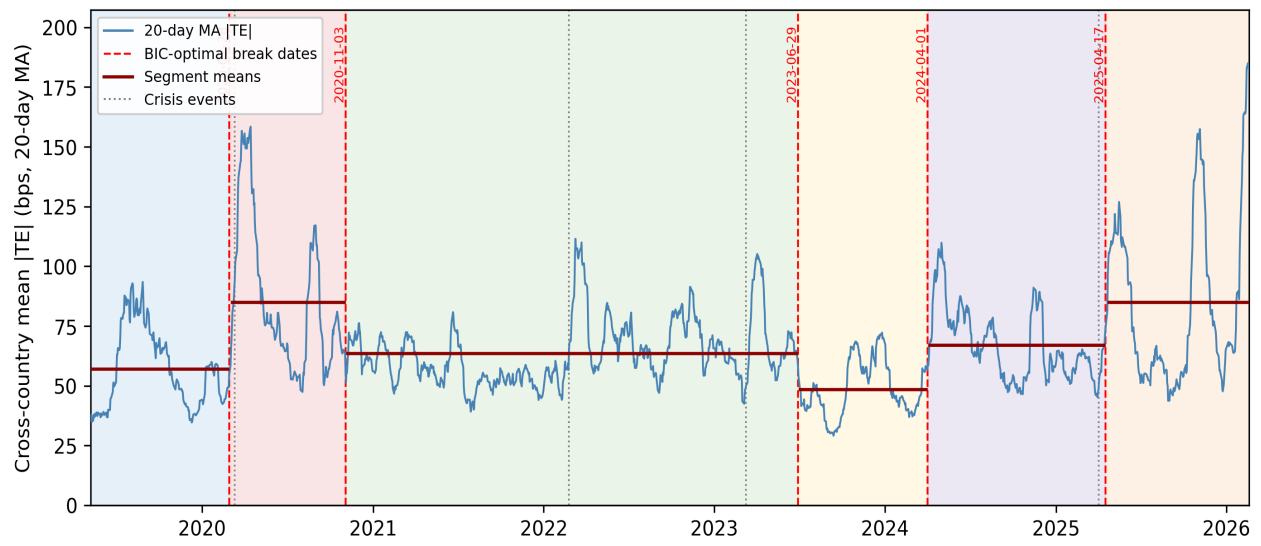


Fig. 5. Cross-Country Mean $|TE|$ with Structural Breaks

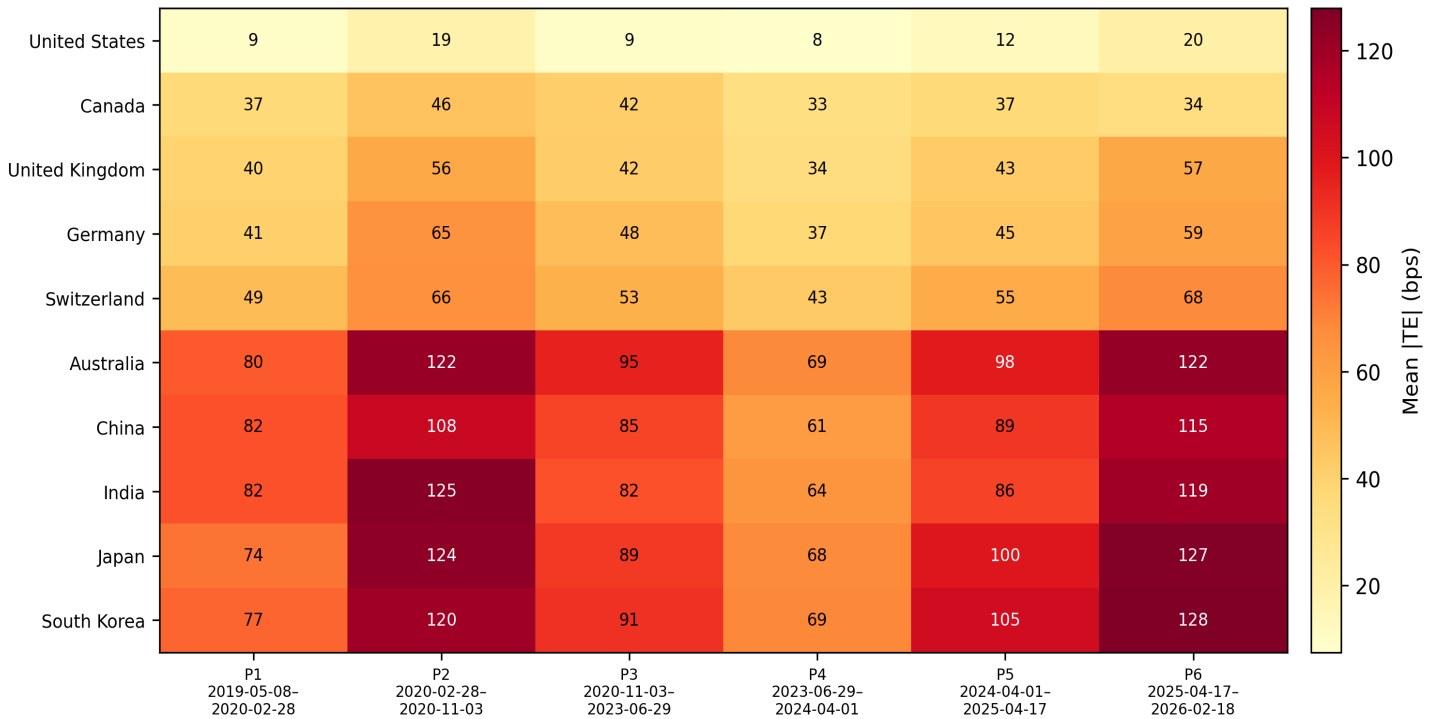


Fig. 6. Segment Mean $|TE|$ Heatmap

4.5 Synthesis across frameworks

The four frameworks were estimated independently, and their agreement is itself evidence. The Markov switching and Bai-Perron approaches test different kinds of instability, recurrent probabilistic transitions in one case and permanent level shifts in the other, yet they converge on the same crisis catalogue, with COVID-19 and the April 2025 tariff shock as the two most consequential disruptions. The connectedness analysis shows that the structure of transmission is itself regime-dependent, with the US receiver position deepening to -32.3 percent at the COVID peak. And the panel's heterogeneous factor loadings, the connectedness network's transmitter-receiver ordering, and the break analysis's amplitude hierarchy all sort the 10 markets identically. A single economic account fits all four: tracking errors are generated where arbitrage capacity is thin, propagate through correlated currencies, flows, and arbitrageur positioning, are absorbed where arbitrage capacity is deep, and everywhere intensify when volatility stress impairs the creation and redemption mechanism, with an amplitude set by each market's structural position.

5 Discussion

Three findings deserve emphasis for what they add to, and in one case overturn in, the existing literature.

The regime results give the liquidity-tracking mechanism of Bae and Kim (2020) its missing time dimension. Their evidence establishes that illiquidity causes tracking errors; the TVTP estimates here show that the mapping is state-dependent, with volatility stress raising both the probability of entering a high-error state and, more strongly, the probability of staying in it. The persistence margin matters practically, because it implies that tracking risk during stress is not a collection of independent bad days but an absorbing episode with a typical one-week duration and a heavy right tail. Conventional tracking error statistics are computed as regime-blind averages, and the averaging is costly: at the cross-country mean, the full-sample average of roughly 66 basis points understates stress-regime tracking risk of 85 basis points by more than a fifth, and the heavy right skew of the crisis regimes makes the understatement considerably larger in the tails.

The directional reversal in the connectedness network is the paper's most surprising result and, we would argue, its most informative one. The hypothesis that the US would dominate as a transmitter followed naturally from its share of gold ETF assets and price discovery, yet the tracking error network runs the other way, from periphery to centre, on virtually every day of the sample. The resolution is that price discovery and tracking-shock generation are different phenomena with opposite geographies: information originates where markets are deep, but tracking innovations originate where arbitrage is thin. Single-market evidence corroborates the distinction from the price-discovery side: in India, price formation runs from spot gold and gold futures to ETF prices rather than the reverse (Kaur & Singh, 2020), yet India in our network is a net transmitter of tracking shocks, so the same thin-participation segment that follows informationally is the one that exports replication noise. The finding extends the connectedness literature (Han, 2024; Shrestha et al., 2023) with a variable, the tracking error, whose network topology inverts the return-based intuition, and it complements the information-propagation evidence of Xu et al. (2020): the US can lead the system informationally while absorbing, rather

than exporting, replication noise. For portfolio and risk applications, the implication is that diversifying gold exposure across listing venues does not diversify tracking risk, because nearly two-thirds of that risk is a common, imported component, and it is most common precisely in the markets where levels are highest.

The structural results discipline the policy conversation. The three-tier hierarchy is not a ranking of financial development in general, as Australia's bottom position demonstrates; it is a ranking of gold-specific arbitrage capacity, shaped by authorised participant competition, in-kind creation and redemption mechanics, market access of the kind Kim et al. (2026) identify, and proximity to the locus of price discovery. The negative EPU coefficient adds a policy nuance: uncertainty that channels safe-haven flow through adequate infrastructure improves tracking by deepening liquidity, so the same shock can be tracking-compressing in Tier 1 and tracking-widening in Tier 3, another face of the multiplicative interaction. For regulators in markets at earlier stages of gold ETF development, the actionable levers the evidence points to are authorised participant entry, physical delivery logistics, and product competition, rather than market size per se.

The multiplicative interaction between structure and stress is, to our knowledge, new to the tracking literature, and it carries the paper's central message. The markets whose investors have the fewest alternative hedging instruments, and the episodes in which hedging is most valuable, are jointly where gold ETF tracking fidelity degrades most. A product class purchased as crisis insurance delivers its weakest replication in crises, and does so most severely where institutional substitutes are scarcest. The failure stacks on conditionality the literature has already documented: gold's own safe-haven performance depends on what drives the downturn (Ryan et al., 2024), and bullion hedges turbulence better than the funds that replicate it even before tracking failure is considered (Ji et al., 2026).

6 Conclusion

This study characterised the dynamics, transmission, and structural determinants of gold ETF tracking errors across 10 national markets through a crisis-dense seven-year window. Tracking errors switch between calm and stress regimes whose volatilities differ by multiples, with the VIX driving both regime entry and regime persistence and the great shocks of the sample producing synchronised, measurable elevations in stress-state probability. Nearly two-thirds of tracking error variance is a system-level phenomenon transmitted across borders, flowing from Switzerland and the high-error Asia-Pacific periphery to a North American centre that absorbs rather than exports shocks, in a reversal of the natural hypothesis. Beneath the dynamics sits a persistent structural hierarchy spanning nearly an order of magnitude, from 11.3 basis points in the US to 98.3 in Australia, that is decisively significant, maps onto arbitrage infrastructure and market access, and interacts multiplicatively with crisis exposure.

The practical implications are specific to each constituency. Investors should treat the domicile of a gold ETF as a risk parameter whose importance is itself state-dependent, and should recognise that constant-tracking-error risk models understate stress-state exposure in Tier 3 markets by a wide margin. Fund sponsors and authorised participants should read the persistence results as evidence that stress-state tracking degradation is an episode to be managed, not a sequence of independent shocks. And policymakers seeking to develop gold ETF markets have,

in the fixed effect hierarchy and its identified channels, an empirical benchmark for prioritising authorised participant access, in-kind mechanics, and delivery logistics.

The question this paper set out to answer was not how well gold ETFs track, but when, where, and why they fail. The answer has an uncomfortable symmetry: they fail together, they fail hardest where arbitrage is thinnest, and they fail most when the case for holding them is strongest. For investors and regulators alike, that is the tracking statistic that matters.

References

- Antonakakis, N., Chatziantoniou, I., & Gabauer, D. (2020). Refined measures of dynamic connectedness based on time-varying parameter vector autoregressions. *Journal of Risk and Financial Management*, 13(4), Article 84. <https://doi.org/10.3390/jrfm13040084>
- Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58(2), 277-297. <https://doi.org/10.2307/2297968>
- Bae, K., & Kim, D. (2020). Liquidity risk and exchange-traded fund returns, variances, and tracking errors. *Journal of Financial Economics*, 138(1), 222-253. <https://doi.org/10.1016/j.jfineco.2019.02.012>
- Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1), 1-22. <https://doi.org/10.1002/jae.659>
- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4), 1593-1636. <https://doi.org/10.1093/qje/qjw024>
- Broman, M. S. (2016). Liquidity, style investing and excess comovement of exchange-traded fund returns. *Journal of Financial Markets*, 30, 27-53. <https://doi.org/10.1016/j.finmar.2016.05.002>
- Caldara, D., & Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review*, 112(4), 1194-1225. <https://doi.org/10.1257/aer.20191823>
- Chen, J., & Xu, L. (2023). Do exchange-traded fund activities destabilize the stock market? Evidence from the China securities index 300 stocks. *Economic Modelling*, 127, Article 106450. <https://doi.org/10.1016/j.econmod.2023.106450>
- Chen, M.-P., Lee, C.-C., & Hsu, Y.-C. (2017). Investor sentiment and country exchange traded funds: Does economic freedom matter? *North American Journal of Economics and Finance*, 42, 285-299. <https://doi.org/10.1016/j.najef.2017.07.015>
- Cheng, W.-H., Chen, C.-D., & Lai, H.-P. (2020). Revisiting the roles of gold: Does gold ETF matter? *North American Journal of Economics and Finance*, 54, Article 100891. <https://doi.org/10.1016/j.najef.2018.12.003>
- Chudik, A., & Pesaran, M. H. (2015). Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics*, 188(2), 393-420. <https://doi.org/10.1016/j.jeconom.2015.03.007>

- Dannhauser, C. D. (2017). The impact of innovation: Evidence from corporate bond exchange-traded funds (ETFs). *Journal of Financial Economics*, 125(3), 537-560. <https://doi.org/10.1016/j.jfineco.2017.06.002>
- Demirer, M., Diebold, F. X., Liu, L., & Yilmaz, K. (2018). Estimating global bank network connectedness. *Journal of Applied Econometrics*, 33(1), 1-15. <https://doi.org/10.1002/jae.2585>
- Diebold, F. X., & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting*, 28(1), 57-66. <https://doi.org/10.1016/j.ijforecast.2011.02.006>
- Diebold, F. X., & Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1), 119-134. <https://doi.org/10.1016/j.jeconom.2014.04.012>
- 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>
- Han, S. (2024). Hedging strategies for U.S. factor and sector exchange-traded funds during geopolitical events. *Finance Research Letters*, 65, Article 105502. <https://doi.org/10.1016/j.frl.2024.105502>
- Hattori, T., & Yoshida, J. (2023). The impact of Bank of Japan's exchange-traded fund purchases. *Journal of Financial Stability*, 65, Article 101102. <https://doi.org/10.1016/j.jfs.2023.101102>
- Ji, H., Naeem, M., Jiang, J., & Ahmed, S. (2026). A study of dynamic dependence of gold and gold ETFs on energy ETFs. *Applied Economics*. Advance online publication. <https://doi.org/10.1080/00036846.2026.2647115>
- Kaur, P., & Singh, J. (2018). Testing the tracking efficiency of commodity exchange traded funds (ETFs) across the bearish and bullish market regimes: Evidence from Indian gold ETFs. *Journal of Asia-Pacific Business*, 19(3), 182-202. <https://doi.org/10.1080/10599231.2018.1490593>
- Kaur, P., & Singh, J. (2020). Price formation in Indian gold market: Analysing the role of gold exchange traded funds (ETFs) against spot and futures markets. *IIMB Management Review*, 32(1), 59-74. <https://doi.org/10.1016/j.iimb.2019.07.017>
- Kaviani, M., Fakhrehosseini, S. F., & Najafpoor, S. (2024). The efficiency of tracking gold exchange-traded funds in the Tehran Stock Exchange: Analysis of tracking error and double beta model. *Journal of Decisions and Operations Research*, 9(3), 747-759. <https://doi.org/10.22105/dmor.2024.475908.1867>
- Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1-2), 1-22. [https://doi.org/10.1016/0304-4076\(94\)90036-1](https://doi.org/10.1016/0304-4076(94)90036-1)
- Kim, J., Cho, H., & Seok, S. (2026). How trading barriers in underlying markets impact ETF trading and characteristics. *Research in International Business and Finance*, 81, Article 103186. <https://doi.org/10.1016/j.ribaf.2025.103186>

- Koop, G., & Korobilis, D. (2014). A new index of financial conditions. *European Economic Review*, 71, 101-116. <https://doi.org/10.1016/j.euroecorev.2014.07.002>
- Lee, C.-C., & Chen, M.-P. (2020a). Do natural disasters and geopolitical risks matter for cross-border country exchange-traded fund returns? *North American Journal of Economics and Finance*, 51, Article 101054. <https://doi.org/10.1016/j.najef.2019.101054>
- Lee, C.-C., & Chen, M.-P. (2020b). Happiness sentiments and the prediction of cross-border country exchange-traded fund returns. *North American Journal of Economics and Finance*, 54, Article 101254. <https://doi.org/10.1016/j.najef.2020.101254>
- Lee, C.-C., & Chen, M.-P. (2021). The effects of investor attention and policy uncertainties on cross-border country exchange-traded fund returns. *International Review of Economics and Finance*, 71, 830-852. <https://doi.org/10.1016/j.iref.2020.10.015>
- Li, Z., & Jiang, Y. (2025). ETF connectedness and its applications: Evidence from RCEP member countries. *Journal of Multinational Financial Management*, 78, Article 100908. <https://doi.org/10.1016/j.mulfin.2025.100908>
- Luo, X., Qin, S., & Ye, Z. (2016). The information content of implied volatility and jumps in forecasting volatility: Evidence from the Shanghai gold futures market. *Finance Research Letters*, 19, 105-111. <https://doi.org/10.1016/j.frl.2016.06.012>
- Marszk, A., & Lechman, E. (2024). What drives sustainable investing? Adoption determinants of sustainable investing exchange-traded funds in Europe. *Structural Change and Economic Dynamics*, 69, 63-82. <https://doi.org/10.1016/j.strueco.2023.11.018>
- Marta, T., & Riva, F. (2025). Do ETFs increase the comovements of their underlying assets? Evidence from a switch in ETF replication technique. *Journal of Banking & Finance*, 170, Article 107333. <https://doi.org/10.1016/j.jbankfin.2024.107333>
- Orlando, G. (2025). Exchange traded products: Taxonomy, risk and mitigations. *International Review of Financial Analysis*, 101, Article 103969. <https://doi.org/10.1016/j.irfa.2025.103969>
- Osterhoff, F., & Kaserer, C. (2016). Determinants of tracking error in German ETFs: The role of market liquidity. *Managerial Finance*, 42(5), 417-437. <https://doi.org/10.1108/MF-04-2015-0105>
- Pagano, M., Sánchez Serrano, A., & Zechner, J. (2019). *Can ETFs contribute to systemic risk?* (Reports of the Advisory Scientific Committee No. 9). European Systemic Risk Board. <https://doi.org/10.2849/45983>
- Papathanasiou, S., Syriopoulos, T., Kenourgios, D., & Koutsokostas, D. (2025). Sailing through uncertainty: Shipping's role in financial shock transmission and hedging strategies. *Global Finance Journal*, 67, Article 101159. <https://doi.org/10.1016/j.gfj.2025.101159>
- Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4), 967-1012. <https://doi.org/10.1111/j.1468-0262.2006.00692.x>

Pesaran, M. H. (2021). General diagnostic tests for cross-sectional dependence in panels. *Empirical Economics*, 60(1), 13-50. <https://doi.org/10.1007/s00181-020-01875-7>

Pesaran, M. H., & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics Letters*, 58(1), 17-29. [https://doi.org/10.1016/S0165-1765\(97\)00214-0](https://doi.org/10.1016/S0165-1765(97)00214-0)

Purohit, H., & Malhotra, N. (2015). Pricing efficiency & performance of exchange traded funds in India. *The IUP Journal of Applied Finance*, 21(3), 16-35. <https://ssrn.com/abstract=2671810>

Ryan, M., Corbet, S., & Oxley, L. (2024). Is gold always a safe haven? *Finance Research Letters*, 64, Article 105438. <https://doi.org/10.1016/j.frl.2024.105438>

Shrestha, K., Naysary, B., & Philip, S. S. S. (2023). Price discovery in carbon exchange traded fund markets. *International Review of Financial Analysis*, 89, Article 102814. <https://doi.org/10.1016/j.irfa.2023.102814>

Xu, L., Xu, L., Zhao, J., & Zhao, Y. (2020). Information-based trading and information propagation: Evidence from the exchange traded fund market. *International Review of Financial Analysis*, 70, Article 101495. <https://doi.org/10.1016/j.irfa.2020.101495>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [OnlineAppendix.docx](#)