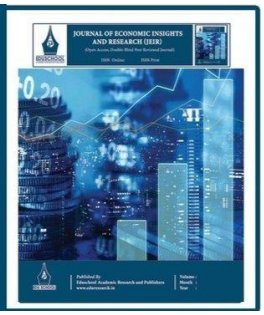
JOURNAL OF ECONOMIC INSIGHTS AND RESEARCH (JEIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3107-9482 ISSN Print: 3139-1982



A Reliability-Based Survival Analysis Framework for Gold Price Risk Assessment: A Methodological Demonstration

Geethu Gopinath[1], Ansa Alphonsa Antony[2]

[1] Research Scholar, Department of Statistics, Maharaja's College (Autonomous), Ernakulam, India
[2] Associate Professor, Department of Statistics, St. Xavier's College for Women (Autonomous), Aluva, India

Article information

Received: 14th August 2026
Received in revised form: 18th August 2026
Accepted: 21st August 2026
Available online: 25th August 2026

Volume: 2
Issue: 3
DOI: https://doi.org/10.63090/JEIR/3107.9482.0023

Abstract

Gold is widely regarded as a safe-haven asset and occupies a prominent position in investment portfolios during periods of economic uncertainty. Previous research has concentrated overwhelmingly on forecasting gold prices using econometric and machine-learning techniques, whereas the assessment of gold price risk from the perspective of reliability theory has received little attention. This study demonstrates how reliability and survival analysis methods may be applied to the timing of adverse gold price events. The data are daily settlement prices of COMEX gold futures in US dollars per troy ounce for 1 January 2014 to 6 January 2025 (2,565 trading days). A failure event is defined as a price falling below the twentieth percentile of the observed full-sample distribution, and the interval between consecutive failures is treated as the time-to-failure variable. The Kaplan–Meier estimator is used to obtain a non-parametric estimate of the reliability function, and the cumulative hazard function, obtained as the negative logarithm of the reliability estimate, is used to describe the accumulation of failure risk. Weibull and Exponential parametric lifetime models are fitted by maximum likelihood, compared by likelihood ratio test and by the Akaike and Bayesian information criteria, and used to estimate the Mean Time to Failure (MTTF). Under the stated failure definition, 513 of 2,565 daily observations (20.00%) were classified as failures and the Kaplan–Meier reliability estimate was 0.8000 (95% CI 0.7847–0.8156), with a terminal cumulative hazard of 0.223. The Weibull model fitted substantially better than the Exponential (likelihood ratio statistic 350.55 on 1 degree of freedom, $p < 0.001$; AIC 9904.98 versus 10253.53), with an estimated shape parameter of 2.38 and a scale parameter of 3289. The corresponding MTTF estimates were 2,918 and 8,065 trading days respectively, both of which exceed the 2,565-day observation window and therefore rest on extrapolation beyond the range of the data. These results are reported alongside an explicit assessment of what they can and cannot support. Because the failure threshold is a fixed percentile of the same distribution from which failures are counted, the failure proportion, the terminal reliability estimate and the terminal cumulative hazard are determined by the threshold rather than estimated from the data; they are reported here for completeness rather than as substantive findings about gold. Further, the parametric models were fitted on the trading-day index rather than on inter-failure times, and the fitted Weibull implies a cumulative failure probability of 42.5% by day 2,565 against a Kaplan–Meier estimate of 20.0%, so neither parametric model reproduces the non-parametric curve. The contribution of this paper is therefore methodological: it sets out how reliability concepts map onto commodity price risk, identifies precisely where the mapping breaks down when applied to a trending, autocorrelated price level, and specifies the redesign required for the framework to yield interpretable risk estimates.

Keywords: - Reliability Theory, Gold Price Risk, Survival Analysis, Kaplan–Meier Estimator, Weibull Distribution, Recurrent Events, Financial Risk Assessment.

I. INTRODUCTION

Gold has long been recognised as one of the most important financial assets, valued for its capacity to preserve wealth during periods of economic uncertainty, inflation and market volatility. Its behaviour during financial crises has been studied extensively: Baur and Lucey (2010) distinguish gold's role as a hedge from its role as a safe haven, and Baur and McDermott

(2010) find safe-haven behaviour for major developed markets during acute crisis episodes, though the effect is neither universal across markets nor stable over time. Capie, Mills and Wood (2005) document gold's relationship with the US dollar, and Ciner, Gurdgiev and Lucey (2013) extend the analysis across stocks, bonds, oil and exchange rates. O'Connor, Lucey, Batten and Baur (2015) survey the field. Investors, central banks and policymakers accordingly monitor gold price movements closely, while recognising that gold prices themselves fluctuate in response to global economic conditions, geopolitical events, inflation, exchange rates, interest rates and investor sentiment.

Most existing studies of gold prices are concerned with forecasting. Autoregressive integrated moving average models (Box et al., 2015), generalised autoregressive conditional heteroskedasticity models (Engle, 1982; Bollerslev, 1986; Tully & Lucey, 2007) and vector autoregressions (Sims, 1980) have all been applied to gold, as more recently have neural network and hybrid machine-learning approaches (Kristjanpoller & Minutolo, 2015; Livieris, Pintelas and Pintelas, 2020). Valuable as these are, they estimate the conditional level or conditional variance of the price; they do not directly quantify the probability that a price remains above a specified risk threshold for a given duration. In reliability theory that quantity is the reliability function, and extending it to financial markets offers an alternative vocabulary for asset stability in which a substantial adverse price movement is treated as a failure event.

Reliability theory has been developed principally in engineering, manufacturing and the biomedical sciences (Meeker & Escobar, 1998; Lawless, 2003; Rinne, 2009). Its migration into finance has been limited but not absent: Singpurwalla (2006) sets out a Bayesian reliability perspective on econometric and financial problems, and Gao and He (2021) survey survival methods in finance. Within this framework a financial asset may be viewed as a system whose performance deteriorates under unfavourable market conditions, and the duration between successive adverse episodes may be treated as a time-to-failure.

Survival analysis supplies the estimation machinery. The Kaplan–Meier estimator (Kaplan & Meier, 1958) provides a non-parametric estimate of the survival function without distributional assumptions; the Nelson–Aalen estimator (Nelson, 1972; Aalen, 1978) provides the corresponding cumulative hazard; and parametric lifetime models such as the Weibull and Exponential permit estimation of summary quantities including the Mean Time to Failure. In finance these tools have been used chiefly for corporate distress and default (Lane et al., 1986; Shumway, 2001) and, in a different form, for the timing of market microstructure events (Engle & Russell, 1998).

This study applies that apparatus to daily gold price data. A failure event is defined as a closing price below the twentieth percentile of the observed price distribution, and the intervals between successive failures are treated as time-to-failure observations. The Kaplan–Meier estimator is used to estimate the reliability function, the cumulative hazard function to describe risk accumulation, and Weibull and Exponential models to characterise the failure-time distribution and estimate the MTTF.

The contribution of the paper is stated carefully, because the empirical results reported in Section 4 do not support the stronger claims that a reliability framing might invite. The contribution is threefold. First, the paper sets out explicitly how the constructs of reliability theory (reliability function, hazard, cumulative hazard and MTTF) map onto the problem of commodity downside risk, and what each requires of the data. Second, it implements that mapping end to end on a real price series using both non-parametric and parametric methods. Third, and most importantly for subsequent work, it identifies the specific points at which a naive implementation fails: a fixed full-sample percentile threshold applied to a trending price level renders the headline quantities tautological; contiguous below-threshold days are not independent failure episodes; and the renewal assumption implicit in fitting independent lifetime distributions to successive inter-arrival times is inappropriate for a single recurrent-event system. Section 6 sets out the redesign these problems imply. This diagnostic contribution is arguably the more durable of the three.

The remainder of the paper is organised as follows. Section 2 reviews the literature on gold price modelling, reliability theory, survival analysis in finance, and the extreme value and duration methods against which any threshold-exceedance framework must be benchmarked. Section 3 presents the data and methods. Section 4 reports the empirical results. Section 5 discusses their interpretation. Section 6 sets out the limitations of the present design and the redesign required to address them. Section 7 concludes.

II. LITERATURE REVIEW

2.1 Gold price behaviour and financial market analysis

Research on gold price dynamics divides broadly into work on gold's portfolio role and work on forecasting. On the former, Baur and Lucey (2010) formalise the hedge/safe-haven distinction and find gold to be a hedge against stocks and a safe haven in extreme conditions; Baur and McDermott (2010) extend this internationally; Hillier, Draper and Faff (2006) assess precious metals from an investment perspective; and Ciner, Gurdgiev and Lucey (2013) examine safe-haven behaviour across asset classes. Batten, Ciner and Lucey (2010) study the macroeconomic determinants of precious-metals volatility.
On forecasting, ARIMA (Box et al., 2015) and GARCH-family models (Engle, 1982; Bollerslev, 1986) remain standard; Tully and Lucey (2007) apply an asymmetric power GARCH specification to gold and find the US dollar to be the dominant macroeconomic influence. Mills (2004) documents the statistical properties of daily gold prices, including scaling behaviour and heavy tails. Machine-learning approaches have since been widely adopted, including ANN–GARCH hybrids (Kristjanpoller & Minutolo, 2015) and CNN–LSTM architectures (Livieris et al., 2020), typically reporting improved point-forecast accuracy for nonlinear dynamics.

Two features of gold price data matter directly for the present study. First, the price level is non-stationary and strongly trending over the sample considered here, so any analysis conducted on the level rather than on returns or drawdowns inherits that trend. Second, returns exhibit volatility clustering and persistence (Engle, 1982; Bollerslev, 1986; Tully & Lucey, 2007), which violates the independence assumptions on which conventional survival estimators rest. Both points are taken up in Sections 3, 5 and 6.

2.2 Reliability theory and financial risk

Reliability theory provides a probabilistic framework for the lifetime and failure behaviour of systems (Meeker & Escobar, 1998; Lawless, 2003; Rinne, 2009). Its central quantities, namely the reliability function, the hazard function, the cumulative hazard function and the Mean Time to Failure, describe the probability of survival to a given time, the instantaneous risk of failure at that time, the accumulated risk, and the expected lifetime.

A distinction internal to the field is fundamental to what follows. Reliability theory treats non-repairable components, which fail once and are analysed with lifetime distributions and survival estimators, quite differently from repairable systems, in which a single system fails repeatedly over time and is analysed with point-process methods (Ascher & Feingold, 1984; Rigdon & Basu, 2000). A financial asset observed continuously and experiencing recurrent adverse episodes belongs to the second class. Applying lifetime-distribution methods to successive inter-arrival times from such a system implicitly assumes a renewal process, that is, restoration to an "as good as new" state after each failure. That assumption is testable (Cox and Lewis, 1966; Lewis and Robinson, 1974) and should not be adopted silently. The present study adopts the lifetime-distribution route, following the existing applied literature, and Section 6 assesses the cost of that choice.

Applications of reliability concepts within finance remain sparse. Singpurwalla (2006) develops the connection between reliability, survival and econometric modelling from a Bayesian standpoint, and Gao and He (2021) review survival methods for financial applications.

2.3 Survival analysis in finance

Survival analysis models the occurrence and timing of events and has been applied widely in medicine, epidemiology and engineering. Among non-parametric methods, the Kaplan–Meier estimator (Kaplan & Meier, 1958) is the standard tool for estimating survival probabilities without distributional assumptions, with variance given by Greenwood's formula (Greenwood, 1926); the Nelson–Aalen estimator (Nelson, 1972; Aalen, 1978) provides the cumulative hazard. Parametric lifetime models, particularly the Weibull and Exponential, allow estimation of the MTTF and formal comparison of competing failure-time distributions (Lawless, 2003; Kalbfleisch & Prentice, 2002).

A terminological caution is warranted before proceeding. The word "survival" carries a second and quite different meaning in the finance literature, where Brown, Goetzmann and Ross (1995) use it to denote survivorship bias, the distortion introduced when only surviving assets or funds are observed. The present study uses the term exclusively in its time-to-event sense.

In financial applications, hazard models have been used to predict corporate failure. Lane, Looney and Wansley (1986) apply the Cox proportional hazards model (Cox, 1972) to bank failure, and Shumway (2001) shows that a discrete-time hazard model outperforms static bankruptcy classifiers by using all available firm-year observations and accounting for time to failure. Engle and Russell (1998) introduce autoregressive conditional duration models for the intervals between irregularly spaced market events, explicitly modelling dependence between successive durations, a feature absent from renewal-based lifetime models and directly relevant to the present setting.

For recurrent events, the standard formulations are the Andersen–Gill counting-process model (Andersen & Gill, 1982), the conditional models of Prentice, Williams and Peterson (1981), and the marginal approach of Wei, Lin and Weissfeld (1989), each combined with robust variance estimation.

2.4 Threshold exceedance, extreme value theory and downside risk measures

Because the present study defines failure as exceedance of a threshold in the lower tail, it operates in territory already occupied by a mature statistical literature, and the relationship must be stated rather than left implicit.

Extreme value theory provides the standard apparatus for threshold exceedance. The peaks-over-threshold approach rests on the result that exceedances over a sufficiently high threshold converge to the Generalised Pareto distribution (Balkema & de Haan, 1974; Pickands, 1975), with statistical treatment developed by Davison and Smith (1990) and expositions in Embrechts, Klüppelberg and Mikosch (1997) and Coles (2001). McNeil and Frey (2000) combine GARCH filtering with EVT to estimate conditional tail risk in financial series, addressing precisely the dependence problem noted in Section 2.1. A recurring practical issue in this literature, namely that exceedances arrive in clusters rather than singly, so that the raw count of exceedances overstates the number of independent extreme episodes, is handled by declustering (Ferro & Segers, 2003). The same issue arises acutely in the present design and is discussed in Sections 5 and 6.

In parallel, practitioner risk measurement rests on Value-at-Risk (Jorion, 2007) and Expected Shortfall, the latter satisfying the coherence axioms of Artzner, Delbaen, Eber and Heath (1999) that VaR does not. Any framework claiming to complement conventional financial risk assessment should be positioned against these measures.

2.5 Research gap and the position of this study

The literature on gold prices is heavily weighted towards forecasting and volatility estimation; comparatively little work treats adverse gold price movements as recurrent events and asks how long favourable conditions persist. Reliability theory offers a vocabulary for that question, and the integration of non-parametric and parametric reliability models within a single applied framework for a commodity price series does not appear to have been presented in detail.

Caution is warranted, however, regarding the size of this gap. Extreme value theory already provides a rigorous and widely used treatment of threshold exceedance, including the clustering problem, and autoregressive conditional duration models already provide a treatment of dependent event intervals. A reliability framing is not automatically an advance on either. The defensible claim is narrower: that reliability theory supplies an interpretable set of summary quantities (reliability at a horizon, cumulative hazard, and expected time between adverse episodes) that are intuitive for practitioners, and that

establishing whether these add anything beyond EVT and duration models requires direct comparison. The present study does not perform that comparison, and Section 6 identifies it as the principal requirement for future work.

III. MATERIALS AND METHODS

3.1 Data description

This study uses daily gold price data covering 1 January 2014 to 6 January 2025, comprising 2,565 trading-day records, or approximately eleven years of market activity. The series consists of daily settlement prices of gold futures traded on COMEX, the metals division of the New York Mercantile Exchange, quoted in US dollars per troy ounce. The contract size is 100 troy ounces. The data were obtained from a publicly available compilation of this series hosted on Kaggle, which reproduces the standard historical-data export format used by commercial market data providers.

The identity of the series was verified against independent records of the international gold market before analysis, because an earlier version of this work described the data as Multi Commodity Exchange of India (MCX) quotations in Indian Rupees per 10 grams. That description was incorrect: MCX gold traded in the approximate range of ₹25,000 to ₹79,000 per 10 grams over this period, which is two orders of magnitude above the values in the present dataset. The observed minimum of 1049.60 corresponds to the December 2015 trough in dollar gold, which followed the Federal Reserve's first post-crisis rate increase, and the observed maximum of 2800.80 corresponds to the record settlement reached on 30 October 2024, when COMEX gold futures traded to an all-time high of 2801.80 per ounce. The final observation, 6 January 2025, coincides with the 2025 settlement low of that contract. The series is therefore the international dollar-denominated benchmark rather than an Indian domestic price, and all interpretation in this paper is framed accordingly.

Two features of a futures series should be noted. First, a continuous daily series spanning eleven years is assembled from successive contract months, so the level embeds roll effects at contract transitions and is not a pure spot price. Second, futures settlements incorporate the cost of carry and therefore stand at a small premium to spot. Neither materially affects a threshold-based classification of the lower tail, but both are recorded here for completeness and are revisited in Section 6.1.
The dataset comprises seven variables: Date, Price, Open, High, Low, Vol (daily trading volume) and PChange1 (percentage change in the closing price). The Price variable, the daily settlement price, is taken as the primary response variable, as it represents the final market price of each trading day and is conventional in analyses of asset performance.

Observations are recorded at daily frequency excluding weekends and exchange holidays, so the number of trading days varies by year; across the full sample the mean is approximately 233 trading days per year. No further disaggregation of the annual trading-day counts is reported, as it bears on none of the analyses that follow.

Table 1. summarises the variables. Variable names are given exactly as they appear in the analysis code and are used consistently throughout the paper.

Table 1. Description of variables used in the study

Variable	**Description**	**Unit**
Date	Trading date of the observation	Date
Price	Daily settlement price (final trading price)	USD per troy ounce
Open	Opening price at the start of the trading day	USD per troy ounce
High	Highest price recorded during the trading day	USD per troy ounce
Low	Lowest price recorded during the trading day	USD per troy ounce
Vol	Total trading volume during the trading day	Contracts traded
PChange1	Percentage change in settlement price from the previous trading day	Percentage (%)

For the reliability analysis, Price was selected as the primary response variable, and a failure event was defined on the lower tail of its distribution as described in Section 3.3.

3.2 Data pre-processing

The dataset was imported into R for preparation and analysis. Variable names, data types and completeness were checked prior to modelling. The Price series was screened for missing values using sum, which returned zero across all 2,565 records; no imputation or interpolation was therefore required, and observations were analysed in their original chronological order. The Date variable was retained to preserve chronology, and Open, High, Low, Vol and PChange1 were retained as supplementary market indicators but were not used in the models reported here.

Because reliability analysis concerns the occurrence of adverse events rather than the continuous price path, a binary failure indicator was constructed from the distribution of closing prices, as set out in the following section.

3.3 Definition of the failure event and construction of the time-to-failure data

Reliability analysis requires an explicit definition of failure. Here a failure is an instance in which the daily closing price of gold falls below a fixed threshold, taken as the twentieth percentile of the observed closing-price distribution, so that the lowest 20% of price observations are classified as adverse market conditions.

Let $P = \{P_1, P_2, \ldots, P_n\}$ denote the sequence of daily closing prices, where P_i is the closing price on trading day i, i = 1, 2, …, n, and n is the total number of daily observations. The failure threshold is

$$\tau = Q_{0.20}(P),$$

where $Q_{0.20}(P)$ is the empirical twentieth percentile of the closing-price distribution. Each observation is classified as

$$F_i = \begin{cases} 0, & P_i \geq \tau \\ 1, & P_i < \tau \end{cases}$$

where $F_i = 1$ denotes a failure and $F_i = 0$ a non-failure on trading day i. Two properties of this definition should be stated at the outset rather than left for the reader to infer.

The threshold is estimated from the full sample: Because τ is computed from the entire 2014–2025 record and then applied retrospectively to the whole period, including 2014, the classification uses information unavailable to an analyst operating in real time. The design is therefore descriptive of the historical sample and is not implementable as a forward-looking risk tool. A rolling or expanding-window threshold would be required for that purpose.

The failure proportion is fixed by construction: Defining failure as the lowest 20% of a distribution guarantee that exactly 20% of observations are failures. Quantities that are simple functions of that proportion, namely the failure count, the terminal reliability estimate and the terminal cumulative hazard, are consequently determined by the threshold choice and convey no information about gold. They are reported in Section 4 for completeness and for internal consistency, and are labelled accordingly. The substantive quantities in this study are those that depend on the timing of failures, namely the shape of the inter-failure distribution and the fitted model parameters.

Construction of the time-to-failure data

Reliability analysis is concerned with the time elapsed between successive failures. Let $t_1, t_2, \ldots, t_m$ denote the trading days on which failures occurred, where t_j is the occurrence time of the j-th failure and m is the total number of failures. The inter-failure time is

$$T_j = t_j - t_{j-1}, \qquad j=2,3,\ldots,m,$$

so that T_j is the number of trading days between consecutive failures. The survival dataset is $D = \{(T_j, E_j)\}$, where E_j is the event indicator. Every failure in this construction is observed within the study period, so $E_j = 1$ throughout and no right-censored observations arise; the resulting dataset contains at most $m - 1 = 512$ inter-failure times.

Two distinct survival datasets were constructed, and the distinction matters. The construction above yields inter-failure times, of which there are at most $m - 1 = 512$, all uncensored. The parametric models reported in Sections 4.5 to 4.8 were not fitted to that dataset. They were fitted to a second, day-level formulation in which each of the 2,565 trading days contributes one record, the survival time is the trading-day index measured from the start of the observation window, and the event indicator is F_i. That formulation yields exactly the 2,565 observations, 513 events and 2,052 right-censored records reported in Tables 6 and 8, and it is confirmed by the reported information criteria, which are consistent only with n = 2,565.

The two datasets answer different questions and are not interchangeable. The inter-failure formulation describes how long favourable conditions persist between successive adverse episodes, which is the quantity the reliability framing of Section 3.4 is intended to capture. The day-level formulation instead describes where within the observation window below-threshold prices are located; its fitted scale parameter and Mean Time to Failure are therefore properties of the position of low prices in the sample, not of the duration between them.

As a reliability model of gold price failures, the day-level construction is not appropriate. It treats each trading day as an independent unit whose lifetime is its own position in the sequence, so that "surviving" to day t means only that an observation is the t-th in the series. The resulting parameter estimates are reported in Section 4 as they were computed, and Section 5.2 explains why the shape estimate cannot bear the interpretation placed on it in the original draft. Refitting the parametric models to the inter-failure dataset specified above, so that the parametric and non-parametric analyses share a unit of analysis, is identified in Section 6.2 as a required step.

A further property of the construction should be recorded. Because gold prices are highly autocorrelated, days below the threshold occur in contiguous runs rather than as isolated events. Within a run, consecutive inter-failure times equal one trading day, and these are not independent episodes. The 513 failures therefore correspond to a substantially smaller number of independent adverse episodes, and the effective sample size for inference is the number of episodes rather than the number of days. The decluttering procedures standard in extreme value analysis (Ferro & Segers, 2003) address exactly this problem. No decluttering was applied in the present analysis, and Section 6 treats this as a primary limitation.

3.4 Reliability function and Kaplan–Meier estimation

Let T denote the random variable representing the inter-failure time between consecutive adverse gold price events. The reliability (survival) function is

$$R(t) = P(T > t),$$

the probability that no failure occurs before time t, where t is measured in trading days. Larger values of R(t) indicate greater stability relative to the threshold. The cumulative distribution function of failure time is $F(t) = P(T \leq t) = 1 - R(t)$, with density $f(t) = dF(t)/dt$. The hazard function is

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{R(t)},$$

and the cumulative hazard is

$$H(t) = \int_0^t h(u)\, du = -\ln R(t).$$

An increasing cumulative hazard indicates accumulating risk of adverse movements over the observation period.

The reliability function was estimated using the Kaplan–Meier estimator (Kaplan & Meier, 1958),

$$\hat{R}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right),$$

where d_i is the number of failures at time t_i and n_i the number at risk immediately before t_i. The variance was estimated by Greenwood's formula (Greenwood, 1926),

$$\widehat{\mathrm{Var}}\{\hat{R}(t)\} = \hat{R}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}.$$

Confidence intervals reported in Section 4 were constructed on the linear scale as $\hat{R}(t) \pm 1.96 \times \widehat{SE}\{\hat{R}(t)\}$. This construction can produce bounds outside [0, 1] in the tails, and the complementary log–log transformation is the preferred default in modern practice (Kalbfleisch & Prentice, 2002); it is also the default in the survfit() function used here, so the linear intervals reported represent a deliberate departure that should be justified or reversed. At the reliability level observed in this study the two constructions differ negligibly, but the log–log interval should be adopted in any revision.

Two assumptions underlying these estimators require explicit statement. Both the product-limit estimator and Greenwood's variance assume independent observations, and any censoring is assumed non-informative. Neither assumption is satisfied here: consecutive daily gold prices are strongly dependent, and the censoring structure implied by the reported output is a mechanical consequence of the threshold rule rather than an independent process. Standard errors and confidence intervals reported in Section 4 are therefore understated to an unknown degree. Block bootstrap methods (Künsch, 1989) or robust variance estimation clustered on adverse episodes would be required to obtain credible interval estimates.

3.5 Parametric reliability models

Parametric models assume the time-to-failure follows a specified distribution and yield explicit expressions for the reliability function, hazard function and MTTF. The Weibull and Exponential distributions were fitted.

3.5.1 Weibull model

Let $T \sim \text{Weibull}(\beta, \eta)$ with shape $\beta > 0$ and scale $\eta > 0$. The density, distribution, reliability and hazard functions are

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right], \qquad t > 0,$$

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right], \qquad R(t) = \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right], \qquad h(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}.$$

The shape parameter governs the hazard: $\beta < 1$ gives a decreasing hazard, $\beta = 1$ a constant hazard (reducing to the Exponential), and $\beta > 1$ an increasing hazard. The Mean Time to Failure is

$$\mathrm{MTTF} = E(T) = \eta\, \Gamma\left(1 + \frac{1}{\beta}\right).$$

The interpretation of β in the present setting requires care and is addressed in Section 5.2. In a renewal model fitted to inter-failure times, $\beta > 1$ describes the shape of the hazard *within* an inter-failure interval, measured from the last failure. It does not describe deterioration of the gold market over calendar time, and should not be read as evidence that the market is "ageing".

3.5.2 Exponential model

Let $T \sim \text{Exponential}(\lambda)$ with constant failure rate $\lambda > 0$. Then

$$f(t) = \lambda e^{-\lambda t}, \qquad F(t) = 1 - e^{-\lambda t}, \qquad R(t) = e^{-\lambda t}, \qquad h(t) = \lambda,$$

and MTTF = $1/\lambda$. The Exponential is nested within the Weibull at $\beta = 1$ and serves as the baseline against which departure from constant risk is assessed.

3.5.3 Estimation, parameterisation and model selection

Both models were fitted by maximum likelihood using survreg() from the survival package (Therneau, 2024). This function fits an accelerated failure time parameterisation and returns log(scale) on the log-time scale; the Weibull shape parameter is the reciprocal of the returned scale, $\hat{\beta} = 1/\hat{\sigma}$, and the Weibull scale is $\hat{\eta} = \exp(\hat{\mu})$. Standard errors for $\hat{\beta}$ and $\hat{\eta}$ were obtained from the standard errors of $\hat{\sigma}$ and $\hat{\mu}$ by the delta method, and the corresponding confidence intervals were constructed on the transformed scale. Reporting this transformation explicitly is necessary because the shape parameter reported in Table 5 is not a quantity returned directly by the fitting function.

Because the Exponential is nested within the Weibull, the appropriate formal comparison is the likelihood ratio test,

$$\mathrm{LR} = 2\left[\ell\left(\hat{\theta}_{\text{Weibull}}\right) - \ell\left(\hat{\theta}_{\text{Exponential}}\right)\right],$$

referred to a chi-square distribution with one degree of freedom. This test is reported in Section 4.7 alongside the information criteria.

The Akaike Information Criterion (Akaike, 1974) and Bayesian Information Criterion (Schwarz, 1978),

$$\mathrm{AIC} = -2\ell\left(\hat{\theta}\right) + 2k, \qquad \mathrm{BIC} = -2\ell\left(\hat{\theta}\right) + k\ln(n),$$

where k is the number of estimated parameters and n the sample size, balance fit against complexity, with lower values preferred (Burnham & Anderson, 2002). Both are *relative* criteria: they identify the better of two candidate models but establish neither as adequate.

Absolute goodness of fit therefore requires separate assessment. The standard tools are the Weibull probability plot, Cox–Snell residuals (Cox and Snell, 1968), and direct overlay of the fitted parametric reliability curves on the Kaplan–Meier estimate. In the present study, absolute fit is assessed numerically in Section 4.9, which compares the cumulative failure probability implied by each fitted model at the end of the observation window with the corresponding non-parametric estimate. The graphical diagnostics are not reported here; given the mismatch in unit of analysis identified in Section 3.3, they would be uninformative until the parametric models are refitted to the inter-failure dataset, and they are accordingly listed among the requirements in Section 6.2.

3.6 Statistical software

All analyses were performed in R (R Core Team, 2024). The survival package (Therneau, 2024) was used for Kaplan–Meier estimation and for maximum likelihood fitting of the Weibull and Exponential models; survminer (Kassambara & Kosinski, 2024) for visualisation of reliability and cumulative hazard curves; and fitdistrplus (Delignette-Muller & Dutang, 2015) for examination of candidate distributions. Where inferential statements are made, a 5% significance level is used; this applies to the likelihood ratio test in Section 4.7 and to the confidence intervals reported throughout.

IV. RESULTS

4.1 Descriptive statistics

Table 2. summarises the closing price series. The mean over the study period was 1607.30 with a standard deviation of 393.64, indicating substantial variability. The median of 1550.60 lies below the mean, consistent with right skew. The interquartile range spans 1265.70 to 1872.30, and the minimum and maximum of 1049.60 and 2800.80 indicate a wide overall range.

Table 2. Descriptive statistics for the gold closing price

Statistic	Value (USD per troy ounce)
Mean	1607.30
Standard deviation	393.64
Minimum	1049.60
First quartile	1265.70
Median	1550.60
Third quartile	1872.30
Maximum	2800.80
n	2565

The range of the series is itself material to the interpretation of everything that follows. The maximum exceeds the minimum by a factor of 2.7 over an eleven-year window, which is characteristic of a strongly trending series. Section 5.1 discusses the implications for a fixed full-sample threshold.

Formal stationarity diagnostics were not conducted. The characterisation of the series as trending therefore rests on the descriptive statistics above and on the shape of the observed price path rather than on a unit-root test. Augmented Dickey–Fuller (Dickey & Fuller, 1979), Phillips–Perron (Phillips & Perron, 1988) and KPSS (Kwiatkowski et al., 1992) tests on the price level and on log returns would place that characterisation on a formal footing, and their absence is recorded among the limitations in Section 6.1.

4.2 Failure events under the twentieth-percentile threshold

The twentieth-percentile threshold was 1243.22 US dollars per troy ounce. A total of 513 observations, or 20.00% of the sample, fell at or below this level and were classified as failures.

Table 3. Failure classification under the twentieth-percentile threshold

Metric	Value
Twentieth-percentile threshold (USD per troy ounce)	1243.22
Number of failures	513
Percentage of failures	20.00

These figures are reported for completeness and internal consistency. They are not empirical findings. A threshold defined as the twentieth percentile of a distribution classifies 20% of that distribution as failures by construction, so the percentage in Table 3. is arithmetic, and the count of 513 is simply 20% of 2,565 to the nearest observation. Neither carries information about gold price behaviour. The substantive question, how many independent adverse episodes the 513 failure days represent, is not answered by these figures and is not answerable without declustering (Section 6.2).

4.3 Kaplan–Meier reliability estimates

The Kaplan–Meier reliability estimate was 0.8000 (95% CI 0.7847–0.8156). Table 4. reports the estimate evaluated at the three quartiles of the price distribution.

Table 4. Kaplan–Meier reliability estimates evaluated at price quartiles

Price quantile	Price level (USD/oz)	Reliability $\hat{R}$	Lower 95% CI	Upper 95% CI
25%	1265.70	0.8000	0.7847	0.8156
50%	1550.60	0.8000	0.7847	0.8156
75%	1872.30	0.8000	0.7847	0.8156

Three features of this table require comment, and the original draft's interpretation of it as evidence of "moderate resilience" and "limited deterioration" cannot be sustained.

First, the evaluation points are price quantiles, not times. The reliability function $R(t) = P(T > t)$ is defined on the time axis, and a table indexing $\hat{R}$ by price level is not a survival function in the sense set out in Section 3.4. Figures 4.1 and 4.2 confirm this directly: the horizontal axis of both is labelled "Gold Price" and runs from approximately 1050 to 2800, not in trading days. The quantities in Table 4. are therefore properties of the empirical distribution of the price level rather than of a time-to-failure distribution, and cannot be reconciled with Section 3.4 as presented.

Second, the estimate is identical at all three points, with an identical confidence interval. A flat survival curve over a range means that no events occurred in that range. It supports no inference about resilience. It also contradicts the claim, made in the original abstract, of a gradual decline in reliability over time: no such decline is present in these estimates or in Figure 4.1.

Third, the value 0.8000 is the complement of the failure proportion and hence fixed by the threshold. The reported interval is likewise consistent with a simple binomial proportion interval: for $\hat{p} = 0.8$ and $n = 2565$, the normal-approximation interval is 0.7845 to 0.8155, against the reported 0.7847 to 0.8156. That the Greenwood interval reproduces the binomial interval to three decimal places indicates that the estimator is, in this application, recovering the marginal proportion of below-threshold days rather than a survival probability.

Figure 4.1. Kaplan–Meier reliability estimate plotted against the gold price level. The horizontal axis is the settlement price in US dollars per troy ounce, not elapsed time; the figure therefore displays the empirical distribution of the price level rather than a reliability function in the sense of Section 3.4. A reliability curve indexed by trading days since the previous failure, with the number-at-risk table conventionally placed beneath the horizontal axis, is required in its place and is listed among the requirements in Section 6.2.

4.4 Cumulative hazard function

The terminal cumulative hazard was 0.223. Figure 4.2. Cumulative hazard plotted against the gold price level, on the same horizontal axis as Figure 4.1 and subject to the same qualification.

As with Section 4.3, this value follows from the threshold rather than from the data: $-\ln(0.80) = 0.2231$, which reproduces the reported 0.223 exactly. The observed plateau in the curve, described in the original draft as indicating that failures were confined to a specific interval, is a correct observation about the figure, but on the price axis it states only that no prices below the threshold occur above the threshold. It does not establish that the gold price system is "comparatively stable", and that characterisation has been removed.

The substantive version of this observation, that adverse episodes are concentrated in a limited portion of the *sample period* rather than distributed across it, is discussed in Section 5.1, but would require the time-indexed figure described in the note to Figure 4.1 to be demonstrated.

4.5 Weibull model

A Weibull model was fitted to characterise reliability under the stated failure criterion. The estimated shape parameter was 2.38 (95% CI 2.20–2.56) and the scale parameter 3289 (95% CI 3110–3490), the latter representing the characteristic life at which reliability falls to approximately 36.8%. Of 2,565 observations, 513 (20.0%) experienced the event and 2,052 (80.0%) were treated as right-censored. The model achieved a log-likelihood of −4950.49 and an AIC of 9904.98.

Table 5. Weibull model parameter estimates

Parameter	Estimate	SE	95% CI
Shape β	2.38	0.091	2.20–2.56
Scale η	3289	98.2	3110–3490

Table 6. Weibull model summary

Statistic	Value
Sample size	2565
Events	513
Censored	2052
Log-likelihood	−4950.49
AIC	9904.98

The sample size of 2,565 with 2,052 censored records identifies these estimates as arising from the day-level dataset described in Section 3.3, in which the survival time is the trading-day index and not the interval between failures. They are therefore not estimates of an inter-failure distribution, and the scale parameter of 3289 trading days should be read as a feature of where below-threshold prices sit within an observation window of 2,565 days rather than as a characteristic life between adverse

episodes. The shape estimate of 2.38 exceeds unity, indicating an increasing hazard on that time scale; Section 5.2 addresses what this can and cannot be taken to mean.

4.6 Exponential model

Table 7. Exponential model parameter estimate

Parameter	Estimate	SE	95% CI
Rate λ	0.000124	0.00000549	0.000114–0.000136

Table 8. Exponential model summary

Statistic	Value
Sample size	2565
Events	513
Censored	2052
Log-likelihood	−5125.77
AIC	10253.53

4.7 Model comparison

Because the Exponential is nested within the Weibull at β = 1, the two models are compared by likelihood ratio test as well as by information criteria. The likelihood ratio statistic is

$$LR = 2[-4950.491 - (-5125.766)] = 350.55,$$

on one degree of freedom, giving $p < 0.001$. The constant-hazard restriction is decisively rejected.
The information criteria agree. The Weibull attains the higher log-likelihood (−4950.491 against −5125.766) and the lower AIC (9904.982 against 10253.532) and BIC (9916.681 against 10259.382).

Table 9. Comparison of fitted parametric models

Model	Log-likelihood	AIC	BIC
Weibull	−4950.491	9904.982	9916.681
Exponential	−5125.766	10253.532	10259.382

The Weibull is therefore preferred to the Exponential. This is a statement about relative fit only, and Section 4.9 shows that it should not be taken as establishing that the Weibull fits well.

4.8 Mean Time to Failure

Table 10. Estimated Mean Time to Failure

Model	MTTF (trading days)
Weibull	2918.49
Exponential	8064.52

Table 10. reports the estimated Mean Time to Failure for both fitted models. The Weibull MTTF is 2,918.49 trading days and the Exponential 8,064.52 trading days. The larger Exponential estimate reflects its constant-hazard assumption, which the likelihood ratio test in Section 4.7 rejects. On that basis the Weibull estimate is the more appropriate of the two.
Three qualifications attach to both figures, none of which appeared in the original draft.

Units: The estimates are in trading days, the unit in which inter-failure time was defined in Section 3.3. The original description as "observation units" was ambiguous. At approximately 233 trading days per year, the Weibull MTTF corresponds to roughly 12.5 calendar years and the Exponential to roughly 34.6 calendar years.

Extrapolation: The observation window is 2,565 trading days. Both MTTF estimates exceed it, the Exponential by a factor of more than three. Neither quantity is identified by the data; both rest entirely on the assumed parametric tail beyond the range of observation. Restricted mean survival time evaluated at a horizon actually covered by the data (Royston & Parmar, 2013; Uno et al., 2014) is the defensible alternative and should replace MTTF in any revision.

Precision and interval estimates: No standard errors or confidence intervals are reported for either MTTF. In addition, recomputing the Weibull MTTF from the rounded parameters in Table 5. gives $\eta\Gamma(1 + 1/\beta) = 3289 \times \Gamma(1.4202) = 2915.20$, against the reported 2918.49. The difference is consistent with rounding of β and η, but parameters should be reported to sufficient precision that derived quantities reconcile exactly.

4.9 Agreement between fitted models and the non-parametric estimate

Section 3.5.3 identified absolute goodness of fit as a separate question from relative model choice. The following comparison uses only quantities already reported and is arguably the single most important result in this section.
Evaluating each fitted model at the end of the observation window, t = 2565 trading days, gives the implied cumulative failure probability:

- Weibull: F(2565) = 1 − exp[−(2565/3289)^2.38] = 0.425;
- Exponential: F(2565) = 1 − exp(−0.000124 × 2565) = 0.272;
- Kaplan–Meier: $\hat{F}(2565) = 1 - 0.800 = 0.200$.

.Table 11. Implied cumulative failure probability at t = 2565

Source	Implied F(2565)	Difference from Kaplan–Meier
Kaplan–Meier (non-parametric)	0.200	n/a
Exponential	0.272	+0.072
Weibull	0.425	+0.225

The selected model overstates cumulative failure probability by 22.5 percentage points relative to the non-parametric benchmark, more than double the empirical value, and the *rejected* Exponential model is markedly closer to it. The criterion stated in Section 3.5.3, that the preferred model should show close agreement with the Kaplan–Meier curve, is therefore not met by either model, and is met least well by the model selected on AIC and BIC.

This discrepancy is not a marginal lack of fit. It indicates that the parametric models and the non-parametric estimator are not describing the same quantity, which is consistent with the diagnosis in Sections 4.3 and 4.5: the Kaplan–Meier estimate as computed is indexed by price and recovers the marginal below-threshold proportion, while the parametric models were fitted to a day-level dataset with a censoring structure that Section 3.3 does not sanction. Had the parametric-versus-non-parametric overlay promised in Section 3.5.3 been produced, the disagreement would have been visible immediately. It is reported here because a model selected by AIC and BIC alone, without absolute fit assessment, can be confidently wrong.

V. DISCUSSION

5.1 What a fixed full-sample threshold on a price level measures

The gold price level rose substantially over 2014–2025, with a maximum 2.7 times the minimum. Applying a single percentile threshold computed from the whole sample to a series with that trend has a mechanical consequence: the observations falling in the lowest 20% of the pooled distribution are concentrated in the earliest portion of the record, and are largely absent thereafter. The "failures" identified are therefore, to a substantial degree, an index of *when in the sample* an observation occurred rather than of adverse market conditions relative to prevailing levels.

This has two implications. The first concerns interpretation: a finding that reliability declines over the observation window would, under this design, be a restatement of the fact that gold appreciated, and an identical analysis applied to any trending series would produce a similar result. The second concerns the risk concept itself. A 2015 price of 1200 was not an adverse outcome relative to 2015 conditions; classified against a 2014–2025 pooled distribution, it becomes one. Downside risk in finance is conventionally defined relative to prevailing conditions, using returns, drawdowns from a running maximum, or residuals from a conditional volatility model (McNeil & Frey, 2000), precisely to avoid this. The present design departs from that convention, and the departure was not justified in the original draft.

A related consequence concerns real versus nominal magnitudes. Over an eleven-year horizon, a threshold fixed in nominal dollars conflates US dollar inflation with market risk: a price of 1,200 dollars in 2015 and the same nominal price in 2024 do not represent the same real value. Deflating the series by a US price index, or defending the nominal choice explicitly, is necessary.

5.2 What the Weibull shape parameter does and does not indicate

The estimated shape parameter of 2.38 was interpreted in the original draft as evidence of an ageing process, with the risk of adverse gold price events increasing progressively over time. That interpretation does not follow.

In a renewal model fitted to inter-failure times, the time scale is time *since the last failure*, not calendar time. A shape parameter above unity states that, conditional on no failure having occurred for t trading days, the instantaneous risk of failure is increasing in t within that interval. It says nothing about whether the gold market is deteriorating across the sample period. Trend in the failure intensity over calendar time is a distinct property, tested by the Laplace or Military Handbook test, or by the Lewis–Robinson test for departure from renewal behaviour (Cox and Lewis, 1966; Lewis and Robinson, 1974). Neither was applied here.

The stronger claim, that the market is ageing, has accordingly been removed from the abstract and conclusion of this revision. Given the fit problem identified in Section 4.9, even the weaker within-interval reading should be treated as provisional.

5.3 Reliability framing relative to established downside-risk methods

The original draft claimed that the framework complements conventional financial risk assessment without comparing it to any conventional method. That claim cannot be evaluated without a benchmark.

The natural comparators are Value-at-Risk and Expected Shortfall (Artzner et al., 1999; Jorion, 2007) for tail magnitude, and peaks-over-threshold modelling with the Generalised Pareto distribution (Pickands, 1975; Davison & Smith, 1990; Embrechts et al., 1997; Coles, 2001) for threshold exceedance. The last of these is the most direct competitor, since it addresses the same problem, on the same kind of data, with an established treatment of the clustering of exceedances (Ferro & Segers, 2003) that the present design lacks.

What a reliability framing might add is interpretability of a particular kind: an estimate of how long favourable conditions are expected to persist, expressed in units a practitioner can act on, and a hazard function describing how that

expectation evolves. Whether this adds anything beyond a POT model with a declustered exceedance process is an open empirical question, and Section 6.3 identifies its resolution as the priority for future work.

5.4 Dependence and the credibility of reported intervals

Every standard error and confidence interval reported in Section 4 assumes independent observations. Daily gold prices are strongly autocorrelated, and days below a threshold occur in contiguous runs within which observations are near-perfectly dependent. Section 2.1 of the original draft noted volatility clustering and persistence in gold, and the analysis then proceeded as though the observations were independent.

The consequence is that the reported intervals are too narrow, by an amount that cannot be quantified without re-analysis. The Weibull shape interval of 2.20–2.56 and the Kaplan–Meier interval of 0.7847–0.8156 should both be regarded as lower bounds on the true uncertainty. A block bootstrap (Künsch, 1989) or robust variance estimation clustered on adverse episodes is the minimum required correction.

VI. LIMITATIONS AND REQUIREMENTS FOR FURTHER WORK

The limitations below are stated at length because several of them bear directly on how the results in Section 4 should be read. They are ordered by severity.

6.1 Limitations of the present design

Threshold definition. Failure is defined on a non-stationary price level using a threshold computed from the full sample. This confounds the failure classification with the price trend (Section 5.1) and introduces look-ahead bias, since the threshold could not have been known in real time. The design is descriptive of the historical sample and is not implementable as a forward-looking risk instrument.

Tautological summary quantities. The failure proportion (20.00%), the terminal reliability estimate (0.8000) and the terminal cumulative hazard (0.223) are determined by the threshold rather than estimated from the data (Sections 4.2-4.4).

Non-independence of failure days. No declustering was applied. The 513 failure days correspond to an unknown but substantially smaller number of independent adverse episodes, so the effective sample size is overstated and all reported intervals are too narrow (Sections 3.3, 5.4).

The parametric models were fitted on the wrong time scale. Section 3.3 specifies inter-failure times, of which there are at most 512 and all uncensored, but the models in Sections 4.5 to 4.8 were fitted to the day-level dataset of 2,565 records with the trading-day index as survival time. The parametric and non-parametric analyses therefore do not share a unit of analysis, and the reported scale parameter and MTTF do not describe intervals between adverse episodes.

Absence of absolute goodness-of-fit assessment. Model selection rested on relative criteria. The comparison in Section 4.9 shows that the selected Weibull model implies a cumulative failure probability more than double the non-parametric estimate.

Renewal assumption untested. Fitting independent lifetime distributions to successive inter-arrival times from a single continuously observed system assumes renewal (Ascher & Feingold, 1984; Rigdon & Basu, 2000). This was neither stated nor tested (Section 5.2).

MTTF extrapolation. Both MTTF estimates lie beyond the observation window and are not identified by the data (Section 4.8).

Single asset, single threshold, no covariates. The analysis covers one asset at one threshold, with no sensitivity analysis and no macroeconomic covariates. Inflation, exchange rates, crude oil prices, interest rates and geopolitical events are excluded, so the estimates reflect historical price behaviour alone.

Trading-day time scale. Inter-failure time is measured in trading days, so a gap of one may span one calendar day or three across a weekend, and longer across exchange holidays. The resulting distortion is unquantified.

Currency perspective. The series is denominated in US dollars, so the risk quantified here is that faced by a dollar-based investor. An investor holding gold in Indian Rupees, or in any currency other than the dollar, is additionally exposed to exchange-rate movements that this analysis does not capture, and for whom the timing of adverse episodes may differ materially.

Futures rather than spot prices. The series is a continuous futures series assembled across contract months, so the level embeds roll effects and a cost-of-carry premium over spot (Section 3.1). A threshold placed on the lower tail is unlikely to be sensitive to this, but the point is untested.

Stationarity not formally assessed. No unit-root or stationarity tests were conducted (Section 4.1), so the characterisation of the price level as trending is descriptive rather than tested.

6.2 Redesign required for interpretable risk estimates

- Resolve the identity and units of the price series.
- Redefine failure on a stationary quantity: a daily return below its conditional quantile, a drawdown exceeding a specified percentage from a running maximum, or an exceedance in GARCH-filtered residuals (McNeil & Frey, 2000). Justify the choice economically.
- Use an expanding- or rolling-window threshold to eliminate look-ahead bias.
- Decluster contiguous below-threshold days into episodes under an explicit rule (Ferro & Segers, 2003), and report the number of independent episodes as the effective sample size.
- Report stationarity diagnostics (Dickey & Fuller, 1979; Phillips and Perron, 1988; Kwiatkowski et al., 1992) and a time-series plot with the threshold marked.

- Test the renewal assumption using the Laplace and Lewis–Robinson tests; if trend in intensity is present, fit a non-homogeneous Poisson process rather than a renewal model. If covariates are introduced, use Andersen–Gill, PWP or WLW formulations with robust variance (Andersen & Gill, 1982; Prentice et al., 1981; Wei et al., 1989).
- Report absolute goodness of fit: Weibull probability plot, Cox–Snell residuals, and the Kaplan–Meier-versus-parametric overlay.
- Replace MTTF with restricted mean survival time over a horizon supported by the data, with interval estimates for all summary quantities.
- Conduct sensitivity analysis across thresholds (5th, 10th, 15th, 25th percentiles) and across declustering rules.
- Benchmark against Value-at-Risk, Expected Shortfall and a POT/GPD model, and demonstrate what the reliability framing adds.

6.3 Priority for future work

Of the above, item 10 is the most consequential for the standing of the approach. The methodological case for a reliability framing of commodity downside risk rests on its supplying interpretable quantities that established methods do not. Until that is demonstrated against a declustered peaks-over-threshold benchmark on the same data, the case remains argued rather than shown.

VII. CONCLUSION

This study set out to apply reliability theory to the analysis of downside risk in gold prices, treating adverse price movements as failure events and analysing their timing with survival methods. Daily settlement prices of COMEX gold futures in US dollars per troy ounce, spanning January 2014 to January 2025, were used, with failure defined as a price below the twentieth percentile of the observed distribution.

The Kaplan–Meier reliability estimate was 0.8000 (95% CI 0.7847–0.8156) with a terminal cumulative hazard of 0.223. Among the parametric models, the Weibull was decisively preferred to the Exponential by likelihood ratio test (LR = 350.55, 1 df, $p < 0.001$) and by AIC and BIC, with an estimated shape parameter of 2.38 and scale parameter of 3289, and an MTTF of 2,918 trading days.

These figures should not, however, be read as risk estimates for gold. The reliability and cumulative hazard values are fixed by the threshold definition rather than estimated from the data; the fitted Weibull model implies a cumulative failure probability of 42.5% at the end of the observation window against a non-parametric estimate of 20.0%, so relative model selection has not produced an adequately fitting model; and the reported intervals assume an independence that daily gold prices do not exhibit. The parametric models were also fitted on the trading-day index rather than on inter-failure times, so they and the non-parametric estimator do not share a unit of analysis, a mismatch that must be corrected by refitting before the parametric results can carry a reliability interpretation.

What the study does establish is the shape of the mapping between reliability theory and commodity risk, and the specific conditions under which that mapping fails. A fixed full-sample percentile threshold on a trending price level cannot support the reliability interpretation; contiguous below-threshold days are not independent episodes; and a single continuously observed asset is a recurrent-event system rather than a population of non-repairable components. Section 6.2 sets out the redesign these conclusions imply, centred on a stationary failure definition, a real-time threshold, explicit declustering, point-process methods, absolute goodness-of-fit assessment, and benchmarking against extreme value methods.

Reliability theory offers a genuinely useful vocabulary for the timing and accumulation of downside risk, and quantities such as the reliability function, the cumulative hazard and the expected interval between adverse episodes are readily interpretable by investors and portfolio managers. Establishing that this vocabulary adds to what extreme value and duration models already provide is the task that remains.

REFERENCES

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics, 6*(4), 701–726.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics, 10*(4), 1100–1120.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance, 9*(3), 203–228.

Ascher, H., & Feingold, H. (1984). *Repairable systems reliability: Modeling, inference, misconceptions and their causes*. Marcel Dekker.

Balkema, A. A., & de Haan, L. (1974). Residual life time at great age. *The Annals of Probability, 2*(5), 792–804.

Batten, J. A., Ciner, C., & Lucey, B. M. (2010). The macroeconomic determinants of volatility in precious metals markets. *Resources Policy, 35*(2), 65–71.

Baur, D. G., & Lucey, B. M. (2010). Is gold a hedge or a safe haven? An analysis of stocks, bonds and gold. *The Financial Review, 45*(2), 217–229.

Baur, D. G., & McDermott, T. K. (2010). Is gold a safe haven? International evidence. *Journal of Banking & Finance, 34*(8), 1886–1898. https://doi.org/10.1016/j.jbankfin.2009.12.008

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

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

Brown, S. J., Goetzmann, W. N., & Ross, S. A. (1995). Survival. *The Journal of Finance, 50*(3), 853–873. https://doi.org/10.1111/j.1540-6261.1995.tb04039.x

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.

Capie, F., Mills, T. C., & Wood, G. (2005). Gold as a hedge against the dollar. *Journal of International Financial Markets, Institutions and Money, 15*(4), 343–352.

Ciner, C., Gurdgiev, C., & Lucey, B. M. (2013). Hedges and safe havens: An examination of stocks, bonds, gold, oil and exchange rates. *International Review of Financial Analysis, 29*, 202–211.
Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer.
Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B, 34*(2), 187–220.
Cox, D. R., & Lewis, P. A. W. (1966). *The statistical analysis of series of events*. Methuen.
Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B, 30*(2), 248–275.
Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society, Series B, 52*(3), 393–442.
Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software, 64*(4), 1–34.
Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431.
Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Springer.
Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987–1007.
Engle, R. F., & Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica, 66*(5), 1127–1162.
Ferro, C. A. T., & Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society, Series B, 65*(2), 545–556.
Gao, F., & He, X. (2021). Survival analysis: Theory and application in finance. In C.-F. Lee & J. C. Lee (Eds.), *Handbook of financial econometrics, mathematics, statistics, and machine learning* (pp. 1–28). World Scientific. https://doi.org/10.1142/9789811202391_0016
Greenwood, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects, 33*, 1–26. His Majesty's Stationery Office.
Hillier, D., Draper, P., & Faff, R. (2006). Do precious metals shine? An investment perspective. *Financial Analysts Journal, 62*(2), 98–106.
Jorion, P. (2007). *Value at risk: The new benchmark for managing financial risk* (3rd ed.). McGraw-Hill.
Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Wiley.
Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452
Kassambara, A., & Kosinski, M. (2024). *survminer: Drawing survival curves using ggplot2*. R package.
Kristjanpoller, W., & Minutolo, M. C. (2015). Gold price volatility: A forecasting approach using the artificial neural network–GARCH model. *Expert Systems with Applications, 42*(20), 7245–7251.
Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics, 17*(3), 1217–1241.
Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54*(1–3), 159–178.
Lane, W. R., Looney, S. W., & Wansley, J. W. (1986). An application of the Cox proportional hazards model to bank failure. *Journal of Banking & Finance, 10*(4), 511–531.
Lawless, J. F. (2003). *Statistical models and methods for lifetime data* (2nd ed.). Wiley.
Lewis, P. A. W., & Robinson, D. W. (1974). Testing for a monotone trend in a modulated renewal process. In F. Proschan & R. J. Serfling (Eds.), *Reliability and biometry* (pp. 163–182). SIAM.
Livieris, I. E., Pintelas, E., & Pintelas, P. (2020). A CNN–LSTM model for gold price time-series forecasting. *Neural Computing and Applications, 32*, 17351–17360.
McNeil, A. J., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance, 7*(3–4), 271–300.
Meeker, W. Q., & Escobar, L. A. (1998). *Statistical methods for reliability data*. Wiley.
Mills, T. C. (2004). Statistical analysis of daily gold price data. *Physica A: Statistical Mechanics and Its Applications, 338*(3–4), 559–566. https://doi.org/10.1016/j.physa.2004.02.003
Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics, 14*(4), 945–966.
O'Connor, F. A., Lucey, B. M., Batten, J. A., & Baur, D. G. (2015). The financial economics of gold – A survey. *International Review of Financial Analysis, 41*, 186–205. https://doi.org/10.1016/j.irfa.2015.07.005
Phillips, P. C. B., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika, 75*(2), 335–346.
Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics, 3*(1), 119–131.
Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika, 68*(2), 373–379.
R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Rigdon, S. E., & Basu, A. P. (2000). *Statistical methods for the reliability of repairable systems*. Wiley.
Rinne, H. (2009). *The Weibull distribution: A handbook*. CRC Press.
Royston, P., & Parmar, M. K. B. (2013). Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology, 13*, 152.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/AOS/1176344136
Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business, 74*(1), 101–124.
Sims, C. A. (1980). Macroeconomics and reality. *Econometrica, 48*(1), 1–48.
Singpurwalla, N. D. (2006). Reliability and survival in econometrics and finance. In *Reliability and risk: A Bayesian perspective* (pp. 251–272). John Wiley & Sons.
Therneau, T. M. (2024). *A package for survival analysis in R*. R package. https://CRAN.R-project.org/package=survival
Tully, E., & Lucey, B. M. (2007). A power GARCH examination of the gold market. *Research in International Business and Finance, 21*(2), 316–325.
Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., & Wei, L. J. (2014). Moving beyond the hazard ratio in quantifying the clinical benefit–risk profile of therapeutic trials. *Journal of Clinical Oncology, 32*(22), 2380–2385.
Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association, 84*(408), 1065–1073.