

Trend Following in Modern Futures Markets

Separating opportunity decay, signal conversion, and implementation

Pratik Bang

2 September 2026

A 75-market study of the decline in time-series momentum, the conditions under which trend signals still convert movement into return, and the portfolio choices that preserve the remaining edge after trading costs and liquidity constraints.

Abstract

This paper asks why a transparent time-series momentum strategy earned much less after 2008 and which changes improve the portfolio without disguising that deterioration. I study a frozen universe of 75 global futures from 1990 through August 2026 and use the monthly 12/1 construction of Moskowitz, Ooi, and Pedersen (2012) as the sole benchmark: the sign of the preceding 12-calendar-month return determines direction, and the position is held during the next month. Gross Sharpe falls from 1.64 before 2009 to 0.52 after 2008. Under one prespecified cost model using actual two-leg Norgate delivery transitions, net Sharpe falls from 1.51 to 0.42. The market-level 12-month continuation coefficient simultaneously declines from 0.128 to 0.038, while path efficiency remains strongly associated with subsequent trend payoff. A 252-day linear-slope direction improves the modern net Sharpe to 0.59 but not decisively enough to replace the benchmark; a 50/50 sign-slope book delivers 0.53 and a shallower drawdown. Across 12 pre-specified lookback and holding-period rules, the largest decay occurs when the whole signal vintage is replaced each month. Combining all 12 rules into one equal-weight, internally netted book raises post-2008 net Sharpe to 0.62, lowers maximum drawdown from 18.8% to 12.8%, and reduces annual cost drag from 1.20% to 0.75%. A correlation-only optimizer does not improve that result. In the final 74-market funded simulation, the equal-weight multi-horizon book has the highest Sharpe above cash at 0.64 and a 9.1% maximum drawdown, whereas the sign-slope ensemble has the highest compound annual growth rate (CAGR) at 6.74%. The evidence favors simple diversification across trend definitions and signal vintages, internal trade netting, and explicit liquidity controls rather than parameter selection or additional covariance optimization.

Keywords: time-series momentum, trend following, futures, path efficiency, transaction costs, collateral yield, horizon diversification, liquidity constraints

1. Introduction

Trend following has an unusually simple premise: when a market moves persistently, take the same direction and scale the position so that no single contract dominates risk. The effect is documented across major futures groups and over samples extending well beyond the modern electronic-trading era (Moskowitz, Ooi, and Pedersen 2012; Hurst, Ooi, and Pedersen 2017). That historical breadth makes the weak years after the global financial crisis more, not less, interesting. A long record establishes that the strategy existed; it does not reveal whether today's weaker outcome comes from a smaller forecasting opportunity, poorer signal conversion, or the cost of implementation.

Those mechanisms imply different responses. If own-return continuation has weakened, changing an optimizer cannot recreate the missing premium. If large moves still occur but unfold through less efficient paths, signal shape and the rate at which positions are replaced become relevant. If several reasonable trend rules fail at different times, combining them may be more robust than choosing a recent winner. If gross returns survive but turnover and contract rolls consume the edge, execution is the binding constraint. The analysis therefore proceeds in that order: benchmark deterioration, underlying predictability, path quality, signal design, horizon design, portfolio combination, and finally funded implementation.

The empirical starting point follows Moskowitz, Ooi, and Pedersen (2012): direction is determined from a market's own past return rather than its rank against other markets. Baltas and Kosowski (2020) motivate treating the volatility estimator, trading rule, turnover, and cross-market dependence as separate design choices. I use that component-by-component discipline but do not replicate their continuous signal or correlation-scaling rule. The linear-slope test and the correlation-aware multi-horizon allocator introduced later are deliberately simpler, auditable experiments tailored to the present question.

The results lead to one connected conclusion rather than a search for a single winning indicator. The 12-month continuation premium is smaller, so the historical benchmark now begins with a lower gross return. Costs then consume a larger share of that thinner opportunity. Two forms of diversification improve how the residual premium is harvested: combining distinct definitions of a one-year trend, and combining lookback and holding periods so that the whole book is not forced to reverse on one rebalance date. The larger and more reliable improvement comes from horizon diversification. Once those targets are combined before execution, turnover and drawdown fall together.

My interpretation. I do not read the evidence as saying that trend following has stopped working. I read it as saying that a weaker continuation premium leaves less room for avoidable implementation loss and less justification for concentrated model choice. The benchmark remains the reference point; every proposed improvement must beat it after the same costs, actual delivery selection, whole-contract sizing, collateral treatment, and liquidity rule.

2. Data and strategy construction

2.1 Market universe and sample

The source universe contained 112 continuous futures series from Norgate Data. The research universe was frozen at 75 markets before strategy performance was examined, preserving representation across six economic groups and requiring sufficient history and metadata for volatility scaling, foreign-exchange (FX) conversion, and delivery-chain analysis. The sample begins in January 1990 and ends on 28 August 2026; strategy returns begin after the required lookback and volatility warm-up. The final funded implementation uses 74 markets because 30-Day Federal Funds futures fails the later exact-delivery liquidity screen. That exclusion is applied identically to every candidate portfolio and is unrelated to the contract's historical return (Norgate Data 2026).

Table 1. Composition of the frozen futures universe

Asset class	Markets	Share of universe
Stock indices	17	22.7%
Interest rates	19	25.3%
Currencies	9	12.0%
Energy	7	9.3%
Metals	5	6.7%
Agriculture/livestock	18	24.0%
Total	75	100.0%

Note. The 75-market universe is fixed for all research comparisons. Nine FX series convert local profit and loss into U.S. dollars and are not tradable markets. Section 10 reports the liquidity-screened 74-market funded implementation.

2.2 Monthly time series momentum benchmark

The sole benchmark follows Moskowitz, Ooi, and Pedersen (2012). The notation 12/1 means a 12-calendar-month lookback and a one-month holding period. At the final available observation of month t , each market receives a direction equal to the sign of its total price change over the preceding 12 calendar months. A positive return produces a long position and a negative return a short position. The signal and target are fixed throughout month $t + 1$, so no holding-month information enters the trade.

Each instrument is normalized to a 40% ex ante annualized volatility contribution before the risk budget is divided equally across eligible markets. This 40% input is an instrument-level normalization inherited from the benchmark construction; it is not a 40% portfolio-volatility target. The volatility estimate uses only information available at formation. Contract point value and explicit daily FX ratios translate local futures profit and loss into U.S. dollars. The number of eligible markets rises from 44 at the beginning of the return sample to the complete frozen universe as histories become available.

The benchmark is intentionally transparent. It does not estimate expected returns, optimize relative market weights, forecast correlations, or use discretionary regime filters. These choices make it an auditable starting point for testing whether modern performance differs from the historical record.

2.3 Return construction, costs and inference

The completed benchmark sample contains 426 non-overlapping calendar months from February 1991 through July 2026. Gross returns are futures excess returns before execution costs. Net returns deduct USD 2.50 per contract per transaction, liquidity-conditioned slippage, monthly resizing, and both legs of 13,859 observed delivery transitions in the completed sample. The outgoing roll leg is priced from the delivery's lagged five-session volume; the incoming leg uses the five-session volume of the exact successor delivery active on the transition date. Sections 3 through 9 report excess returns. Section 10 adds Treasury collateral, recursive account sizing, whole-contract rounding, and a market-level volume-participation cap.

The prespecified comparison separates months before January 2009 from months beginning in January 2009. Changes in mean return use heteroskedasticity-and-autocorrelation-consistent (HAC) standard errors (Newey and West 1987). Sharpe-ratio differences use the stationary bootstrap of Politis and Romano (1994), preserving short-run dependence through an expected 12-month block length. Market-level regressions use two-way clustering by market and calendar month (Cameron,

Gelbach, and Miller 2011). Holm's (1979) sequential adjustment controls the family-wise error rate across prespecified horizon families. Confidence intervals (CIs) are 95% intervals unless stated otherwise.

My interpretation. Gross and net are two views of the same monthly portfolio. Gross return identifies the economic opportunity available to the rule. Net return shows how much survives the single prespecified execution model. Reporting both prevents a weak signal from being blamed on costs and prevents an expensive implementation from being mistaken for a strong strategy.

3. Performance deterioration in the monthly benchmark

The benchmark remains profitable in both eras, but the reward per unit of risk is substantially lower after 2008. Gross CAGR falls from 19.6% before 2009 to 5.5% after 2008, while annualized volatility remains near 11.7%. Gross Sharpe therefore falls from 1.64 to 0.52. With the identical positions and the prespecified execution model, net CAGR falls from 17.9% to 4.3%, net Sharpe from 1.51 to 0.42, and maximum drawdown deepens from 11.6% to 18.8%.

Table 2. Monthly Moskowitz 12/1 benchmark before and after 2008

Period	Return layer	CAGR	Volatility	Sharpe ratio	Maximum drawdown
Full sample	Gross	12.4%	11.7%	1.07	-16.5%
Full sample	Net	10.9%	11.6%	0.95	-18.8%
Pre-2009	Gross	19.6%	11.4%	1.64	-11.2%
Pre-2009	Net	17.9%	11.4%	1.51	-11.6%
Post-2008	Gross	5.5%	11.7%	0.52	-16.5%
Post-2008	Net	4.3%	11.7%	0.42	-18.8%

Note. Gross and net results use identical monthly positions. Net results use the single prespecified commission and slippage schedule and the actual two-leg delivery transitions.

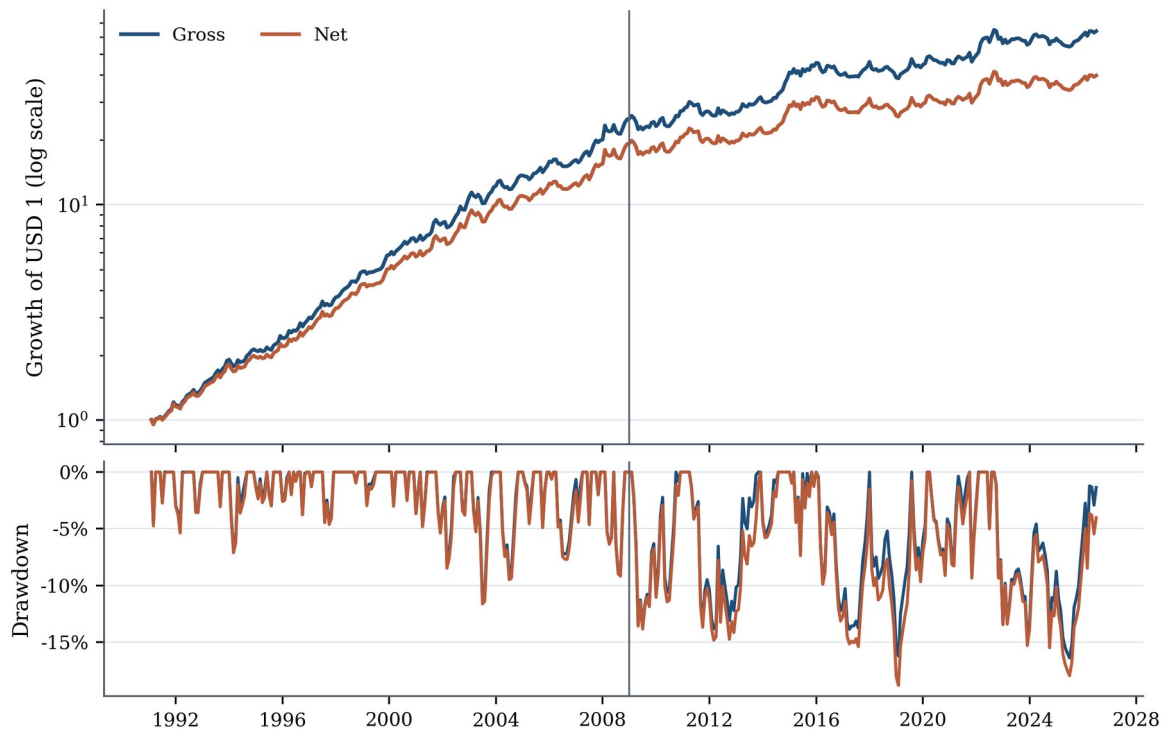


Figure 1. Gross and net wealth and drawdown for the monthly Moskowitz 12/1 benchmark

Note. The vertical line marks the prespecified January 2009 reporting split. Wealth is shown on a logarithmic scale. Both series contain completed months through July 2026.

My interpretation. Continued compounding after 2008 means the strategy did not cease to work. The important change is the slope of the wealth curve: almost unchanged realized volatility earned far less return. For a contemporary hurdle, the post-2008 net Sharpe of 0.42 is more informative than the full-sample value of 0.95; the latter averages the strong early regime with the weaker modern one.

3.1 Statistical evidence for the change

The gross annualized mean return falls by 12.65 percentage points after 2008. The HAC standard error is 3.44 percentage points, the t-statistic is -3.68 , and the p-value is 0.00027. The 95% CI ranges from -19.40 to -5.91 percentage points. This is an arithmetic mean-return comparison, so it is not numerically identical to the 14.1-point gap between the two gross CAGRs. The gross Sharpe change is -1.13 , with a stationary-bootstrap 95% CI of $[-1.69, -0.62]$.

Table 3. Formal change tests for the monthly benchmark

Test	Estimate	Uncertainty	p-value	Inference
Gross annual mean change	-12.65 pp	95% CI: -19.40 to -5.91 pp	0.00027	Supported
Gross Sharpe change	-1.13	95% bootstrap CI: -1.69 to -0.62	<0.001	Supported

Note. Changes are post-2008 minus pre-2009. The mean test uses Newey-West HAC inference; the Sharpe-ratio test uses a stationary bootstrap.

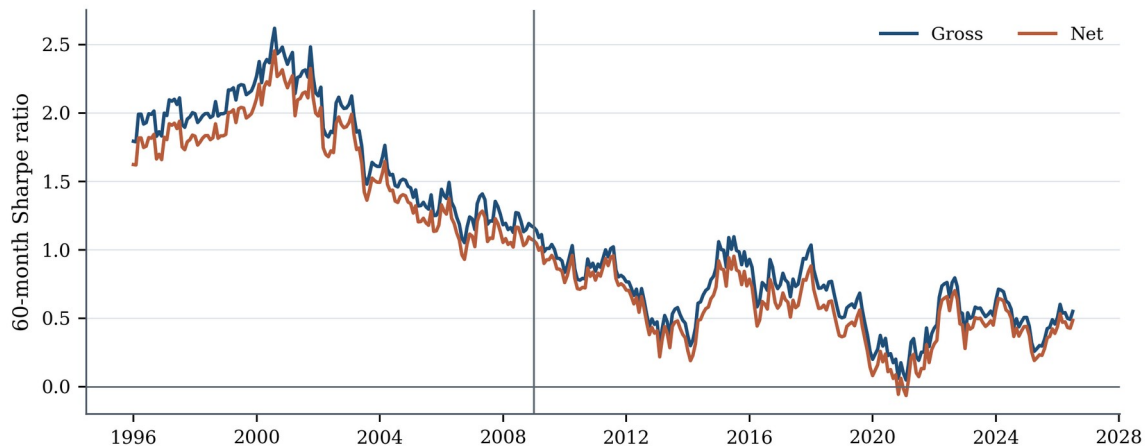


Figure 2. Rolling 60-month gross and net Sharpe ratios for the monthly benchmark

Note. Each point uses the preceding 60 completed monthly returns. Because adjacent estimates share 59 months, the plot documents persistent weakness but cannot identify the date or form of a structural break.

My interpretation. The decline is too large to dismiss as ordinary sampling variation. The rolling plot shows that weakness persisted across many windows, but its overlap mechanically smooths the path. I therefore use January 2009 as a prespecified reporting boundary, not as evidence that the market changed permanently in one month.

4. Did the underlying trend opportunity weaken?

Portfolio performance can deteriorate even when the underlying market effect is unchanged. The next test removes the portfolio and asks directly whether a market's own past direction still predicts its next return.

4.1 Market level continuation regressions

The opportunity panel contains 28,834 market-month records across all 75 markets and 439 calendar months; regression counts decline at longer horizons as additional history is required. For each formation horizon, the sign of a market's own past return predicts its next volatility-standardized monthly return. A positive coefficient indicates continuation. The coefficient is not a portfolio return: it is a direct measure of the raw forecasting opportunity available to a broad class of time-series trend rules.

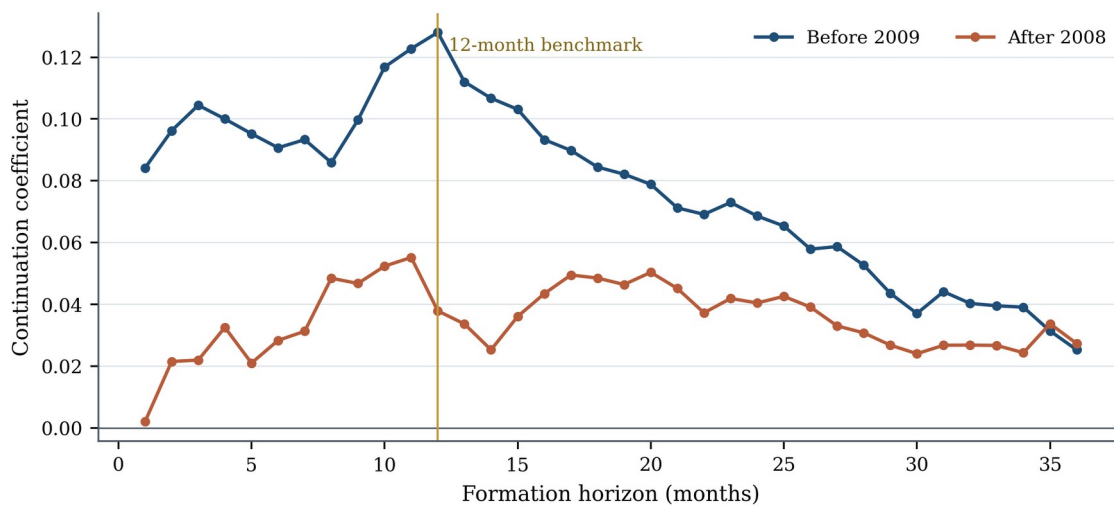
The strongest confirmed change occurs at the benchmark's 12-month horizon. The coefficient falls from 0.128 before 2009 to 0.038 after 2008. The change is -0.090 , the two-way-clustered t-statistic is -3.17 , and the Holm-adjusted p-value is 0.018. The modern coefficient remains positive but its own p-value is 0.072. The evidence therefore supports compression of the continuation effect, not a confident claim that it vanished or reversed.

Table 4. Selected sign continuation coefficients

Formation horizon	Before 2009	After 2008	Change	t-statistic	Holm-adjusted p
1 month	0.084	0.002	-0.082	-2.66	0.078
3 months	0.104	0.022	-0.082	-2.75	0.066
6 months	0.091	0.028	-0.062	-2.16	0.144
12 months	0.128	0.038	-0.090	-3.17	0.018
24 months	0.069	0.040	-0.028	-0.97	1.000
36 months	0.025	0.027	+0.002	+0.07	1.000

Note. Standard errors are clustered by market and calendar month. Holm-adjusted p-values refer to the complete prespecified horizon family, not only the six rows displayed.

Figure 3 shows the complete horizon profile. Before 2009, continuation rises toward an intermediate peak near 12 months. After 2008, the profile is lower at short and intermediate horizons and converges with the earlier profile at long horizons. It does not become systematically negative. The deterioration is therefore a flattening of the profitable continuation hump rather than a clean migration from momentum to long-horizon reversal.

**Figure 3. Own-return continuation across formation horizons**

Note. Coefficients are estimated separately before 2009 and after 2008 with two-way-clustered inference. The vertical reference marks the benchmark's 12-month lookahead.

My interpretation. This is the clearest evidence that part of the portfolio deterioration begins in the opportunity set itself. An optimizer cannot manufacture the missing 12-month continuation premium. Signal design can still improve conversion, but its realistic task is to harvest a smaller residual opportunity more consistently rather than assume the historical edge is waiting to be recovered in full.

4.2 Persistence, reversal, and asset class heterogeneity

A weaker 12-month coefficient does not imply that every trend became shorter. Coefficients from 13 through 36 months remain positive, so the aggregate data do not show a universal continuation-to-reversal crossover. A secondary reversal-hazard model nevertheless estimates a 3.35-percentage-point increase in the probability of a next-month sign reversal after 2008 ($t = 2.55$, $p = 0.011$). Local reversals became more common even though long-horizon continuation did not invert.

The 12-month decline is uneven across asset classes. Figure 4 reports exploratory estimates and 95% CIs. The coefficient change is -0.200 for stock indices and -0.150 for currencies; both unadjusted p-values are about 0.002. Metals and energy also weaken, but their intervals include zero. Agriculture/livestock and interest rates are more resilient. Because these class-level tests were not protected by a separate multiplicity adjustment, they locate where the aggregate change is concentrated but do not justify tactical sector weights.

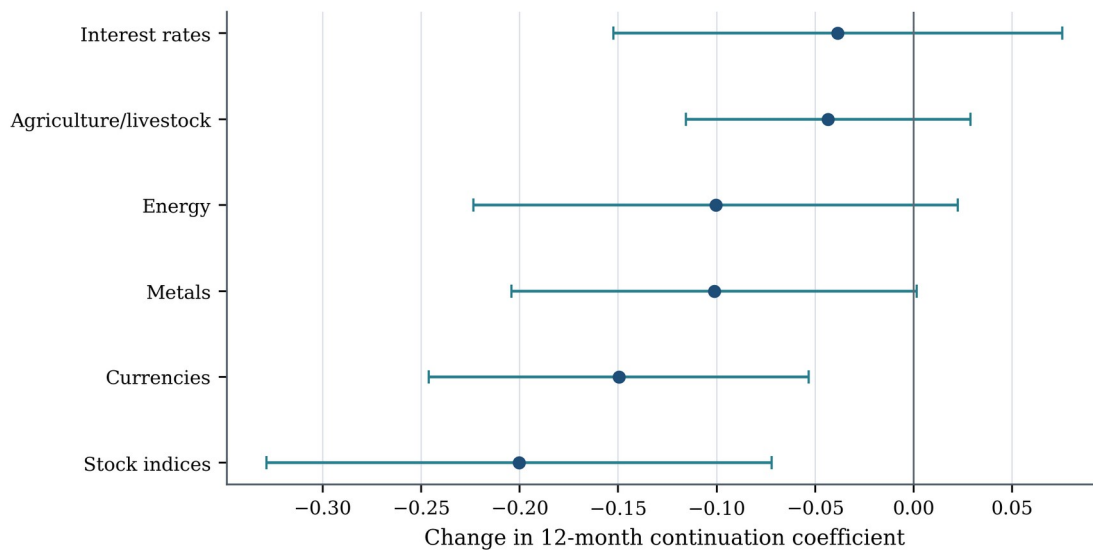


Figure 4. Change in the 12-month continuation coefficient by asset class

Note. Points show the post-2008 coefficient minus the pre-2009 coefficient; bars are 95% CIs. This decomposition is exploratory and is not used as an allocation rule.

My interpretation. I would not shorten every rule or remove equities and currencies after seeing these estimates. The absence of a universal reversal crossover argues for testing several speeds. The class heterogeneity is better used as a robustness challenge: an improvement should not depend only on the two groups that deteriorated most.

5. Path quality and signal conversion

The opportunity tests establish that continuation weakened. They do not explain why the same amount of movement can be easy or difficult for a trend rule to monetize. That requires a measure of the path, not only its volatility.

5.1 Measuring path efficiency

Path efficiency follows the logic of Kaufman's efficiency ratio (Kaufman 1995). It divides the absolute 252-day net price change by the sum of absolute daily price changes over the same window. A smooth directional path approaches one. A market that travels far but repeatedly reverses and finishes near its starting point approaches zero. The statistic therefore separates directionality from movement magnitude.

$$\text{Efficiency}_{i,t} = |P_{i,t} - P_{i,t-252}| / \sum_{k=1}^{252} |P_{i,t-k+1} - P_{i,t-k}| \quad (1)$$

The daily source panel contains 595,894 observations. After a five-year expanding warm-up used to form lagged states, the forward sample contains 22,860 market-month observations. The primary model explains next-month trend payoff using standardized lagged volatility, standardized lagged path efficiency, their interaction, signal strength, and market fixed effects. Standard errors are clustered by market and calendar month.

5.2 Regression evidence

Path efficiency has a positive coefficient of 0.194 risk units ($z = 5.34$; Holm-adjusted $p < 0.001$). Volatility has a coefficient of -0.025 (Holm-adjusted $p = 0.217$), and the volatility-by-efficiency interaction is -0.014 (Holm-adjusted $p = 0.217$). Once path quality is measured directly, volatility alone does not provide a statistically supported explanation of trend payoff.

Table 5. Primary path quality regression and monotonicity tests

Quantity	Estimate	Test statistic	Holm-adjusted p	Economic reading
Path efficiency	+0.194 risk units	$z = 5.34$	<0.001	Strong positive association
Volatility	-0.025 risk units	$z = -1.61$	0.217	Not supported
Volatility $\times$ path efficiency	-0.014 risk units	$z = -1.29$	0.217	Not supported
High minus low efficiency; low volatility	+10.5 pp annualized	$z = 4.12$	<0.001	Supported
High minus low efficiency; middle volatility	+8.7 pp annualized	$z = 4.45$	<0.001	Supported
High minus low efficiency; high volatility	+6.7 pp annualized	$z = 3.63$	0.001	Supported

Note. Annualized contrasts are expressed at a 10% volatility target. Holm correction is applied within the prespecified path-quality test families.

Figure 5 reports raw state-conditioned averages. Moving from low to high path efficiency raises annualized gross return from 0.7% to 7.7% in the low-volatility row, from 0.5% to 5.8% in the middle row, and from 0.8% to 4.2% in the high row. Table 5 reports larger regression-adjusted high-minus-low contrasts of 10.5, 8.7, and 6.7 percentage points because it holds the model's other covariates constant; Figure 5 does not. Both estimands show the same monotonic relation across path quality, and neither shows a reliable monotonic benefit from higher volatility.

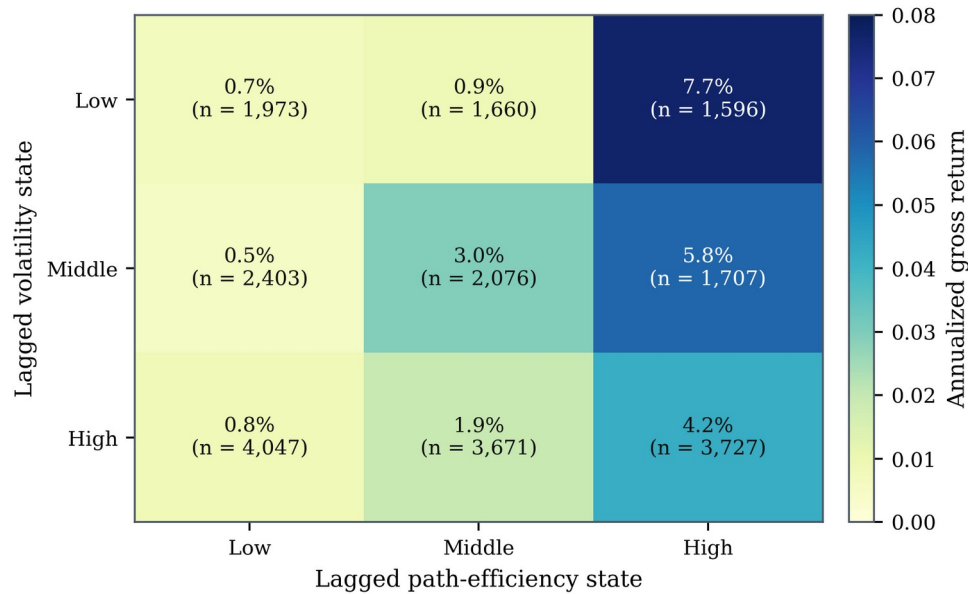


Figure 5. Trend payoff across lagged volatility and path-efficiency states

Note. Cells are raw annualized gross-return averages at a 10% volatility target; observation counts appear in parentheses. Lagged expanding thresholds define the states.

My interpretation. The payoff relation is tied to the cleanliness of a move, not to large daily fluctuations by themselves. This test identifies environments in which a trend rule is more or less effective; it does not establish that low-efficiency environments became more prevalent after 2008. I would therefore treat path efficiency as an overlay candidate, not a production filter, until an expanding-window out-of-sample test measures both saved whipsaw costs and the abrupt trends such a filter would miss.

5.3 Directional capture and the whipsaw falsification

The directional-capture test asks how much of the next month's absolute move is captured when the prior signal points in the realized direction. Figure 6 groups next-month moves into quintiles and compares three lookback speeds. Before 2009, capture rises sharply in the largest-move quintile. After 2008, those moves still produce positive payoff, but capture is lower at every speed. In the largest quintile it falls from 17.3% to 6.0% for the 63-day signal, from 13.4% to 7.6% for the 126-day signal, and from 18.1% to 7.3% for the 252-day signal.

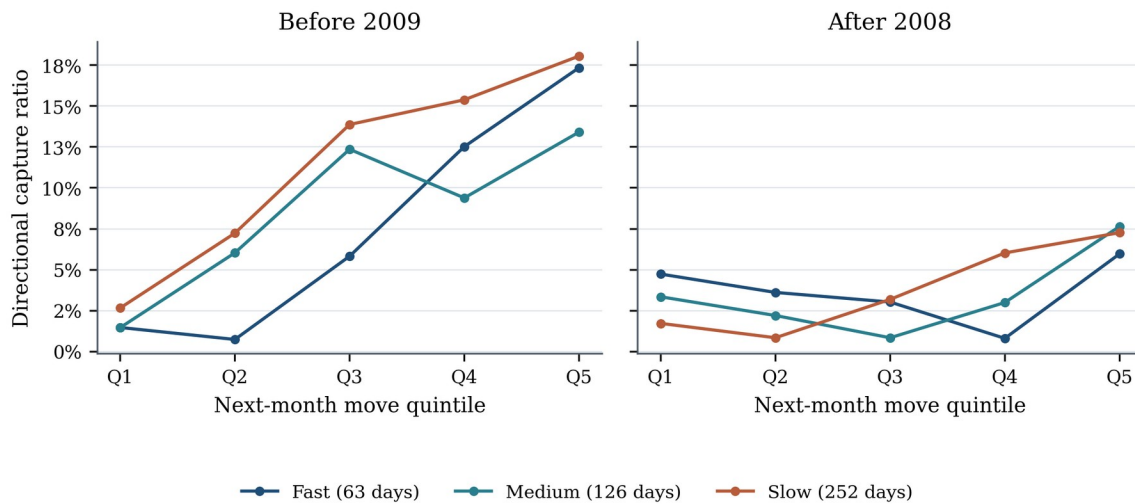


Figure 6. Directional capture by move magnitude and signal speed

Note. The capture ratio divides aligned trend payoff by the absolute next-month move. This is a descriptive signal-conversion diagnostic, not a separately traded strategy.

A tempting explanation is that more frequent whipsaws caused the entire deterioration. The falsification evidence is weaker than that story. Flip frequency rises by 0.059 flips per market-month, but its bootstrap CI includes zero. Average loss per flip falls rather than rises and is not statistically different across eras. Average recovery time becomes 3.55 days shorter, with a CI that excludes zero.

Table 6. Whipsaw evidence before and after 2008

Metric	Before 2009	After 2008	Change	95% bootstrap CI
Flips per market-month	0.443	0.502	+0.059	[−0.007, 0.128]
Average loss per flip (daily risk units)	5.179	4.885	−0.294	[−0.771, 0.184]
Average recovery time (days)	38.33	34.78	−3.55	[−6.71, −0.27]

Note. Stationary bootstrap with an expected 12-month block length and 5,000 replications. Recovery time is an average across market-month observations.

My interpretation. The claim that modern markets simply flip more often and punish every flip more severely is not supported. The narrower reading is that weaker continuation and less efficient paths reduce the fraction of large moves captured by the monthly rule. That motivates testing continuous and multi-speed signals, but it does not validate a generic anti-whipsaw filter.

6. Transaction costs and actual roll implementation

Gross and net returns have been reported together from the outset; this section isolates the implementation wedge before changing the signal. Each contract transaction is charged USD 2.50 in commission. Routine slippage is 0.25 tick when lagged 63-session average volume is at least 100,000 contracts, 0.5 tick between 20,000 and 99,999, and 1 tick below 20,000. The schedule is applied identically to entries, exits, monthly resizing, and each leg of a roll. The paper reports this single prespecified institutional base case rather than presenting a menu of cost scenarios.

Roll dates come from the active contract identifier in the Norgate continuous series, not from an assumed annual frequency (Norgate Data 2026). Expiry identifiers are normalized to contract month because the source format changes for part of 2026. The audit identifies 13,875 forward transitions, of which 13,859 occur in the completed benchmark sample. The outgoing leg uses lagged five-session volume and the incoming leg the exact successor delivery's transition-day five-session volume; neither leg required a missing-volume fallback. This distinction matters because 958 transitions skip at least one listed delivery. Consistent with the implementation logic emphasized by Baltas and Kosowski (2020), the adjacent-contract price spread remains futures carry embedded in the return series rather than being misclassified as an execution cost.

Table 7. Effect of execution costs on the monthly Moskowitz benchmark

Period	Gross CAGR	Net CAGR	Net volatility	Net Sharpe	Net maximum drawdown
Full sample	12.4%	10.9%	11.6%	0.95	-18.8%
Pre-2009	19.6%	17.9%	11.4%	1.51	-11.6%
Post-2008	5.5%	4.3%	11.7%	0.42	-18.8%

Note. Net returns deduct commission, liquidity-tier slippage, monthly resizing, and exact two-leg delivery transitions. Arithmetic annual cost drag is 1.35 percentage points over the full sample and 1.20 points after 2008.

My interpretation. Costs matter, but they are not the primary source of deterioration. Post-2008 gross CAGR is already only 5.50%; the execution model reduces it to 4.26%. The main problem is a thinner gross opportunity, after which costs consume a larger fraction of what remains. The benchmark is economically positive but not compelling as a stand-alone institutional allocation.

7. Does a second trend definition improve the modern portfolio?

The evidence now narrows the design problem: continuation weakened, path quality affects capture, and costs amplify rather than cause the decline. The first candidate changes only the definition of a one-year trend, replacing the binary 12-month price difference with the sign of an ordinary-least-squares slope fitted to 252 daily closes. Both signals share timing, volatility scaling, FX conversion, roll dates, and costs, so the comparison isolates signal definition rather than speed. This directional slope is deliberately simple and is not the continuous trend-strength rule of Baltas and Kosowski (2020).

The signals agree in 88.0 percent of market months. Their 12.0 percent disagreement supplies the diversification test: can two closely related measurements of the same phenomenon behave differently when the price path is noisy?

Table 8. Net performance by one-year trend definition

Period	Trend definition	CAGR	Volatility	Sharpe ratio	Maximum drawdown
Full sample	Return sign	10.9%	11.6%	0.95	-18.8%
Full sample	Linear slope	10.0%	11.0%	0.92	-20.1%
Post-2008	Return sign	4.3%	11.7%	0.42	-18.8%
Post-2008	Linear slope	6.3%	11.4%	0.59	-20.1%

Note. The post-2008 linear-minus-sign annual mean difference is 1.93 percentage points (HAC $p = 0.120$). Its Sharpe-ratio advantage is 0.177 with a stationary-bootstrap 95% CI of $[-0.022, 0.383]$.

My interpretation. The slope signal is more attractive in the modern sample, but the paired evidence is not strong enough to declare it the superior stand-alone rule. It gives up substantial pre-2009 performance and its full-sample maximum drawdown is 20.13%, 1.30 percentage points deeper than the sign benchmark. Its defensible role is as a second measurement of trend, not as a recent winner selected after observing the split.

7.1 A 50/50 signal ensemble

The ensemble combines half of each market's sign target and half of its slope target before whole-contract rounding and cost calculation. Opposing targets therefore net rather than trade against one another; signal disagreement cancels the position in 12.0% of eligible market-months. This single implementable book tests signal diversification without double-counting offsetting turnover.

Table 9. Net performance of the 50/50 sign-slope ensemble

Period	Portfolio	CAGR	Volatility	Sharpe ratio	Maximum drawdown
Full sample	Return-sign benchmark	10.9%	11.6%	0.95	-18.8%
Full sample	Linear-slope rule	10.0%	11.0%	0.92	-20.1%
Full sample	50/50 sign-slope ensemble	10.6%	11.0%	0.98	-17.5%
Post-2008	Return-sign benchmark	4.3%	11.7%	0.42	-18.8%
Post-2008	Linear-slope rule	6.3%	11.4%	0.59	-20.1%
Post-2008	50/50 sign-slope ensemble	5.4%	11.2%	0.53	-17.5%

Note. The post-2008 ensemble-minus-benchmark Sharpe difference is 0.112, with a paired stationary-bootstrap 95% CI of [0.016, 0.212]. The annual mean-return advantage is 1.06 percentage points (HAC $p = 0.088$). Internal target netting adds approximately 0.09 percentage point per year after 2008.

The ensemble does not beat the slope signal on post-2008 return or Sharpe. Its advantage is balance. It gives up some of the slope signal's modern return, retains more of the stronger early sign record, and produces the smallest maximum drawdown of the three books. Over the full sample its Sharpe is 0.98, compared with 0.95 for the sign benchmark and 0.92 for the slope rule.

My interpretation. The result supports a measured portfolio change. The post-2008 Sharpe improvement over the benchmark is statistically credible and the drawdown improvement is economically visible, but the absolute gain is modest. I would combine the two definitions rather than replace the benchmark with the recent winner. A post-2008 net Sharpe of 0.53 remains far below the benchmark's pre-2009 value of 1.51.

8. Does deterioration depend on trend horizon?

The grid contains 12 combinations: lookbacks of 3, 6, 12, and 24 calendar months and holdings of 1, 3, and 6 months. The 12/1 rule remains the sole benchmark. Holdings longer than one month are overlapping vintages: a six-month rule holds six one-sixth-sized cohorts formed in consecutive months. Each month only the oldest cohort expires and one new cohort enters. This avoids dependence on an arbitrary starting month and separates how quickly new information enters from how abruptly the old position is discarded.

All cells use the same 75 markets, lagged volatility estimate, equal market-risk allocation, daily FX conversion, whole-contract sizing, cost model, and 13,859 actual delivery transitions. The common sample runs from July 1992 through July 2026. Mean changes use Newey-West inference with 12 lags and Holm adjustment across the 12 cells. Sharpe changes use 5,000 stationary-bootstrap samples with an expected 12-month block length.

Table 10. Net performance deterioration across lookback and holding periods

Lookback (months)	Holding (months)	Pre-2009 net Sharpe	Post-2008 net Sharpe	Sharpe change	Post-2008 net CAGR	Sharpe-change 95% CI
3	1	1.15	0.23	-0.91	2.1%	[-1.45, -0.42]
3	3	0.82	0.42	-0.41	3.7%	[-0.96, 0.16]
3	6	0.63	0.77	+0.14	6.0%	[-0.41, 0.68]
6	1	0.91	0.33	-0.59	3.1%	[-1.12, -0.06]
6	3	0.81	0.54	-0.27	5.4%	[-0.85, 0.26]
6	6	0.86	0.67	-0.19	6.2%	[-0.79, 0.35]
12	1	1.51	0.42	-1.10	4.3%	[-1.66, -0.59]
12	3	1.31	0.50	-0.81	5.2%	[-1.38, -0.31]
12	6	1.06	0.59	-0.47	6.2%	[-1.04, 0.04]
24	1	0.85	0.47	-0.38	5.0%	[-0.87, 0.11]
24	3	0.72	0.43	-0.29	4.4%	[-0.80, 0.22]
24	6	0.53	0.26	-0.28	2.3%	[-0.83, 0.26]

Note. The shaded row is the 12/1 benchmark. Only the 3/1 and 12/1 annual mean declines remain significant after Holm adjustment, with p -values of 0.0096 and 0.0008. Bootstrap CIs describe each Sharpe change and are not Holm-adjusted.

The 12/1 benchmark suffers the largest Sharpe decline in the grid: 1.51 before 2009 to 0.42 after 2008, a change of -1.10 . The 3/1 rule also deteriorates sharply, from 1.15 to 0.23. By contrast, extending the holding period improves the modern result for 3-, 6-, and 12-month lookbacks. The clearest case is 3/6: post-2008 net Sharpe is 0.77 and net CAGR is 6.02%, compared with 0.23 and 2.08% for 3/1. Its Sharpe change is $+0.14$, but the CI includes zero; the defensible conclusion is preservation, not evidence that its expected return increased.

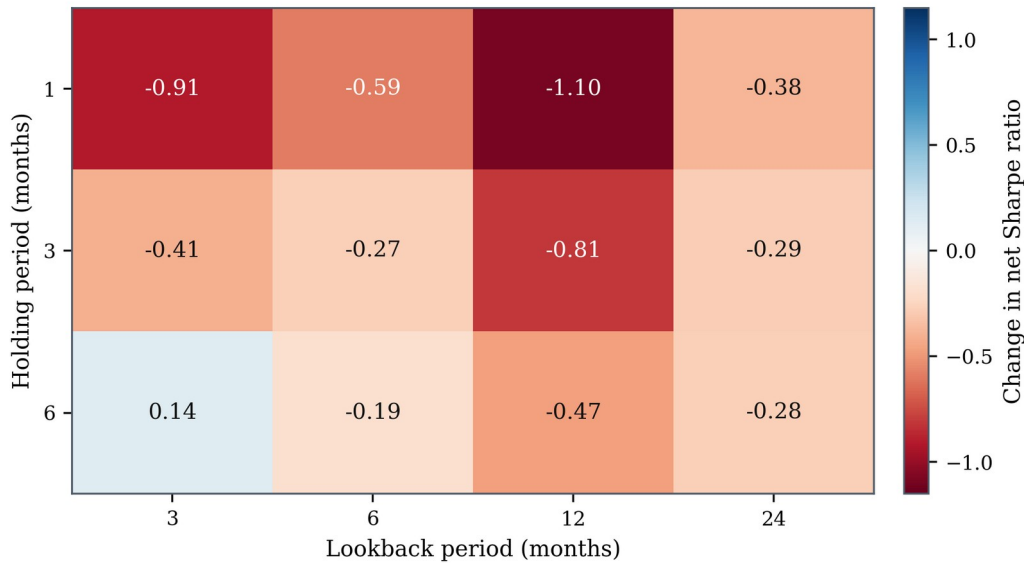


Figure 7. Change in net Sharpe after 2008 across lookback and holding periods

Note. Each cell is post-2008 net Sharpe minus pre-2009 net Sharpe. Positive values indicate preservation or improvement; negative values indicate deterioration.

This is not merely a transaction-cost result. Moving from one to six holding months lowers post-2008 annual cost drag from 1.31% to 0.63% for the 3-month signal, from 1.25% to 0.85% for the 6-month signal, and from 1.20% to 0.91% for the 12-month signal. Gross Sharpe also rises in each comparison. Overlapping vintages therefore improve both implementation efficiency and signal conversion. The exception is the already-slow 24-month lookback: carrying it for six months leaves a post-2008 net Sharpe of only 0.26, consistent with stale information rather than useful smoothing.

My interpretation. Fast and slow are not single labels. A rule can admit new information quickly while replacing the existing book gradually. The 3/6 rule does exactly that: it adds a fresh three-month view every month but removes only one-sixth of its prior vintages. It reacts without allowing one reversal to flip the entire position. The useful insight is the separation of signal speed from replacement speed. This is not a claim that 3/6 is a universal optimum.

9. Does horizon diversification survive as one portfolio?

9.1 The equal-weight multi-horizon portfolio

The horizon grid shows that decay differs across lookback and holding periods, but it does not yet show that an investor can capture that pattern in one portfolio. Selecting 3/6 because it has the highest post-2008 Sharpe would use the modern sample twice: first to discover the result and then to declare a winner. The primary portfolio therefore retains every pre-specified rule rather than selecting any cell.

The equal-weight multi-horizon book assigns one-twelfth of the risk budget to each combination of 3-, 6-, 12-, and 24-month lookbacks with 1-, 3-, and 6-month holdings. For every market and month, all 12 targets are weighted and summed before whole-contract rounding. Opposing views offset inside the book, and commission and slippage are charged only on the resulting net trade. This is one implementable portfolio, not an after-the-fact average of 12 separately traded return series.

Table 11. Performance of the equal-weight multi-horizon book combining 12 rules

Period	Net CAGR	Net volatility	Net Sharpe	Maximum drawdown	Annual cost drag
Full sample	7.51%	9.08%	0.84	-12.84%	0.79%
Pre-2009	11.38%	9.68%	1.17	-10.55%	0.85%
2009 to 2019	4.18%	8.52%	0.52	-12.84%	0.93%
2020 to July 2026	6.52%	8.83%	0.76	-9.05%	0.44%
Post-2008	5.05%	8.62%	0.62	-12.84%	0.75%

Note. The common portfolio-comparison sample is July 1997 through July 2026. Net returns deduct commission, liquidity-conditioned slippage, resizing, and both legs of actual delivery transitions. Calendar spreads remain part of futures carry. The 2020 to July 2026 subperiod is descriptive and is not used for model selection or formal inference.

The book remains profitable in every reported subperiod, but Table 11 prevents an overly optimistic reading. Net Sharpe falls from 1.17 before 2009 to 0.52 during 2009 to 2019. It improves to 0.76 from 2020 through July 2026, with a 6.52% net CAGR and a 9.05% maximum drawdown. Horizon diversification therefore mitigates deterioration and participates in the more favorable recent environment; it does not restore the pre-2009 edge.

The mechanism is timing diversification. Short lookbacks recognize new directions earlier, long lookbacks remain committed to established moves, and overlapping holding vintages prevent the entire book from changing on one rebalance date. Because targets are combined by market before trading, disagreement reduces the external order rather than creating offsetting transactions. The strategy diversifies decision timing and execution timing at once.

Relative to the monthly 12/1 benchmark, the equal-weight multi-horizon book improves the modern result in three dimensions. Post-2008 net Sharpe rises from 0.42 to 0.62, maximum drawdown falls from 18.83% to 12.84%, and annual cost drag falls from 1.20% to 0.75%. Net CAGR rises from 4.26% to 5.05% even though annualized volatility declines from 11.66% to 8.62%. Annual contract turnover falls from approximately 181,674 to 113,482 contracts because opposing horizon targets are netted before execution.

Paired inference confirms a risk-adjusted improvement rather than a reliably higher mean return. The post-2008 Sharpe difference between the equal-weight book and the benchmark is +0.20, with a stationary-bootstrap 95% CI of [0.04, 0.36]. The book also suffers 0.43 less Sharpe deterioration, with a CI of [0.12, 0.72]. Its annual mean-return advantage is only 0.45 percentage point, however, with a Holm-adjusted p-value of 1.00. The evidence supports more consistent harvesting of the remaining trend premium. It does not show that combination raises expected return or recreates the old opportunity set.

My interpretation. I regard the equal-weight multi-horizon book as a main result, not merely an input to an optimizer.

It replaces the fragile choice of one apparently superior cell with an implementable portfolio whose advantage comes from diversified reaction speeds, overlapping vintages, and internal trade netting. That is a more credible design improvement than choosing the best modern parameter after observing the grid.

9.2 Does correlation-aware weighting improve equal weighting?

Only after establishing the equal-weight book as a complete portfolio do I ask whether historical sleeve correlations can improve it. Beginning in July 1997, after a 60-month estimation window, the allocator uses expanding gross-return history available only through the previous month. The correlation matrix is shrunk 50% toward the identity; weights are long-only and capped at 20% per sleeve; and the objective minimizes variance without an expected-return forecast. This is a prespecified correlation-only experiment, not the correlation-dependent portfolio-leverage adjustment proposed by Baltas and Kosowski (2020). Universe, risk scaling, FX conversion, whole-contract implementation, costs, and roll events are identical to the equal-weight test.

Table 12. Equal-weight and correlation-aware multi-horizon portfolios

Portfolio	Pre-2009 net Sharpe	Post-2008 net Sharpe	Sharpe change	Post-2008 net CAGR	Post-2008 maximum drawdown	Post-2008 annual cost drag
12/1 benchmark	1.39	0.42	−0.98	4.26%	−18.83%	1.20%
Equal-weight multi-horizon book	1.17	0.62	−0.55	5.05%	−12.84%	0.75%
Correlation-aware multi-horizon book	1.15	0.58	−0.57	4.56%	−12.24%	0.73%

Note. The common allocation sample is July 1997 through July 2026. It begins later than the July 1992 horizon-grid sample because the optimizer requires 60 prior months. This common sample explains why the benchmark's pre-2009 Sharpe here is 1.39 rather than 1.51 in earlier tables.

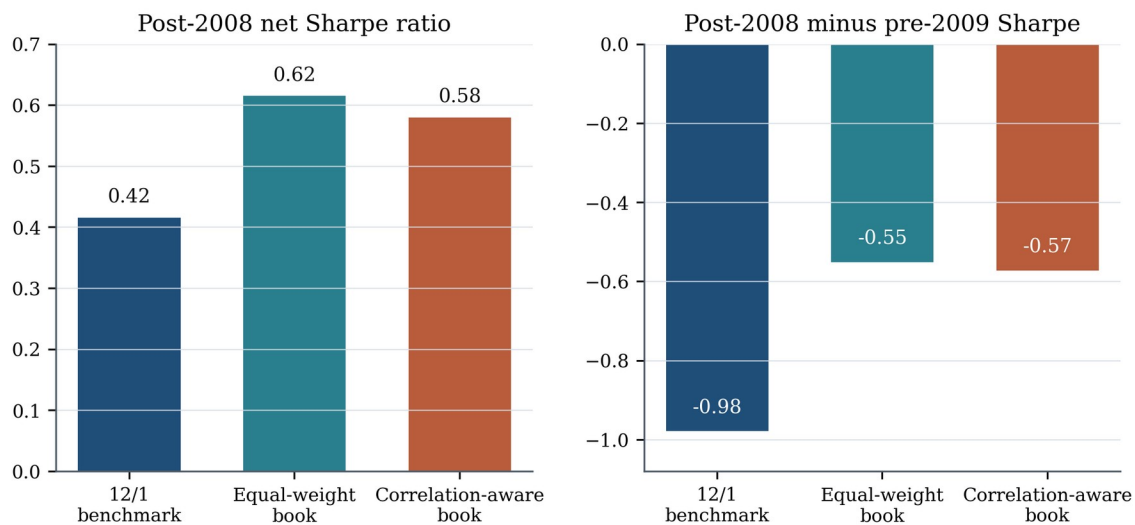


Figure 8. Modern net Sharpe and deterioration across portfolio constructions

Note. The left panel reports post-2008 net Sharpe. The right panel reports post-2008 minus pre-2009 net Sharpe; values nearer zero indicate less deterioration.

Correlation-aware allocation reduces post-2008 volatility slightly further, from 8.62% to 8.29%, and maximum drawdown from 12.84% to 12.24%. Those small risk reductions do not improve total performance. Net CAGR falls from 5.05% to 4.56% and Sharpe from 0.62 to 0.58. Relative to equal weighting, the Sharpe difference is -0.04 with a 95% CI of $[-0.10, 0.03]$, and the annual mean-return difference is -0.50 percentage point with a Holm-adjusted p-value of 0.40. There is no statistical evidence that the optimizer improves the diversified book.

The weight pattern explains why. On average the optimizer assigns 18.5% to 3/1 and 15.5% to 24/6, but only 3.1% to 12/3. It rewards the first two sleeves because their return patterns are comparatively distinctive, not because their expected returns are strong. Yet 3/1 suffers one of the sharpest modern declines and 24/6 is weakened by stale signals. A correlation-only objective cannot distinguish a genuinely diversifying return source from an unusual but weak one.

My interpretation. I would not adopt the correlation-aware weights in their current form. The optimizer solves the problem it was given, but that problem ignores deterioration in expected return and places precision on noisy distinctions among related sleeves. The negative result strengthens the case for equal weighting: simple horizon diversification captures the useful timing differences without over-allocating to weak rules merely because they look different.

10. Funded implementation after the exact-delivery liquidity screen

The preceding sections study excess returns from fixed-capital target portfolios. The final test asks what survives when account equity compounds, collateral earns interest, and thin contracts are not allowed to absorb the full risk target. The simulation begins with USD 100 million in January 2009, sizes each month from beginning equity, earns the lagged three-month Treasury constant-maturity yield on 90% of capital, and retains 10% as an operational and margin reserve. The yield is the Federal Reserve's DGS3MO series, lagged one calendar day to avoid look-ahead (Board of Governors of the Federal Reserve System 2026). Whole-contract rounding and the same execution model apply throughout.

A 5% volume-participation cap limits each market position. Away from rolls, the reference is lagged 63-session average volume. Near a roll, the constraint uses the smaller five-session volume of the outgoing delivery and the exact Norgate successor. Constrained risk is left unused rather than reallocated. The audit identifies 30-Day Federal Funds futures as an exception: even after capping, its volatility-scaled contract demand remains disproportionate to delivery liquidity. It is excluded identically from all three final portfolios before sizing. The ex-ZQ roll panel contains 13,420 transitions through July 2026, of which 7,270 occur during the January 2009 to July 2026 funded window. Calendar spreads remain carry and both execution legs are charged.

Table 13. Funded performance of the final 74-market portfolios

Portfolio	Implementation	Funded CAGR	Volatility	Sharpe above cash	Maximum drawdown	Annual cost drag	Average risk retained
Monthly Moskowitz 12/1 benchmark	Uncapped	5.77%	11.63%	0.42	-16.02%	0.79%	100.0%
Monthly Moskowitz 12/1 benchmark	5% participation cap	5.65%	9.76%	0.47	-14.02%	0.67%	83.3%
50/50 sign-slope ensemble	Uncapped	7.05%	11.19%	0.54	-17.77%	0.69%	100.0%
50/50 sign-slope ensemble	5% participation cap	6.74%	9.37%	0.59	-14.03%	0.58%	83.1%
Equal-weight multi-horizon book	Uncapped	6.39%	8.57%	0.60	-11.13%	0.48%	100.0%
Equal-weight multi-horizon book	5% participation cap	6.05%	7.37%	0.64	-9.12%	0.42%	85.2%

Note. Funded Sharpe above cash subtracts the full Treasury return even though the portfolio earns interest on only 90% of capital; this is conservative. Capped risk is not redistributed. The sample contains 211 months from January 2009 through July 2026.

The cap improves efficiency mainly by refusing contract-intensive risk. For the benchmark, annual cost drag falls from 0.79% to 0.67%, volatility from 11.63% to 9.76%, and maximum drawdown from 16.02% to 14.02%. For the sign-slope ensemble, cost drag falls from 0.69% to 0.58% and drawdown from 17.77% to 14.03%. The multi-horizon book is already more efficient because its targets are netted before execution; its cost drag falls from 0.48% to 0.42% and drawdown from 11.13% to 9.12% while retaining 85.2% of intended risk.

Among capped implementations, the sign-slope ensemble earns the highest funded CAGR at 6.74%. The multi-horizon book has the strongest risk-adjusted outcome: 0.64 Sharpe above cash, 7.37% volatility, and a 9.12% maximum drawdown. Its Sharpe advantage over the capped benchmark is +0.18 with a stationary-bootstrap 95% CI of [0.01, 0.36]. The capped ensemble's +0.13 advantage also has a CI above zero, [0.06, 0.21]. The +0.05 difference between the two candidate portfolios is unresolved because its CI [-0.11, 0.21] includes zero.

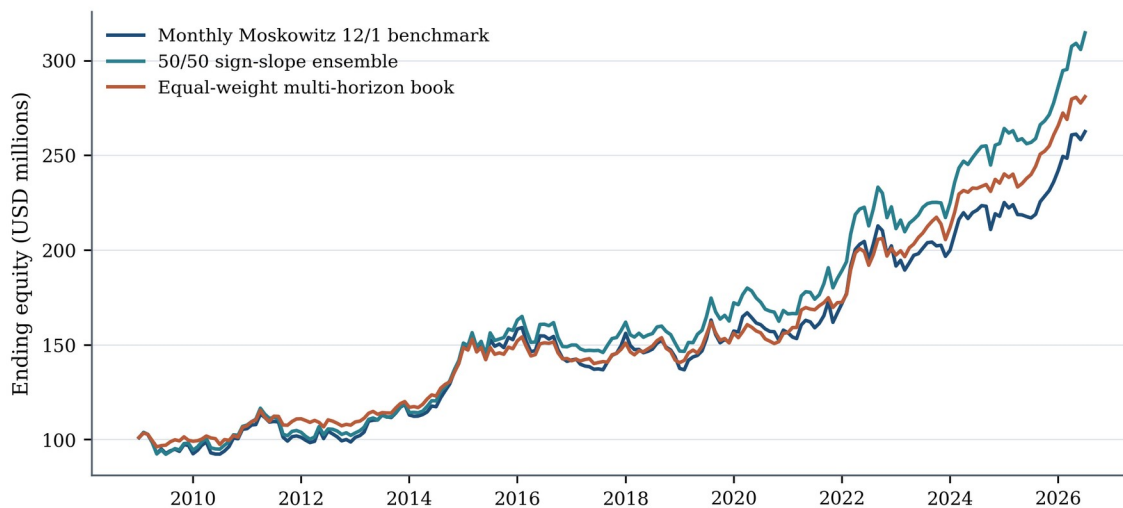


Figure 9. Funded wealth of the final 74-market portfolios after the 5% volume-participation cap

Note. All series begin with USD 100 million in January 2009. Returns include recursive sizing, 90% Treasury collateral, the common execution model, exact two-leg delivery transitions, and the same market-level liquidity rule.

My interpretation. The cap and common market exclusion do not add a forecast. They improve the conversion of a weak trend premium by refusing exposures whose contract demand is disproportionate to observed liquidity. I view the multi-horizon book as the leading risk-adjusted design and the sign-slope ensemble as the higher-CAGR comparison. Their Sharpe difference is not statistically resolved, so CAGR alone should not decide between them.

Removing the binding contract changes the implementation screen materially. In the final capped books, 95% of transactions use no more than 0.79% to 1.63% of the relevant reference volume, and only 0.72% to 1.33% exceed 5%. The worst market-specific 95th percentile is 14.66% for the benchmark, 13.54% for the ensemble, and 9.85% for the multi-horizon book. The portfolio ranking therefore does not depend on retaining the market that created the original bottleneck.

Table 14. Liquidity diagnostics for the final 74-market portfolios

Portfolio	Average risk retained	Mean constrained markets per month	95th percentile participation	Transactions above 5% of reference volume	Worst-market 95th percentile	Maximum participation
Monthly Moskowitz 12/1 benchmark	83.3%	16.4	1.63%	1.33%	Feeder Cattle (14.66%)	29.12%
50/50 sign-slope ensemble	83.1%	14.6	1.45%	1.29%	Feeder Cattle (13.54%)	30.98%
Equal-weight multi-horizon book	85.2%	12.9	0.79%	0.72%	Canola (9.85%)	37.74%

Note. Results refer to capped USD 100 million portfolios. Participation is contracts traded divided by lagged 63-session volume for routine trades or the relevant five-session delivery volume near rolls. Feeder Cattle and Canola are the largest remaining market-level 95th-percentile burdens.

The exclusion effect is economically meaningful but not performance-selected. Relative to the same capped books including the contract, annual cost drag falls by 0.09 to 0.12 percentage point, funded CAGR rises by 0.09 to 0.15 point, and Sharpe above cash rises by 0.01 to 0.02. Average intended risk retained also rises by roughly 0.7 to 0.8 percentage point because the cap no longer collides repeatedly with the same contract-intensive market.

10.1 What the liquidity screen can and cannot establish

The residual tail remains material. Maximum single-event participation is 29.12% for the benchmark, 30.98% for the ensemble, and 37.74% for the multi-horizon book. Those events cannot credibly be described as one-day executions at the modeled slippage. They would require staged orders, execution discretion, or a lower capital allocation.

The screen is therefore comparative, not a hard capacity estimate. Daily settlement data cannot observe intraday depth, bid-ask spread, order-book resilience, or nonlinear market impact; historical volume may also overstate liquidity during stress. Exchange and clearing fees beyond the stated commission are not modeled. The results support the relative implementability and ranking of the candidate books at USD 100 million, but not a guarantee that every order can be executed at the base-case cost.

My interpretation. Volatility scaling is necessary but not sufficient portfolio construction. A contract can appear small in risk terms while requiring an excessive number of contracts. I would treat liquidity limits as part of the portfolio definition, leave constrained risk unused, and escalate large residual roll events to a dedicated execution process rather than hide them inside an average cost assumption.

11. Conclusion

The evidence forms a coherent chain. The 12-month continuation opportunity weakened, and the monthly Moskowitz 12/1 benchmark remains profitable after exact delivery-level costs but produces only a 0.42 post-2008 net Sharpe. A 252-day slope offers a complementary trend definition, although its stand-alone advantage is not decisive. The horizon experiment then shows that decay is concentrated in rules that replace the complete signal vintage every month. Combining all 12 lookback and holding-period rules in one equal-weight, internally netted book raises net Sharpe to 0.62 and reduces maximum drawdown to 12.8% without choosing the best cell after observing the data. Correlation-aware weighting does not improve that simple portfolio.

The funded analysis changes the scale of the question but not its economic meaning. After the common liquidity screen, the multi-horizon book produces a 0.64 Sharpe above cash and a 9.1% maximum drawdown; the sign-slope ensemble produces the highest funded CAGR at 6.74%. Both have statistically positive Sharpe advantages over the capped benchmark. Most transaction participation is low, but isolated events between 29% and 38% remain outside what daily data can validate as one-session execution. The recommendation is therefore a portfolio design plus an execution constraint, not an unrestricted capacity claim.

My interpretation. I would retain the monthly 12/1 rule as the auditable benchmark, use the equal-weight multi-horizon book as the leading risk-adjusted design, and keep the sign-slope ensemble as the higher-CAGR comparison. The evidence does not support further covariance optimization. The useful gains come from diversifying signal timing, netting offsetting targets, and refusing exposures whose contract demand is inconsistent with observable liquidity. That conclusion is simpler and more defensible than searching for one superior trend rule.

References

- Baltas, N., and R. Kosowski. 2020. "Demystifying Time-Series Momentum Strategies: Volatility Estimators, Trading Rules and Pairwise Correlations." In *Market Momentum: Theory and Practice*, edited by S. Satchell and A. Grant, 30-67. Wiley.
<https://doi.org/10.1002/9781119599364.ch3>.
- Board of Governors of the Federal Reserve System (US). 2026. Market Yield on U.S. Treasury Securities at 3-Month Constant Maturity, Quoted on an Investment Basis [DGS3MO]. FRED, Federal Reserve Bank of St. Louis. Accessed 2 September 2026.
<https://fred.stlouisfed.org/series/DGS3MO>.
- Cameron, A. C., J. B. Gelbach, and D. L. Miller. 2011. "Robust Inference with Multiway Clustering." *Journal of Business & Economic Statistics* 29 (2): 238-249. <https://doi.org/10.1198/jbes.2010.07136>.
- Holm, S. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65-70.
<https://doi.org/10.2307/4615733>.
- Hurst, B., Y. H. Ooi, and L. H. Pedersen. 2017. "A Century of Evidence on Trend-Following Investing." *Journal of Portfolio Management* 44 (1): 15-29. <https://doi.org/10.3905/jpm.2017.44.1.015>.
- Kaufman, P. J. 1995. *Smarter Trading: Improving Performance in Changing Markets*. New York: McGraw-Hill.
- Moskowitz, T. J., Y. H. Ooi, and L. H. Pedersen. 2012. "Time Series Momentum." *Journal of Financial Economics* 104 (2): 228-250.
<https://doi.org/10.1016/j.jfineco.2011.11.003>.
- Newey, W. K., and K. D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (3): 703-708. <https://doi.org/10.2307/1913610>.
- Norgate Data. 2026. Futures Data package: individual delivery histories and continuous contracts. Accessed 2 September 2026.
<https://norgatedata.com/futurespackage.php>.
- Politis, D. N., and J. P. Romano. 1994. "The Stationary Bootstrap." *Journal of the American Statistical Association* 89 (428): 1303-1313.
<https://doi.org/10.1080/01621459.1994.10476870>.

Appendix A. Reproducibility record

Table A1. Validated research packages

Package	Purpose	Primary output
Research2_24	Monthly Moskowitz 12/1 benchmark	Gross returns, targets, and inference
Research2_25	Commission, slippage, and actual delivery transitions	Net returns and transaction audit
Research2_26	252-day linear-slope comparison	Paired performance and inference
Research2_27	50/50 sign-slope ensemble	Netted targets, returns, and inference
Research2_28	Lookback and holding-period grid	Horizon deterioration and paired inference
Research2_29	Combined 12-rule target book	Equal weighting versus correlation-aware allocation
Research2_31	Exact delivery-liquidity audit	Outgoing and successor delivery volumes
Research2_32	Funded liquidity-cap analysis	Participation and capacity diagnostics
Research2_33	Common exclusion of the failed-liquidity contract	Final 74-market funded comparison

Note. Research2_24 establishes the monthly gross benchmark; Research2_25 validates its net implementation with zero roll-leg volume fallbacks. Research2_26 through Research2_29 test signal and horizon diversification. Research2_31 through Research2_33 add exact delivery liquidity, funded recursive sizing, the 5% volume-participation cap, and the common 74-market implementation screen. Versioned code, inputs, output tables, and run-audit hashes are archived and available with the paper.