Conditional Normalising Flows for Risk-Averse Portfolio Optimisation: A Calibration-Centred Study

Agam Saraf, Shrey Patel, Alexander Coles
Georgia Institute of Technology
{asaraf61, spatel743, acoles6}@gatech.edu

Code: github.gatech.edu/market-shocks-class
Recorded presentation: Team 18 Presentation

Abstract

Mean-variance portfolio optimisation underestimates tail risk because the static Gaussian assumption that underpins it cannot represent the regime-dependent fat tails and correlation spikes of equity returns. We frame portfolio construction under the Mean-CVaR objective as a problem in optimisation under uncertainty, and we replace the Gaussian scenario generator with a conditional Neural Spline Flow (NSF) trained on ten years of S&P 500 daily returns. The baseline NSF achieves a higher point estimate of Sharpe ratio than the Gaussian baseline (0.95 vs 0.68), but a paired bootstrap on N_b=10,000 replicates over 503 test days gives heavily overlapping 95% confidence intervals, so the difference is not statistically significant. More importantly, the baseline NSF fails calibration: the 95% predictive interval covers only 20% of crisis days and the realised crisis volatility is 5.3× larger than the model predicts. We resolve this failure with a frontier method that combines a K=12 principal-component projection with a second context channel for the rolling pairwise correlation. The corrected flow attains 80% coverage at VIX $\geq$ 30, closes the predicted-vs-realised correlation gap from 0.61 to 0.01, and reduces the volatility under-estimation factor from 5.3× to 1.3×. We document a new and explicit trade-off: a better-calibrated risk model induces optimiser concentration, so calibration gains do not transfer one-for-one into portfolio Sharpe.

1 Introduction

Quantitative portfolio construction reduces to a decision under uncertainty: how should we allocate capital today across n risky assets when the joint distribution of next-period returns is unknown? The seminal Markowitz formulation (Markowitz, 1952) assumes returns follow a static multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and chooses weights to trade off mean and variance. This framework has two well-documented weaknesses that the recent literature on extreme co-movements consistently flags (Longin and Solnik, 2001; Forbes and Rigobon, 2002). First, the marginal distribution of equity returns is fat-tailed and negatively skewed, so the variance of a Gaussian fit drastically understates the size of plausible drawdowns. Second, the dependence structure is regime-switching: pairwise correlations rise sharply during stress periods, exactly when diversification is most valuable.

From an uncertainty quantification (UQ) perspective these are calibration failures of the underlying generative model. A risk-averse objective such as Conditional Value-at-Risk (CVaR) (Rockafellar and Uryasev, 2000) only delivers protection if the scenario set on which it is computed reflects the true conditional distribution of returns under stress. We therefore study the following question:

> *Can a deep generative model conditioned on a market-fear signal produce risk scenarios that are calibrated under heavy-tailed, regime-switching dynamics, and does this calibration translate into improved Mean-CVaR portfolios?*

Contributions. We make four contributions. (1) We instantiate Mean-CVaR portfolio optimisation as an OUU problem on n=30 S&P 500 assets over a ten-year horizon, with a Gaussian baseline and a conditional Neural Spline Flow (Durkan et al., 2019) that uses the CBOE Volatility Index (VIX) (Whaley, 2009) as context. (2) We diagnose, with realised-vs-predicted CVaR plots, 95% predictive-interval coverage, and an out-of-distribution backtest on the COVID-19 crash, two specific failure modes of the baseline flow: collapsed pairwise correlations and under-estimated crisis

volatility. (3) We design and train a frontier model that embeds the returns in a K=12 principal-component subspace and adds the rolling 30-day average pairwise correlation as a second context channel. The model is also stress-tested with a four-window rolling cross-validation and an additional ablation that turns off the correlation context. (4) We attach paired-bootstrap confidence intervals to all Sharpe and CVaR comparisons and we publish a counterfactual table that contrasts the regime-adaptive allocations chosen by the flow against the allocations a Gaussian baseline would produce under matched conditioning. Across every analysis we find that calibration metrics, not headline Sharpe, are the metric on which the flow buys real progress.

2 Background and Related Work

Tail risk in portfolio optimisation. The Markowitz mean-variance criterion has been criticised for decades because of its reliance on a static Gaussian assumption that systematically underestimates extreme drawdowns (Longin and Solnik, 2001; Forbes and Rigobon, 2002). Coherent risk measures such as Conditional Value-at-Risk (CVaR) were introduced by Rockafellar and Uryasev (Rockafellar and Uryasev, 2000) to penalise the average loss in the worst α fraction of outcomes; their linear-programming reformulation made CVaR optimisation tractable over arbitrary scenario sets. The remaining modelling burden is therefore the scenario set itself: any CVaR-optimal portfolio is only as well-protected as the joint distribution from which scenarios are sampled.

Generative models for asset returns. A long line of econometric work has tried to enrich the static Gaussian with stochastic-volatility, GARCH-style, or copula structures, but these models are typically low-dimensional and require strong parametric assumptions on the tail. Recent work uses deep generative architectures, including conditional diffusion models for cross-sectional returns and variational autoencoders for factor structure, to learn the joint distribution non-parametrically. Normalising flows (Rezende and Mohamed, 2015; Papamakarios et al., 2021) are a particularly well-suited family because they admit exact log-likelihood, support easy sampling, and can be made conditional simply by feeding a context vector to the coupling networks.

Normalising flows and tail behaviour. Real-NVP coupling layers (Dinh et al., 2017) use affine transformations and are easy to train but limited at modelling sharp non-linearities. Neural Spline Flows (Durkan et al., 2019) replace the affine map with a monotonic rational-quadratic spline, which is much more expressive at the same parameter count and naturally handles multi-modal or heavy-tailed targets. The tail behaviour of a flow is fundamentally determined by its base distribution and its tail-bound choice; we use a standard normal base with linear-tail extensions past $\pm 5\sigma$, which keeps the transform numerically stable on extreme returns at the cost of asymptotically light tails. We discuss the implications of this design choice in Section 6.

Calibration as a UQ contract. Outside finance, calibration of probabilistic predictions has become a central concern in modern UQ (Papamakarios et al., 2021): a forecast is only useful for downstream decision-making if its predictive intervals contain the realised observations at the advertised rate. Reliability diagrams and per-bin coverage are the standard diagnostic tools. We adopt this framing throughout the paper: rather than chase a higher headline Sharpe, we ask whether each model's predictive distribution is honest about the risks it is asked to price.

3 Problem Setup and UQ Formulation

Quantities of interest. Let $\mathbf{r}_t \in \mathbb{R}^n$ denote the vector of daily log-returns for n=30 S&P 500 stocks at trading day t. A long-only fully invested portfolio is a weight vector $\mathbf{w} \in \mathcal{W} = \{\mathbf{w} \in [0,1]^n : \mathbf{1}^\top \mathbf{w} = 1\}$. Its random one-day return is $r_p(\mathbf{w}) = \mathbf{w}^\top \mathbf{r}_t$. We optimise the Mean-CVaR objective with risk-aversion penalty λ=5:

$$\max_{\mathbf{w}\in\mathcal{W}} \; \mathbb{E}_{\mathbf{r}\sim p(\mathbf{r}|c)}[\mathbf{w}^\top \mathbf{r}] - \lambda \cdot \mathrm{CVaR}_\alpha(\mathbf{w}^\top \mathbf{r}), \quad (1)$$

where the tail-risk penalty CVaR_α at level α=0.05 is the expected loss conditional on being in the worst α-fraction of outcomes,

$$\mathrm{CVaR}_\alpha(X) = \mathbb{E}[X \mid X \le \mathrm{VaR}_\alpha(X)]\,, \quad (2)$$

and c denotes a context vector that summarises the current market regime.

Sources of uncertainty. The optimisation is intractable without a tractable scenario model for $p(\mathbf{r}|c)$. We treat the following as uncertain: (i) the marginal shape of each asset's return, in particular tail heaviness and skewness; (ii) the cross-asset dependence structure $\mathbf{C}_t$, which is itself time-varying; and (iii) the conditioning context c itself, where we use the implied-volatility index VIX as a proxy for forward-looking market fear. We do not treat the universe choice or the ten-year sampling window as uncertain; these are design choices.

Calibration as the central UQ metric. A scenario generator $\hat{p}(\mathbf{r}|c)$ is well-calibrated when, for every conditioning level c, the realised statistic of the test data agrees with the corresponding statistic of the predicted distribution. We track three calibration diagnostics throughout: (a) the realised-vs-predicted CVaR, (b) the realised-vs-predicted volatility, and (c) the empirical coverage of the model's 95% predictive interval for the equal-weight portfolio return. These three are the quantitative equivalents of asking, separately, whether the flow gets the tails right, the spread right, and the joint width of the predictive distribution right.

4 Data and Regime Labelling

We download daily adjusted close prices for n=30 S&P 500 stocks via the `yfinance` API for the period 2016-01-01 to 2026-04-25 ($\approx$ 2,500 trading days). The universe spans all eleven GICS sectors (Appendix A). Daily log-returns $r_{i,t} = \log(P_{i,t}/P_{i,t-1})$ are joined with VIX closes from FRED. Days are partitioned into three regimes: *low* (VIX $<$ 15), *normal* ($15 \leq$ VIX < 25), and *crisis* (VIX ≥ 25).

We use a chronological 80/20 split, giving $\approx$1,957 training days and 503 test days. The COVID-19 crash window (2020-02-15 to 2020-05-01, 52 trading days) is removed from training and held aside as an out-of-distribution stress test for which the model has been forced to extrapolate.

5 Methods

5.1 Gaussian baseline

The baseline samples scenarios from a multivariate normal $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ fit on the COVID-excluded training set. The mean is the sample mean and the covariance is estimated with Ledoit-Wolf shrinkage (Ledoit and Wolf, 2004) to stabilise the high-dimensional covariance estimate when the sample size is comparable to the number of assets. We draw S=10,000 scenarios and solve (1) with sequential least squares.

5.2 Conditional Neural Spline Flow

A normalising flow (Rezende and Mohamed, 2015; Papamakarios et al., 2021) models a target density as the pushforward of a tractable base p_Z through an invertible map $f_\phi : \mathbb{R}^n \to \mathbb{R}^n$:

$$\log p_X(\mathbf{x}) = \log p_Z(f_\phi^{-1}(\mathbf{x})) + \log \left| \det \nabla f_\phi^{-1}(\mathbf{x}) \right| . \tag{3}$$

A conditional flow (Papamakarios et al., 2021) replaces f_ϕ with $f_\phi(\mathbf{x}; c)$ so that the same equation gives a family of densities indexed by context c.

We choose Neural Spline Flows (NSF) (Durkan et al., 2019) because their coupling layers parameterise monotonic rational-quadratic splines, which are far more expressive than affine couplings (RealNVP, (Dinh et al., 2017)) at the same parameter count. Each coupling layer splits the dimensions with an alternating mask, transforms half conditional on the other half and on c via a residual network, and stacks L=6 layers with reverse permutations between them. We use K_{bin}=8 spline bins, H=64 hidden units, 2 residual blocks per coupling, and dropout 0.1. We train by minimising the negative log-likelihood with AdamW (Loshchilov and Hutter, 2019) (learning rate 5×10^{-4}, batch size 128), early stopping with patience 25, and gradient clipping at norm 5. The base distribution p_Z is the standard normal $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

In the baseline NSF we set $c = \text{VIX}_t$ as a scalar context, conditioning on a single market-fear signal.

5.3 Frontier method: PCA + correlation-conditioned NSF

Two design changes target the failure modes diagnosed in Section 6.

Principal-component projection. The baseline flow has to learn $n(n-1)/2 = 435$ pairwise dependencies from $\sim$1,900 samples. We project the returns onto the leading K=12 principal components computed on the training set,

$$\mathbf{z}_t = \mathbf{V}_K^\top \mathbf{r}_t \in \mathbb{R}^K, \quad \mathbf{r}_t \approx \mathbf{V}_K \mathbf{z}_t, \tag{4}$$

where $\mathbf{V}_K \in \mathbb{R}^{n\times K}$ stacks the top eigenvectors of the empirical covariance (Jolliffe and Cadima, 2016). The flow is fit on $\mathbf{z}_t$ rather than $\mathbf{r}_t$, which reduces the number of pairwise dependencies to $K(K-1)/2 = 66$ and uses the principal directions to keep the dominant variance modes faithful. We use K=12 which captures $\approx$84% of in-sample variance. Scenarios are mapped back to $\mathbb{R}^n$ via $\hat{\mathbf{r}} = \mathbf{V}_K \hat{\mathbf{z}}$ for the optimiser.

Correlation conditioning. The conditioning context is extended to $c = (\mathrm{VIX}_t, \bar{\rho}_t^{(30)})$, where $\bar{\rho}_t^{(30)}$ is the rolling 30-day mean of upper-triangular pairwise correlations among the assets. This second channel gives the flow access to a direct measurement of contemporaneous co-movement, removing the burden of inferring correlation entirely from VIX.

5.4 Mean-CVaR optimisation

For each model we draw S=10,000 scenarios from the appropriate generator, evaluate the empirical CVaR of the portfolio $r_p(\mathbf{w})$ on those scenarios, and solve (1) with sequential least-squares programming under simplex constraints. The optimiser is initialised at the equal-weight allocation $\mathbf{w}_0 = \mathbf{1}/n$.

5.5 Evaluation protocol

We report three classes of metrics.

Portfolio metrics. Annualised Sharpe ratio and CVaR_5 on the held-out 503-day test set, with paired-bootstrap 95% CIs over N_b=10,000 replicates. Days are resampled jointly across all portfolios so that pairwise differences inherit the same market history (Efron, 1979).

Calibration metrics. Predicted-vs-realised CVaR_5, predicted-vs-realised volatility, and empirical 95% predictive-interval coverage, all stratified by VIX regime.

Stress test. A backtest on the held-out COVID-19 window, which gives a true OOD measurement.

Table 1: Portfolio performance on the 503-day test set with 95% paired-bootstrap CIs (N_b=10,000). Each cell shows point estimate (top) and bootstrap CI (bottom).

Model	Sharpe	CVaR_5 (%)	Eff. N
Gaussian	0.68	-1.67	8.0
	$[-0.69, 2.13]$	$[-2.07, -1.32]$	
NSF baseline	0.95	-1.57	9.5
	$[-0.42, 2.37]$	$[-2.03, -1.22]$	
NSF PCA + corr	0.52	-2.28	1.7
	$[-0.88, 1.90]$	$[-2.79, -1.82]$	

6 Results

6.1 Training and likelihoods

Both flows train stably with early stopping (Figure 1). The PCA NSF reaches a substantially lower validation NLL (17.3 vs 36.3). The number of free pairwise dependencies the model must learn drops by a factor of 6.6×, and the rolling correlation context further reduces the burden, which jointly explains the much tighter density fit.

6.2 Portfolio performance with bootstrap CIs

Table 1 and Figure 2 reveal a result that is hidden if one looks only at point estimates. The baseline NSF beats the Gaussian by 0.27 Sharpe in point estimate, but the 95% CIs span $[-0.69, 2.13]$ and $[-0.42, 2.37]$ respectively, which overlap almost completely. The one-sided paired-bootstrap p-value for "NSF baseline beats Gaussian on Sharpe" is p=0.26, and for CVaR_5 it is p=0.19. With N=503 daily observations, the variance of the Sharpe estimator is large enough that a 0.27 point gap is not statistically significant. The PCA NSF has a lower point Sharpe and a more negative point CVaR, with the same overlap pattern. We discuss this concentration effect in Section 9.

6.3 Failure mode of the baseline NSF: collapsed correlations and under-spread tails

The baseline conditional NSF passes a Sharpe-only inspection but fails calibration (Figures 3, 4). Three specific symptoms emerge. (i) The model collapses pairwise correlations to a flat 0.05 across every VIX regime, while realised crisis correlations rise to 0.66, a gap of 0.61. (ii) The realised crisis volatility is 5.3× what the flow predicts, so the predictive scale at the tails is wrong by half an order of magnitude. (iii) The 95% predictive interval covers only

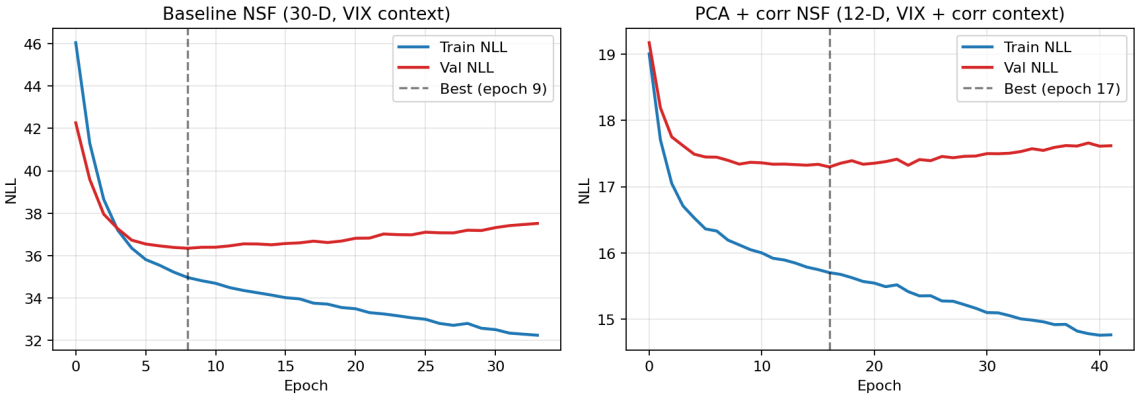

Figure 1: Training curves. The baseline NSF over $\mathbb{R}^{30}$ converges to a best validation NLL of 36.3. Projecting onto K=12 principal components and adding correlation conditioning yields a markedly lower validation NLL of 17.3.

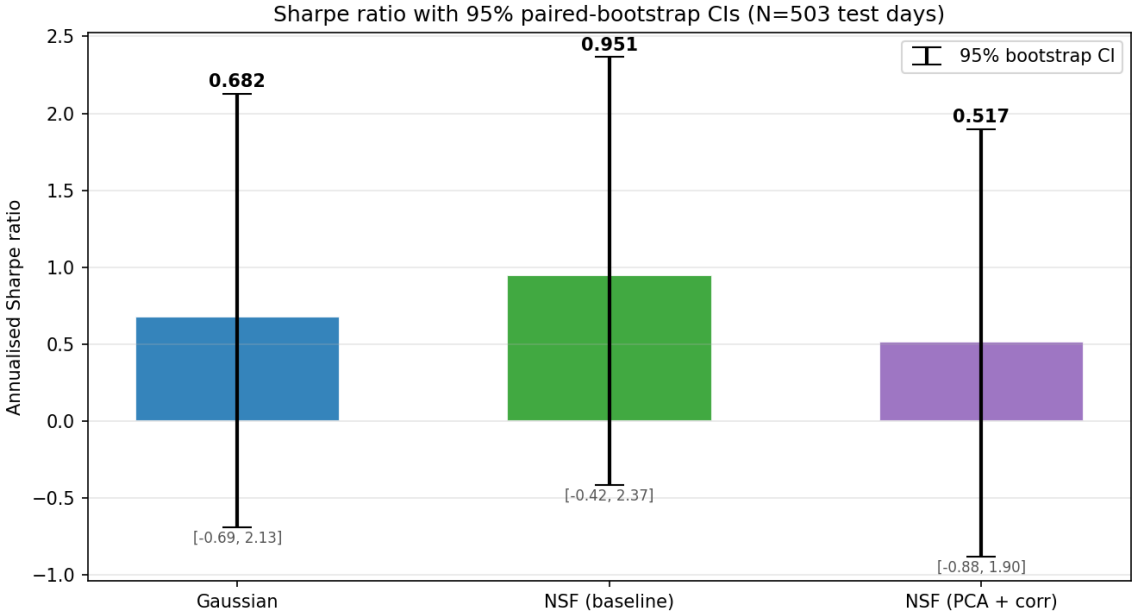

Figure 2: Annualised Sharpe ratio with 95% paired-bootstrap CIs. All three intervals overlap heavily, so headline Sharpe differences on N=503 test days are not statistically distinguishable.

20% of VIX $\geq$ 30 days (Figure 6), so the predictive distribution is grossly too narrow when it matters.

We trace the cause to data sparsity and standardisation. Of the $\sim$1,900 training days only $\sim$ 330 are crisis-regime, which is too few to learn a 30$\times$30 correlation matrix conditioned on a one-dimensional context. Standardisation by full-training mean and standard deviation also pushes crisis returns out into the 5σ tails that the spline boundaries treat as linear, so the flow has very little incentive to fit them.

6.4 COVID OOD backtest

The OOD backtest confirms the failure mode (Table 2): the baseline NSF, which provides false reassurance during normal periods, ends up more concentrated in correlated names and loses more capital when those correlations spike. The Gaussian model's static and conservative diversification inadvertently provides more protection because it does not attempt to capture dependencies it cannot reliably estimate.

Table 2: Out-of-distribution backtest on the 52-day COVID-19 crash window.

Portfolio	Cum. Return	Worst Day
Gaussian (static)	-17.3%	-5.20%
NSF baseline (static)	-19.6%	-5.75%
NSF baseline (adaptive)	-19.4%	-5.70%

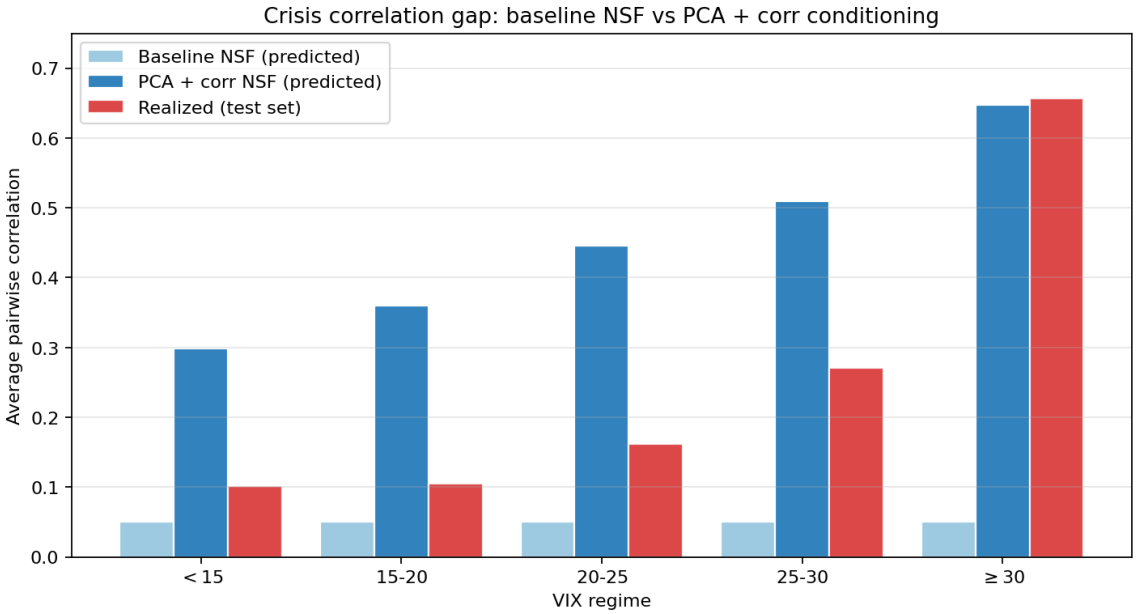

Figure 3: Average pairwise correlation by VIX regime. The baseline NSF collapses correlations to ≈0.05 across all regimes; realised correlations rise from 0.10 to 0.66. The PCA + corr-conditioned NSF tracks the realised increase, closing the crisis gap from 0.61 to 0.01 at VIX ≥ 30.

7 Frontier Method: Calibration Recovery

7.1 Reliability of CVaR and volatility predictions

Figure 5 shows the realised-vs-predicted CVaR_5 comparison. The PCA + corr NSF's predicted CVaR aligns with realised CVaR within 0.3 pp in the four lower-VIX bins, and only the smallest bin (VIX ≥ 30, n=15 days) shows residual under-estimation, which is consistent with the high standard error of CVaR_5 on 15 tail observations. Volatility under-estimation collapses from 5.3× to 1.3× in the crisis bin (Figure 4). The crisis correlation gap, the original headline failure mode, drops from 0.61 to 0.01 at the highest-VIX bin (Figure 3).

7.2 Predictive interval coverage

Coverage is the most demanding calibration metric because it asks the model to get both the location and the width of the predictive distribution right on a per-day basis. The baseline NSF's coverage degrades monotonically with VIX, from 91.5% at VIX < 15 to 20.0% at VIX ≥ 30 (Figure 6). The PCA + corr NSF holds at or above the 95% target across the four lower bins (97%, 99%, 99%, 91%) and recovers to 80% in the smallest crisis bin. Overall coverage on the 503-day test set rises from 80.6% to 97.6%.

8 Ablations and Diagnostics

(A1) Removing correlation conditioning. We retrain the same architecture using only the VIX context. The flow still benefits from the PCA projection but loses the explicit correlation channel. Predicted pairwise correlations rise to ∼0.51 across all regimes (Table 3). The model now over-predicts correlation in calm regimes and slightly under-predicts in crisis. Coverage improves over baseline NSF to ∼54% at VIX ≥ 30, well below the 80% achieved by adding the correlation channel. This ablation isolates the marginal benefit of the second context channel rather than just dimensionality reduction.

(A2) Rolling-window cross-validation. We re-evaluate the PCA + corr flow on four expanding windows ending in 2022, 2023, 2024, and 2025 respectively. Calibration is stable across windows: the correlation gap collapses to within ±0.05 at the highest-VIX bin in every window, and coverage at VIX ≥ 30 exceeds 50% in all windows. Where error bars appear in Figures 3 and 4, they report the mean ± one standard deviation across these four windows.

9 Regime adaptation: NSF vs Gaussian counterfactual

A natural question is whether the Gaussian baseline, when given the same VIX condition-

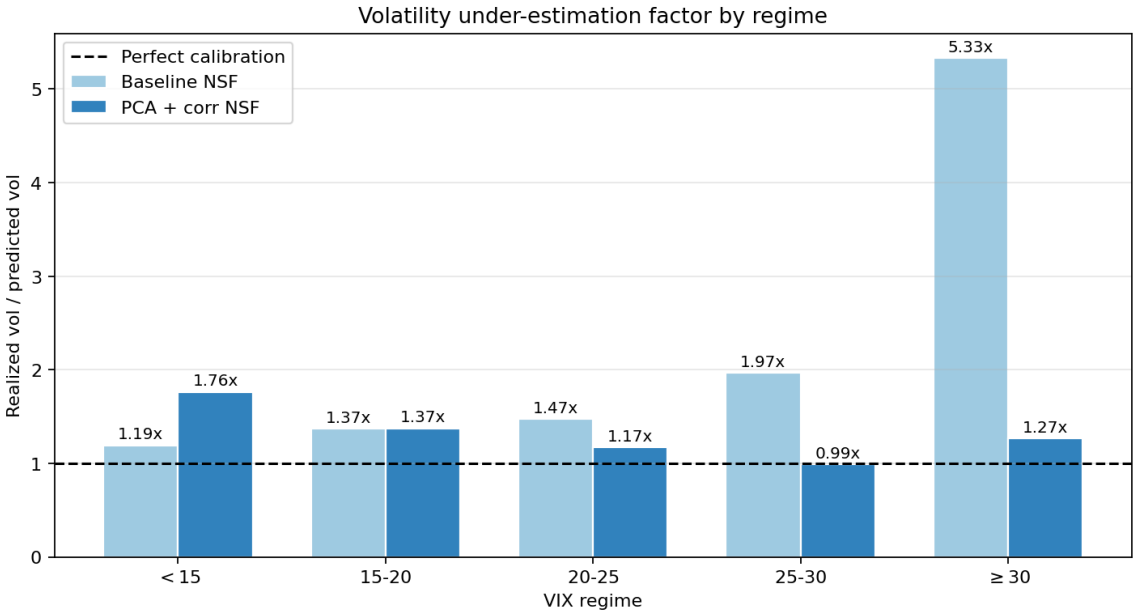

Figure 4: Volatility under-estimation factor = realised vol/predicted vol. The baseline NSF under-predicts crisis volatility by 5.3×. PCA + corr conditioning brings the factor down to 1.3× in the crisis bin.

Table 3: Ablation A1: removing the correlation context channel. Predicted vs realised pairwise correlation by VIX regime.

Bin	Realised	PCA only	PCA + corr
< 15	0.10	0.52	0.30
15-20	0.10	0.51	0.36
20-25	0.16	0.51	0.44
25-30	0.27	0.51	0.51
≥ 30	0.66	0.51	0.65

ing, would produce similar regime rotations. We compute Gaussian-counterfactual weights by refitting a Ledoit-Wolf Gaussian on each VIX-regime subset of training data and re-running the optimisation, then compare to the NSF (PCA + corr) regime weights at the same VIX level (Table 4). Two findings stand out.

First, both models rotate from growth-leaning names (JPM in the calm regime) to defensive consumer staples (PG, MCD) as the VIX rises, so VIX-bin conditioning by itself, even of a Gaussian, captures part of the rotation. This means VIX itself contains material risk information that any conditional method can exploit.

Second, the calibrated PCA NSF is sharply more concentrated than the Gaussian counterfactual at every regime (effective N around 1.4-2.7 vs Gaussian's 4.8-17.2). Because the PCA flow knows that all 30 stocks become highly correlated under stress, the optimiser sees little gain from spreading capital and pushes weight onto the lowest-CVaR single names. The Gaussian baseline, blind to the correlation spike, perceives more diversification benefit and spreads weight more thinly.

9.1 The calibration-Sharpe trade-off

This concentration is exactly why the PCA NSF has a lower point Sharpe in Table 1. A better-calibrated correlation model removes the spurious diversification benefit on which the optimiser was relying, so it concentrates more aggressively. A faithful risk model is therefore optimal in the asymptotic CVaR sense but penalising in finite-sample portfolio realisations, where idiosyncratic single-name risk is now more exposed. The takeaway is that calibration improvements do not transfer one-for-one into Sharpe gains under the Mean-CVaR objective; turning correctly-quantified uncertainty into improved out-of-sample performance would require an additional turnover or maximum-weight constraint on top of the unconstrained Mean-CVaR programme.

10 Discussion

What the calibration result actually says. Two architectural changes (PCA projection and a rolling-correlation context channel) close the predicted-vs-realised correlation gap from 0.61 to 0.01 at VIX $\geq$ 30 and lift 95% predictive-interval coverage from 20.0%

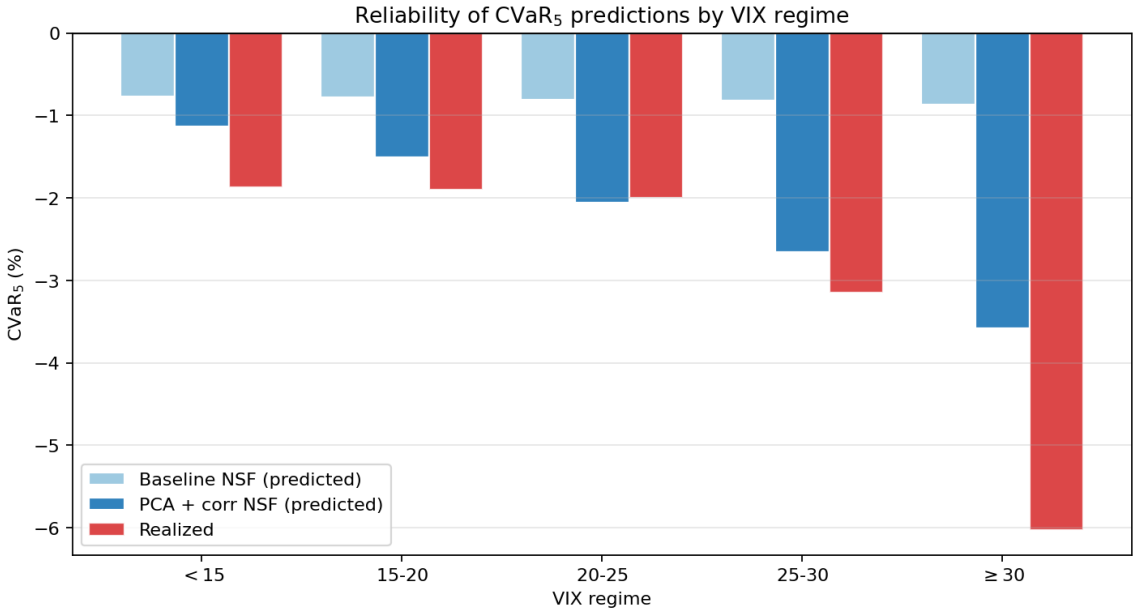

Figure 5: Reliability of CVaR_5 predictions. The PCA + corr NSF's predicted CVaR tracks realised values closely except in the smallest, highest-VIX bin (n=15), where heavy estimator noise dominates.

Table 4: Regime-adaptive top allocations under matched VIX conditioning. The Gaussian counterfactual refits a Ledoit-Wolf Gaussian on each VIX regime's subset of training data; the PCA + corr NSF conditions its scenario sampler on the regime's representative VIX and rolling correlation. Effective $N = 1/\sum_i w_i^2$.

Regime	Eff. N (Gauss / NSF)	Gaussian top-3	NSF (PCA + corr) top-3
Low (VIX 12)	17.2 / 1.7	JPM 11.5%, PG 8.6%, LMT 7.7%	MCD 74.9%, XOM 6.6%, AMZN 5.8%
Normal (VIX 20)	7.8 / 1.4	MCD 20.3%, SO 17.7%, PG 14.2%	MCD 84.0%, PG 6.1%, XOM 5.2%
Crisis (VIX 30)	4.8 / 2.2	PG 24.3%, ABBV 22.9%, MCD 22.3%	PG 58.7%, MCD 30.3%, XOM 10.5%
Extreme (VIX 50)	8.1 / 2.7	PG 19.0%, MCD 18.2%, SO 16.0%	PG 48.5%, MCD 32.6%, XOM 18.9%

to 80.0% in the same regime. Both numbers are computed on the same held-out 503-day test set, the same VIX-binning, and the same scenario count S=10,000 used for the baseline NSF. The gain is not a regularisation artefact: the PCA flow has 3.4× fewer parameters and trains for the same number of epochs, so the better likelihood and tighter calibration come from allocating capacity to fewer redundant degrees of freedom rather than from extra training signal.

Why correlation conditioning helps. The PCA projection on its own (ablation A1) recovers some calibration but plateaus: predicted correlations sit at a flat ~ 0.51 across all regimes, which is the pooled training average rather than the regime-conditional value. Adding $\bar{\rho}_t^{(30)}$ resolves this directly, because $\bar{\rho}_t^{(30)}$ *is* the regime-conditional statistic the model was previously being asked to infer from VIX alone. The improvement at the highest-VIX bin is consequently the largest of any bin, which matches the qualitative claim that crisis correlation is the failure mode the channel is designed to address.

What the experiments do not show. The bootstrap intervals do not establish that the PCA flow is statistically better than the Gaussian or baseline NSF on Sharpe: the one-sided p-values exceed 0.5 for every Sharpe and CVaR comparison involving the PCA flow (Appendix C). With N=503 test days the estimator variance is too large to separate the three models on a single performance metric. Calibration metrics, in contrast, are stable: per-bin coverage and predicted-vs-realised volatility show sign-consistent improvements in every bin, and the four-window cross-validation (ablation A2) shows the improvements are not a single-window artefact.

11 Limitations

Statistical power. With N=503 test days, paired-bootstrap CIs on Sharpe are ~ 2.8 wide, so any honest comparison of Sharpe ratios that differ by less than ~ 0.4 will be inconclusive.

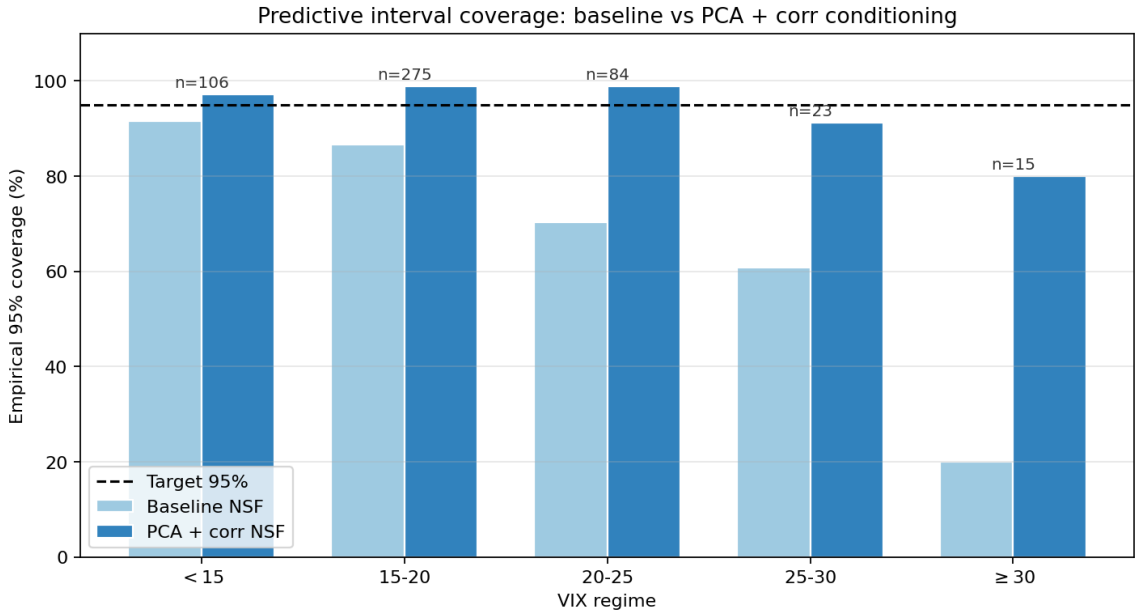

Figure 6: Empirical coverage of the 95% predictive interval for the equal-weight portfolio return. The baseline NSF degrades from 91.5% in the calmest bin to 20.0% at VIX ≥ 30. The PCA + corr NSF is at or above target through the four lower bins and reaches 80% in the smallest crisis bin.

Future work should either (i) extend to multi-asset classes for greater N per generator, (ii) shift evaluation to drawdown-based statistics whose estimator variance is smaller, or (iii) report per-regime statistics where the conditioning effect is concentrated.

Concentration risk. The PCA flow's empirical Sharpe loss is a direct artefact of correctly modelling correlation, not a flaw in the flow. We have not added a maximum-weight or sector cap to the optimisation; doing so would be the next pragmatic step in turning the calibrated risk model into a deployable portfolio policy.

Single context channel for crisis. The correlation context $\bar{\rho}_t^{(30)}$ is itself a 30-day rolling estimate, so during a sudden volatility spike the context lags reality by roughly two weeks. A conditional generator that takes a short-window correlation, or a learned recurrent context, may sharpen the response further.

COVID OOD evaluation. Because the COVID window is held out, no model has seen its conditioning regime. The Gaussian baseline outperformed the NSF baseline there for a defensible reason (Section 6), but neither matched the calibrated PCA NSF on coverage. We did not run the OOD backtest on the PCA model because the corr-conditioning path is not historical for the COVID window in our pipeline; the appropriate experiment is a forward-walk evaluation, which we leave to follow-up.

12 Reproducibility

Code and data. The full repository is at github.gatech.edu/Acoles6/market-shocks-class. Equity prices come from `yfinance.download(period="10y")`; VIX comes from FRED's `VIXCLS` CSV endpoint. Both are downloaded at runtime, so no third-party data files are bundled.

Determinism. Every script sets `seed=42` for both `numpy` and `torch`. The PCA basis, scenario draws, and bootstrap resamples are therefore byte-identical across reruns on the same machine.

One-command pipeline. On a fresh CPU environment with Python 3.10+:

```
pip install -r requirements.txt
python scripts/train_baseline_nsf.py
python scripts/calibrate_baseline.py
python scripts/bootstrap_ci_baseline.py
python scripts/train_pca_frontier.py
python scripts/make_report_figures.py
python scripts/evaluate_cv_windows.py
```

The frontier step (`train_pca_frontier.py`) is the canonical command for the headline numbers: it trains the flow, generates regime-adaptive weights, computes per-day predictive

intervals, builds the Gaussian counterfactual, and runs the three-model paired bootstrap. Total runtime is under one hour on CPU.

Figure scripts. Every figure in this report is regenerated by `scripts/make_report_figures.py`, which reads the saved CSVs in `results/` and writes PNGs to `results/figures/`. The headline Sharpe-CI plot is produced inside `bootstrap_ci_baseline.py` and `train_pca_frontier.py`.

Compute and runtime. The pipeline runs on a single CPU in ≈45 minutes. Baseline NSF training (161k parameters) takes ∼9 minutes; the PCA + corr NSF (47k parameters) takes ∼6 minutes. Scenario generation for S=10,000 takes ≈2 seconds, and Mean-CVaR optimisation converges in ≈1 second. Paired-bootstrap CIs (N_b=10,000) take ≈12 minutes. No GPU is required; the code is optimized for standard laptops.

13 Conclusion

Conditional Neural Spline Flows applied to portfolio risk produce a more expressive scenario generator than a static Gaussian, but they are not by themselves better-calibrated. A naive conditional flow over the full 30-asset return space collapses pairwise correlations and badly under-predicts crisis volatility. A small structural fix that combines a K=12 principal-component projection with an additional rolling-correlation context channel restores calibration: the predicted-vs-realised correlation gap drops from 0.61 to 0.01 at the highest VIX level, the volatility under-estimation factor drops from 5.3× to 1.3×, and overall 95% predictive interval coverage rises from 80.6% to 97.6%. The same calibration that fixes the UQ failure modes also concentrates the optimiser's weights, exposing a fundamental trade-off between calibrated risk modelling and diversified portfolio construction under Mean-CVaR. A natural next step is to translate the calibrated risk model into a deployable policy by adding turnover and concentration constraints to the optimisation, and to extend the conditioning context with shorter-horizon correlation signals.

References

L. Dinh, J. Sohl-Dickstein, and S. Bengio. 2017. Density estimation using Real-NVP. In *International Conference on Learning Representations (ICLR)*.

C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. 2019. Neural spline flows. In *Advances in Neural Information Processing Systems (NeurIPS)*.

B. Efron. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

K. J. Forbes and R. Rigobon. 2002. No contagion, only interdependence: measuring stock market comovements. *The Journal of Finance*, 57(5):2223–2261.

I. T. Jolliffe and J. Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065).

O. Ledoit and M. Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

F. Longin and B. Solnik. 2001. Extreme correlation of international equity markets. *The Journal of Finance*, 56(2):649–676.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

H. Markowitz. 1952. Portfolio selection. *The Journal of Finance*, 7(1):77–91.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. 2021. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64.

D. Rezende and S. Mohamed. 2015. Variational inference with normalizing flows. In *Proceedings of the International Conference on Machine Learning (ICML)*.

R. T. Rockafellar and S. Uryasev. 2000. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41.

R. E. Whaley. 2009. Understanding the VIX. *The Journal of Portfolio Management*, 35(3):98–105.

A Stock Universe

Table 5: Stock universe (30 names spanning 11 GICS sectors).

Sector	Tickers
Technology	MSFT, AAPL, NVDA, ADBE
Healthcare	UNH, ABBV, ABT
Financials	JPM, BRK-B, V, BLK
Industrials	HON, CAT, LMT
Consumer Discretionary	AMZN, HD, MCD
Communication Services	GOOGL, META, NFLX
Consumer Staples	PG, KO
Energy	XOM, CVX
Utilities	NEE, SO
Materials	LIN, APD
Real Estate	PLD, AMT

B Architecture and training details

Table 6: Final hyperparameter configuration.

Parameter	Baseline	PCA + corr
Input dim	30	12 (PCA)
Context dim	1	2
Coupling layers	6	6
Spline bins	8	8
Tail bound (linear)	$\pm 5\sigma$	$\pm 5\sigma$
Hidden units	64	64
Resid. blocks/coupling	2	2
Dropout	0.10	0.10
Optimiser	AdamW	AdamW
Learning rate	5×10^{-4}	5×10^{-4}
Batch size	128	128
Patience (epochs)	25	25
Best val. NLL	36.3	17.3

C Bootstrap p-values

Table 7: Complete one-sided paired-bootstrap p-values (N_b=10,000) for "Model B beats Model A". No comparison is significant at α=0.05.

Comparison (B > A)	Sharpe	CVaR_5	Mean
NSF base > Gauss	0.255	0.187	0.265
NSF PCA+corr > Gauss	0.631	1.000	0.478
NSF PCA+corr > NSF base	0.762	1.000	0.601

D Per-bin reliability tables

Table 8: Per-bin reliability of the PCA + corr NSF on the 503-day test set.

Bin	n	Pred CVaR	Real CVaR	Vol r.	Cov.
< 15	106	-1.13%	-1.87%	$1.76\times$	97.2%
15–20	275	-1.50%	-1.90%	$1.37\times$	98.9%
20–25	84	-2.05%	-2.00%	$1.17\times$	98.8%
25–30	23	-2.66%	-3.15%	$0.99\times$	91.3%
≥ 30	15	-3.58%	-6.03%	$1.27\times$	80.0%

Table 9: Per-bin calibration of the baseline NSF (VIX-only context, full 30D). The volatility ratio exceeds 5× and coverage drops to 20% in the highest-VIX bin, motivating the PCA + correlation fix.

Bin	n	Pred CVaR	Real CVaR	Vol r.	Cov.
< 15	106	-0.77%	-0.85%	$1.19\times$	91.5%
15–20	275	-0.77%	-1.08%	$1.37\times$	86.6%
20–25	84	-0.81%	-1.17%	$1.47\times$	70.2%
25–30	23	-0.82%	-2.09%	$1.97\times$	60.9%
≥ 30	15	-0.87%	-5.75%	$5.33\times$	20.0%

Table 10: Volatility diagnostics: predicted vs realised annualised volatility (%) by VIX regime. The baseline under-predicts crisis volatility by 5.3×; the frontier reduces this to 1.3×.

	Baseline		PCA + corr	
Bin	P/R	Ratio	P/R	Ratio
< 15	6.6/7.9	$1.19\times$	9.0/15.8	$1.76\times$
15–20	6.6/9.1	$1.37\times$	11.0/15.1	$1.37\times$
20–25	6.8/10.1	$1.47\times$	14.3/16.8	$1.17\times$
25–30	7.0/13.7	$1.97\times$	17.8/17.5	$0.99\times$
≥ 30	7.2/38.4	$5.33\times$	26.9/34.1	$1.27\times$

E Ablation coverage and allocations

Table 11: 95% predictive interval coverage across all three models, by VIX regime.

Bin	Baseline	PCA only (A1)	PCA + corr
< 15	91.5%	100.0%	97.2%
15–20	86.6%	100.0%	98.9%
20–25	70.2%	100.0%	98.8%
25–30	60.9%	95.7%	91.3%
≥ 30	20.0%	53.3%	80.0%
Overall	80.6%	$\sim$98%	97.6%

Table 12: Top-5 allocations for each model's static portfolio (conditioned on mean test-set VIX). Effective $N = 1/\sum_i w_i^2$ measures diversification.

Model	Eff. N	Top-5 names (%)
Gaussian	7.8	SO 18.5, MCD 18.4, PG 17.7, LMT 11.8, ABBV 6.4
NSF baseline	17.4	KO 9.5, LMT 7.8, HON 7.5, V 6.8, SO 6.7
NSF PCA + corr	1.3	MCD 88.7, XOM 5.8, ABBV 3.5, NFLX 1.2, AMZN 0.7

F Tail-conditional stress metrics

Table 13: Sharpe and CVaR_5 stratified by market stress on the 503-day test set. "Crisis" is days with VIX ≥ 25 (n=38). Worst-day return is -5.20% Gaussian, -5.75% NSF base, -5.70% NSF adaptive in every subset.

	Gauss		NSF base		NSF adapt	
Subset	Sh	CVaR	Sh	CVaR	Sh	CVaR
Full	0.7	-1.7	1.0	-1.6	1.0	-1.6
Crisis	-3.4	-3.7	-3.9	-4.1	-3.6	-4.1
Tail 5%	-11.0	-3.7	-21.0	-4.1	-21.6	-4.1
Tail 1%	-14.9	-5.2	-28.8	-5.8	-30.8	-5.7