Sequence-of-Returns Risk Revisited: Can Time-Series Foundation Models Detect Danger Zones?

Harry Vadalkar[1]

Abstract: Sequence-of-returns risk (SoRR) is the dominant driver of retirement portfolio failure, yet planning practice relies on static probability-of-ruin estimates rather than prospective early-warning signals. This study tests whether a time-series foundation model (TSFM)—Amazon's Chronos—can detect SoRR danger zones, benchmarked against five traditional volatility indicators. Using 395 months of U.S. data (1993–2025) with actual inflation and a real-bond series, we construct danger-zone labels from 15-year, 4% real-withdrawal retirement simulations and evaluate each indicator's detection accuracy and retirement-outcome impact under a defensive-response strategy. As a danger-zone detector, Chronos forecasts are almost perfectly inverted: a naive reading (low forecast as danger) yields an ROC-AUC of 0.022, while reading it contrarily (high forecast as danger) yields 0.978, exceeding every traditional indicator (VIX = 0.654; composite = 0.678). The model's optimism is itself the warning. However, this near-perfect ranking power does not translate into an actionable real-time trigger: an expanding-window contrarian rule fires too rarely to protect portfolios, and both Chronos-based defensive strategies reduce worst-case wealth. Traditional indicators improve point-estimate worst-case (5th-percentile) terminal wealth by 13–51%, but moving-block-bootstrap confidence intervals include zero in every case, so the benefit is not statistically distinguishable from chance. The contribution is a corrected interpretation of TSFM output in financial planning, a rigorously benchmarked comparison, and candid guidance for practitioners on both the promise and the limits of foundation models for retirement risk.

Recommended Citation

Vadalkar, H. (2026). Sequence-of-returns risk revisited: Can time-series foundation models detect danger zones?. *Financial Services Review,* Article e009. https://doi.org/10.61190/et356m79

I. INTRODUCTION

Retirement planning rests on a deceptively simple question: will the money last? The dominant analytical tool for answering it—Monte Carlo simulation—has served wealth managers for three decades, translating uncertain returns into survival probabilities and terminal wealth distributions (Bengen, 1994; Cooley et al., 1998). Yet the greatest threat to a decumulating portfolio is not average return but the order in which returns are realized. Sequence-of-returns risk (SoRR), first quantified by Bengen (1994) and developed by Pfau (2015, 2019), Kitces and Pfau (2015), and Milevsky (2006), shows that early-retirement cohorts who encounter poor markets in their first decade fare dramatically worse than those who experience the same average return in a more favorable order. Pfau (2015) estimated that roughly 77% of a portfolio's final outcome is explained by the first ten years of returns alone.

Despite this vulnerability, the standard planning response remains retrospective: stress-test against historical worst cases, or report a static probability of ruin. Neither provides a forward-looking signal a planner could act on before damage materializes. We ask whether a modern forecasting tool—a time-series foundation model (TSFM)—can serve as a prospective early-warning detector for SoRR danger zones.

[1] Corresponding author (hvadalka@ttu.edu). Professor, Texas Tech University.

TSFMs are pretrained transformer models that learn temporal patterns from massive multi-domain corpora and forecast unseen series without task-specific training (Das et al., 2024; Ansari et al., 2024; Rasul et al., 2023). Goel et al. (2024) found that a fine-tuned TSFM outperforms GARCH for Value-at-Risk, raising the possibility that such models capture nonlinear dependencies traditional indicators miss. We test this in the SoRR context using Amazon's Chronos, benchmarked against rolling realized volatility, the VIX, GARCH(1,1) conditional volatility, the rolling Sharpe ratio, and a composite. We construct danger-zone labels from 156 retirement start dates, evaluate detection with precision, recall, the false-positive rate, and ROC-AUC, and measure the retirement-outcome impact of acting on each signal.

Our central finding reframes an apparent failure into a more interesting result. Read naively, the Chronos forecast is a near-useless danger-zone signal; read contrarily, it is a near-perfect one. The model systematically forecasts strong returns precisely ahead of the cohorts that go on to suffer the worst sequence risk—so its optimism is the warning. We show, however, that this striking ranking power does not survive translation into an actionable real-time rule, and that even the best traditional indicators deliver tail protection whose statistical precision is limited by the small number of independent danger-zone episodes in the historical record.

II. LITERATURE REVIEW

A. Sequence-of-Returns Risk and the Early-Warning Problem

Bengen (1994) demonstrated that portfolio survival in retirement depends not on average return but on the order in which returns occur, establishing the 4% rule from U.S. historical data. The mechanism is arithmetic but consequential: withdrawals during an early drawdown crystallize losses by forcing sales when prices are depressed, permanently shrinking the capital base that must later compound. Pfau (2015) quantified the magnitude, attributing roughly 77% of the final outcome to the first decade of returns, and characterized a retirement-date risk zone spanning the years immediately before and after retirement (Pfau, 2019). Finke et al. (2013) sharpened the practical stakes, showing the 4% rule's failure rate rising from 6% under historical averages to 57% under low-yield conditions—evidence that the danger is not sequence risk in the abstract but the specific case of beginning decumulation in an unfavorable regime that the static plan does not anticipate.

The literature's responses to SoRR all presume, implicitly, that the danger can be anticipated or cushioned. Pfau and Kitces (2014) showed that a rising equity glide path—conservative at the outset, rising thereafter—reduces both the probability and the magnitude of failure, and Kitces and Pfau (2015) linked the adjustment to market valuations via the cyclically adjusted price-earnings ratio. Guyton and Klinger (2006) developed guardrail rules in which spending responds to portfolio performance. Blanchett (2015) established that optimal glide paths are conditional on initial market conditions. Milevsky and Robinson (2005) provided an analytical ruin-probability framework, and Milevsky (2016) argued that binary ruin measures are too coarse, favoring richer sustainability metrics. What this body of work has not done is define and test a prospective detector: a real-time signal that flags, before the fact, that a given retirement start date is entering a danger zone. The present study supplies and evaluates exactly such a signal, adopting the danger-zone construct as its detection target and using the signal-detection metrics standard in the early-warning literature.

B. Regime Detection in Financial Markets

If markets alternate between distinct return-and-volatility regimes, then detecting a shift to an adverse regime early is the formal analogue of the SoRR early-warning problem. Hamilton (1989) introduced the Markov regime-switching model, the standard framework for capturing such shifts. Ang and Bekaert (2002) applied it to international portfolio choice and showed that ignoring regimes costs investors the equivalent of 2–3% of initial wealth, in part because cross-asset correlations rise in high-volatility regimes—negating diversification precisely when retirees most need it. Ang and Timmermann (2012) surveyed the broad evidence that regime-switching models capture fat tails, volatility clustering, and asymmetric dependence, and Guidolin and Timmermann (2007) demonstrated materially different optimal allocations across a four-state stock-bond model. This literature motivates danger-zone detection but has not been connected to retirement decumulation, leaving open whether regime signals translate into better retirement outcomes.

C. *Traditional Volatility Indicators: GARCH and the VIX*

The benchmark indicators in this study rest on a mature volatility literature. Engle (1982) introduced the ARCH model and Bollerslev (1986) generalized it to GARCH(1,1), which became the standard for conditional volatility estimation (Bollerslev, Chou, & Kroner, 1992). GARCH is a backward-looking econometric estimate: it infers current conditional variance from the recent history of squared returns. The VIX, by contrast, is forward-looking, derived from S&P 500 option prices and characterized by Whaley (2000) as the investor fear gauge. The two are conceptually distinct—an econometric estimate of realized variance versus a risk-neutral market expectation of future variance—and Bekaert and Hoerova (2014) showed that the squared VIX decomposes into a conditional-variance component and a variance premium, with the components carrying different predictive content for stock returns and financial instability. This distinction matters here: GARCH conditional volatility is, by construction, highly persistent and tracks recent realized volatility closely, so it adds little information beyond a rolling realized-volatility measure, whereas the VIX embeds anticipatory information from option markets. We document this empirically through the indicator correlation structure.

D. *Time-Series Foundation Models*

TSFMs extend the pretraining paradigm of large language models to numerical sequences. TimesFM (Das et al., 2024) is a decoder-only transformer pretrained on a large time-series corpus, achieving competitive zero-shot accuracy across domains. Chronos (Ansari et al., 2024), the model used here, tokenizes scaled-and-quantized series and trains language-model architectures on the resulting tokens; critically for this application, it produces probabilistic forecasts by sampling multiple trajectories, yielding both a forecast mean and a forecast-dispersion estimate. Lag-Llama (Rasul et al., 2023) and MOMENT (Goswami et al., 2024) demonstrate related zero-shot and multi-task capabilities, the latter including anomaly detection. The first financial application of a TSFM to risk, Goel et al. (2024), found that a fine-tuned foundation model (Google's TimesFM) beat GARCH for Value-at-Risk on S&P 100 data—but emphasized that the gains required fine-tuning, with zero-shot performance more modest. A recognized limitation across this literature is that pretraining corpora are dominated by energy, weather, retail, and transportation data whose statistical properties—trend persistence, seasonality, comparatively benign tails—differ markedly from financial returns, which approximate a near-random walk with volatility clustering, fat tails, and regime shifts. Whether zero-shot transfer succeeds for a financial early-warning task is therefore an open empirical question, and one this study answers.

E. *Machine-Learning Early-Warning Systems and the TSFM Distinction*

A parallel literature applies supervised machine learning to financial-crisis prediction. Samitas, Kampouris, and Kenourgios (2020) combined network analysis with machine learning to predict contagion; Bluwstein et al. (2021) showed that tree-based models outperform logistic regression for crisis prediction using credit and yield-curve inputs; Reimann (2024) found Random Forest and Extremely Randomised Trees superior by AUROC across 13 countries and 150 years; and Fouliard, Howell, and Rey (2021) argued that nonlinear methods detect subtle early-warning features that linear models miss. In asset pricing more broadly, Gu, Kelly, and Xiu (2020) documented large economic gains from machine-learning return forecasts, tracing the improvement to nonlinear predictor interactions. These approaches differ fundamentally from a TSFM applied zero-shot: supervised models are trained on labeled financial outcomes and learn task-specific, in-domain relationships, whereas a zero-shot TSFM brings only patterns learned from non-financial pretraining and is never shown a single labeled danger zone. This distinction frames our central question—whether out-of-domain pretrained knowledge transfers to a financial early-warning task—and supplies the evaluation framework (precision, recall, false-positive rate, AUROC) we adopt.

F. *Research Gap*

No prior study applies a time-series foundation model as a prospective early-warning detector for sequence-of-returns risk, benchmarks it against traditional volatility indicators using signal-detection metrics, and evaluates the retirement-outcome impact of acting on its signals. This study addresses all three gaps and, in doing so, surfaces a methodological lesson about how foundation-model output must be interpreted before it can be used in financial planning.

III. METHODOLOGY

A. Data and Sample

The sample comprises 395 monthly observations from February 1993 through December 2025. U.S. equity returns are proxied by the SPDR S&P 500 ETF (SPY). Bond returns combine the iShares Core U.S. Aggregate Bond ETF (AGG) from 2003 onward with the Vanguard Total Bond Market Index Fund (VBMFX) before 2003; both track the U.S. Aggregate Bond universe, so the spliced series preserves genuine bond-return volatility throughout—including the 1998–2002 span where danger zones cluster. Inflation is measured with actual monthly U.S. CPI (CPIAUCSL, Federal Reserve Bank of St. Louis). The portfolio is 60% equity and 40% bonds; real returns are computed as (1 + nominal) / (1 + inflation) − 1. Over the sample, mean monthly real return is 0.50% (SD = 2.73%) and mean VIX is 19.6 (SD = 7.6).

TABLE I
DESCRIPTIVE STATISTICS

Statistic	60/40 monthly real return	VIX (monthly level)
Mean	0.50%	19.6
Standard deviation	2.73%	7.6
Observations	395	395
Sample period	Feb 1993–Dec 2025	Feb 1993–Dec 2025

Note: Monthly observations over February 1993–December 2025. The real return is that of the 60% equity / 40% bond portfolio, deflated by actual monthly CPI; the VIX is the monthly index level. Indicator-level correlations appear in Table 2 and detection and retirement outcomes in Table 3.

B. Danger-Zone Construction

The simulation is conducted entirely in real terms: a constant real withdrawal of $3,333 per month (4% of an initial $1,000,000, divided by 12) is applied against real returns, so purchasing power is held constant across the horizon. For each start month t having at least 180 subsequent months, we simulate a 15-year decumulation path and record terminal wealth. A start month is labeled a danger zone if its terminal wealth falls at or below the 25th percentile of all evaluable start dates. Applying a single evaluation set—start dates with both a full 180-month forward window and at least 60 months of prior history—yields 156 evaluable start dates and 39 danger-zone months (25.0%), a single sample size used consistently throughout. Danger zones span February 1998 to June 2001, corresponding to cohorts entering decumulation just before the dot-com collapse.

C. Indicators and One-Month Lagging

Five traditional indicators are constructed: rolling 12-month realized volatility, the monthly VIX level, GARCH(1,1) conditional volatility (estimated on an expanding window with a one-step-ahead forecast), the rolling 12-month Sharpe ratio, and an equal-weight composite z-score of the four. To eliminate any look-ahead advantage, every backward-looking indicator (realized volatility, VIX, Sharpe) is lagged one month, so the signal for month t uses only information through month t − 1; the GARCH and Chronos forecasts are one-step-ahead by construction and therefore already use only past data. Chronos-T5-Small receives the trailing 60 months of real returns as context and generates 20 sampled 12-month-ahead trajectories, from which we extract the forecast mean. A 60-month context is the longest the data permit: because all danger zones fall in 1998–2001, a longer context (e.g., 120 months) pushes the model's first forecast past the danger-zone window entirely, leaving no positive labels to evaluate. We report this as a data constraint rather than a free design choice.

D. Signal Rules and the Contrarian Specification

Each indicator is converted to a binary signal using expanding-window percentile thresholds, so no signal uses future information. Volatility indicators fire above their expanding 80th percentile; the Sharpe ratio fires below its 20th percentile. Because an ROC-AUC near zero denotes a near-perfectly inverted classifier rather than a failed one, we evaluate the Chronos forecast under two single-condition specifications: Chronos-Low fires when the mean forecast is below its expanding 20th percentile (the original, naive hypothesis), and Chronos-Contrarian fires when the mean forecast is above its expanding 80th percentile (the inverted reading). A single-condition rule is applied to every indicator, so no indicator mechanically fires more often than another by construction.

E. Evaluation Metrics

Detection is evaluated with four metrics, each defined here so the tables stand alone. Precision is the share of fired signals that coincide with true danger zones; recall is the share of true danger zones the signal captures; the false positive rate is the share of non-danger months that nonetheless fire. The area under the receiver operating characteristic curve (ROC-AUC) measures ranking quality using each indicator's continuous score (with the sign convention noted above), where 0.5 is uninformative, 1.0 is perfect, and values near 0 indicate a near-perfect inversion. Binary metrics use the 0/1 signal; ROC-AUC uses the continuous score.

F. Retirement Simulation and Statistical Precision

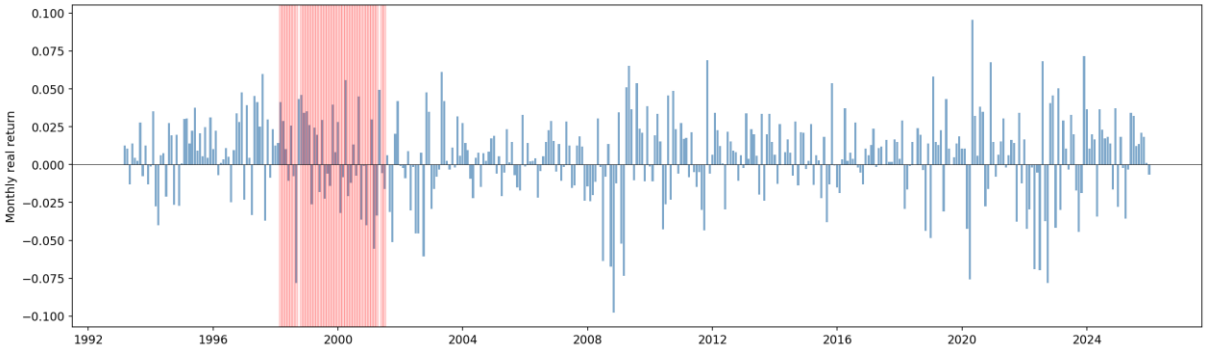
To measure the retirement-outcome impact of acting on each signal, a defensive-response strategy shifts the portfolio from 60/40 to 30/70 for 12 months whenever the signal fires, then reverts. For each of the 156 start dates we record terminal wealth under each strategy and report the ruin probability, median, and 5th-percentile (worst-case) terminal wealth. Because the start dates use heavily overlapping forward windows, naive comparison overstates precision; we therefore compute 95% confidence intervals for the difference in 5th-percentile wealth versus the no-signal baseline using a moving block bootstrap (block length 24 months, 1,000 resamples) that preserves the serial dependence created by the overlap. Robustness is assessed across danger-zone thresholds (20th/25th/30th percentile), horizons (15 and 20 years), withdrawal rates (3.5%/4.0%/4.5%), signal percentiles (70th–85th), and defensive designs (equity weight 20–40%, duration 6–18 months).

IV. RESULTS

A. Danger-Zone Geography

Figure 1 plots monthly real returns with danger-zone months shaded. All 39 danger-zone start dates fall between February 1998 and June 2001: these are the cohorts whose 15-year decumulation begins just before the dot-com collapse, locking in early losses. No post-2001 start date qualifies as a danger zone within the 15-year horizon, reflecting the strong subsequent U.S. market. This concentration is central to interpreting the results: the historical record contains essentially one extended danger-zone episode, which limits the statistical precision attainable for any detector.

Figure 1
Monthly Real Portfolio Returns and Sequence-Risk Danger Zones, 1993–2025



Note: Bars show monthly real returns of the 60/40 portfolio. Red shading marks danger-zone start months—those whose 15-year, 4% real-withdrawal terminal wealth falls in the bottom 25th percentile of evaluable start dates. Danger zones span February 1998 to June 2001.

B. Indicator Redundancy

Table 2 reports the correlation structure of the (lagged) indicators. GARCH conditional volatility correlates 0.81 with rolling realized volatility, consistent with GARCH's high persistence, which makes its forecast largely a restatement of recent realized volatility; it adds little independent information. The VIX is more weakly correlated with realized volatility (0.53), consistent with its forward-looking, option-implied content.

TABLE II
INDICATOR CORRELATION MATRIX (ONE-MONTH-LAGGED LEVELS)

	Realized Vol	**VIX**	**GARCH**	**Sharpe**	**Composite**
Realized Vol	1.00	0.53	0.81	−0.53	0.80

	Realized Vol	VIX	GARCH	Sharpe	Composite
VIX	0.53	1.00	0.68	−0.50	0.79
GARCH(1,1)	0.81	0.68	1.00	−0.65	0.88
Sharpe 12m	−0.53	−0.50	−0.65	1.00	−0.65
Composite	0.80	0.79	0.88	−0.65	1.00

Note: Pearson correlations of one-month-lagged indicator levels over the 156-date evaluation sample. The 0.81 correlation between GARCH(1,1) conditional volatility and 12-month realized volatility indicates substantial redundancy between the two.

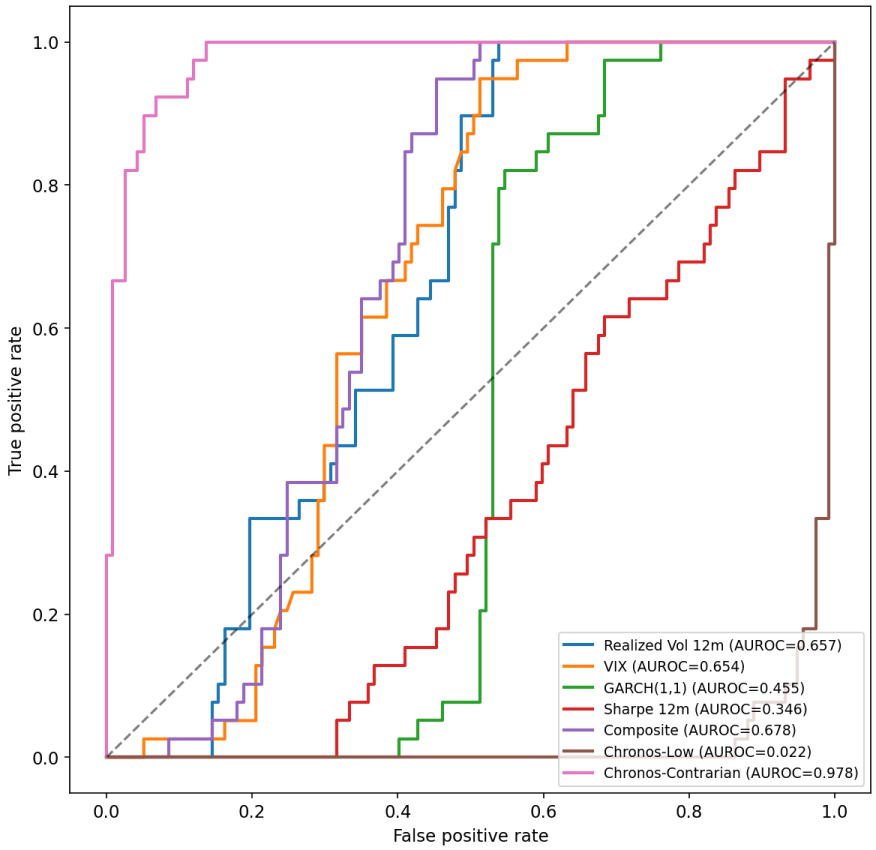
C. Signal Detection and the Chronos Inversion

Table 3 reports detection metrics. Among traditional indicators, the VIX has the strongest recall (0.62) and a representative ROC-AUC of 0.65; realized volatility and the composite achieve ROC-AUCs of 0.66 and 0.68. GARCH and the lagged Sharpe ratio underperform once properly lagged (ROC-AUCs of 0.46 and 0.35), the Sharpe ratio's sub-0.5 value indicating that, in this single episode, a low trailing Sharpe was a weak and sometimes contrary signal.

The Chronos result is the study's centerpiece. Read naively, Chronos-Low achieves an ROC-AUC of 0.022—far below the 0.5 of an uninformative classifier. A score this close to zero does not denote noise; it denotes a near-perfect inversion. Reading the same forecast contrarily, Chronos-Contrarian achieves an ROC-AUC of 0.978, exceeding every traditional indicator. The interpretation is economic, not mechanical: pretrained on non-financial series in which recent strength tends to persist, Chronos extrapolates the late-1990s bull run and forecasts continued high returns—precisely for the cohorts that go on to suffer the worst sequence risk. The model's optimism is the danger signal.

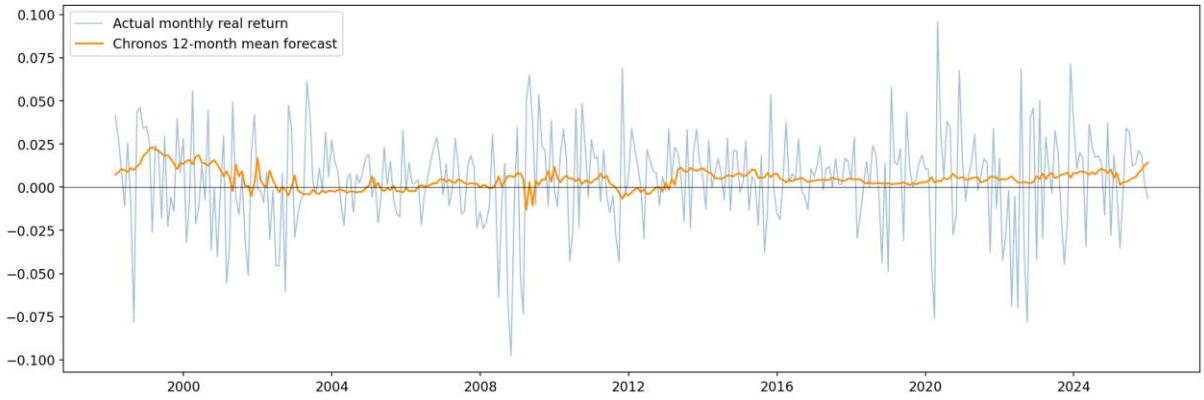
Figure 2 shows the ROC curves: the Chronos-Contrarian curve hugs the upper-left corner while Chronos-Low traces its mirror image along the lower-right. Figure 3 makes the mechanism visible—the Chronos forecast is a smooth, persistently positive line that rises with recent returns rather than anticipating drawdowns. Figure 4 plots the timing of every indicator's signals against the realized danger zones, showing that the traditional indicators concentrate their warnings within the 1998–2001 window but also fire repeatedly outside it, whereas the Chronos-Contrarian signal fires only rarely.

Figure 2
ROC Curves for Danger-Zone Detection by Indicator



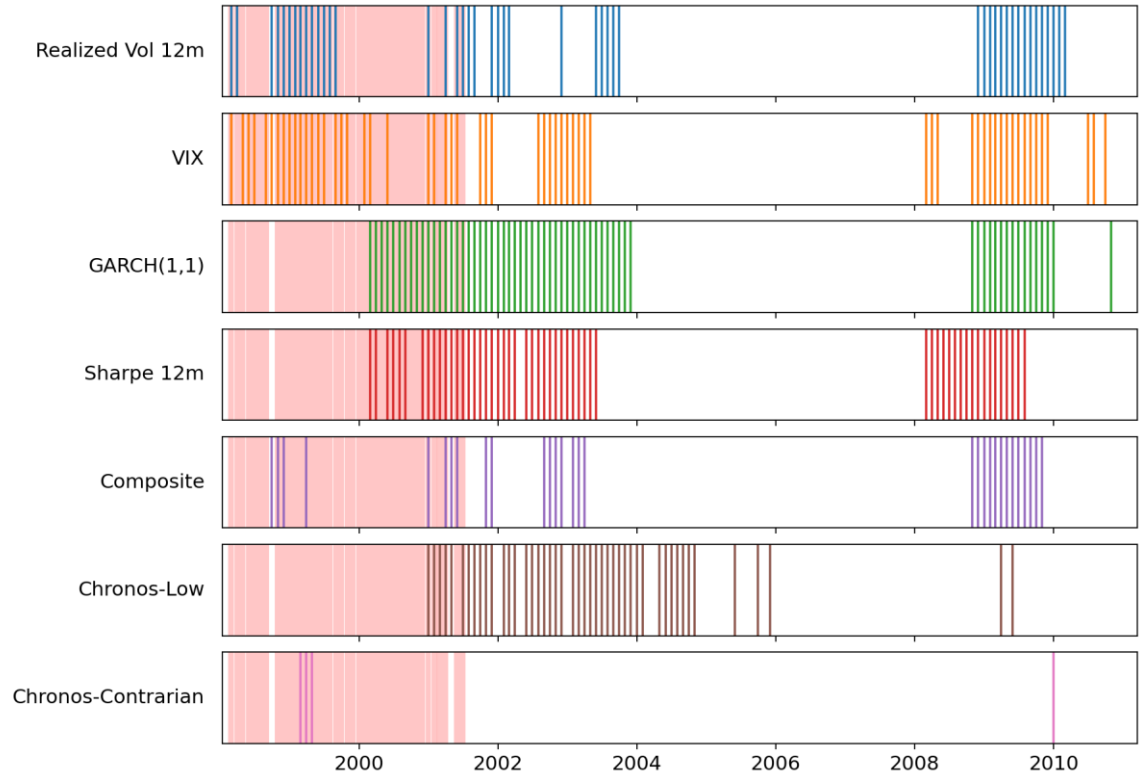
Note: Curves use each indicator's continuous score on the 156-date evaluation sample. Chronos-Contrarian (ROC-AUC = 0.978) dominates all traditional indicators; Chronos-Low (0.022) is its near-perfect inversion. The dashed diagonal denotes an uninformative classifier (ROC-AUC = 0.50).

Figure 3
Chronos Mean Forecast Versus Realized Returns (60-Month Context)



Note: The Chronos 12-month mean forecast (solid line) is smooth and persistently positive, tracking recent strength rather than anticipating drawdowns in realized monthly returns (light line). This pattern produces the inverted detection result in Table 3.

Figure 4
Signal Timing Versus Realized Danger Zones



Note: The figure covers the evaluation sample only (February 1998–January 2011, 156 start dates), which begins after the 60-month minimum-history requirement and ends at the last start date with a full 180-month forward window; signal counts therefore match those in Table 3. Vertical marks indicate months in which each indicator fires; red shading marks realized danger zones (February 1998–June 2001). Traditional indicators cluster signals within the danger zone but also fire repeatedly outside it.

Figure 4 also shows the traditional indicators firing heavily around the 2007–2008 financial crisis, a period that carries no danger-zone label. This is a consequence of how the labels are constructed, and it exposes the evaluation's central constraint. A start month can be labeled only if a full 180-month forward window exists, so no start after December 2010 is evaluable; and the label marks the worst quartile of terminal outcomes, not every crisis. Cohorts entering decumulation around 2007–2008 absorbed the crisis early but were then carried by the 2009–2020 bull market to terminal wealth above the worst-quartile cutoff, so they escape the label even though the alarms were economically reasonable when they fired. Because all 39 danger-zone months fall within a single episode (February 1998–June 2001), the detection metrics effectively measure each indicator's ability to flag that one episode: 2008-era signals are scored as false positives despite being defensible warnings. The precision and false-positive rates in Table 3 should therefore be read as episode-specific, and the indicator rankings as conditional on a historical record containing one realized danger episode; Section 5.4 returns to this limitation.

D. From Detection to Action: The Contrarian Signal Does Not Translate

A near-perfect detector is not automatically a useful planning tool. Table 3 reports retirement outcomes. No strategy produces ruin in this favorable sample, so the comparison rests on terminal-wealth distributions, particularly the 5th-percentile worst case. The traditional indicators improve point-estimate worst-case wealth: the Sharpe-triggered strategy raises 5th-percentile wealth from $609,594 to $920,013 (+50.9%), with VIX (+33.7%) and GARCH (+36.5%) also positive. Both Chronos strategies, by contrast, reduce worst-case wealth (Chronos-Low −15.9%, Chronos-Contrarian −12.1%).

The pairing of the Sharpe signal's weak detection score (ROC-AUC = 0.346) with the largest point-estimate outcome gain (+50.9%) deserves explanation, because the two metrics reward different things. The ROC-AUC asks how well an indicator's continuous score ranks the 156 start-date cohorts by danger-zone membership—a labeling exercise anchored entirely to the single 1998–2001 episode. The simulated outcome instead depends on where the defensive windows fall relative to realized

drawdowns anywhere in the sample. The lagged Sharpe signal ranks the labeled cohorts poorly, but a number of its 54 firings coincide with momentum breaks that precede large realized drawdowns—most visibly 2000–2002 and 2008—so its defensive windows overlap the worst return sequences even though many of those months carry no danger-zone label. Because the 5th-percentile statistic is determined precisely by those worst sequences, a signal can improve the simulated tail while ranking the labels poorly. Section 4.5 supplies the rest of the explanation: the +50.9% point estimate carries a bootstrap interval that includes zero, so part of the apparent paradox is sampling noise from heavily overlapping windows rather than genuine signal.

Two forces explain the contrarian signal's failure to help despite its detection power. First, the expanding-window contrarian rule fires only four times over the evaluation sample, because a forecast must exceed the 80th percentile of all forecasts seen to date—a high bar reached too rarely and often too late to mount a defense. Second, when it does fire, the 12-month defensive shift sacrifices growth without reliably avoiding the relevant drawdowns. The detection power is real but is expressed in the continuous ranking, not in a tradable threshold rule under realistic, causal calibration.

TABLE III
DANGER-ZONE DETECTION AND RETIREMENT OUTCOMES BY STRATEGY

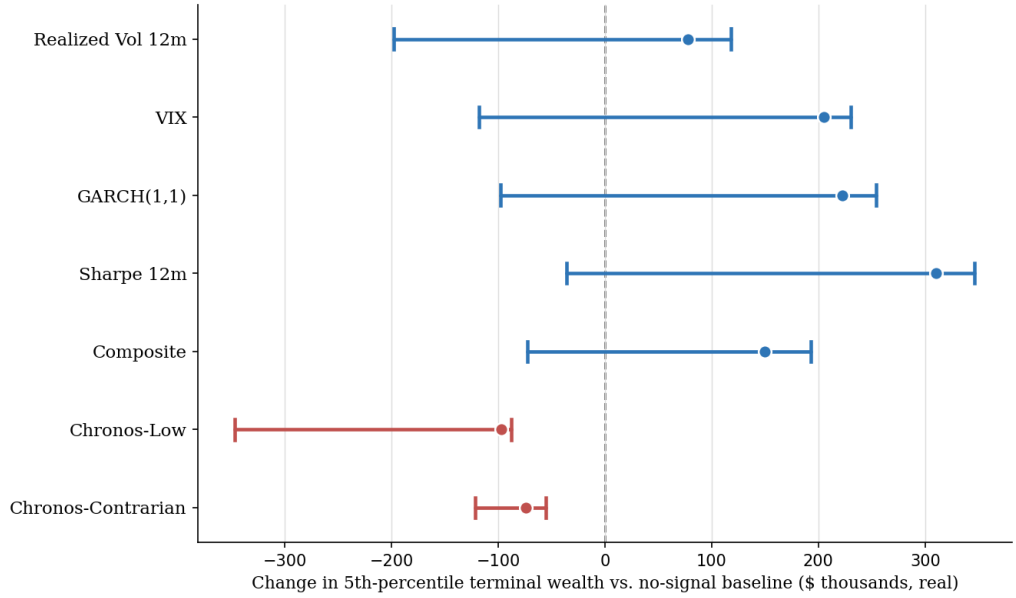
Strategy	ROC-AUC	Precision	Recall	Signals	FPR	Median TW	P5 TW	ΔP5 vs. Base
No Signal (60/40)	—	—	—	—	—	$1,111,859	$609,594	—
Realized Vol 12m	0.657	0.37	0.44	46	0.25	$975,915	$687,197	+12.7%
VIX	0.654	0.41	0.62	59	0.30	$1,039,234	$814,980	+33.7%
GARCH(1,1)	0.455	0.26	0.41	62	0.39	$1,045,111	$831,877	+36.5%
Sharpe 12m	0.346	0.24	0.33	54	0.35	$1,119,527	$920,013	+50.9%
Composite	0.678	0.20	0.15	30	0.21	$1,106,381	$759,665	+24.6%
Chronos-Low	0.022	0.11	0.13	46	0.35	$699,854	$512,689	−15.9%
Chronos-Contrarian	0.978	0.75	0.08	4	0.01	$962,673	$535,643	−12.1%

Note: Detection metrics (ROC-AUC, precision, recall) are computed on the 156 evaluation start dates (39 danger zones); the No-Signal baseline is not a detector, so its detection cells are blank. Terminal wealth (TW) is in real dollars; P5 is the 5th-percentile (worst-case) outcome. No strategy produces ruin (0% in every case). Signals is the number of months each indicator fires over the evaluation sample, and FPR is its false-positive rate. 95% moving-block-bootstrap confidence intervals for ΔP5 appear in Figure 5.

E. Statistical Precision of the Tail-Protection Estimates

Because the 156 start dates use heavily overlapping windows—and because the historical record contains essentially one danger-zone episode—the point-estimate improvements must be read with caution. Figure 5 plots the change in 5th-percentile wealth versus baseline (ΔP5) with 95% moving-block-bootstrap confidence intervals. Every traditional indicator's interval includes zero (e.g., Sharpe ΔP5 = +$310,419, 95% CI [−$36,119, +$346,481]), so although the point estimates are uniformly positive, the tail-protection benefit is not statistically distinguishable from chance at conventional levels. The two Chronos strategies are the only ones whose intervals exclude zero, and they do so on the wrong side—confirming a reliable reduction in worst-case wealth. We therefore present the traditional-indicator improvements as economically suggestive but statistically imprecise, not as performance guarantees.

Figure 5
Change in Worst-Case (5th-Percentile) Terminal Wealth Versus Baseline, with 95% Confidence Intervals

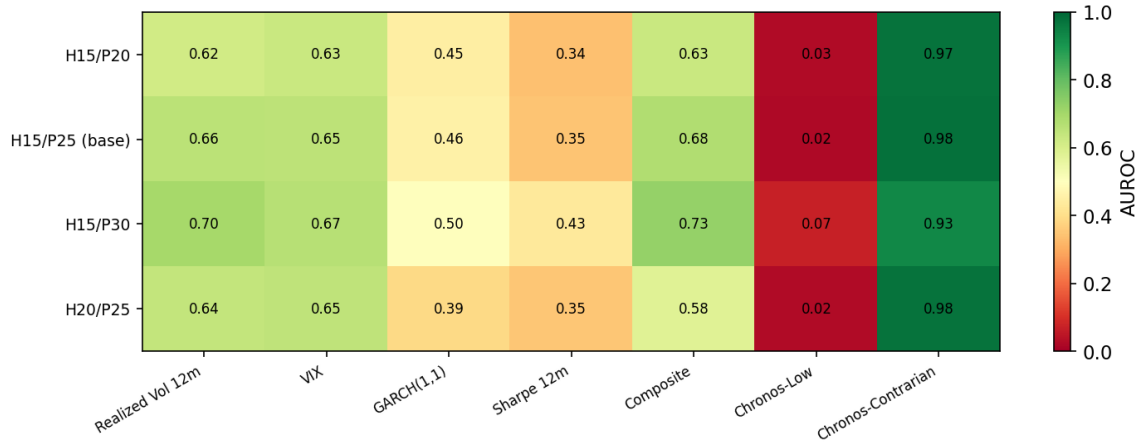


Note: Points show the difference in 5th-percentile terminal wealth (ΔP5) between each strategy and the no-signal 60/40 baseline, in thousands of real dollars; horizontal whiskers are 95% moving-block-bootstrap confidence intervals (block length = 24 months, 1,000 resamples). Intervals crossing the dashed zero line indicate tail protection not statistically distinguishable from chance; both Chronos intervals lie entirely below zero.

F. Robustness

The qualitative findings are stable across specifications. Figure 6 shows that the indicator ranking by ROC-AUC is preserved across danger-zone thresholds (20th/25th/30th percentile) and horizons (15 and 20 years): Chronos-Contrarian remains near 0.93–0.98, Chronos-Low near 0.02–0.07, and the traditional indicators in the 0.34–0.73 band throughout. Varying the signal percentile (70th–85th) traces the expected precision-recall trade-off—e.g., VIX recall falls from 0.82 at the 70th percentile to 0.49 at the 85th—without overturning the comparison; the Chronos-Contrarian rule continues to fire too rarely (2–10 times) to be actionable at any threshold tested. Across defensive designs, a shorter, more conservative shift (20% equity for 6 months) gives the best worst-case Sharpe-triggered outcome ($997,490), and the baseline's vulnerability rises with the withdrawal rate (5th-percentile wealth of $716,217, $609,594, and $498,130 at 3.5%, 4.0%, and 4.5%), with the Sharpe-triggered strategy improving the point estimate at every rate.

Figure 6
ROC-AUC Stability Across Danger-Zone Definitions



Note: Cell values are ROC-AUCs for each indicator under alternative horizon and percentile danger-zone definitions (H = horizon in years; P = percentile threshold). The contrarian Chronos reading remains near-perfect (0.93–0.98) and the overall indicator ranking is preserved throughout.

V. DISCUSSION

A. How Foundation-Model Output Must Be Interpreted

The headline lesson is interpretive. A practitioner who fed recent returns to Chronos and acted on its forecasts at face value would do precisely the wrong thing: the model's confidence is highest exactly when sequence risk is greatest. The information is present—indeed, in ranking terms it is the strongest of any indicator we examined—but its sign is inverted relative to naive intuition. This is consistent with the domain-mismatch concern in the TSFM literature: a model pretrained on series in which momentum persists extrapolates strength forward, which in equity markets coincides with elevated valuations and subsequent sequence risk. The practical implication is not that foundation models are useless for finance but that their raw output cannot be trusted directionally without domain-aware calibration.

B. Why the Detector Does Not Become a Tool

The gap between Chronos-Contrarian's near-perfect ROC-AUC and its failure to improve retirement outcomes is itself instructive. Ranking power measured with full-sample continuous scores does not guarantee a usable real-time rule: an expanding-window threshold that must be exceeded causally fires too seldom and too late, and a single defensive instrument (a temporary allocation shift) is a blunt response to a signal that is informative only in relative terms. Bridging this gap would require either a calibrated, possibly continuous, position-sizing response or an explicitly inverted and smoothed signal—both of which risk overfitting to the single danger-zone episode in the record.

C. Implications for Wealth Managers

For practitioners, three messages follow. First, do not deploy off-the-shelf time-series foundation models as retirement risk monitors; their forecasts are directionally misleading in this setting. Second, among transparent traditional indicators, the VIX and rolling realized volatility offer the most reliable danger-zone detection, and a defensive shift triggered by them improves worst-case outcomes in point-estimate terms—while recognizing that the historical record cannot establish this benefit with statistical confidence. Third, the practitioner implications section that follows translates these findings into a concrete monitoring rule and decision flow.

D. Limitations

Several limitations bound the conclusions. The U.S. sample of 1993–2025 contains essentially one extended danger-zone episode (1998–2001), so all detectors are effectively evaluated against a single regime, and the bootstrap intervals are correspondingly wide. Because a start month can be labeled only if a full 15-year forward window exists, no start after December 2010 is evaluable, and later stress episodes such as 2007–2008 cannot enter the label set. The favorable market produces no ruin events, shifting the comparison onto terminal-wealth tails. We evaluate one TSFM (Chronos-T5-Small) in zero-shot mode; fine-tuned or larger models may differ, as Goel et al. (2024) suggests. The 60-month context is imposed by the timing of the danger zones rather than

chosen freely. Transaction costs and taxes from the defensive shift are not modeled, though for broad-market ETFs the former are small and the latter are minimized in tax-advantaged accounts. Extending the analysis to international markets with less favorable histories, and to fine-tuned foundation models, are the natural next steps.

VI. PRACTITIONER IMPLICATIONS

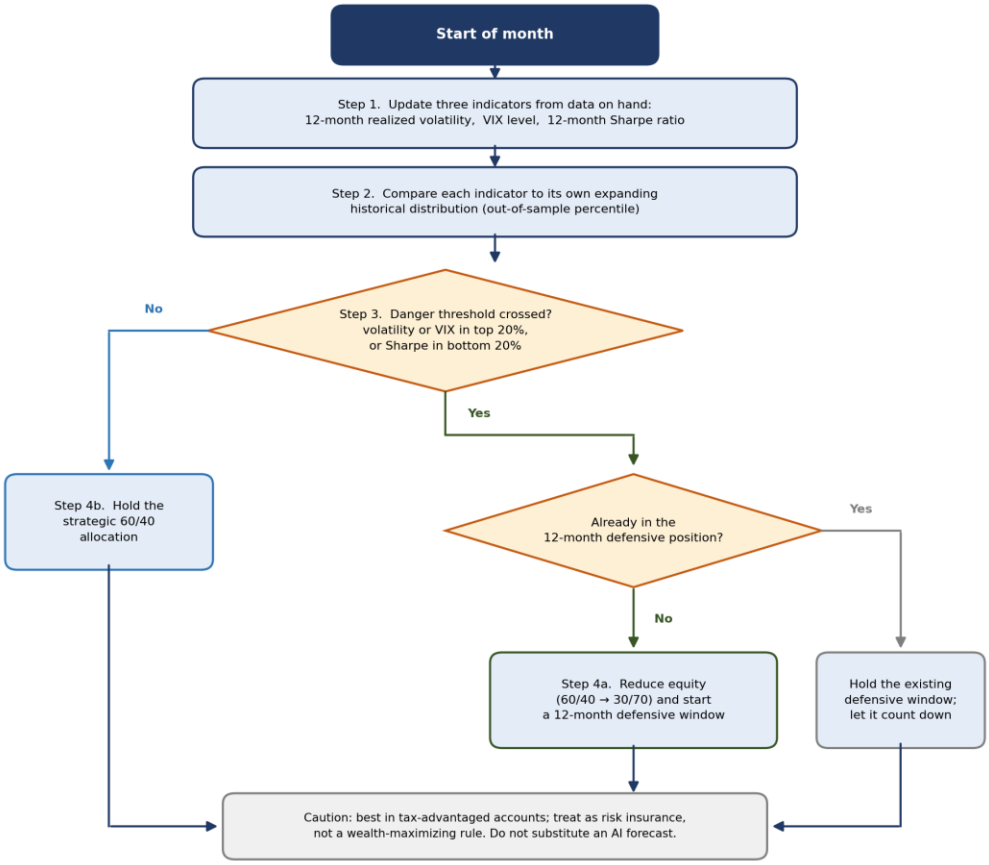
The findings reduce to a simple, transparent monitoring protocol that requires no proprietary technology. Each month, a planner computes two quantities from data already on hand: the trailing 12-month realized volatility of the retiree's portfolio and the current VIX level, each compared against its own history. When either measure rises into the top fifth of its historical range—or, equivalently for the Sharpe-based variant, when the trailing 12-month Sharpe ratio falls into its bottom fifth—the indicator is in its danger zone. The recommended response is a temporary, partial de-risking: shift the equity allocation down (in our tests, from 60% to 30%) for roughly twelve months, then revert. The monitoring is mechanical, the trigger is observable in real time, and the action is reversible.

Two cautions accompany the protocol. It is most appropriate for assets held in tax-advantaged accounts, where the defensive shift incurs no capital-gains tax; in taxable accounts the tax cost may outweigh the benefit. And the planner should treat the strategy as risk insurance with an uncertain premium rather than as a wealth-maximizing rule: in our data it raised worst-case outcomes on average but not with statistical certainty, and it modestly reduced upside. Crucially, the planner should not substitute an artificial-intelligence forecasting tool for this protocol. Our results show that a leading foundation model, read at face value, points in exactly the wrong direction.

A practical decision flow follows four steps each month: (1) update trailing 12-month realized volatility, the VIX, and the trailing Sharpe ratio; (2) compare each to its expanding historical distribution; (3) if a danger threshold is crossed and the portfolio is not already defensive, reduce equity exposure for twelve months; (4) otherwise, hold or revert to the strategic allocation as the twelve-month window elapses.

Figure 7 summarizes the protocol as a one-page decision flowchart. Two implementation parameters deserve explicit statement. First, the twelve-month defensive duration shown in the flowchart is not the only configuration tested: the Section 4.6 sensitivity grid crosses defensive durations of 6, 12, and 18 months with defensive allocations of 20/80, 30/70, and 40/60, and the most conservative short configuration—20% equity held for six months—produced the best worst-case outcome ($997,490). Second, the expanding-percentile thresholds require history before they are meaningful: throughout the analysis no signal may fire until at least 60 months of prior data exist, and practitioners should likewise regard roughly five years of monthly history as the minimum before the percentile comparisons in Steps 2 and 3 of the flowchart are reliable.

Figure 7
Monthly Decision Protocol for Sequence-of-Returns-Risk Monitoring



Note: A plain-language decision flow for practitioners. Each month the planner updates three indicators (12-month realized volatility, the VIX, and the 12-month Sharpe ratio), compares each to its own expanding history, and de-risks the portfolio from 60/40 to 30/70 for twelve months when a danger threshold is crossed, reverting thereafter. The protocol is best suited to tax-advantaged accounts and should be treated as risk insurance rather than a wealth-maximizing rule.

VII. CONCLUSION

This study asked whether a time-series foundation model can detect sequence-of-returns danger zones before they materialize. The answer is subtle. Measured by ranking power, Chronos is the single best detector we examined—but only when its forecast is read contrarily, because the model systematically forecasts strength ahead of the cohorts most exposed to sequence risk. Its optimism is the warning. Yet this near-perfect detection does not translate into an actionable real-time rule, and the traditional indicators that do offer transparent, implementable triggers deliver tail protection whose statistical precision is limited by the single danger-zone episode in the historical record.

The contributions are threefold. First, we provide the first empirical test of a TSFM for SoRR detection and, in correcting an initial misreading, show that foundation-model output requires domain-aware, directionally aware interpretation before it can inform financial planning. Second, we benchmark traditional indicators rigorously—with proper lagging, real data, and bootstrap-based inference—and find their tail-protection benefits economically suggestive but statistically imprecise. Third, we translate the evidence into a concrete, cautious monitoring protocol for practitioners. Future work should test fine-tuned foundation models and international samples containing more than one danger-zone episode, the binding constraint on what any detector can demonstrate from the historical record.

AUTHOR DISCLOSURE STATEMENT FOR AI USE

Amazon's Chronos foundation model (chronos-t5-small) is the subject of the empirical analysis. Grammarly AI was used to improve the writing and readability of the manuscript. The literature review, research design, data

analysis, interpretation, and conclusions are entirely the author's own work. The code and data pipeline are available upon request. The author declares no conflicts of interest, and no external funding was received.

REFERENCES

Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. Review of Financial Studies, 15(4), 1137–1187.

Ang, A., & Timmermann, A. (2012). Regime changes and financial markets. Annual Review of Financial Economics, 4(1), 313–337.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research.*

Bekaert, G., & Hoerova, M. (2014). The VIX, the variance premium and stock market volatility. Journal of Econometrics, 183(2), 181–192.

Bengen, W. P. (1994). Determining withdrawal rates using historical data. Journal of Financial Planning, 7(4), 171–180.

Blanchett, D. (2015). Initial conditions and optimal retirement glide paths. Journal of Financial Planning, 28(9), 52–60.

Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Şimşek, Ö. (2021). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach (ECB Working Paper Series No. 2614). European Central Bank.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327.

Bollerslev, T., Chou, R. Y., & Kroner, K. F. (1992). ARCH modeling in finance: A review of the theory and empirical evidence. Journal of Econometrics, 52(1–2), 5–59.

Cooley, P. L., Hubbard, C. M., & Walz, D. T. (1998). Retirement savings: Choosing a withdrawal rate that is sustainable. AAII Journal, 20(2), 16–21.

Das, A., Kong, W., Sen, R., & Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235, 10148–10167.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. Econometrica, 50(4), 987–1008.

Finke, M., Pfau, W. D., & Blanchett, D. (2013). The 4 percent rule is not safe in a low-yield world. Journal of Financial Planning, 26(6), 46–55.

Fouliard, J., Howell, M., & Rey, H. (2021). Answering the Queen: Machine learning and financial crises (BIS Working Papers No. 926). Bank for International Settlements.

Goel, A., Pasricha, P., & Kanniainen, J. (2024). Time-series foundation AI model for Value-at-Risk forecasting. arXiv preprint arXiv:2410.11773.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., & Dubrawski, A. (2024). MOMENT: A family of open time-series foundation models. Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235, 16115–16152.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. Review of Financial Studies, 33(5), 2223–2273.

Guidolin, M., & Timmermann, A. (2007). Asset allocation under multivariate regime switching. Journal of Economic Dynamics and Control, 31(11), 3503–3544.

Guyton, J. T., & Klinger, W. J. (2006). Decision rules and maximum initial withdrawal rates. Journal of Financial Planning, 19(3), 48–58.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57(2), 357–384.

Kitces, M. E., & Pfau, W. D. (2015). Retirement risk, rising equity glide paths, and valuation-based asset allocation. Journal of Financial Planning, 28(3), 38–48.

Milevsky, M. A. (2006). The calculus of retirement income: Financial models for pension annuities and life insurance. Cambridge University Press.

Milevsky, M. A. (2016). It's time to retire ruin (probabilities). Financial Analysts Journal, 72(2), 8–12.

Milevsky, M. A., & Robinson, C. (2005). A sustainable spending rate without simulation. Financial Analysts Journal, 61(6), 89–100.

Pfau, W. D. (2015). The lifetime sequence of returns: A retirement planning conundrum (SSRN Working Paper No. 2544637). https://ssrn.com/abstract=2544637

Pfau, W. D. (2019). The four approaches to managing retirement income risk (SSRN Working Paper No. 3501673). https://doi.org/10.2139/ssrn.3501673

Pfau, W. D., & Kitces, M. E. (2014). Reducing retirement risk with a rising equity glide path. Journal of Financial Planning, 27(1), 38–45.

Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Darvishi Bayazi, M. J., Adamopoulos, G., Riachi, R., Hassen, N., Biloš, M., Garg, S., Schneider, A., Chapados, N., Drouin, A., Zantedeschi, V., Nevmyvaka, Y., & Rish, I. (2023). Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278.*

Reimann, C. (2024). Predicting financial crises: An evaluation of machine learning algorithms and model explainability for early warning systems. Review of Evolutionary Political Economy, 5(1), 51–83.

Samitas, A., Kampouris, E., & Kenourgios, D. (2020). Machine learning as an early warning system to predict financial crisis. International Review of Financial Analysis, 71, 101507.

Whaley, R. E. (2000). The investor fear gauge. Journal of Portfolio Management, 26(3), 12–17.