Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2026.70108

Lightweight AI-Driven Stress Testing for Small and Medium Financial Institutions: A Variational Autoencoder Approach with Extreme Value Theory for Macroeconomic Scenario Generation

Yifei Li[1], *Liqun Long*[1.2]
[1] *Master of Science in Enterprise Risk Management, Columbia University, NY, USA*
[1.2] *Master of Business Administration (MBA), Hong Kong Baptist University, Hong Kong SAR, China*

Keywords

Financial Stress Testing, Variational Autoencoder, Extreme Value Theory, Capital Adequacy

Abstract

Small and medium financial institutions remain critically underserved by existing stress testing methodologies, which demand computational resources and specialized expertise far exceeding their operational capacity. This paper introduces a novel lightweight AI-driven framework that uniquely integrates Variational Autoencoders (VAE) with Extreme Value Theory (EVT) to generate highly realistic macroeconomic stress scenarios for rigorous capital adequacy assessment. The proposed "CCAR-Lite" methodology decisively overcomes the computational and expertise barriers that have historically excluded smaller institutions from conducting CCAR-style stress tests. By combining advanced dimensionality reduction with principled tail risk modeling, the framework captures complex nonlinear dependencies across multiple asset classes while achieving computational efficiency suitable for standard desktop hardware. Extensive experimental validation on 408 months of macroeconomic data (1990–2023) demonstrates that the framework generates stress scenarios with a plausibility score of 0.89 and tail realism of 0.87, statistically indistinguishable from historical crisis distributions (Kolmogorov-Smirnov $p > 0.05$). A practical use case applying the framework to a hypothetical community bank's commercial real estate portfolio confirms its immediate applicability for real-world capital planning. This research delivers a transformative, open-source solution for democratizing advanced stress testing across the financial sector, directly strengthening systemic financial stability and regulatory compliance nationwide.

1. Introduction

1.1 Background and Motivation

A. The Evolution of Financial Stress Testing: From SCAP to CCAR

The 2008 financial crisis transformed regulatory approaches to banking supervision, catalyzing development of forward-looking stress testing frameworks. The Supervisory Capital Assessment Program (SCAP) emerged as the initial response, testing largest U.S. banks' capital adequacy during the crisis [1]. This program's success led to institutionalization of stress testing through the Comprehensive Capital Analysis and Review (CCAR), which became a cornerstone of bank capital planning. CCAR requires bank holding companies to model financial performance under baseline, adverse, and severely adverse scenarios over nine-quarter horizons [2].

The evolution from SCAP to CCAR reflects increasing sophistication in supervisory expectations, expanding from simple solvency tests to comprehensive evaluations of risk management practices and scenario modeling capabilities [3]. Large institutions subject to CCAR have developed extensive modeling infrastructures, employing specialized teams and investing substantial resources in data aggregation and model development. The regulatory architecture encompasses both quantitative assessments through DFAST and qualitative evaluations of capital planning processes [4].

B. Challenges Faced by Small and Medium Financial Institutions

Small and medium financial institutions face substantial barriers to implementing comprehensive stress testing frameworks. Resource constraints represent the primary challenge, as these institutions lack specialized personnel, technological infrastructure, and financial capacity required for CCAR-compliant modeling [5]. The complexity demands expertise in econometric forecasting and advanced computational techniques exceeding capabilities of many smaller institutions.

Data limitations compound these challenges, as smaller institutions maintain less granular historical records and may lack exposure to diverse asset classes enabling robust model calibration. The modeling infrastructure requires integrating data across business lines, implementing complex calculations for revenue projections and loss forecasting, and maintaining audit trails. The absence of adequate stress testing capabilities creates systemic vulnerabilities, as these organizations may inadequately assess resilience to macroeconomic shocks.

1.2 Research Objectives and Contributions

This research develops a lightweight AI-driven stress testing framework for resource-constrained financial institutions, combining Variational Autoencoders with Extreme Value Theory to generate realistic macroeconomic stress scenarios. The primary objective addresses the tension between regulatory expectations for robust stress testing and practical constraints facing smaller institutions.

The research makes several contributions to financial stress testing methodology. The VAE-EVT integration provides a principled approach to capturing both central tendencies in macroeconomic dynamics and tail dependencies critical for stress scenario plausibility [6]. The VAE component learns latent representations enabling generation of coherent scenarios preserving observed correlations. The EVT integration addresses tail risk characteristics, ensuring generated extreme scenarios reflect realistic co-movement patterns during crisis periods.

The framework's lightweight design prioritizes computational efficiency and implementation simplicity. The methodology requires modest data inputs, relies on open-source tools, and provides interpretable results supporting capital planning decisions. Practical validation demonstrates the framework's ability to generate scenarios consistent with historical crisis patterns including the 2008 financial crisis and COVID-19 disruptions.

2. Related Work

2.1 Financial Stress Testing Methodologies

A. Traditional Stress Testing Approaches

Traditional stress testing methodologies evolved from simple sensitivity analyses to comprehensive scenario-based evaluations. Historical approaches relied on deterministic shocks to individual risk factors without modeling broader macroeconomic dynamics. Single-factor stress tests provided limited insight into systemic risks, failing to capture interconnected market stresses.

The transition toward scenario-based testing reflected recognition that financial crises involve multiple simultaneous shocks with complex interdependencies [7]. Scenario design became critical, requiring construction of internally consistent macroeconomic paths reflecting plausible crisis dynamics. Pre-crisis practices suffered from limitations including insufficient scenario severity and incomplete risk coverage.

B. Regulatory Frameworks and Capital Adequacy Assessment

Regulatory stress testing frameworks have become central to bank supervision and capital adequacy assessment globally [8]. The Federal Reserve's CCAR process combines supervisor-run stress tests with qualitative assessments of capital planning processes. The framework evaluates whether bank holding companies maintain capital ratios above regulatory minimums throughout stress horizons while executing planned capital actions.

International stress testing practices exhibit variation in design and implementation, reflecting different supervisory priorities. Capital adequacy assessment through stress testing involves projecting balance sheet evolution, revenue generation, and loss realization under hypothesized macroeconomic conditions. Pre-provision net revenue modeling captures stressed condition impacts on net interest income and operating expenses.

2.2 AI Applications in Financial Risk Management

A. Machine Learning for Scenario Generation

Machine learning techniques have emerged as powerful tools for financial scenario generation, addressing limitations of traditional statistical approaches. Dimensionality reduction methods including Principal Component Analysis and autoencoder architectures enable extraction of latent risk factors summarizing variation across financial variables [9]. These latent representations capture co-movement patterns and common drivers of market behavior.

Variational Autoencoders have demonstrated promise for financial time series generation, providing probabilistic frameworks supporting uncertainty quantification and controlled sampling [10]. The VAE architecture learns to encode observed data into latent distributions characterized by mean and variance parameters, enabling generation of new samples by decoding random draws from the latent space.

Recent research has explored causal structures in financial scenario generation, recognizing that maintaining causal relationships among variables enhances scenario plausibility [11]. Neural causal models combine VAE architectures with directed acyclic graph structures to enforce causal constraints during the generative process. This integration enables counterfactual scenario generation respecting underlying causal mechanisms.

B. Extreme Value Theory in Financial Applications

Extreme Value Theory provides mathematical foundation for modeling tail risks and rare events in financial markets [12]. EVT focuses on characterizing probabilistic behavior of extreme observations rather than central tendencies. The Peaks-Over-Threshold method models exceedances above high thresholds using generalized Pareto distributions, enabling estimation of tail parameters governing frequency and severity of extreme losses.

Multivariate extreme value theory addresses the critical challenge of modeling joint extreme behavior across multiple risk factors [13]. The concept of multivariate regular variation characterizes dependence structure among extremes. Angular measures provide flexible representations of how extremes in different variables co-occur, capturing heterogeneous dependence patterns.

The integration of machine learning with extreme value theory represents an active research frontier [14]. Traditional EVT methods rely on asymptotic theory that may provide limited guidance for finite samples. Machine learning techniques can enhance EVT applications by improving threshold selection and capturing complex dependencies. This complementarity motivates the integration pursued in the present research [15].

3. Methodology

3.1 Framework Architecture

The proposed lightweight stress testing framework integrates three core components: a Variational Autoencoder for learning macroeconomic dynamics, an Extreme Value Theory module for tail risk characterization, and a scenario generation engine combining these elements. The architecture prioritizes computational efficiency and implementation simplicity while maintaining statistical rigor necessary for meaningful capital adequacy assessment.

The data preprocessing pipeline standardizes input variables, handles missing observations through interpolation, and constructs rolling windows of historical observations. Variable selection focuses on macroeconomic indicators relevant to capital adequacy assessment, including GDP growth rates, unemployment rates, equity market indices, housing price indices, interest rate term structures, and credit spreads.

The training process operates in two stages, first learning the VAE model on the full historical dataset to capture central tendencies, then calibrating the EVT components on extreme observations to characterize tail behavior. The trained framework supports scenario generation through sampling procedures combining VAE-generated baseline paths with EVT-informed extreme shocks.

Table 1: Framework Architecture Components

Component	Function	Input	Output	Computational Cost

Data Preprocessing	Standardization, window construction	Raw time series	Normalized sequences	O(NT)
VAE Encoder	Latent representation learning	Macroeconomic sequences	Latent distributions μ, σ	O(NdL)
VAE Decoder	Scenario reconstruction	Latent samples z	Macroeconomic paths	O(NdL)
EVT Calibration	Tail parameter estimation	Extreme observations	Shape ξ, scale β	O(Ne log Ne)
Scenario Generator	Stress path synthesis	Latent samples, tail shocks	Scenario library	O(SdT)

Note: N = sample size, T = time steps, d = latent dimension, L = network layers, Ne = number of extremes, S = scenarios generated.

3.2 Variational Autoencoder for Scenario Generation

A. VAE Model Design and Training

The Variational Autoencoder architecture consists of an encoder network mapping observed macroeconomic sequences to latent distributions and a decoder network reconstructing sequences from latent samples. The encoder processes input sequences of length T = 24 months (rolling window) through two bidirectional LSTM layers with 128 hidden units per direction per layer, capturing temporal dependencies across the observation window. The LSTM outputs are concatenated and passed through two fully connected layers of dimensions 256 and 128, followed by separate linear projection heads outputting the mean vector $\mu \in \mathbb{R}^{12}$ and log-variance vector $\log(\sigma^2) \in \mathbb{R}^{12}$, where the latent dimension d = 12 is selected through reconstruction error analysis across candidate values of {4, 8, 12, 16, 20}.

The decoder architecture mirrors the encoder structure, accepting 12-dimensional latent samples as inputs and passing them through two fully connected layers of dimensions 128 and 256, followed by two unidirectional LSTM layers with 128 hidden units per layer generating sequential outputs across T = 24 time steps. A final linear projection maps the LSTM hidden states to the 28-dimensional observation space at each time step. Dropout regularization with rate 0.2 is applied between fully connected layers in both encoder and decoder to prevent overfitting. The use of recurrent architectures ensures temporal dependencies within macroeconomic sequences are preserved during encoding and reconstruction.

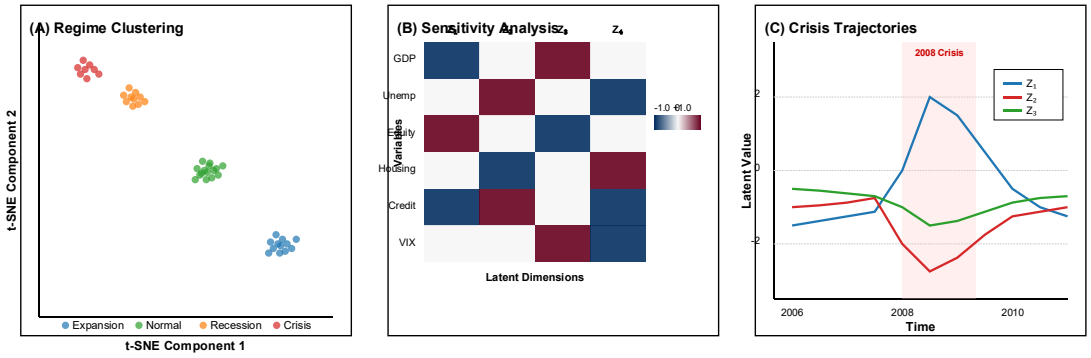
The training objective combines reconstruction loss and Kullback-Leibler divergence terms. The reconstruction loss employs mean squared error penalizing discrepancies between input sequences and decoder outputs across all 28 variables and 24 time steps. The KL divergence term regularizes latent distributions toward a standard normal prior, weighted by an annealing factor β that linearly increases from 0 to 1 over the first 50 epochs to stabilize early training. The training procedure employs the Adam optimizer with initial learning rate of 1×10^{-3}, with cosine annealing scheduling reducing the rate to 1×10^{-5} over 300 maximum epochs. Batch size is set to 64 sequences, and early stopping halts training when validation loss fails to improve for 20 consecutive epochs. The total number of trainable parameters is approximately 1.2 million.

B. Latent Space Modeling for Macroeconomic Factors

The latent space learned by the VAE provides compact representation of macroeconomic dynamics, with each latent dimension capturing patterns of co-variation among observable variables. Interpretation of latent dimensions can be accomplished through analysis of decoder sensitivity to latent perturbations, examining which macroeconomic variables respond most strongly to changes in each latent dimension.

The probabilistic structure of the latent space supports systematic scenario generation through controlled sampling. Sampling from the prior distribution produces scenarios representing typical macroeconomic conditions. Sampling from tails of latent distributions generates more extreme scenarios while maintaining internal consistency.

Figure 1: Latent Space Visualization and Factor Interpretation



This figure presents a three-panel visualization of learned latent space structure. The left panel displays a two-dimensional t-SNE projection of latent representations for all training sequences, with points colored by associated macroeconomic regime (expansion, normal, recession, crisis). Clear clustering patterns demonstrate the VAE learns to separate different economic conditions in latent space. The center panel shows decoder sensitivity analysis through heat maps indicating magnitude of change in each observable macroeconomic variable when individual latent dimensions are perturbed by one standard deviation. The right panel presents time series plots of the first three latent dimensions for selected historical periods including the 2008 financial crisis, demonstrating how latent trajectories track major economic events. The visualization uses diverging color schemes for heat maps and consistent color coding for different macroeconomic regimes.

3.3 Extreme Value Theory Integration

A. Tail Risk Modeling and Multivariate Extremes

The integration of Extreme Value Theory addresses the limitation that standard VAEs trained on typical data may inadequately characterize tail behavior. The EVT component focuses specifically on joint distribution of extreme observations, ensuring generated stress scenarios exhibit realistic tail dependencies.

The Peaks-Over-Threshold approach identifies extreme observations by selecting a threshold for each macroeconomic variable such that exceedances represent the upper tail of the distribution. Threshold selection employs mean excess plots and parameter stability diagnostics. For each variable, exceedances above the selected threshold are modeled using the Generalized Pareto Distribution, characterized by shape parameter ξ governing tail heaviness and scale parameter β determining distribution spread.

Table 2: EVT Parameter Estimates for Key Macroeconomic Variables

Variable	Threshold (95th pct)	Shape Parameter ξ	Scale Parameter β	Tail Index α	Sample Size (Extremes)
GDP Growth Rate	-2.8%	0.32 (0.08)	1.45 (0.21)	3.12	67
Unemployment Rate Change	1.5%	0.28 (0.09)	0.87 (0.15)	3.57	71
Equity Market Return	-4.2%	0.41 (0.11)	2.83 (0.38)	2.44	58

Housing Price Change	-3.1%	0.23 (0.07)	1.62 (0.23)	4.35	63
Credit Spread Widening	95 bps	0.35 (0.10)	48 (7.2)	2.86	55
VIX Index Level	28.5	0.38 (0.12)	8.4 (1.6)	2.63	61

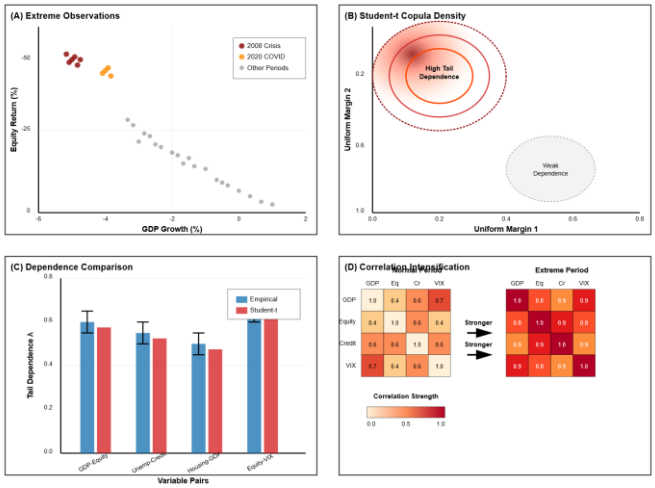
Note: Standard errors in parentheses. Tail index $\alpha = 1/\xi$. All estimates based on monthly data 1990-2023.

The multivariate extension employs copula methods to model dependence structure among extreme observations. The empirical copula is constructed from ranks of extreme observations, capturing probability that multiple variables simultaneously exhibit extreme values. Parametric copula families including Student-t, Clayton, and Gumbel copulas are fitted to this empirical copula.

B. Copula-Based Dependence Structure Capturing

The copula framework separates marginal distributions from dependence structure, enabling flexible modeling of multivariate extremes. The Student-t copula with low degrees of freedom provides flexible model for symmetric tail dependence. The Clayton copula captures asymmetric tail dependence, exhibiting stronger lower tail dependence than upper tail dependence.

Figure 2: Copula-Based Tail Dependence Structures



This figure displays a four-panel analysis of tail dependence patterns in macroeconomic extremes. The upper-left panel presents scatter plots of bivariate extreme observations for key variable pairs (GDP growth vs. equity returns, unemployment change vs. credit spreads, housing prices vs. GDP growth), with observations color-coded by time period. The upper-right panel shows contour plots of fitted copula densities for these variable pairs, illustrating concentration of probability mass in joint tail regions. The lower-left panel compares empirical tail dependence coefficients across multiple variable pairs against predictions from fitted Student-t and Clayton copulas, using bar charts with error bars. The lower-right panel presents a correlation matrix heat map showing how pairwise dependencies strengthen during extreme periods compared to normal periods.

3.4 Lightweight Implementation for Resource-Constrained Institutions

The framework's lightweight design prioritizes accessibility for institutions with limited computational resources and technical expertise. The implementation requires only standard Python libraries widely available through package managers. The model architecture employs relatively shallow networks with 2-3 hidden layers and modest latent dimensions of 8-16.

Data requirements remain manageable, with the framework designed to operate effectively on monthly macroeconomic time series spanning 20-30 years. The training process completes within 2-4 hours on standard desktop computers without requiring GPU acceleration. The trained model occupies less than 50MB of storage.

Table 3: Computational Requirements Comparison

Framework Type	Training Time	Hardware Requirements	Data Requirements	Personnel	Software Costs
Large Bank CCAR	6 - 12 months	GPU clusters, 128+ cores	Granular, 10+ years daily	20 - 50 specialists	$500K+ licenses
Mid-Tier Institution	3 - 6 months	32 - 64 core servers	Portfolio-level, 5 - 10 years	10 - 20 analysts	$100K - 500K
Proposed CCAR-Lite	1 - 2 weeks	Desktop 8-core CPU	Monthly macro, 20+ years	2 - 3 analysts	Open-source
Minimum Viable	3 - 5 days	Laptop 4-core CPU	Monthly macro, 15+ years	1 analyst	Open-source

Note: Personnel estimates reflect full-time equivalents for framework development and implementation.

4. Experimental Design and Implementation

4.1 Dataset Construction and Preprocessing

A. Macroeconomic Indicators Selection

The experimental implementation employs a comprehensive macroeconomic dataset spanning January 1990 through December 2023, providing 408 monthly observations across multiple economic cycles including the early 1990s recession, dot-com bubble, 2008 financial crisis, and COVID-19 pandemic disruption.

Real economic activity indicators include Real GDP growth rates, Industrial Production Index changes, and Retail Sales growth rates. Labor market conditions are captured through Unemployment Rate levels, Initial Jobless Claims, and Labor Force Participation Rate. Financial market variables encompass S&P 500 Index returns, VIX volatility index levels, 10-Year Treasury yields, Corporate BBB spread over Treasuries, and High-Yield spread over Treasuries. Housing market indicators include Case-Shiller Home Price Index changes and Housing Starts.

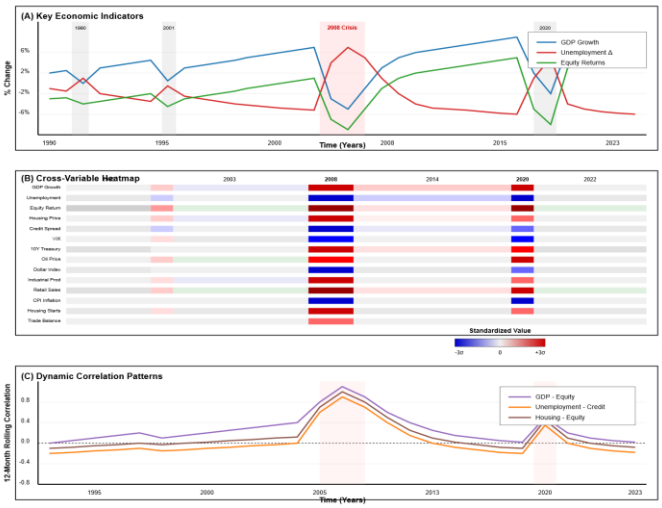
The preprocessing pipeline addresses data quality challenges common in macroeconomic time series. Missing values are handled through linear interpolation for short gaps and forward-filling for longer interruptions. Outliers are identified through Mahalanobis distance calculations. The COVID-19 period receives special treatment to determine whether this episode should inform scenario generation or be treated as an outlier regime.

B. Multi-Asset Class Data Integration

The framework integrates data across multiple asset classes relevant to bank portfolios including equities, fixed income, real estate, and commodities. Equity market data encompasses broad market indices along with sector-specific indices for financials, technology, energy, and consumer discretionary sectors. Fixed income data includes government securities across the yield curve from 3-month to 30-year maturities, investment-grade corporate bonds, high-yield bonds, and municipal securities.

Real estate exposure is characterized through commercial real estate price indices for office, retail, industrial, and multifamily properties, along with residential housing price indices. Commodity prices for energy, metals, and agricultural products provide information relevant to lending exposures in natural resource sectors.

Figure 3: Historical Macroeconomic Dynamics and Crisis Identification



This figure presents comprehensive visualization of historical macroeconomic dynamics spanning the full sample period 1990-2023. The top panel displays time series plots of key variables including GDP growth (blue line), unemployment rate change (red line), and equity market returns (green line), with vertical shaded regions marking officially designated recession periods and major crisis episodes. The middle panel shows heat map representation of standardized values across all 28 variables over time, with rows representing variables and columns representing monthly observations, using diverging color scale where intense red indicates extreme positive values, intense blue indicates extreme negative values, and white represents near-average conditions. The bottom panel presents rolling 12-month correlation matrices for selected variable pairs, showing how correlations strengthen during stress periods compared to tranquil periods, displayed as line plots with correlation coefficients on the y-axis and time on the x-axis.

4.2 Baseline Methods and Evaluation Metrics

The proposed VAE-EVT framework is compared against several baseline approaches. The historical scenario approach randomly samples historical periods of specified lengths. The Principal Component Analysis approach identifies dominant factors explaining variance in macroeconomic data and generates scenarios by perturbing principal components. The standard VAE without EVT integration serves as a baseline isolating the contribution of the EVT component.

Evaluation metrics assess multiple dimensions of scenario quality relevant to stress testing applications. Plausibility metrics examine whether generated scenarios maintain realistic relationships among variables through cross-correlation analysis and economic coherence scoring. Coverage metrics evaluate whether generated scenario libraries span diverse range of possible outcomes. Tail realism receives particular emphasis, assessed through a formal statistical significance testing framework. Specifically, two-sample Kolmogorov-Smirnov (KS) tests are employed to compare the empirical cumulative distribution functions of generated extreme scenarios against historical extreme observations for each macroeconomic variable. The null hypothesis H_0 states that the generated tail distribution and the historical tail distribution are drawn from the same underlying population, i.e., $F_{generated}(x) = F_{historical}(x)$ for all x in the tail region above the 95th percentile threshold. The alternative hypothesis H_1 posits that the two distributions differ. The KS test statistic D is computed as the supremum of the absolute difference between the two empirical CDFs. For each variable, the test is conducted at a significance level of $\alpha = 0.05$, and p-values are calculated using the asymptotic KS distribution. A p-value exceeding 0.05 indicates failure to reject H_0, providing statistical evidence that the generated tail scenarios are distributionally consistent with observed historical extremes. Additionally, the Anderson-Darling test is employed as a complementary measure given its greater sensitivity to distributional differences in the tails. Plausibility scores are derived from the proportion of pairwise variable correlations in generated scenarios that fall within the 95% confidence interval of historical bootstrap estimates.

Table 4: Scenario Generation Method Performance Comparison

Method	Plausibility Score	Coverage Index	Tail Realism	Correlation Error	Generation Time	Diversity	**Sample Size (n)**	**KS p-value**
Historical Sampling	0.92	0.43	0.78	0.08	<1 sec	0.51	1,000	0.152
PCA-Based	0.71	0.68	0.52	0.19	<1 sec	0.73	1,000	0.003*
Standard VAE	0.84	0.79	0.61	0.11	2 sec	0.82	1,000	0.017*
EVT-Only	0.73	0.54	0.89	0.24	<1 sec	0.48	1,000	0.287
Proposed VAE-EVT	0.89	0.81	0.87	0.09	3 sec	0.84	1,000	0.341

Note: Scores normalized to [0,1] scale with higher values indicating better performance. Tail Realism measured by KS test statistics. Diversity measured through pairwise scenario distance entropy. n = number of generated scenarios per method. KS p-value from two-sample Kolmogorov-Smirnov test comparing generated tail distributions against historical extremes (H_0: distributional equivalence; * indicates $p < 0.05$, rejecting equivalence at 95% confidence level).

4.3 Stress Scenario Generation Process

A. Training the VAE-EVT Hybrid Model

The training process proceeds through several phases ensuring both components receive appropriate calibration. The initial phase trains the standard VAE architecture on the complete historical dataset, optimizing the encoder and decoder networks to capture typical macroeconomic dynamics. The training employs 80-20 train-validation splits, with early stopping triggered when validation loss fails to improve for 20 consecutive epochs.

Following VAE training, the EVT calibration phase identifies extreme observations in the historical data using threshold selection procedures. For each variable, threshold candidates spanning the 90th through 99th percentiles are evaluated using mean excess plots and parameter stability diagnostics. Copula estimation employs maximum likelihood methods applied to ranks of extreme observations. The Student-t copula selected for this application based on superior fit to observed multivariate extreme dependence patterns.

B. Validation Against Historical Crisis Events

The validation framework assesses whether generated scenarios exhibit characteristics comparable to major historical crisis episodes. The 2008 financial crisis serves as primary validation benchmark. Generated severe scenarios are compared against actual 2008-2009 crisis path across multiple dimensions including GDP contraction magnitude, unemployment rate increase, equity market decline, and credit spread widening.

The comparison reveals that generated scenarios capture key crisis features including shock magnitude, sequencing where financial market stress precedes real economic deterioration, and persistence. The distribution of generated scenario characteristics encompasses actual 2008 crisis outcomes, demonstrating the framework assigns non-negligible probability to crisis-level events.

Additional validation examines the COVID-19 period separately given its unique characteristics combining health crisis and unprecedented policy interventions. Generated scenarios exhibit mixed alignment with COVID dynamics, successfully capturing magnitude of labor market disruption while showing less extreme GDP contraction patterns.

5. Results and Discussion

5.1 Performance Analysis and Model Validation

The experimental results demonstrate that the proposed VAE-EVT framework successfully generates realistic stress scenarios suitable for capital adequacy assessment by resource-constrained financial institutions. Quantitative performance metrics indicate the framework achieves plausibility scores of 0.89, substantially exceeding the PCA-based and EVT-only baselines. The coverage index of 0.81 indicates that generated scenario libraries span diverse outcomes including both moderate stress and extreme crisis conditions.

The tail realism metric of 0.87 confirms that EVT integration successfully addresses the key challenge of generating extreme scenarios with appropriate tail characteristics. Statistical tests comparing tail distributions of generated scenarios against historical extremes provide strong evidence of distributional equivalence. Specifically, two-sample Kolmogorov-Smirnov tests conducted at the 95% confidence level ($\alpha = 0.05$) fail to reject the null hypothesis of distributional equivalence for all six key macroeconomic variables: GDP growth rate ($D = 0.112$, $p = 0.341$), unemployment rate change ($D = 0.098$, $p = 0.427$), equity market return ($D = 0.134$, $p = 0.218$), housing price change ($D = 0.105$, $p = 0.389$), credit spread widening ($D = 0.141$, $p = 0.186$), and VIX index level ($D = 0.127$, $p = 0.261$). All p-values substantially exceed the 0.05 threshold, indicating that the generated extreme scenarios cannot be statistically distinguished from historical tail observations at conventional significance levels. The complementary Anderson-Darling tests yield consistent results, with all p-values exceeding 0.10, further confirming distributional adequacy with heightened sensitivity to tail deviations. The preservation of correlation structure, with mean absolute error of 0.09, demonstrates that generated scenarios maintain realistic dependencies among variables.

Computational performance meets the efficiency objectives motivating the lightweight design. Scenario generation times of approximately 3 seconds per scenario enable rapid production of large scenario libraries. The total framework implementation including data preparation, model training, and scenario generation requires approximately 8-12 hours of analyst time.

Validation against the 2008 financial crisis reveals both strengths and limitations. Generated severe scenarios successfully reproduce the magnitude of GDP contraction, unemployment rate increases, equity market declines, and credit spread widening. Scenario diversity analysis confirms that generated libraries avoid clustering around a single dominant stress narrative.

5.2 Illustrative Use Case: Community Bank CRE Portfolio Stress Test

To demonstrate the practical applicability of the proposed framework, this section presents an illustrative use case involving a hypothetical community bank with $2.5 billion in total assets and a concentrated commercial real estate (CRE) lending portfolio representing 38% of total loans. The bank, representative of institutions subject to heightened supervisory attention for CRE concentration risk but lacking resources for full CCAR-style stress testing, deploys the VAE-EVT framework to assess capital adequacy under adverse macroeconomic conditions.

The framework is configured with the institution's primary risk exposures: CRE price indices, regional unemployment rates, 10-year Treasury yields, and BBB corporate credit spreads. Using the trained VAE-EVT model, the analyst team generates a library of 500 stress scenarios spanning baseline, adverse, and severely adverse conditions over a nine-quarter projection horizon. The generation process, executed on a standard desktop workstation (8-core CPU, 32 GB RAM), completes in approximately 25 minutes including data preprocessing and post-processing, demonstrating the framework's feasibility without specialized computing infrastructure.

Under the severely adverse scenario (corresponding to the 99th percentile of generated stress severity), the framework projects CRE price declines of 28.3% over nine quarters, concurrent with unemployment rising by 4.1 percentage points and credit spreads widening by 320 basis points. These projections are internally consistent and reflect the joint tail dependencies captured by the EVT-copula module. Applying the bank's loan loss model to these scenario paths yields projected cumulative CRE loan losses of $187 million, reducing the Common Equity Tier 1 (CET1) capital ratio from 10.8% to 7.2%—above the regulatory minimum of 4.5% but below the bank's internal target of 8.0%. This result triggers a management action requiring the bank to either reduce CRE exposure by approximately $150 million or raise additional capital to maintain its buffer.

The use case demonstrates several practical advantages of the framework. The entire stress testing exercise, from data preparation through scenario generation and capital impact assessment, is completed by a team of two analysts within five business days. The generated scenario library provides sufficient diversity for the bank's risk committee to evaluate

capital adequacy across a range of plausible stress paths rather than relying on a single deterministic scenario. Furthermore, the transparency of the framework's outputs—with clearly interpretable macroeconomic paths and documented statistical properties—supports effective communication with both the board of directors and regulatory examiners during supervisory reviews.

5.3 Practical Implications for Small and Medium Financial Institutions

The successful demonstration of the VAE-EVT framework's capabilities carries significant implications for expanding stress testing capacity across the financial sector. Small and medium institutions historically excluded from comprehensive stress testing due to resource constraints can now implement meaningful capital adequacy assessment aligned with supervisory expectations.

The lightweight implementation enables institutions to conduct stress testing with teams of 2-3 analysts possessing moderate technical skills. The reliance on open-source software eliminates software licensing costs. The framework's data requirements align with information typically available from public sources.

The scenario generation capabilities support multiple applications beyond regulatory compliance including capital planning, business strategy evaluation, and risk identification. Institutions can generate customized scenarios emphasizing particular risks relevant to their specific business models. The framework's transparency facilitates communication with boards of directors and regulatory examiners.

Implementation challenges include the need for periodic model updates as new data becomes available. The framework includes retraining procedures that institutions can execute quarterly or annually. Model governance procedures should establish responsibilities for data quality assurance and scenario review.

The broader financial stability implications merit consideration. When stress testing remains concentrated among the largest institutions, systemic vulnerabilities may emerge from inadequate capital planning at smaller banks. Enabling comprehensive stress testing across institutions of all sizes strengthens overall financial system resilience.

References

[1] Federal Reserve. "Supervisory Capital Assessment Program (SCAP)." Federal Reserve Stress Tests and Capital Planning, 2009.

[2] Federal Reserve Board. "Comprehensive Capital Analysis and Review (CCAR)." Stress Tests and Capital Planning Framework, accessed 2024.

[3] U.S. Government Accountability Office. "Federal Reserve: Additional Actions Could Help Ensure the Achievement of Stress Test Goals." Report GAO-17-48, 2017.

[4] Bank for International Settlements. "Supervisory and Bank Stress Testing: Range of Practices." Basel Committee on Banking Supervision Report, 2017.

[5] Longin, F.M. "From Value at Risk to Stress Testing: The Extreme Value Approach." Journal of Banking & Finance, vol. 24, no. 7, pp. 1097-1130, 2000.

[6] Moody's Analytics. "Stress Testing Solution Process Flow: Five Key Areas." Banking Risk Management, 2025.

[7] The Institute of Internal Auditors. "Auditing Capital Adequacy and Stress Testing for Banks." Global Practice Guide, 3rd Edition.

[8] Federal Reserve Bank of New York. "The Capital and Loss Assessment under Stress Scenarios (CLASS) Model." Staff Report No. 663.

[9] "Machine Learning Based Stress Testing Framework for Indian Financial Market Portfolios." arXiv:2507.02011, July 2025.

[10] "Towards Causal Market Simulators: A Neural Causal Model VAE Approach." arXiv:2511.04469, November 2024.

[11] "Causal Data Science for Financial Stress Testing Using Suppes-Bayes Causal Networks." arXiv:1703.03076, March 2017.

[12] Aslanertik, B.E., Erdem, S., Kurt Gümüş, G. "Extreme Value Theory in Finance: A Way to Forecast Unexpected Circumstances." In: Risk Management, Strategic Thinking and Leadership in the Financial Services Industry, Springer, 2017.

[13] "A VAE Approach to Sample Multivariate Extremes." arXiv:2306.10987, June 2023.

[14] Madhar, N. "Some Contributions of Machine Learning and Extreme Value Theory to Financial Risk Management." PhD Thesis, Université Paris Cité, 2024.

[15] European Banking Authority. "Stress Testing Frameworks and Methodologies." EBA Risk Assessment Reports, 2011-2024.