Title page:

Can dynamic Bayesian networks improve daily expected shortfall forecasts? Evidence from the S&P 500 index

Names and institutional affiliations of the authors:

Eden Gross

School of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, South Africa

eden.gross@wits.ac.za

ORCiD: 0000-0002-7647-8890

Ryan Kruger

Department of Finance and Tax, University of Cape Town, Cape Town, South Africa

ryan.kruger@uct.ac.za

ORCiD: 0009-0000-3185-1789

Francois Toerien

Department of Finance and Tax, University of Cape Town, Cape Town, South Africa

francois.toerien@uct.ac.za

ORCiD: 0000-0002-6051-1094

CAN DYNAMIC BAYESIAN NETWORKS IMPROVE DAILY EXPECTED SHORTFALL FORECASTS? EVIDENCE FROM THE S&P 500 INDEX

ABSTRACT

Expected shortfall (ES) is a central tool for measuring and managing extreme losses at banks, asset managers, hedge funds, and other market participants. We tested whether dynamic Bayesian networks (BNs) improve daily 97.5% ES forecasts relative to six traditional models, using S&P 500 index returns from 7 August 2007 to 30 June 2026 and five BN structure-learning algorithms across three calibration periods (252, 504, and 1,008 trading days). Forecasts are assessed using the traffic light, conditional, unconditional, minimally biased, and Du-Escanciano backtests. Calibration period length materially shapes model performance. The ARCH(1) model is significantly more accurate at the shortest calibration period, but significantly less accurate at the two longer periods, a reversal not observed for any other model. The EGARCH(1,1) model under the normal distribution is the most accurate forecasting model across all calibration periods, while the skewed Student's t distribution and empirical distribution models consistently underperform the normal distribution. Although the traffic light test results are all favorable, all models fail the remaining statistical backtests. Overall, we conclude that distribution and calibration period length specifications significantly influence model performance, and the incorporation of BN-generated forecasts alone into market risk management, devoid of any other risk management information offered by the BN, does not improve ES forecasting.

Keywords:

Market risk forecasting; Tail risk, Autoregressive models, Backtesting; Machine learning

JEL codes:

G17, G21, G23, G32, C51, C52, C53

Statements and Declarations:

Competing Interests and Funding

The authors declare that they are not aware of any competing interests that exist with regards to this research. This research did not receive any specific funding from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of the use of generative artificial intelligence in scientific writing

During the preparation of this work, the author(s) used Copilot and Claude to improve the flow and grammar of the manuscript and to improve some of the methodology employed. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

1. Introduction

Extreme losses remain a central concern for financial institutions, asset managers, hedge funds, financial regulators, and other market participants. Among the various measures used to quantify market risk, expected shortfall (ES) has received increasing attention due to its ability to capture tail losses beyond a specified threshold, thereby providing information that is not available from value at risk (VaR) alone (Acerbi & Tasche, 2002; Yamai & Yoshiba, 2005). Therefore, ES has become an important tool for both risk measurement and decision-making in financial markets, with applications extending well beyond regulatory capital frameworks.

Despite its widespread use, accurately forecasting ES remains challenging. The literature on ES forecasting is smaller than the corresponding VaR literature, and frequently focuses on specialized approaches incorporating extreme value theory, filtering techniques, and quantile-based methods (Novales & Garcia-Jorcano, 2019; Storti & Wang, 2022). At the same time, recent studies have emphasized that forecasting accuracy alone may not be sufficient when evaluating ES models. Zhao (2026), for example, argues that model robustness and performance during periods of market stress should also be considered, while Souffargi and Boubaker (2025) demonstrate that structural breaks and regime changes can materially impact VaR and ES forecasts. These findings suggest that both model specification and estimation design may play important roles in evaluating ES forecasting performance.

Advances in machine learning have created new opportunities to incorporate forward-looking information into financial forecasting models. Bayesian networks (BNs) provide one such method. By modeling dependency structures among variables, BNs are capable of representing complex causal relationships, allowing them to generate probabilistic forecasts. Recent applications have demonstrated their usefulness in areas such as systemic risk modeling (Chan, Chu, & So, 2023), credit risk prediction, and causal machine learning (Liu, Zhang, & Xiong, 2024). These characteristics suggest that BNs may provide valuable information for market risk forecasting, particularly in environments where relationships between financial variables evolve over time.

Within market risk forecasting, however, empirical applications of BNs remain comparatively limited. Early studies by Shenoy and Shenoy (2000) and Demirer, Mau, and Shenoy (2006) proposed the use of BNs in portfolio and market risk management. More recently, Apps (2020) demonstrated how BN-generated forecasts could be incorporated into VaR estimation, while Gross, Kruger, and Toerien (2025) extended this framework to forecast 10-day 99% VaR and

stressed VaR in a regulatory banking context. Evidence regarding the usefulness of BNs for forecasting ES remains comparatively sparse, creating an opportunity to evaluate whether forward-looking BN forecasts improve tail risk estimation relative to traditional approaches.

This study investigates whether dynamic BNs (referred to simply as BNs in this study) can improve one-day 97.5% ES forecasts. Using the Standard & Poor's (S&P) 500 index, we compare five BN structure-learning algorithms with several traditional ES forecasting models across multiple calibration period lengths and alternative distributional assumptions. In addition to evaluating BN-based forecasts, we assess the importance of calibration period choices and distributional assumptions in determining forecasting model performance. This broader perspective allows us to examine not only whether BN-generated forecasts improve ES estimation, but also how forecasting performance is influenced by key modeling choices recently highlighted in the ES forecasting literature.

This study contributes to the literature in three ways. First, it provides an extensive comparison of traditional ES forecasting models across multiple calibration periods (with out-of-sample periods ranging between 3,242 and 3,998 trading days in length) and distributional assumptions, contributing to the growing literature on robustness in ES forecasting by showing how calibration period length and distributional choice (and their combination) affect forecast performance. Second, it extends the application of BNs from VaR forecasting to ES forecasting and evaluates five alternative BN structure-learning algorithms. Third, it provides new evidence regarding the interaction between forward-looking machine learning forecasts and traditional tail risk estimation techniques, offering insights relevant to both academics developing forecasting methodologies and practitioners responsible for market risk measurement and management.

The remainder of this article is structured as follows. Section 2 provides a literature review and the technical context underpinning this paper. Section 3 then discusses the data and methodology, while Section 4 discusses and analyses our results. Section 5 concludes this paper.

2. Literature Review and Methodological Framework

ES is often regarded as a substitute for, or complement to, VaR, since VaR does not quantify the selected quantile of the profit and loss distribution. This limitation is one of the primary motivations behind the transition from VaR-based to ES-based market risk measurement frameworks mandated by the Basel Committee on Banking Supervision (BCBS) under its

Fundamental Review of the Trading Book (FRTB) (Basel Committee on Banking Supervision, 2019a). Moreover, ES is a coherent risk measure, leading many authors to regard it as theoretically preferable to VaR, which does not satisfy the full set of coherence properties (see, for example, Acerbi and Tasche, 2002). The adoption of ES under the FRTB framework reflects the growing regulatory emphasis on the accurate measurement and management of extreme tail risk.

ES, also known as conditional VaR or expected tail loss, measures the average loss conditional on losses exceeding the corresponding VaR threshold (Yamai & Yoshiba, 2005). Therefore, the calculation of ES requires that we first calculate VaR. Let X be the random variable representing the profit and loss distribution of a bank's equities trading desk. We define the h-day α-quantile profit and loss value at time t to be $x_{ht,\alpha}$, such that

$$\Pr\left[X_{ht} < x_{ht,\alpha}\right] < \alpha \tag{1}$$

The corresponding ES figure is then given by the expected loss condition on losses exceeding this threshold. Following Acerbi and Tasche (2002), ES at confidence level $100\alpha\%$ is then

$$ES^{(\alpha)}(X) = -\frac{1}{\alpha}\left[E\left[X\mathbf{1}_{\{X\leq x^{(\alpha)}\}}\right] - x^{(\alpha)}\left(F\left(x^{(\alpha)}\right) - \alpha\right)\right] \tag{2}$$

where $\mathbf{1}_{\{X\leq x^{(\alpha)}\}}$ is an indicator variable equal to 1 if $X \leq x^{(\alpha)}$ or 0 if $X > x^{(\alpha)}$; and $x^{(\alpha)} = \sup\{x|\ \Pr[X \leq x\} \leq \alpha\}$.

We consider several widely used models to forecast ES and refer to these as traditional models. These include the historical simulation, the delta-normal, the autoregressive conditional heteroscedasticity (ARCH(1)), the generalized autoregressive conditional heteroscedasticity (GARCH(1,1)), the exponential generalized autoregressive conditional heteroscedasticity (EGARCH(1,1)), and the RiskMetrics models.

The historical simulation model estimates tail risk metrics directly from the empirical distribution of historical returns and, therefore, makes no explicit distributional assumptions for the return distribution. In contrast, the delta-normal model assumes that returns are normally distributed and estimates tail risk metrics using the mean and variance of historical returns. Owing to their relative computational simplicity and widespread use in practice (Berkowitz & O'Brien, 2002; Pérignon & Smith, 2010), both models remain common benchmark approaches in market risk forecasting, with the historical simulation model, in particular, continuing to be favored despite its documented limitations (García-Risueño, 2025). Despite the growth of

machine learning approaches for forecasting tail risk, research into autoregressive models remains highly active, with recent studies continuing to develop new volatility specifications for ES forecasting (Wang & Gerlach, 2024), a point we return to in Section 4.

The ARCH(1) model captures volatility clustering by modeling current volatility as a function of past squared innovations, and is defined as follows.

$$\sigma_{t+1}^2 = \omega + \alpha\varepsilon_t^2, \qquad \omega > 0, \qquad \alpha \geq 0, \qquad \varepsilon_t = \sigma_t z_t \tag{3}$$

where σ_t denotes return volatility and z_t is a series of independently and identically distributed standard normal random variables (Bollerslev, 2007). The GARCH(1,1) model extends the ARCH framework by incorporating both lagged volatility and lagged squared residuals, thereby allowing volatility shocks to persist through time (Engle, 2001). The model is specified as follows.

$$\sigma_{t+1}^2 = \omega + \alpha\sigma_t^2\varepsilon_t^2 + \beta\sigma_t^2, \qquad \alpha, \beta, \omega > 0, \qquad \alpha + \beta < 1 \tag{4}$$

where ε_t is the residual term and $\sqrt{\omega/[1-(\alpha+\beta)]}$ represents the long-term variance of the process. The EGARCH(1,1) model additionally accounts for leverage effects, allowing positive and negative shocks to have asymmetric effects on volatility, with models these leverage effects outperforming symmetric specifications (Wang & Gerlach, 2024). The EGARCH(1,1) model is specified as follows.

$$\ln(\sigma_{t+1}^2) = \omega + \beta\ln(\sigma_t^2) + \alpha\left(\frac{|\varepsilon_t|}{\sigma_t} - E\left[\frac{|\varepsilon_t|}{\sigma_t}\right]\right) + \gamma\cdot\frac{\varepsilon_t}{\sigma_t} \tag{5}$$

Finally, the RiskMetrics model estimates volatility using an exponentially weighted moving average framework, and is specified as follows.

$$\sigma_{t+1}^2 = \lambda\sigma_t^2 + (1-\lambda)r_t^2 \tag{6}$$

where $\lambda = 0.94$ is the smoothing parameter and r_t is the previous day's return.

The ARCH, GARCH, EGARCH, and RiskMetrics models were calibrated assuming either a normal or a skewed Student's t distribution as the underlying return distribution. In contrast, the historical simulation and delta-normal models are distribution agnostic. Both distributions were tested since the return series exhibits negative skewness and excess kurtosis (see Section 3), motivating a comparison between a distribution that accommodates this asymmetry and one that does not (the latter being the commonly-used underlying distribution for these models).

While there is substantial research on the performance of traditional models in estimating VaR, the literature on ES forecasting remains comparatively smaller, and often focuses on more sophisticated approaches incorporating extreme value theory, filtering techniques, and quantile-based methods (see, for example, Novales & Garcia-Jorcano, 2019, and Storti & Wang, 2022). The literature evaluating the ES forecasting performance of benchmark models such as the historical simulation, RiskMetrics, and GARCH-family models remains comparatively limited.

The regulatory adoption of ES under the FRTB framework has renewed interest in both ES forecasting and ES forecasting evaluation (Grajales & Medina Hurtado, 2023). Recent studies have increasingly examined the performance and robustness of ES forecasting methods under changing market conditions. Zhao (2026), for example, argues that the evaluation of ES forecasts should extend beyond forecasting accuracy alone and should consider robustness, prudence, and performance during conditions of market stress. Similarly, Souffargi and Boubaker (2025) demonstrate that structural breaks and regime changes can materially affect VaR and ES forecasts, suggesting that the quantity of historical information used in model estimation and, therefore, the length of the calibration period, may influence forecasting quality.

Žiković and Filer (2013) compare the performance of various models when calculating VaR and ES for eight developed and eight emerging market equity indices over a ten-year period. Although they consider a broader range of models than those examined in our study, they find the historical simulation and RiskMetrics models to be among the worst-performing approaches for both VaR and ES estimation.

García-Risueño (2025) further highlight the shortcomings of the historical simulation model for ES estimation. The study fitted return data to several heavy-tailed distributions, including the Student's t distribution (although not its skewed variant), and showed that such distributions capture tail risk more effectively than the normal distribution. Using synthetic returns generated from these fitted distributions, García-Risueño (2025) demonstrates that the historical simulation model systematically underestimates ES relative to parametric alternatives and, consequently, strongly argues for a more limited regulatory reliance on this model.

Backtesting is a key component of market risk model evaluation, whereby model forecasts are compared with subsequently observed returns to assess forecasting accuracy (Basel Committee on Banking Supervision, 2019b). A breach (also referred to as an exceedance or an exception)

occurs when the observed loss exceeds the forecasted risk measure, and model performance can then be evaluated by comparing the observed number of breaches with the expected number. A challenge when evaluating ES forecasts is that ES is not individually elicitable in the same manner as VaR (Gneiting, 2011), making the comparison and ranking of competing ES models more difficult. In fact, Diebold, Gunther, and Tay (1998) conclude that ES cannot be directly backtested. However, subsequent studies, such as that by Acerbi and Székely (2014), have developed backtests that either avoid reliance on elicitability or exploit the joint elicitability of VaR and ES.

One of the most widely used regulatory approaches for evaluating ES forecasts is the BCBS traffic light test (Chen, 2018). The test uses a cumulative probability framework to classify models into 'Green zone', 'Yellow zone', or 'Red zone' based on their observed performance. In addition to this test, we also employ three theoretically motivated backtests: the conditional backtest (Acerbi & Székely, 2014), the minimally biased backtest (Acerbi & Székely, 2017), and the Du-Escanciano independence test (Du & Escanciano, 2017).

The backtests proposed by Acerbi and Székely (2014) assume that the independence of ES breaches has been assessed separately from the backtesting procedure itself (Acerbi & Székely, 2014). He, Kou, and Peng (2022) further classify these procedures as indirect ES backtests[1], meaning that they evaluate properties associated with ES rather than the risk measure directly. They argue that the effectiveness of such indirect approaches may deteriorate as portfolio size increases, potentially creating opportunities for the under-reporting of risk (He, et al., 2022). These limitations should, therefore, be borne in mind when interpreting the results of ES backtests.

Given these persistent challenges in evaluating ES forecasts, attention has increasingly turned to machine learning methods capable of incorporating forward-looking information into the forecasting process itself. BNs have become increasingly established within financial risk management applications. Recent studies have successfully applied BNs to areas such as credit

[1] A backtest is said to be indirect if it either: (a) Examines whether the distribution (in full or in part, e.g., the tail) or its properties correspond to the quantities of the true, yet unknown, distribution; or (b) Examines the backtestability of a collection of risk measures, which are elicitable collectively (He, et al., 2022). A direct alternative is the comparative graphical approach of Ziegel, Krüger, Jordan, and Fasciati (2020), who use Murphy diagrams to test whether one ES forecast dominates another under a class of elicitable scoring functions, rather than assessing indirect properties of the forecasting distribution. We do not employ this approach here.

risk prediction, default inference, and causal machine learning (Liu, et al., 2024), while rolling training period BN frameworks have been used to assess systemic risk and evolving interdependencies within financial markets (Chan, et al., 2023). These studies highlight not only the flexibility of BN methodologies in modeling complex, non-linear, and potentially causal relationships in financial systems, but also the importance of allowing network structures to evolve through time as market conditions change. Such adaptability may be particularly valuable in financial markets, where relationships between economic and financial variables are rarely static through time, displaying temporal elements. However, the application of BNs to market risk forecasting remains comparatively under-developed, particularly with respect to modeling evolving causal relationships and temporal dynamics, suggesting the potential value of rolling calibration period approaches in such applications.

Early studies by Shenoy and Shenoy (2000) and Demirer, et al., (2006) proposed the use of BNs in portfolio and market risk management, with the latter outlining a top-down approach to specifying BNs. More recently, Apps (2020) demonstrated how a BN could be used to incorporate directional return forecasts into VaR estimation under multivariate normality assumptions. Apps (2020) makes key methodological choices that can be abstracted and improved upon. First, Apps (2020) forecasts directional return, i.e., profit or loss, and uses this to forecast VaR, instead of using the BN to produce a point estimate return forecast (as opposed to a binary variable of profit or loss). Second, Apps (2020) imposes a parametric return distribution (the use of multivariate normality) as opposed to using the observed returns to calibrate an empirical return distribution. In the present study, we abstract the methodology of Apps (2020) to account for both improvements (i.e., allowing for the use of the empirical return distribution by each structure-learning algorithm to produce actual return forecasts), thereby generalizing the methodology presented by earlier studies and extending the literature by examining whether BNs can improve the forecasting of one-day 97.5% ES relative to a range of traditional risk models. Despite these developments, empirical applications of BNs to the direct forecasting of market risk measures remains relatively sparse.

Collectively, the literature highlights the need for further evidence on BN-based ES forecasting and motivates our present research, which evaluates the performance of BNs in forecasting one-day 97.5% ES and examines the sensitivity of those forecasts to alternative calibration period lengths across a range of traditional models and structure-learning algorithms.

3. Data and Methodology

We use daily closing values of the S&P 500 index from 7 August 2007 to 30 June 2026, obtained from the Bloomberg database, to forecast one-day 97.5% ES forecasts using ten traditional models and five BNs. The S&P 500 index is used as a proxy for the equities market, as done by many others (see, for example, Grinold, 1992), as it tracks a broad portfolio of large capitalization US equities. The study period spans several market regimes, including the 2008 Global Financial Crisis, the European sovereign debt crisis, the COVID-19 market shock, the subsequent inflationary environment and the monetary policy tightening that followed, and the subsequent elevated market volatility observed in the US. Hence, the study provides a robust setting in which to evaluate the performance of alternative ES forecasting models under a wide range of market conditions.

To examine the sensitivity of ES forecasts to the quantity of historical information used in estimation, we employ multiple calibration period lengths. Specifically, we evaluate calibration periods of 252, 504, and 1,008 trading days when constructing ES forecasts, corresponding approximately to one, two, and four years of daily market data, respectively. These, in turn, yielded out-of-sample periods of 3,998, 3,746, and 3,242 days, respectively. For the BNs, a 504-day (two-trading-year) rolling training period was used to learn and update the network structures through time, chosen to ensure adequate coverage of lower-frequency macroeconomic variables within each training window. The use of rolling calibration periods allows both BNs and ES forecasting models to adapt to changing market conditions and evolving relationships between financial variables (Chan, et al., 2023; Souffargi & Boubaker, 2025), while facilitating a direct comparison of the effects of calibration period length on forecast performance.

We proxy the profit and loss account of the equity portfolio using the daily logarithmic returns of the S&P 500 index. These returns were calculated as follows.

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \tag{7}$$

where $\ln(\cdot)$ is the natural logarithm, and P_t is the index's level at time t.

Descriptive statistics for the daily returns indicate a mean return of 0.03%, a standard deviation of 1.26%, and minimum and maximum daily returns of -12.77% and 10.96%, respectively. The return distribution exhibits a skewness coefficient of -0.4656, indicating a longer left tail, consistent with the tendency for downside market moves to be more extreme than

corresponding upside moves, and an excess kurtosis of 12.3723, indicating substantially heavier tails than the normal distribution.

To generate BN-based return forecasts, we employ five structure-learning algorithms: the max-min hill climbing (MMHC) algorithm (Tsamardinos, Brown, & Aliferis, 2006), the genetic (natural encoding particle swarm optimization, or natPSOHO) algorithm[2] (Quesada, Bielza, & Larrañaga, 2021), the Peter and Clark (Stable) (PC (Stable)) algorithm, (Spirtes & Glymour, 1991), the semi-interleaved HITON parents-children (SI-HITON-PC) algorithm (Aliferis, Tsamardinos, & Statnikov, 2003), and the Incremental Association Markov Blanket (IAMB) algorithm (Tsamardinos, Aliferis, & Statnikov, 2003). These algorithms were selected to provide a diverse set of constraint-based, hybrid, and metaheuristic approaches to BN structure learning. Beyond BN-based market risk applications, the MMHC algorithm has established precedent in financial settings, including credit risk scoring (Leong, 2016), constraint-based causal discovery methods related to the PC (Stable) algorithm have recently been applied to factor investing and portfolio diversification (Sadeghi, Gopal, & Fesanghary, 2024), while genetic and particle-swarm-based optimization methods are well established in portfolio optimization more broadly (see, for example, Cura, 2009, and Chang, Yang, and Chang, 2009), even where not applied to BN structure-learning specifically. Precedent for the SI-HITON-PC and the IAMB algorithms in financial applications is comparatively sparse.

The five algorithms were used to forecast the closing values of the S&P 500 index over the BN return out-of-sample period, which were then used to generate one-day 97.5% ES forecasts over the various out-of-sample periods discussed earlier. Network structures were learned using a 504-day rolling training period and were re-estimated throughout the sample daily to allow the learned relationships between variables to evolve over time. For each training iteration, the preferred network structure was selected as the structure that minimized the Akaike Information Criterion (AIC). To facilitate network estimation and forecasting, macroeconomic and financial time series data were collected from 7 August 2007 onward, although the availability of individual variables differed across time series, with the first S&P 500 index forecast being 6 August 2009, i.e., approximately two calendar years after the start of the

[2] During the BN return forecasting process, the natPSOHO algorithm failed to produce a forecast on one trading day when using the 504-day training period. To reflect practical risk management procedures, we carried forward the preceding day's forecast. This affected a single observation and had no material impact on the study's results.

training period. These forecasts were subsequently converted into one-day-ahead return forecasts and incorporated into the return distribution used to calculate ES.

The BNs were trained using financial time series data for 39 macroeconomic and financial variables (detailed in Appendix A). The final predictor set was obtained from an initial collection of approximately 46 variables after removing duplicate and near-target-proxy variables during the data cleaning process. Specifically, one duplicated liquidity variable, five variables exhibiting near-perfect correlations with the S&P 500 index, and one near-duplicate inflation-protected securities variable were removed to reduce redundancy and limit information leakage from variables that closely tracked the target series. These variables were identified through the literature or expert judgement as potentially causally related to the S&P 500 index. Our approach follows the top-down framework proposed by Demirer, et al., (2006). We used the dbnR package, among other resources, to construct and learn the network structures associated with the various BN algorithms. Since structure-learning algorithms require complete datasets (i.e., no missing values), non-daily variables were converted to daily frequency by carrying forward the most recently observed value. This approach reflects practical risk management settings, where only the latest available observation of a variable can be incorporated into forecasting models until a newer value becomes available. Prior to training the BNs, the predictor dataset was screened for variables that could adversely affect the structure-learning process. Specifically, variables exhibiting zero variance within any iteration of the rolling training period (due to carried forward values or otherwise) were removed from that iteration, as they contained no information for learning causal relationships. Moreover, highly collinear variables were filtered using pairwise correlations, with one variable removed whenever the absolute correlation exceeded 0.98 in any given iteration. These procedures were applied dynamically within each iteration of network learning to account for changes in data characteristics through time and to allow the most informative structure to be learned at each iteration.

The BN forecasting procedure utilized the initial 504-day training period, resulting in the first one-day-ahead S&P 500 index closing value forecast being produced on 6 August 2009. These forecasts were converted into forecasted returns using Equation (7). For each calibration period length considered (205, 504, and 1,008 trading days), the BN forecasted return was incorporated into the rolling return distribution as the most recent observation. The resulting distribution was used to produce the one-day 97.5% ES forecasts. Hence, BN-based ES

forecasts differ from those of the traditional models through the inclusion of a forward-looking return estimate within the calibration period.

The traditional models, namely the historical simulation, delta-normal, ARCH(1), GARCH(1,1), EGARCH(1,1), and RiskMetrics models, were calibrated using rolling calibration periods of 252, 504, and 1,008 trading days to generate one-day 97.5% ES forecasts. This produced three sets of forecasts for each model, allowing us to examine the sensitivity of forecast performance to the quantity of historical information used in estimation. For the autoregressive models, ES forecasts were calculated under both the normal and skewed Student's t return distributions, with model parameters estimated by maximum likelihood. In contrast, the historical simulation and delta-normal models are distribution agnostic and, therefore, do not require the specification of an underlying return distribution. Hence, we tested a total of 30 traditional models (three calibration period lengths, with four of the six traditional models being calibrated using two distributions).

We employ several backtesting procedures to assess the statistical accuracy of both the traditional and BN ES forecasts. These include the BCBS regulatory traffic light test, the conditional and unconditional backtests of Acerbi and Székely (2014), the minimally biased backtest of Acerbi and Székely (2017), and the Du-Escanciano backtest (Du & Escanciano, 2017). Collectively, these backtests assess forecast calibration, tail-loss estimation, and the temporal behavior of ES breaches from both regulatory and statistical perspectives. We describe each test in turn, below, beginning with the conditional, unconditional, and minimally biased backtests, before turning to the Du-Escanciano independence test.

The conditional, unconditional and minimally biased backtests of Acerbi and Székely (2014; 2017) evaluate whether the ES forecasts produced by a model correspond to the true, but unobservable, ES values. Under the backtests' null hypothesis, the observed ES forecasts are correctly specified, while the alternative hypothesis states that the forecasts underestimate the true level of tail risk (Acerbi & Székely, 2014; 2017). The conditional backtest can only be applied when at least one corresponding VaR breach is observed, whereas the unconditional and minimally biased backtests do not require this condition (Acerbi & Székely, 2014; 2017).

The conditional backtest (Acerbi & Székely, 2014) rests on the assumption that ES forecasts are produced and tested after their corresponding VaR forecasts have been produced and tested. The following expectation can then be derived.

$$E\left[\frac{X}{ES_{ht,\alpha}} + 1 \middle| X + VaR_{ht,\alpha} < 0\right] = 0 \tag{8}$$

Since the null hypothesis states that $ES_{ht,\alpha} = \widehat{ES}_{ht,\alpha}$ (Acerbi & Székely, 2014), the test statistic, is as follows.

$$\bar{Z}_{CB}(X) = \frac{1}{N_{Breaches}} \sum_{t=1}^{N} \frac{X_t \mathbf{1}_{\{X \leq x^{(\alpha)}\}}}{\widehat{ES}_{ht,\alpha}} + 1 \tag{9}$$

where X is a vector of profit and loss amounts; $\widehat{ES}_{ht,\alpha}$ is a vector of expected shortfall forecasts; $\mathbf{1}_{\{X \leq x^{(\alpha)}\}}$ is as defined for Equation (2); and $N_{Breaches} = \sum_{t=1}^{N} \mathbf{1}_{\{X \leq x^{(\alpha)}\}}$ (Acerbi & Székely, 2014), i.e., $N_{Breaches}$ is the number of VaR breaches observed.

Acerbi and Székely (2014) also propose an unconditional backtesting technique for ES forecasts. Subsequently, Acerbi and Székely (2017) note that this backtest exhibits a significant linear relationship to VaR forecasts (Acerbi & Székely, 2017), meaning that the incorrect VaR calibration may influence its performance and increase the probability of Type I or Type II errors. To complement the conditional and unconditional backtests, Acerbi and Székely (2017) also propose the minimally biased backtest. The test statistic, $\bar{Z}_{MB}$, is as follows.

$$\bar{Z}_{MB}(X) = \frac{1}{N} \sum_{t=1}^{N} \widehat{ES}_{ht,\alpha} - \widehat{VaR}_{ht,\alpha} + \frac{\left(\widehat{VaR}_{ht,\alpha} + \widehat{ES}_{ht,\alpha}\right)\mathbf{1}_{\{X \leq x^{(\alpha)}\}}}{\alpha} \tag{10}$$

where X and $\widehat{ES}_{ht,\alpha}$ are, again, as defined for Equation (8); $\widehat{VaR}_{ht,\alpha}$ is the corresponding VaR forecast; and $\mathbf{1}_{\{X \leq x^{(\alpha)}\}}$ is, again, as defined for Equation (2) (Acerbi & Székely, 2017).

For both backtests, under the null hypothesis, the expected values of the test statistics are zero (where, for the conditional backtest, this expectation is conditional on the corresponding number of VaR breaches being positive, i.e., $E[\bar{Z}_{CB} | N_{Breaches} > 0] = 0$), while under the alternative hypothesis, this expectation is negative (Acerbi & Székely, 2014; 2017). These expected values can be used to evaluate ES forecast performance, with negative values indicating evidence to reject the null hypothesis at the chosen confidence level.

We also employed the Du-Escanciano independence backtest for ES, as it provides a formal statistical assessment of whether ES breaches occur independently through time. This test is analogous to Christoffersen's independence test for VaR forecasts. Du and Escanciano (2017) develop a Portmanteau Box-Pierce conditional backtest for the independence of ES breaches, whose primary objective is to determine whether ES breaches exhibit clustering.

Consider the parametric distribution of the profit and loss account, conditional on some parameter $\Theta \in \mathbb{R}^p$, that unknown parameter being θ_0. Du and Escanciano (2017) show that the associated cumulative breach function, $H_t(\alpha, \theta_0)$, is as follows.

$$H_t(\alpha, \theta_0) = \frac{1}{\alpha}(\alpha - u_t(\theta_0))\mathbf{1}_{\{u_t(\theta_0) \leq \alpha\}} \tag{11}$$

where $u_t(\theta_0)$ is the conditional cumulative distribution function of the profit and loss account of a bank at time t, evaluated at the level u, conditional on the parameter θ_0 and the filtration system $\mathcal{F}_t$, the set of which is termed by Du and Escanciano as generalised errors (Du & Escanciano, 2017).

The conditional test presented in Du and Escanciano (2017) is based on such cumulative breaches, as captured in Equation (11). This conditional test essentially tests whether the terms of the series $\{H_t(\alpha) - \alpha/2\}_{i=1}^{\infty}$ are uncorrelated. The null hypothesis states that, given the filtration system $\mathcal{F}_t$, the expected value of each term of this series is zero, i.e., $E[H_t(\alpha) - \alpha/2|\mathcal{F}_t] = 0$.

Consider the autocovariance function and autocorrelation function of $H_t(\alpha)$ for lag j as $\gamma_{N,j}$ and $\rho_{N,j}$, where $\rho_j = \gamma_{N,j}/\gamma_{N,0}$ as usual. The Du-Escanciano conditional backtest statistic is, then, as follows.

$$\bar{Z}_{DE}(X, n) = N \sum_{j=1}^{n} \hat{\rho}_{T,j}^2 \tag{12}$$

where N is the number of out-of-sample trading days; n is the highest lag of autocorrelations, or 1 as default; and $\hat{\rho}_{N,j}$ is the estimate of $\rho_{N,j}$ as defined above (Du & Escanciano, 2017). Du and Escanciano (2017) show that this test statistic is asymptotically distributed as a chi-squared random variable with n degree of freedom.

The backtesting procedures described above assess the statistical accuracy of ES forecasts. However, from a risk management perspective, forecast efficiency is also important, as excessively conservative forecasts may result in unnecessarily large capital allocations. Therefore, to complement the backtesting results, we also employed three forecasting error measures: The mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). Where we observed returns equal to zero, we employed the symmetric MAPE (SMAPE), as conventional MAPE values are undefined in such cases. Lower values of these measures indicate closer agreement between forecasted and observed

outcomes and, therefore, more accurate forecasts. These measures are particularly useful when multiple models satisfy (regulatory) backtests or when statistical backtests provide limited discrimination between competing forecasting approaches.

Since multiple forecasting models are evaluated in this study, differences in forecasting error measures alone do not indicate whether one model is statistically more accurate than another. Hence, we employed the Diebold-Mariano (DM) statistic (Diebold & Mariano, 1995) to assess the relative predictive accuracy of competing forecasting models. The null hypothesis of the DM test states that the two models under comparison possess equal predictive accuracy, while the alternative hypothesis states that one model (the first one stated in this study) is more accurate than the other (the second one stated). Hence, the test provides a formal statistical assessment of whether observed differences in forecasting performance are likely to reflect genuine differences in predictive ability rather than sampling variation.

The first step in the DM test is to calculate the loss differential, d_i, together with its average, $\bar{d}_i$, and autocovariance at lag k, γ_k, defined as follows.

$$\gamma_k = \frac{1}{N}\sum_{i=k+1}^{N}\left(d_i - \bar{d}\right)\left(d_{i-k} - \bar{d}\right) \tag{13}$$

Using the autocovariance function detailed above to obtain an estimator of the variance of loss differential, the Diebold-Mariano statistic is then as follows.

$$DM = \frac{\bar{d}}{\tilde{\sigma}_{\bar{d}}} \tag{14}$$

where DM is the Diebold-Mariano statistic and $\tilde{\sigma}_{\bar{d}}$ is a consistent estimator of the standard deviation of the average loss differential (Diebold & Mariano, 1995). Under the null hypothesis, the test is asymptotically standard normally distributed. The DM statistic provides information regarding the relative forecasting performance of the two competing models (Diebold, 2015) and is, therefore, used in conjunction with the forecasting error measures discussed above rather than as a standalone measure of forecasting quality.

4. Results and Analyses

Table 1 reports the numbers of one-day 97.5% ES breaches and the corresponding BCBS traffic light classifications for each model and calibration period length. Several patterns emerge. First, the traffic light test provides limited discrimination between competing models. Specifically, 40 of the 45 model-calibration period combinations evaluated were classified as

'Green' models, with 13 of the 15 models consistently achieving this classification across all calibration period lengths. The sole exception is the RiskMetrics model under the skewed Student's t distribution, which consistently achieved a 'Yellow' classification, while the ARCH(1) model under the normal distribution exhibited a non-monotonic Red-Green-Red pattern as the calibration period increases from 252 to 1,008 trading days. Second, the five BN algorithms produce breach counts that are remarkably similar to the historical simulation model, differing by at most two breaches at any calibration period length and converging to the same number of breaches (34) when the 1,008-day calibration period is employed. These results suggest that, from a regulatory traffic light perspective, the ES forecasts incorporating a BN-generated return behave almost identically to those produced by the historical simulation model.

Table 1: One-day 97.5% expected shortfall breaches and BCBS traffic light test classification across calibration periods

Model	**252 days**		**504 days**		**1,008 days**	
	Breaches	**BCBS Zone**	**Breaches**	**BCBS Zone**	**Breaches**	**BCBS Zone**
Historical Simulation	*55*	*Green*	*47*	*Green*	*34*	*Green*
Delta-Normal	*101*	*Green*	*93*	*Green*	*67*	*Green*
ARCH(1) (n)	*140*	*Red*	*74*	*Green*	*138*	*Red*
GARCH(1,1) (n)	*27*	*Green*	*14*	*Green*	*5*	*Green*
EGARCH(1,1) (n)	*10*	*Green*	*13*	*Green*	*6*	*Green*
RiskMetrics (n)	*96*	*Green*	*89*	*Green*	*77*	*Green*
ARCH(1) (sst)	*70*	*Green*	*47*	*Green*	*62*	*Green*
GARCH(1,1) (sst)	*83*	*Green*	*77*	*Green*	*67*	*Green*
EGARCH(1,1) (sst)	*75*	*Green*	*71*	*Green*	*65*	*Green*
RiskMetrics (sst)	*118*	*Yellow*	*121*	*Yellow*	*101*	*Yellow*
MMHC	*55*	*Green*	*47*	*Green*	*34*	*Green*
Genetic	*54*	*Green*	*49*	*Green*	*34*	*Green*
PC (Stable)	*56*	*Green*	*46*	*Green*	*34*	*Green*
SI-HITON-PC	*55*	*Green*	*47*	*Green*	*34*	*Green*
IAMB	*55*	*Green*	*46*	*Green*	*34*	*Green*

Note: This table reports the number of one-day 97.5% expected shortfall (ES) breaches and the corresponding Basel Committee on Banking Supervision (BCBS) traffic light classification for each model and calibration period length. An ES breach occurred when the observed loss exceeded the forecasted ES estimate produced by each of the models in the table. Traffic light classifications are reported as Green, Yellow, or Red according to the observed number of breaches over the relevant out-of-sample period. Results are presented for calibration periods of 252, 504, and 1,008 days, with corresponding out-of-sample periods of 3,998, 3,746, and 3,242 days. The historical simulation and delta-normal models are distribution agnostic, while the ARCH(1), GARCH(1,1), EGARCH(1,1), and RiskMetrics models are reported under both the normal (n) and skewed Student's t (sst) distributions. For the Bayesian network (BN) models, the algorithms considered were the Max-Min Hill Climbing (MMHC), genetic (natural encoding particle swarm optimization, or natPSOHO), Peter and Clark (Stable) (PC (Stable)), the Semi-Interleaved HITON Parents-Children (SI-HITON-PC), and the Incremental Association Markov Blanket

(IAMB) structure-learning algorithms. Breach counts are based on one-day 97.5% ES forecasts generated from S&P 500 index data over the study period.

The breach count further reveals substantial variation in model performance despite the similarity of the resulting traffic light classification. For example, the EGARCH(1,1) model under the normal distribution produced only between 6 and 13 breaches across the three calibration periods, while the delta-normal model produced between 67 and 101 breaches and the RiskMetrics model under the skewed Student's t distribution produced between 101 and 121 breaches. Nevertheless, all three models achieved either 'Green' or 'Yellow' classifications throughout. This result suggests that the BCBS traffic light framework may have limited ability to distinguish between models exhibiting materially different levels of forecasting conservatism and tail-risk estimation performance.

The weak performance of the historical simulation and RiskMetrics models, when compared with the GARCH(1,1) and EGARCH(1,1) models, is broadly consistent with the findings of Žiković and Filer (2013), who identify these approaches to be among the weaker-performing VaR and ES forecasting methods across a range of developed and emerging equity markets. At the same time, the persistence of predominantly 'Green' outcomes contrasts with the more pessimistic assessment of historical simulation presented by García-Risueño (2025), who argues that historical simulation systematically underestimates ES relative to parametric alternatives. Collectively, these findings suggest that regulatory breach-count-based assessments may provide a more favorable view of model performance than comparisons based on statistical accuracy or forecasting efficiency.

While the traffic light test classifications provide a useful regulatory benchmark, they do not indicate whether the underlying ES forecasts are statistically accurate. Hence, we turn our attention to considering statistical backtesting results, which evaluate the calibration, tail-loss estimation, and independence of properties of the forecasts more directly than the BCBS traffic light test.

The statistical backtests provide a significantly different assessment of forecasting performance. The conditional, unconditional, and minimally biased backtests of Acerbi and Székely (2014; 2017) rejected the null hypothesis of correctly specified ES forecasts at the 97.5% confidence level for all models and calibration period length combinations. Similarly, the Du-Escanciano (2017) independence backtest rejected the null hypothesis of independent ES breaches across all traditional model specifications.

The contrast between the favorable BCBS traffic light outcomes and the uniformly negative statistical backtesting results highlights the limitations of relying exclusively on breach counts when evaluating ES forecasts. While the BCBS traffic light backtest focuses primarily on the frequency of breaches, the statistical backtests additionally assess forecast calibration, tail loss estimation, and breach dependence through time. Hence, the 'Green' outcomes reported in Table 1 should not be interpreted as evidence that the forecasts accurately describe the underlying tail risk process, but rather than breach frequencies remained within regulatory tolerance levels.

These findings are generally consistent with concerns raised in the ES backtesting literature regarding the difficulty of accurately forecasting and evaluating tail risk. He, et al., (2022) note that commonly used ES backtests evaluate properties associated with ES rather than the risk measure directly and may, therefore, expose weaknesses in model specification that are not apparent from breach counts alone. Similarly, Zhao (2026) argues that the assessment of ES forecasting methods should extend beyond forecasting accuracy and consider model robustness and performance under changing market conditions. Even recently developed Bayesian realized volatility GARCH frameworks achieve only incremental, rather than transformational, improvements in tail risk forecasting accuracy (Wang & Gerlach, 2024), underscoring the persistent difficulty of modeling extreme losses regardless of model sophistication. Given that our study period spans multiple episodes of severe market stress, including the Global Financial Crisis, the COVID-19 market shock, and the subsequent inflationary environment and monetary policy tightening cycles, the universal rejection of the statistical backtests further underlines the challenges associated with accurately modeling extreme tail behavior in financial markets.

If all models are statistically inaccurate, the natural question that follows is, given that all models tested are statistically inaccurate, which one yields the lowest forecasting error? Table 2 reports the forecasting accuracy of the one-day 97.5% ES forecasts using the MAE and RMSE measures[3]. Several clear trends emerge. First, the EGARCH(1,1) model under the normal distribution produced the most accurate forecasts across all calibration period lengths,

[3] The S/MAPE measures were also calculated. However, these results are not reported because they yielded conclusions that were qualitatively identical to those obtained using the MAE and RMSE measures. To avoid unnecessary duplication, only the MAE and RMSE results are presented. The full S/MAPE results are available from the authors upon request.

achieving MAE values between 0.0212 and 0.0215 and RMSE values of 0.0254 throughout. The GARCH(1,1) and RiskMetrics models under the normal distribution also performed well, consistently ranking among the most accurate forecasting approaches. In contrast, the historical simulation model and the BN-based models produced considerably larger forecasting errors, while the ARCH(1) model under the normal distribution exhibited a substantial deterioration in forecasting accuracy as the calibration period increased in length. These results suggest that accounting for time-varying volatility materially improves ES forecasting performance relative to both historical simulation and BN-augmented approaches.

Table 2: Forecasting accuracy of one-day 97.5% expected shortfall forecasts across calibration periods

Models	**MAE**			**RMSE**		
	252 days	**504 days**	**1,008 days**	**252 days**	**504 days**	**1,008 days**
Historical Simulation	*0.0303*	*0.0328*	*0.0339*	*0.0348*	*0.0362*	*0.0364*
Delta-Normal	*0.0245*	*0.0252*	*0.0255*	*0.0279*	*0.0279*	*0.0278*
ARCH(1) (n)	*0.0246*	*0.0321*	*0.0391*	*0.0329*	*0.0451*	*0.0539*
GARCH(1,1) (n)	*0.0220*	*0.0221*	*0.0220*	*0.0276*	*0.0277*	*0.0274*
EGARCH(1,1) (n)	*0.0212*	*0.0213*	*0.0215*	*0.0254*	*0.0254*	*0.0254*
RiskMetrics (n)	*0.0224*	*0.0226*	*0.0222*	*0.0278*	*0.0280*	*0.0277*
ARCH(1) (sst)	*0.0283*	*0.0270*	*0.0282*	*0.0339*	*0.0309*	*0.0319*
GARCH(1,1) (sst)	*0.0252*	*0.0242*	*0.0246*	*0.0303*	*0.0296*	*0.0293*
EGARCH(1,1) (sst)	*0.0247*	*0.0235*	*0.0245*	*0.0296*	*0.0278*	*0.0285*
RiskMetrics (sst)	*0.0246*	*0.0240*	*0.0239*	*0.0293*	*0.0291*	*0.0283*
MMHC	*0.0303*	*0.0328*	*0.0339*	*0.0348*	*0.0362*	*0.0364*
Genetic	*0.0304*	*0.0329*	*0.0339*	*0.0349*	*0.0362*	*0.0364*
PC (Stable)	*0.0303*	*0.0328*	*0.0339*	*0.0348*	*0.0362*	*0.0364*
SI-HITON-PC	*0.0303*	*0.0328*	*0.0339*	*0.0348*	*0.0362*	*0.0364*
IAMB	*0.0303*	*0.0328*	*0.0339*	*0.0348*	*0.0362*	*0.0364*

Note: This table reports the forecasting accuracy of the one-day 97.5% expected shortfall (ES) forecasts produced by the various models using the mean absolute error (MAE) and the root mean square error (RMSE). Results are presented for calibration periods of 252, 504, and 1,008 days, with corresponding out-of-sample periods of 3,998, 3,746, and 3,242 days. Lower values of either forecasting error measure indicate greater forecasting accuracy. The historical simulation and delta-normal models are distribution agnostic, while the ARCH(1), GARCH(1,1), EGARCH(1,1), and RiskMetrics models are reported under both the normal (n) and skewed Student's t (sst) distributions, For the Bayesian network (BN) models, the algorithms considered were the Max-Min Hill Climbing (MMHC), genetic (natural encoding particle swarm optimization, or natPSOHO), Peter and Clark (Stable) (PC (Stable)), the Semi-Interleaved HITON Parents-Children (SI-HITON-PC), and the Incremental Association Markov Blanket (IAMB) structure-learning algorithms. Breach counts are based on one-day 97.5% ES forecasts generated from S&P 500 index data over the study period.

The superior performance of the GARCH-family models is broadly consistent with the literature on VaR and ES forecasting. Our study finds that the volatility modeling framework

employed by the GARCH-family models is better able to adapt to changing market conditions and generate more accurate ES forecasts. Similarly, the relatively poor performance of the historical simulation is consistent with the criticism offered by both, Žiković and Filer (2013), who identify the historical simulation and RiskMetrics models as among the weaker-performing models for tail risk estimation, and García-Risueño (2025), who argues that the historical simulation underestimates tail risk relative to parametric alternatives. Although the historical simulation model achieved consistently favorable BCBS traffic light outcomes, its forecasting accuracy was significantly lower than that of the best performing volatility-based models.

A particularly notable finding concerns the sensitivity of model performance to calibration period length. Zhao (2026) argues that the evaluation of ES forecasting methods should consider robustness in addition to forecasting accuracy, while Souffargi and Boubaker (2025) demonstrate that structural breaks and regimes can materially affect VaR and ES forecasts. The results presented in Table 2 provide strong support for these arguments. Most models exhibited relatively stable forecasting accuracy as the calibration period increased from 252 to 1,008 trading days. In contrast, the ARCH(1) model under the normal distribution displayed substantial deterioration, with the MAE increasing from 0.0246 to 0.0391 and the RMSE increasing from 0.0329 to 0.0539 as the calibration period lengthened. This decline mirrors the Red-Green-Red pattern observed in Table 1 and suggests that the calibration period length can materially influence forecast performance for some model specifications. A direct comparison against the historical simulation model confirms this pattern statistically. At the 252-day calibration period, the ARCH(1) model is significantly more accurate than the historical simulation model (DM = 2.63, p = 0.0043), whereas at the 504-day and 1,008-day calibration periods, this relationship fully reverses, with the ARCH(1) model becoming significantly less accurate than the historical simulation model (DM = -9.52 and DM = 17.53, respectively, both p-values are approximately equal to 1). A likely explanation is that ARCH(1)'s single-lag structure provides no persistence beyond the most recent shock, which is advantageous when the calibration window is short and volatility is locally homogeneous, but becomes a liability as the window lengthens and increasingly spans multiple volatility regimes that E/GARCH-type persistence terms and the empirical distribution of the historical simulation can accommodate more effectively.

The choice of return distribution also had a meaningful effect on model performance. For all volatility-based models except for the ARCH(1) model, forecasting accuracy deteriorated when

the skewed Student's t distribution replaced the normal distribution. This effect was particularly pronounced for the EGARCH(1,1) model. While the normal distribution EGARCH(1,1) model consistently produced the most accurate forecasts in this study, its skewed Student's t counterpart generated noticeably larger forecasting errors across all calibration periods. Taken together, these results provide little support for the proposition that the additional flexibility afforded by the skewed Student's t distribution necessarily translates into more accurate ES forecasts. This finding is particularly noteworthy given the descriptive statistics reported in Section 3, which indicate both substantial negative skewness and excess kurtosis in S&P 500 index returns over the study period. One possible explanation is that while the skewed Student's t distribution may provide a superior fit to the unconditional return distribution's body, market risk metrics are tail risk metrics, and the focus should be placed on how well the tails fit rather than on the general fit. Hence, a distribution that better describes the overall return distribution does not necessarily produce more accurate tail risk forecasts. For the S&P 500 index and our study period, the normal distribution consistently generated more accurate ES forecasts. This is consistent with Zhu and Galbraith (2011), who similarly find that added distributional flexibility does not guarantee superior tail risk forecasts once estimation uncertainty in the additional shape parameters is taken into account.

The five BN algorithms produced remarkably similar forecasting error measures, with differences generally appearing only in the fourth decimal place. Moreover, the BN results were almost indistinguishable from those of the historical simulation model. This outcome is not unexpected given the construction of the BN-based ES forecasts. In each case, only a single BN-generated return forecast was incorporated into the calibration period return distribution, while the remaining observations consisted of realized historical returns. Hence, the influence of the BN forecast declines as the calibration period length increases. One avenue for strengthening this influence would be an explicit weighting mechanism that gives the BN-generated forecast greater weight than a single historical observation, rather than treating it as equivalent to any other point in the calibration window (see, for example, Taylor, 2019, for a semiparametric approach to weighting forecast information asymmetrically). This observation links directly to the arguments of Zhao (2026) and Souffargi and Boubaker (2025), who emphasize the importance of robustness, structural change, and calibration period design when evaluating ES forecasting models. It is worth noting that Zhao (2026) finds a deep learning approach to be more resilient than GARCH-family models under regime shifts in a Chinese regulatory context. Our results, by contrast, show that BN-generated forecasts do not

achieve comparable outperformance over the traditional models, especially the GARCH(1,1) and EGARCH(1,1) models. This may reflect differences in market structure, forecast horizon, or the specific machine learning mechanisms used to produce forecasts and how we integrated those into the return distribution. Rather than indicating that BN forecasts lack useful information, the results suggest that the effect of forward-looking information depends critically on the quantity of historical data against which it is evaluated. As the calibration period increases from 252 to 1,008 trading days, the relative weight assigned to the BN-generated forecast declines mechanically, reducing its ability to influence the resulting ES forecast. Hence, longer calibration periods may obscure the contribution of forward-looking information, even when that information is informative. This interpretation is supported by the increasing similarity between the BN and historical simulation results as the calibration period lengthens. By the 1,008-day calibration period, all five BN algorithms produced identical breach counts to the historical simulation model and virtually identical forecasting error measures, despite the inclusion of a forward-looking return forecast. This convergence suggests that the influence of the BN forecast becomes increasingly diluted as the amount of historical information incorporated into the ES forecasting process increases.

The findings, therefore, highlight calibration period design as an important determinant of ES forecasting performance. More broadly, the results suggest that future improvements in ES forecasting may depend not only on generating more accurate forecasts, but also on identifying effective mechanisms through which forward-looking information can influence the return distributions used to forecast ES, specifically, and tail risk, generally. Importantly, these findings should not be interpreted as evidence that BN forecasts are without value. Rather, the results suggest that the potential benefits of BNs, including their ability to capture evolving relationships among economic and financial variables, provide interpretable causal structures, and incorporate forward-looking information into the forecasting process, may be constrained by the manner in which such forecasts are incorporated into the ES estimation. Hence, the value of the additional information generated by the BN may not be fully reflected in the resulting ES forecasts or their observed performance in this study. The broader benefits of BNs to the risk management process, including the identification of evolving relationships among economic and financial variables and the provision of forward-looking information, should not be discounted solely on the basis of their forecasting performance in the present framework.

Although Table 2 provides a clear ranking of forecasting performance, differences in forecasting error measures alone do not indicate whether one model is statistically more

accurate than another. Hence, we next consider the Diebold-Mariano test to formally compare the predictive accuracy across competing forecasting approaches. Since the forecasting results showed that the BN algorithms performed almost identically to the historical simulation model, while the delta-normal model consistently outperformed the BN models. Table 3 focuses on comparisons between these benchmark traditional models and the BN algorithms. This approach allows us to determine whether the apparent similarities between the BN and historical simulation models, as well as the apparent superiority of the delta-normal model, are statistically significant rather than the result of sampling variation.

Table 3: Diebold-Mariano test results for selected models versus Bayesian network models across calibration periods

Model	**252 days**		**504 days**		**1,008 days**	
	DM	**p-value**	**DM**	**p-value**	**DM**	**p-value**
HS vs MMHC	*-2.67*	*0.996*	*-2.83*	*0.998*	*-1.81*	*0.965*
HS vs Genetic	*-4.61*	*1.000*	*-4.69*	*1.000*	*-4.71*	*1.000*
HS vs PC (Stable)	*-2.26*	*0.988*	*-3.02*	*0.999*	*-0.89*	*0.813*
HS vs SI-HITON-PC	*-1.94*	*0.974*	*-2.38*	*0.991*	*-1.43*	*0.924*
HS vs IAMB	*-1.38*	*0.917*	*-1.99*	*0.977*	*-0.87*	*0.809*
DN vs MMHC	*-36.51*	*1.000*	*-54.99*	*1.000*	*-73.70*	*1.000*
DN vs Genetic	*-36.76*	*1.000*	*-55.08*	*1.000*	*-73.70*	*1.000*
DN vs PC (Stable)	*-36.52*	*1.000*	*-55.00*	*1.000*	*-73.70*	*1.000*
DN vs SI-HITON-PC	*-36.51*	*1.000*	*-54.99*	*1.000*	*-73.70*	*1.000*
DN vs IAMB	*-36.52*	*1.000*	*-54.99*	*1.000*	*-73.70*	*1.000*

Note: This table reports the Diebold-Mariano (DM) test statistics and corresponding p-values comparing the forecasting accuracy of the historical simulation (HS) and delta-normal (DN) models against five Bayesian network (BN) structure-learning algorithms. Results are presented for calibration periods of 252, 504, and 1,008 days, with corresponding out-of-sample periods of 3,998, 3,746, and 3,242 days. The null hypothesis states that the two models being compared possess equal predictive accuracy, while the alternative hypothesis states that the first model listed in the comparison is less accurate than the second model. Hence, large negative DM statistics accompanied by p-values close to one indicate evidence in favor of the first model being more accurate than the second. The BN algorithms considered were the Max-Min Hill Climbing (MMHC), genetic (natural encoding particle swarm optimization, or natPSOHO), Peter and Clark (Stable) (PC (Stable)), the Semi-Interleaved HITON Parents-Children (SI-HITON-PC), and the Incremental Association Markov Blanket (IAMB) structure-learning algorithms. Breach counts are based on one-day 97.5% ES forecasts generated from S&P 500 index data over the study period.

The results strongly reinforce the patterns observed in Table 1 and Table 2. For the comparison between the historical simulation model and the BN algorithms, the DM statistics are uniformly negative and the corresponding p-values are consistently close to one. Hence, the null hypothesis of equal predictive accuracy is rejected in favor of the historical simulation model being more accurate than each of the BN specifications. However, the magnitudes of the differences remain economically small, reflecting the near-identical breach count and

forecasting error measures reported earlier. In practical terms, the BN algorithms failed to deliver a statistically meaningful improvement over the historical simulation benchmark, despite incorporating forward-looking return forecasts.

The differences become considerably more pronounced when the delta-normal model is compared with the BN algorithms. Across all calibration lengths, the delta-normal model produced extremely large DM statistics accompanied by p values effectively equal to one. These results provide strong statistical evidence that the delta-normal model outperformed every BN algorithm considered in this study. Viewing this in conjunction with the forecasting error results presented in Table 2, these findings highlight the importance of distributional assumptions in ES forecasting. While the BNs rely on empirical return distributions augmented by forward-looking forecasts, the delta-normal model assumes a parametric normal distribution and delivers significantly superior forecasting performance. This result is consistent with the broader finding that the normal distribution models generally outperformed both the skewed Student's t and empirical distribution alternatives. Hence, the results suggest that, for the S&P 500 index and study period considered here, the choice of underlying distribution exerts a greater influence on ES forecasting performance than the inclusion of a single forward-looking BN-generated return forecast.

While Table 3 compares the BN algorithms with selected traditional benchmark models, it does not establish whether meaningful differences exist among the BN algorithms themselves. Table 4 reports pairwise DM tests for the five BN structure-learning algorithms. These results help determine whether any particular BN learning algorithm consistently produced superior ES forecasts.

Table 4: Diebold-Mariano test results for pairwise comparisons of Bayesian network algorithms across calibration periods

Model	**p-value (252 days)**	**p-value (504 days)**	**p-value (1,008 days)**	**Significant at any window?**
MMHC vs Genetic	*1.000*	*1.000*	*1.000*	*All three*
MMHC vs PC (Stable)	*0.886*	*0.979*	*0.508*	*No*
MMHC vs SI-HITON-PC	*0.484*	*0.796*	*0.458*	*No*
MMHC vs IAMB	*0.822*	*0.192*	*0.040*	*1,008-day only*
Genetic vs PC (Stable)	*<0.0001*	*<0.0001*	*<0.0001*	*All three*
Genetic vs SI-HITON-PC	*<0.0001*	*<0.0001*	*<0.0001*	*All three*

Genetic vs IAMB	*<0.0001*	*<0.0001*	*<0.0001*	*All three*
PC (Stable) vs SI-HITON-PC	*0.100*	*0.034*	*0.477*	*504-day only*
PC (Stable) vs IAMB	*0.677*	*0.021*	*0.310*	*504-day only*
SI-HITON-PC vs IAMB	*0.822*	*0.131*	*0.173*	*No*

Note: This table reports the Diebold-Mariano (DM) test statistics and corresponding p-values comparing the forecasting accuracy of the five Bayesian network (BN) structure-learning algorithms. Results are presented for calibration periods of 252, 504, and 1,008 days, with corresponding out-of-sample periods of 3,998, 3,746, and 3,242 days. The null hypothesis states that the two models being compared possess equal predictive accuracy, while the alternative hypothesis states that the first model listed in the comparison is less accurate than the second model. Hence, large negative DM statistics accompanied by p-values close to one indicate evidence in favor of the first model being more accurate than the second. The final column summarizes whether statistically significant differences were observed across any of the calibration periods considered. The BN algorithms considered were the Max-Min Hill Climbing (MMHC), genetic (natural encoding particle swarm optimization, or natPSOHO), Peter and Clark (Stable) (PC (Stable)), the Semi-Interleaved HITON Parents-Children (SI-HITON-PC), and the Incremental Association Markov Blanket (IAMB) structure-learning algorithms. Breach counts are based on one-day 97.5% ES forecasts generated from S&P 500 index data over the study period.

The pairwise comparisons indicate that, with the exception of the genetic algorithm, the BN algorithms produced remarkably similar forecasting performance. Most pairwise comparisons were not statistically significant across any calibration period length, supporting the earlier observation that the forecasting accuracies of the BN models differed only marginally. This reinforces the conclusion that the dominant determinant of forecast performance is the broader BN-ES setting itself rather than the specific structure-learning algorithm employed. Given the limited differences in forecasting performance across most BN algorithms, computational efficiency may become more an important criterion when selecting a learning algorithm in practice.

The genetic algorithm represents the primary exception to this pattern. This result may be explained given the stochastic nature of the natPSOHO algorithm, which may generate greater variability in learned network structures than the more deterministic constraint-based and hybrid approaches considered in this study. Significant differences were observed between the genetic algorithm and the PC (Stable), SI-HITON-PC, and IAMB algorithms across all three calibration periods. In contrast, the MMHC, PC (Stable), SI-HITON-PC, and IAMB algorithms generally exhibited no persistent evidence of differential predictive accuracy. A small number of isolated significant results were observed for specific calibration periods, such as the MMHC-IAMB comparison at the 1,008-day calibration period and the PC (Stable)

comparisons at the 504-day calibration periods. However, these isolated findings were not consistently replicated across calibration period lengths and, therefore, provide limited evidence of systematic forecasting superiority.

Collectively, the DM results complement the forecasting error measures presented in Table 2. Both sets of results indicate that the BN algorithms deliver forecasting performance broadly comparable to that of the historical simulation model while remaining inferior to several traditional volatility-based approaches. More importantly, the results suggest that improvements in E forecasting are unlikely to arise solely from changing the BN structure-learning algorithm. Instead, future improvements may require modifications to the manner in which BN-generated forecasts are incorporated into the return distribution used to forecast ES. Nonetheless, the broader findings of this study indicate that distributional assumptions and calibration period length exert a substantial influence on ES forecasting performance. In particular, the superiority of the normal distribution relative to the skewed Student's t distribution and the pronounced calibration period sensitivity displayed by the ARCH(1) model suggest that model specification and estimation design remain central determinants of ES forecast quality.

5. Conclusion

In this study, we extended the application of BNs to market risk metrics by using five BN structure-learning algorithms to forecast one-day 97.5% ES. Over the period 7 August 2007 to 30 June 2026 and using the S&P 500 index as a proxy for an equity portfolio, we compared the BN algorithms with a range of traditional ES forecasting models across multiple calibration period lengths and distributional assumptions. In doing so, we evaluated the relative importance of model choice, distributional assumptions, calibration period design, and BN-based forecasts in ES estimation.

Several important findings emerged. Although most models achieved favorable BCBS traffic light backtest results, all models failed the statistical backtests considered, highlighting the distinction between regulatory acceptability and statistical model accuracy. We tested 45 ES forecasting models and found that the EGARCH(1,1) model under the normal distribution consistently produced the most accurate ES forecasts. More broadly, the results demonstrate that distributional assumptions exert a substantial influence on forecast performance. Despite the negative skewness and excess kurtosis observed in S&P 500 index returns over the study period, the skewed Student's t distribution consistently produced less accurate ES forecasts

than the normal distribution, suggesting that a superior fit to the overall return distribution does not necessarily translate into superior tail risk forecasts.

A second finding concerns calibration period design. While most models exhibited relatively stable forecasting performance across the three calibration periods considered (252, 504, and 1,008 trading days), the performance of the ARCH(1) model under the normal distribution deteriorated significantly as the calibration period increased from 252 to 1,008 trading days. This finding provides empirical support for Zhao's (2026) argument that ES forecasting models should be evaluated not only on forecasting accuracy, but also on robustness, and for the findings of Souffargi and Boubaker (2025), who demonstrate that structural breaks and changing market conditions can materially influence VaR and ES forecasts. Collectively, these results suggest that the quantity of historical information used in forecasting may be at least as important as the choice of forecasting model itself when evaluating tail risk forecasts.

With respect to the BN models, the five structure-learning algorithms produced remarkably similar forecasting performance and generated ES forecasts broadly comparable to those of the historical simulation model. However, this result should not be interpreted as evidence that BNs are unsuitable for market risk forecasting. Rather, the findings suggest that the influence of BN-generated forecasts depends critically on how those forecasts are incorporated into the ES forecasting models employed. As calibration period length increases, the contribution of the single forward-looking BN-generated forecast in an equal-weight model becomes progressively diluted within the historical return distribution, reducing its influence on the resulting ES forecast. Hence, the limited forecasting improvement observed in this study may reflect the manner in which BN forecasts were incorporated into the ES forecasting model rather than limitations of the BN forecasting methodology itself.

This interpretation is consistent with the broader BN literature. Chan, et al., (2023) demonstrate the usefulness of BNs for modeling evolving financial relationships, while Liu, et al., (2024) highlight the networks' ability to capture complex dependency structures and provide interpretable causal insights. Our findings, therefore, suggest that the potential advantages of BNs may not be fully realized when their forecasts exert only a limited influence on the return distribution used to forecast ES. Rather, the structures learned and strengths of causal relationships offered by the BNs' graphical structures may be more useful to complement the risk management process than the integration of a point-estimate forecast into a much larger empirical return distribution on its own can afford. Overall, the results indicate that

distributional assumptions and calibration period design exert a greater influence on ES forecasting performance than the particular BN structure-learning algorithm employed.

For academics, our findings contribute to the growing literature on ES forecasting by highlighting the importance of robustness, calibration period selection, and model specification. For practitioners, they emphasize the importance of carefully setting distributional assumptions and selecting calibration period lengths when forecasting ES. Future research should, therefore, focus on developing alternative mechanisms through which forward-looking BN forecasts and their accompanying information can be incorporated in the risk management process, particularly through alternative weighting schemes (Taylor 2019; 2020), mixed-frequency methods for integrating macroeconomic and financial variables underlying the BN structures more directly into the forecasting process (Le, 2020), and hybrid frameworks combining BN-generated forecasts with traditional tail risk models.

References

Acerbi, C. & Székely, B., 2014. Backtesting expected shortfall. *Risk,* 27(11), pp. 76-81.

Acerbi, C. & Székely, B., 2017. *General properties of backtestable statistics,* s.l.: SSRN. https://dx.doi.org/10.2139/ssrn.2905109

Acerbi, C. & Tasche, D., 2002. Expected shortfall: A natural coherent alternative to value at risk. *Economic Notes by Banca Monte dei Paschi di Siena SpA, 31*(2), pp. 379-388. https://doi.org/10.1111/1468-0300.00091

Aliferis, C. F., Tsamardinos, I. & Statnikov, A., 2003. *A novel Markov blanket algorithm for optimal variable selection.* Bethesda, s.n., pp. 21-25.

Apps, E., 2020. *Applying a Bayesian network to VaR calculations (Working Paper),* Liverpool: University of Liverpool Management School.

Basel Committee on Banking Supervision, 1996. *Amendment to the Capital Accord to incorporate market risks,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2019a. *Internal models approach: Backtesting and P&L attribution test requirements,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2019b. *Minimum capital requirements for market risk,* Basel: Bank for International Settlements.

Basel Committee on Banking Supervision, 2023. *Calculation of RWA for market risk,* Basel: Bank for International Settlements.

Berkowitz, J. & O'Brien, J., 2002. How accurate are value-at-risk models at commercial banks? *Journal of Finance, 57*(3), pp. 1093-1111. https://doi.org/10.1111/1540-6261.0045

Bollerslev, T., 2007. *Glossary to ARCH (GARCH),* s.l.: Duke University.

Chan, L. S. H., Chu, A. M. Y. & So, M. K. P., 2023. A moving-window Bayesian network model for assessing systemic risk in financial markets. *PLoS ONE, 18*(1), p. e0279888. https://doi.org/10.1371/journal.pone.0279888

Chang, T.-J., Yang, S.-C., & Chang, K.-J., 2009. Portfolio optimization problems in different risk measures using genetic algorithm. *Expert Systems with Applications, 36*(7), pp. 10529-10537. https://doi.org/10.1016/j.eswa.2009.02.062

Chen, J. M., 2018. On exactitude in financial regulation: Value-at-risk, expected shortfall, and expectiles. *Risks, 6*(2), p. 61. https://doi.org/10.3390/risks6020061

Cura, T., 2009. Particle swarm optimization approach to portfolio optimization. *Nonlinear Analysis: Real World Applications, 10*(4), pp. 2396-2406. https://doi.org/10.1016/j.nonrwa.2008.04.023

Demirer, R., Mau, R. R. & Shenoy, C., 2006. Bayesian networks: A decision tool to improve portfolio risk analysis. *Journal of Applied Finance, 16*, pp. 106-119.

Diebold, F. X., 2015. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. *Journal of Business and Economic Stiatistics, 33*(1), p. 1. https://doi.org/10.1080/07350015.2014.983236

Diebold, F. X., Gunther, T. A. & Tay, A. S., 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review, 39*(4), pp. 863-883. https://doi.org/10.2307/2527342

Diebold, F. X. & Mariano, R. S., 2002. Comparing predictive accuracy. *Journal of Business and Economic Statistics, 20*(1), pp. 134-144. https://doi.org/10.1198/073500102753410444

Du, Z. & Escanciano, J. C., 2017. Backtesting expected shortfall: Accounting for tail risk. *Management Science, 63*(4), pp. 940-958. https://doi.org/10.1287/mnsc.2015.2342

Engle, R., 2001. GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives, 15*(4), pp. 157-168. https://doi.org10.1257/jep.15.4.157

García-Risueño, P. 2025. Historical simulation systematically underestimates the expected shortfall. *Journal of Risk and Financial Management, 18*(1), p. 34. https://doi.org/10.3390/jrfm18010034

Gneiting, T., 2011. Making and evaluating point forecasts. *Journal of the American Statistical Association, 106*(494), pp. 746-762. https://doi.org/10.1198/jasa.2011.r10138

Grajales, C. A. & Medina Hurtado, S., 2023. Sensitivities-based method and expected shortfall for market risk under FRTB: Its impact on options risk capital. *Journal of Economics, Finance and Administrative Science, 28*(55), pp. 96-115. https://doi.org/10.1108/JEFAS-12-2021-0268

Grinold, R. C., 1992. Are benchmark portfolios efficient? *Journal of Portfolio Management, 19*(1), pp. 14-21.

Gross, E., Kruger, R. & Toerien, F., 2025. *Standard and stressed value at risk forecasting using dynamic Bayesian networks*, arXiv preprint arXiv:2512.05661 https://doi.org/10.48550/arXiv.2512.05661

He, X. D., Kou, S. & Peng, X., 2022. Risk measures: Robustness, elicitability, and backtesting. *Annual Review of Statistics and its Application, 9*, pp. 141-166. https://doi.org/10.1146/annurev-statistics-030718-105122

Le, T. H., 2020. Forecasting value at risk and expected shortfall with mixed data sampling. *International Journal of Forecasting, 36*(4), pp. 1362-1379. https://doi.org/10.1016/j.ijforecast.2020.01.008

Leong, C. K., 2016. Credit risk scoring with Bayesian network models. *Computational Economics, 47*(3), pp. 423-446. https://doi.org/10.1007/s10614-015-9505-8

Liu, J., Zhang, X. & Xiong, H., 2024. Credit risk prediction on causal machine learning: Bayesian network learning, default inference, and interpretation. *Journal of Forecasting, 43*(5), pp. 1625-1660. https://doi.org/10.1002/for.3080

Novales, A. & Garcia-Jorcano, L., 2019. Backtesting extreme value theory models of expected shortfall. *Quantitative Finance 19*(5), pp. 799-825. https://doi.org/10.1080/14697688.2018.1535182

Pérignon, C. & Smith, D. R., 2010. The level and quality of value-at-risk disclosed by commercial banks. *Journal of Banking and Finance, 34*(2), pp. 362-377. https://doi.org/10.1016/j.jbankfin.2009.08.009

Quesada, D., Bielza, C. & Larrañga, P., 2021. *Structure learning of high-order dynamic Bayesian networks via particle swarm optimization with order invariant encoding.* In: Sanjurjo González, H., Pastor López, I., García Bringas, P., Quintián, H., Corchado, E.

(eds) Hybrid Artificial Systems. HAIS 2021. Lectures Notes in Computer Science, 12886, pp. 158-171. Springer, Cham. https://doi.org/10.1007/978-3-030-86271-8_14

Sadeghi, A., Gopal, A. & Fesanghary, M., 2024. *Causal discovery in financial markets: A framework for nonstationary time-series data*, arXiv preprint arXiv:2312.17375. https://doi.org/10.48550/arXiv.2312.17375

Shenoy, C. & Shenoy, P. P., 2000. Bayesian network models of portfolio risk and return. In: Y. S. Abu-Mostafa, B. LeBaron, A. W. Lo & A. S. Weigned, eds. *Computational Finance.* Cambridge: MIT Press, pp. 87-106. https://hdl.handle.net/1808/161

Souffargi, W. & Boubaker, A., 2025. Modelling value-at-risk and expected shortfall for a small capital market: Do fractionally integrated models and regime shifts matter? *Journal of Risk and Financial Management, 18*(4), p. 203. https://doi.org/10.3390/jrfm18040203

Spirtes, P. & Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review, 9*(1), pp. 62-72. https://doi.org/10.1177/089443939100900106

Storti, G. & Wang, C. 2022. Nonparametric expected shortfall forecasting incorporating weighted quantiles. *International Journal of Forecasting, 38*(1). pp. 224-239. https://doi.org/10.1016/j.ijforecast.2021.04.004

Taylor, J. W., 2019. Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business and Economic Statistics, 37*(1), pp. 121-133. https://doi.org/10.1080/07350015.2017.1281815

Taylor, J. W., 2020. Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting, 36*(2), pp. 428-441. https://doi.org/10.1016/j.ijforecast.2019.05.014

Tsamardinos, I., Brown, L. E. & Aliferis, C. F., 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning, 65*(1), pp. 31-78. https://doi.org/10.1007/s10994-006-6889-7

Tsamardinos, I., Aliferis, C. F. & Statnikov, A. R., 2003. *Algorithms for large scale Markov blanket discovery*. Menlo Park, CA, Association for the Advancement of Artificial Intelligence, pp. 376-381.

Wang, C. & Gerlach, R., 2024. A Bayesian realized threshold measurement GARCH framework for financial tail risk forecasting. *Journal of Forecasting, 43*(1), pp. 40-57. https://doi.org/10.1002/for.3025

Yamai, Y. & Yoshiba, T., 2005. Value at risk versus expected shortfall: A practical perspective. *Journal of Banking and Finance, 29*(4), pp. 997-1015. https://doi.org/10.1016/j.jbankfin.2004.08.010

Zhao, W., 2026. E-backtesting expected shortfall: What defines a "good" forecasting method for Chinese regulators? *Risks, 14*(5), p. 110. https://doi.org/10.3390/risks14050110

Zhu, D. & Galbraith, J. W., 2011. Modeling and forecasting expected shortfall with the generalized asymmetric Student-t and asymmetric exponential power distributions. *Journal of Empirical Finance, 18*(5), pp. 765-778. https://doi.org/10.1016/j.jempfin.2011.05.006

Ziegel, J. F., Krüger, F., Jordan, A. & Fasciati, F., 2020. Robust forecast evaluation of expected shortfall. *Journal of Financial Econometrics, 18*(1), pp. 95-120. https://doi.org/10.1093/jjfinec/nby035

Žiković, S. & Filer, R., 2013. Ranking of VaR and ES models: Performance in developed and emerging markets. *Finance a úvěr: Czech Journal of Economics and Finance, 63*(4), pp. 327-359

Appendix A: Variables used to train the Bayesian networks

Table 5: Variables used to train the Bayesian networks

Variable name	Classification
S&P 500 index (closing value)	*Target Variable*
Australian Dollar/US Dollar	*Currency Exchange Rate*
Average-Weighted US Government Inflation-linked All Maturity Average Duration	*Financial*
Bloomberg Commodities Index	*Financial*
Bloomberg US Government 1-3-year Average Modified Duration	*Financial*
Bloomberg US Government 3-5-year Average Modified Duration	*Financial*
Bloomberg US Government 5-7-year Average Modified Duration	*Financial*
Bloomberg US Government 7-10-year Average Modified Duration	*Financial*
Bloomberg US Treasury Total Return Index	*Financial*
Canadian Dollar/US Dollar	*Currency Exchange Rate*
CBOE Put-Call Ratio	*Financial*
CBOE Volatility Index	*Financial*
Emerging Market Sovereign Debt Total Return Index	*Financial*
Euro/US Dollar	*Currency Exchange Rate*
Federal Reserve Effective Funds Rate	*Economic*
FTSE Developed Markets	*Financial*
FTSE Emerging Markets	*Financial*
Government Securities Liquidity Index	*Financial*
Great British Pound/US Dollar	*Currency Exchange Rate*
Japanese Yen/US Dollar	*Currency Exchange Rate*
S&P 500 Dividend Yield	*Financial*
Secured Overnight Financing Rate	*Financial*
Three-month US Dollar LIBOR	*Financial*
US Breakeven Inflation Rate	*Economic*
US Consumer Price Index	*Economic*
US Corporate AA-rated Ten-year Yield Curve	*Financial*
US Corporate AAA-rated Ten-year Spread	*Financial*
US Corporate BBB-rated Ten-year Yield Curve	*Financial*
US Corporate Bond Index	*Financial*
US Disposable Income Growth	*Economic*
US Dollar Inflation Swap Index	*Financial*
US Economic Policy Uncertainty Index	*Political*
US Economic Surprise Index	*Economic*
US Equity Momentum Index	*Financial*
US M1 Money Supply	*Economic*
US M2 Money Supply	*Economic*

US Non-Farm Payroll	*Economic*
US Real Gross Domestic Product	*Economic*
US Treasury Inflation-Protected Securities Index	*Financial*

Note: This table lists the variables used to train the Bayesian networks (BNs) employed to generate one-day-ahead S&P 500 index forecasts. The variables were obtained from the Bloomberg database over the period 7 August 2007 to 30 June 2026, and comprise a mixture of financial, macroeconomic, and policy-related indicators identified through the literature and expert judgement as potentially related to the performance of the equity market in the United States (US).