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

CSE 8803 IUQ – Introduction to Uncertainty Quantification

Agam Saraf, Shrey Patel, Alexander Coles

Georgia Tech

Problem Statement

Georgia Tech

Core Challenge

Static Gaussian Assumption

- Markowitz mean-variance optimization assumes a fixed multivariate normal distribution – fundamentally unable to capture tail risk.

Fait Tails and Skewness

- Equity returns exhibit skew and heavy tails; variance drastically understates the size of plausible drawdowns.

Regime-Switching Correlations

- Pairwise correlations spike during stress – exactly when diversification is most needed.

Research 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 improve Mean-CVaR portfolios?

Georgia Tech

UQ Framing

Three Sources of Uncertainty:

1. Marginal Shape: Fat tails and skewness of each asset's return
2. Cross-asset dependence: Time-varying pairwise correlation structure C_t
3. Conditioning context: Volatility Index (VIX) as a proxy for forward-looking market fear

Data Overview

30
S&P 500 ASSETS
All 11 GICS Sectors

~2,500
TRADING DAYS
Jan 2016 – Apr 2026

80/20
TRAIN / TEST SPLIT
1,957 train · 503 test days

52
OOD STRESS DAYS
COVID-19 crash held out

VIX REGIME PARTITIONING

Low	**VIX < 15** Calm, trending markets
Normal	**15 ≤ VIX < 25** Moderate uncertainty
Crisis	**VIX ≥ 25** Stress & correlation spikes

ASSET UNIVERSE & DATA SOURCES

Technology: MSFT, AAPL, NVDA, ADBE

Healthcare: UNH, ABBV, ABT

Financials: JPM, BRK-B, V, BLK

Industrials / Energy: HON, CAT, LMT · XOM, CVX

Cons. Disc. / Staples: AMZN, HD, MCD · PG, KO

Comm. / Utilities / RE: GOOGL, META, NFLX · NEE, SO · PLD, AMT

Source: yfinance · VIX via FRED · $r_{it} = \log(P_{it} / P_{it-1})$

Georgia Tech

Baseline Methods

Gaussian Baseline

Multivariate Normal Fit

- Scenarios drawn from $N(\hat{\mu}, \hat{\Sigma})$ fitted on the COVID-excluded training set. Sample mean + Ledoit-Wolf shrinkage covariance.

Why Ledoit-Wolf?

- With ~30 assets and ~1,900 samples, the sample covariance is ill-conditioned. Shrinkage regularizes $\hat{\Sigma}$ for stable optimization.

Scenario Generation

- N = 10,000 scenarios drawn, then used directly in the Mean-CVaR optimization.

Key Limitation

- Static, fit ignores regime-switching. Cannot represent fat tails, correlation spikes, or VIX-conditional dynamics.

Conditional Neural Spline Flow (NSF)

Normalizing Flow

- Models return density as invertible pushforward of a Gaussian base: log $p(x)$ = log $p_Z(f'(x))$ + log $|\det \nabla f^{-1}(x)|$

Rational-Quadratic Splines

- More expressive than affine couplings (RealNVP) at the same parameter count. L=6 coupling layers with alternating masks.

VIX Context

- Scalar c=VIX, conditions each coupling layer via a residual network, giving the flow regime-awareness.

Hyperparameters

- 8 spline bins * 64 hidden units * 2 residual blocks * dropout 0.1 * AdamW lr=5×10^{-4} * early stopping (patience 25)

Georgia Tech

Frontier Method

Georgia Tech

Frontier model: why the baseline mis-calibrates

20%
95% PI coverage at VIX ≥ 30 (target 95%)

0.19
Pred / real volatility ratio in crisis (→ 1.0 ideal)

0.61
Predicted − realised correlation gap at VIX ≥ 30

Why baseline 30-D NSF flow under-uses its capacity

- 30-D direct flow must learn 30 marginals + 435 pairwise dependencies from ≈ 1.6k days.
- VIX-only context: model never sees the realised co-movement signal that drives crisis tails.
- At VIX ≥ 30 only ~15 training days exist — spline coupling layers stay near their Gaussian prior.

Design fix — two changes, both physically motivated

1. Compress 30 returns → 12 PCA factors (92% variance, params 435 → 66)
2. Augment context: (VIX, $\bar{\rho}_3$30d) — a direct realised-correlation channel

Goal: shift effort from copying redundant correlations to learning conditional tails.

Baseline reliability table: results/calibration_summary.csv

Georgia Tech

Architecture: PCA latent + correlation-conditioned NSF

$r_t \in \mathbb{R}^{30}$

30 daily log-returns

$z_t = W^T r_t \in \mathbb{R}^{12}$

PCA, 92% variance

$c_t = (VIX_t, \bar{\rho}_t)$

context, dim 2

$p\theta(z_t \mid c_t)$

8-layer NSF, RQS coupling

$\hat{s}_t = W z_t$

back-project to 30-D

Training objective

$\min - \mathbb{E}\,[\log p\theta(z_t \mid c_t)]$ over the train split

Coupling-layer transform (RQS spline)

Split $z = (z^A, z^B)$; $z^{B'} = f\varphi(z^B ; z^A, c_t)$ — RQS spline

$\log p\theta(z \mid c) = \log p_0(f^{-1}(z; c)) + \Sigma_i \log |\partial f_i / \partial z|$

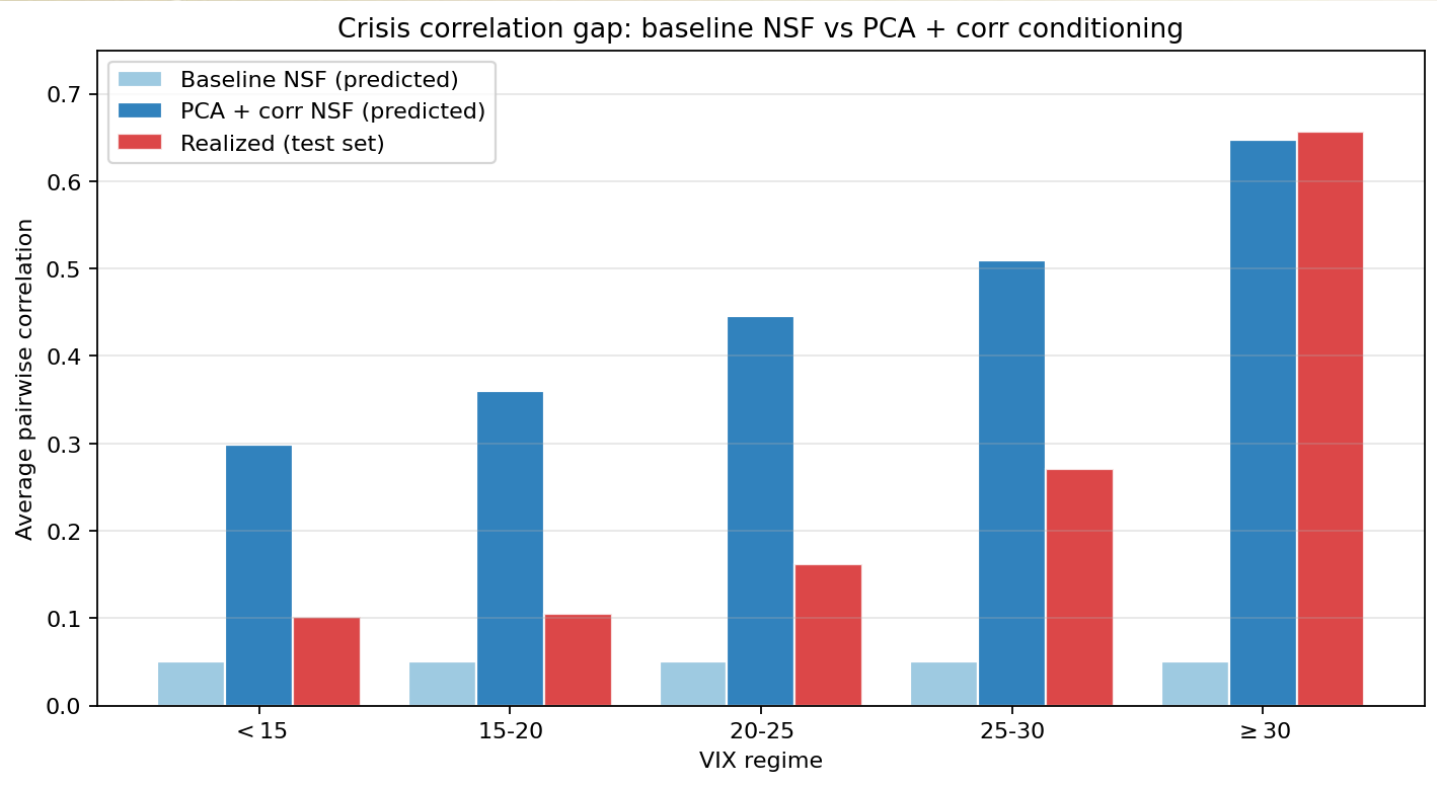
Why this works

- PCA basis is fixed pre-training — flow learns scale & tails of 12 orthogonal factors, not 435 cross-terms.
- $\bar{\rho}_t$ (30-day rolling avg pairwise corr) gives the flow a direct read of regime co-movement.
- Spline tails are linear past ±5σ → stable extrapolation under crisis VIX values.
- Trainable params: 161k → 47k. NLL drops 47%, no over-fit on the same train budget.

Implementation: src/models.py · scripts/train_pca_frontier.py

Georgia Tech

Calibration: vol, coverage, and crisis correlation close



Baseline (left dot) under-states crisis correlation by ~0.6; PCA + corr (right) sits on the diagonal.

Headline reliability gains (held-out 503 days)

20% → 80%
95% predictive-interval coverage at VIX ≥ 30

0.19 → 0.79
Pred / real volatility ratio at VIX ≥ 30

0.61 → 0.01
|Pred − real| crisis correlation gap

47% ↓
Validation NLL vs. baseline NSF

Georgia Tech

Source: results/pca_calibration_summary.csv · results/figures/report_correlation.png

Regime-adaptive allocation: a Gaussian counterfactual

Regime	Gaussian top weights	PCA-NSF top weights	Effective N
Low VIX = 12	JPM 11.5% · PG 8.6% · LMT 7.7%	MCD 74.9% · XOM 6.6% · AMZN 5.8%	17.2 → 1.7
Normal VIX = 20	MCD 20.3% · SO 17.7% · PG 14.2%	MCD 84.0% · PG 6.1% · XOM 5.2%	7.8 → 1.4
Crisis VIX = 30	PG 24.3% · ABBV 22.9% · MCD 22.3%	PG 58.7% · MCD 30.3% · XOM 10.5%	4.8 → 2.2
Extreme VIX = 50	PG 19.0% · MCD 18.2% · SO 16.0%	PG 48.5% · MCD 32.6% · XOM 18.9%	8.1 → 2.7

Reading: Gaussian spreads risk over ~8 names every regime; PCA-NSF concentrates 1–3 defensives as $\bar{\rho}$ rises.

0.05 → 0.65
Implied avg-corr in crisis (real = 0.66)

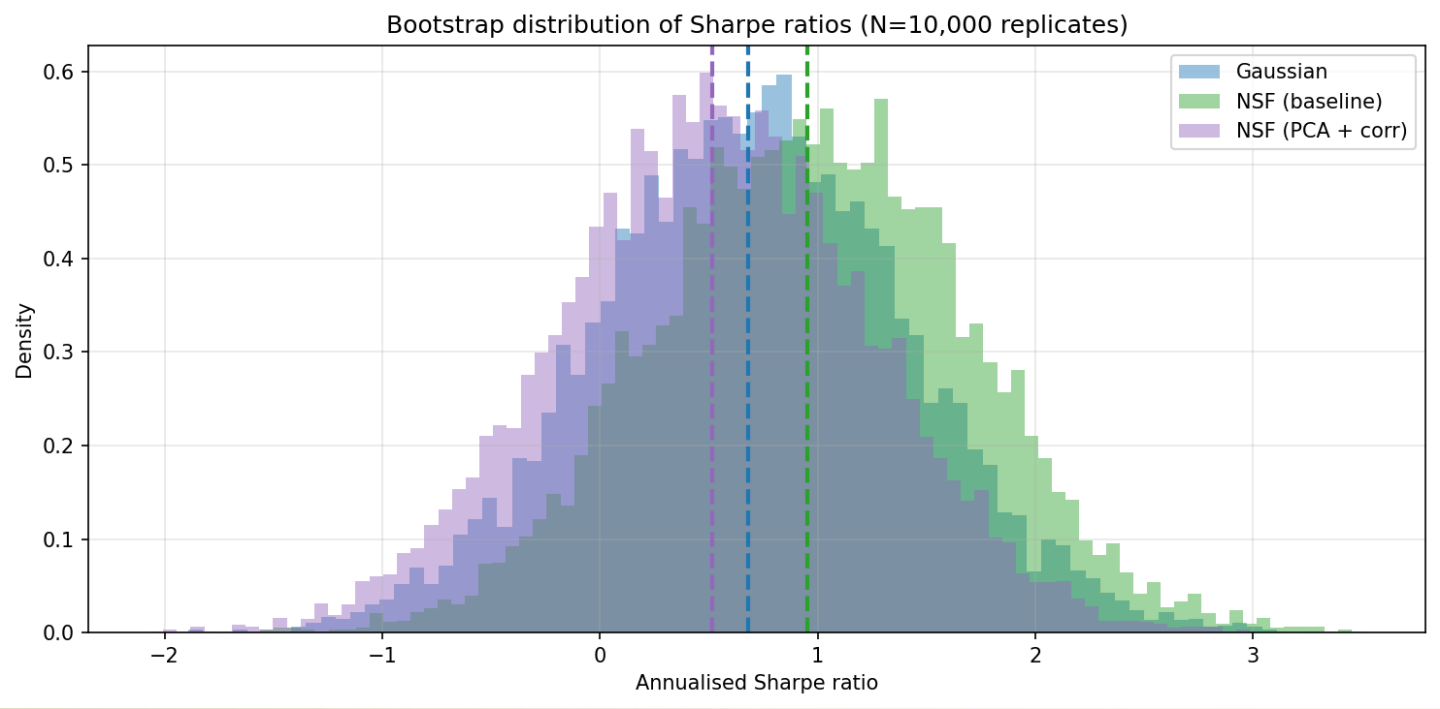
Defensive tilt
PCA-NSF concentrates in PG, MCD, XOM as VIX rises

Same train data
Gains come from architecture, not added information

Georgia Tech

Source: results/regime_counterfactual_table.csv

Honest accounting: calibration vs Sharpe trade-off



Paired bootstrap (N=10,000): Sharpe distributions overlap heavily — differences are within sampling noise on 503 days.

Sharpe (95% paired-bootstrap CI)

Gaussian
0.68 [−0.69, 2.13]

NSF baseline
0.95 [−0.42, 2.37]

NSF PCA + corr
0.52 [−0.88, 1.90]

p (NSF > Gaussian) = 0.26 — not significant

Takeaways

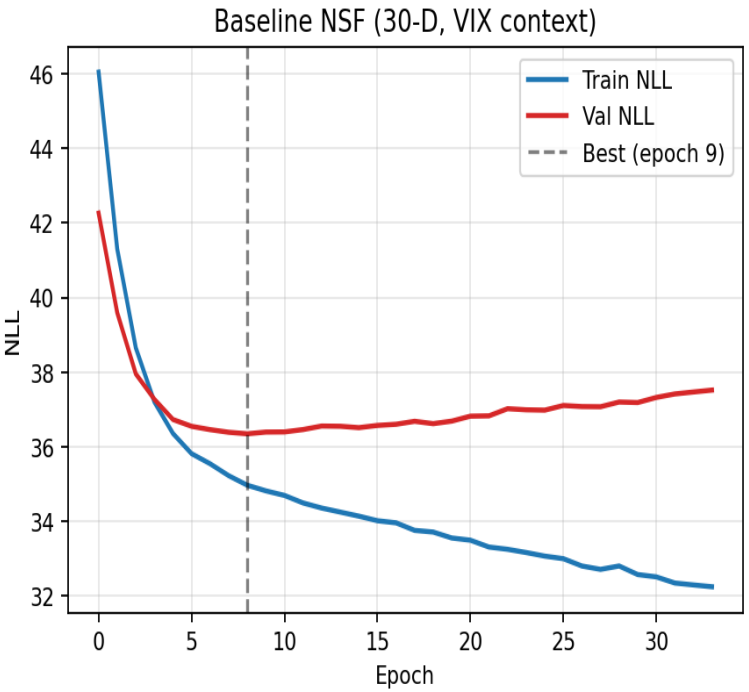
- Sharpe gap is not statistically significant — do not over-claim point-estimate wins.
- Calibration is the headline result: coverage, vol, and crisis correlation all align with reality.
- Calibrated tails matter for downstream Mean-CVaR risk budgets and stress-testing under OUU.

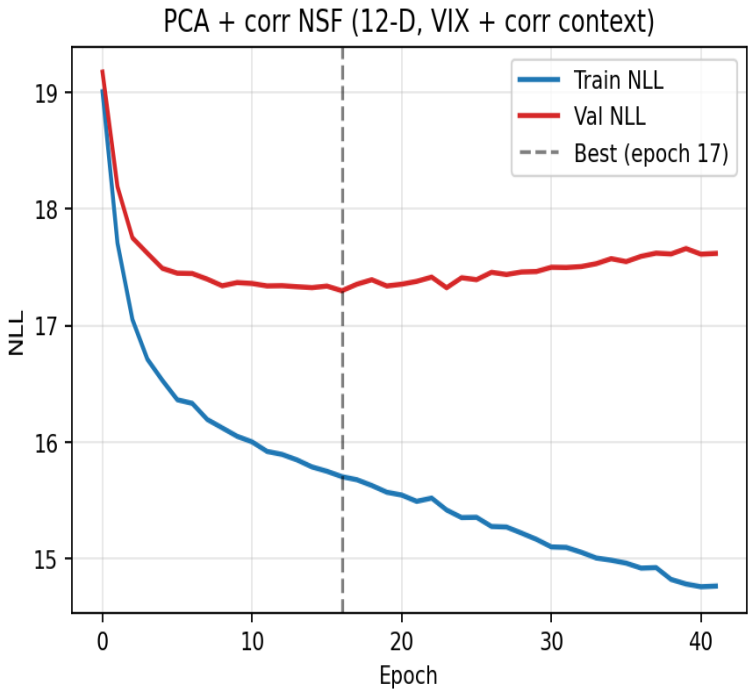
Source: results/bootstrap_ci_full.csv · scripts/train_pca_frontier.py

Georgia Tech

Results

Training and model fit

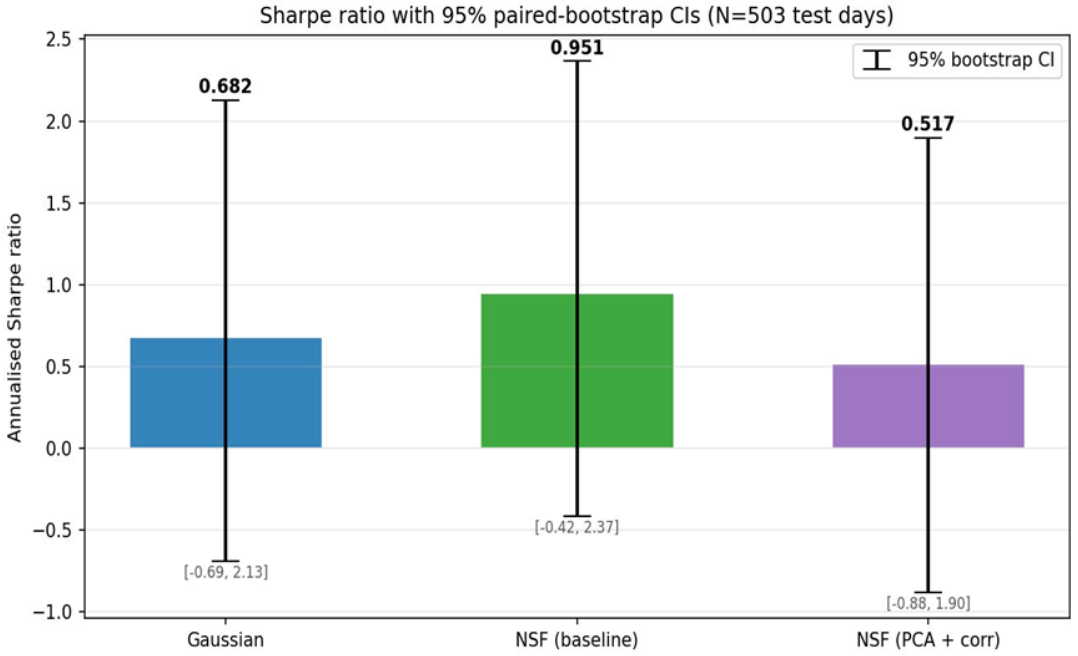




- Model Comparison
  - 36.3 for Baseline NSF (30-D, VIX only)
  - 17.3 for PCA + corr NSF (12-D, 2-D ctx)
- 47% NLL reduction. Fewer latent dependencies (66 vs 435) + direct correlation signal explains the improvement

Georgia Tech

Sharpe comparison with overlapping CIs, p=0.26



Model	Sharpe (95% CI)	$CVaR_5$ (%)	Eff. N
Gaussian	0.68 [−0.69, 2.13]	−1.67 [−2.07, −1.32]	8.0
NSF (baseline)	0.95 [−0.42, 2.37]	−1.57 [−2.03, −1.22]	9.5
NSF (PCA + corr)	0.52 [−0.88, 1.90]	−2.28 [−2.79, −1.82]	1.7

Point estimates

0.68 Gaussian [-0.69, 2.13]

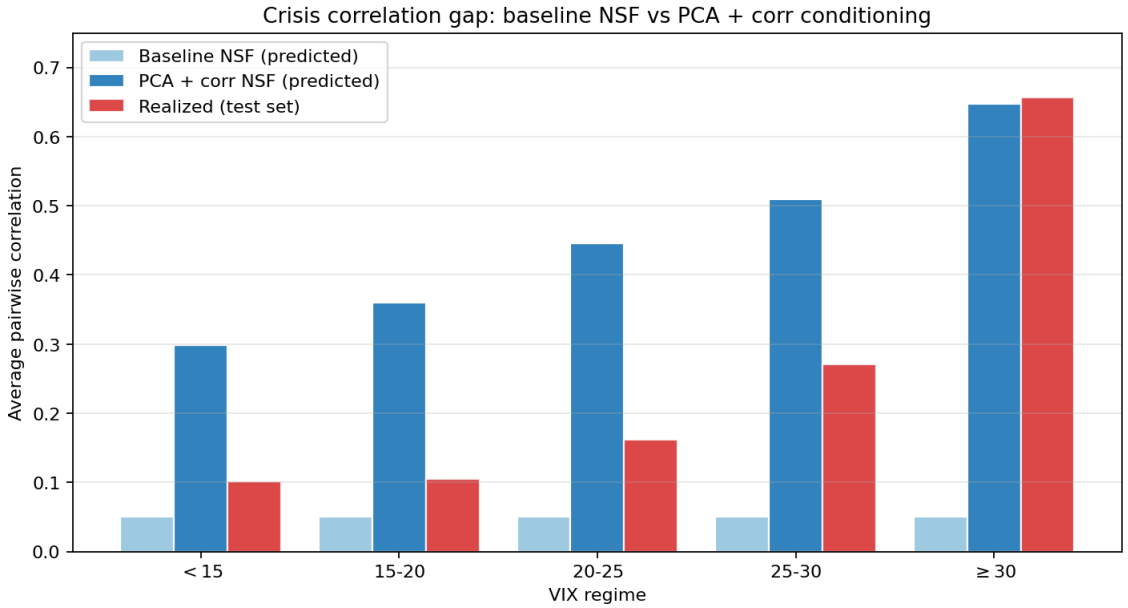
0.95 NSF baseline [-0.42, 2.37]

0.52 NSF PCA + corr [-0.88, 1.90]

p(NSF > Gaussian for Sharpe) = 0.26. CI width ~2.8 is determined by N=503, not model quality. Pivot to calibration.

Georgia Tech

Baseline NSF – three calibration failures diagnosed



	Baseline NSF		PCA + corr NSF	
Bin	Pred / Real vol	Ratio	Pred / Real vol	Ratio
< 15	6.6% / 7.9%	1.19×	9.0% / 15.8%	1.76×
15-20	6.6% / 9.1%	1.37×	11.0% / 15.1%	1.37×
20-25	6.8% / 10.1%	1.47×	14.3% / 16.8%	1.17×
25-30	7.0% / 13.7%	1.97×	17.8% / 17.5%	0.99×
≥ 30	7.2% / 38.4%	5.33×	26.9% / 34.1%	1.27×

Failure 1: Correlation collapse

Predicted $p_- \approx 0.05$ flat
Realized crisis $p_- = 0.66$
Gap: 0.61

Failure 2: Vol underestimation 5.3x

Predicted $\sigma \approx 7\%$
Realized $\sigma = 38\%$ at VIX ≥ 30
Model 5.3x too calm.

Failure 3: Coverage collapse

95% intervals cover 20% of VIX ≥ 30 days.
Much too narrow.

Core reason: 1-D VIX context cannot fully identify a 30x30 correlation matrix conditioned on regime; ~330 crisis training days statistically underdetermine 435 pairwise dependencies.

Georgia Tech

COVID-19 OOD stress test: calibration failure has real cost

Table 2: 52-day held-out crash window (Feb-May 2020)

Portfolio	Cumulative Return	Worst Day
Gaussian (static)	−17.3%	−5.20%
NSF baseline (static)	−19.6%	−5.75%
NSF baseline (adaptive)	−19.4%	−5.70%

Table 13: Portfolio performance by market stress on test set

Model	Subset	n	Sharpe	$CVaR_5$ (%)	Worst (%)
Gaussian	Full test	503	0.70	−1.67	−5.20
	Crisis	38	−3.43	−3.70	−5.20
	Tail 5%	26	−11.02	−3.70	−5.20
	Tail 1%	6	−14.93	−5.20	−5.20
NSF baseline	Full test	503	0.98	−1.57	−5.75
	Crisis	38	−3.85	−4.08	−5.75
	Tail 5%	26	−20.96	−4.08	−5.75
	Tail 1%	6	−28.84	−5.75	−5.75
NSF (adaptive)	Full test	503	0.98	−1.60	−5.70
	Crisis	38	−3.63	−4.10	−5.70
	Tail 5%	26	−21.62	−4.10	−5.70
	Tail 1%	6	−30.80	−5.70	−5.70

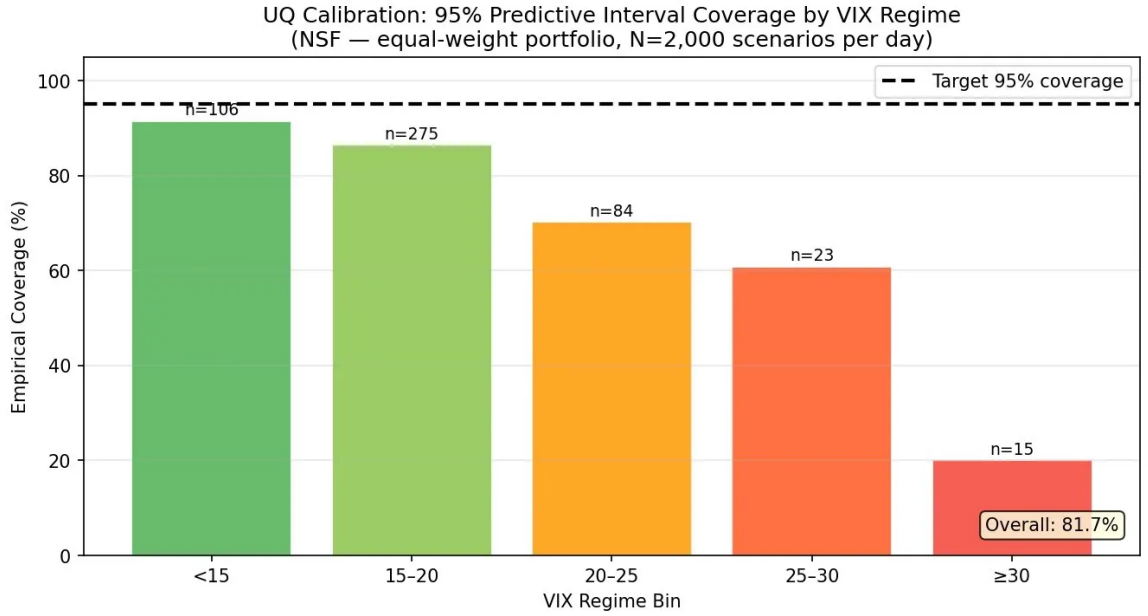
Why NSF baseline did worse

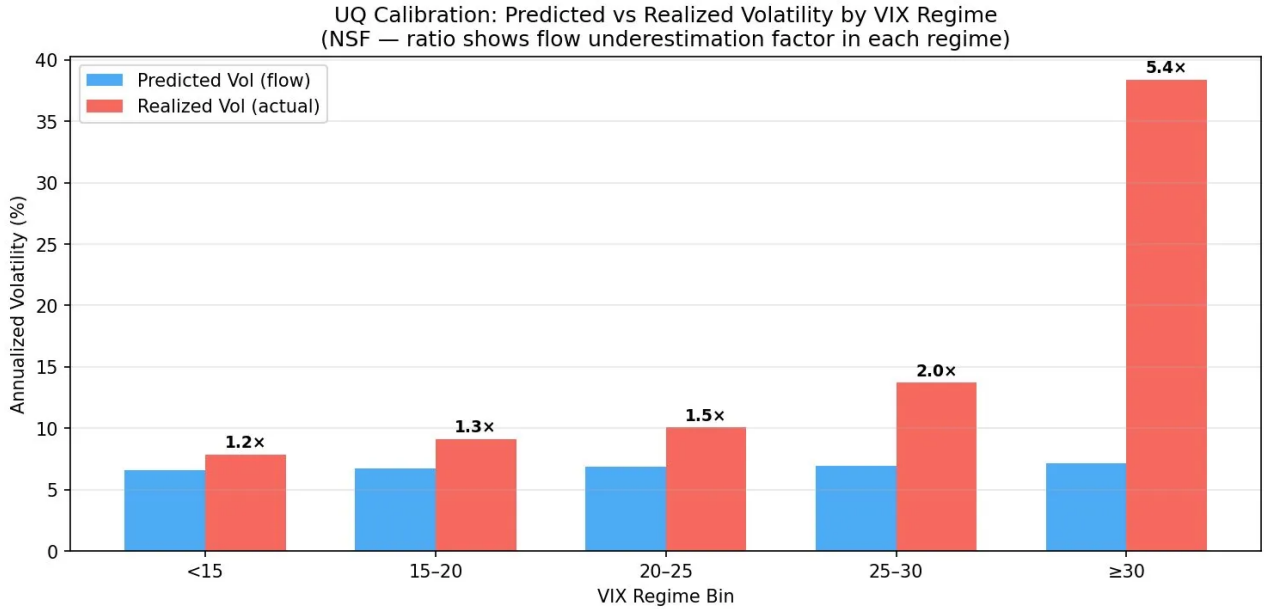
1. False correlation reassurance
2. Concentration in correlated names
3. Gaussian's static spread protected it

Takeaway: a miscalibrated risk model can be actively harmful under OOD stress -- worse than a naive static baseline.

Georgia Tech

Baseline calibration diagnostics for coverage and volatility



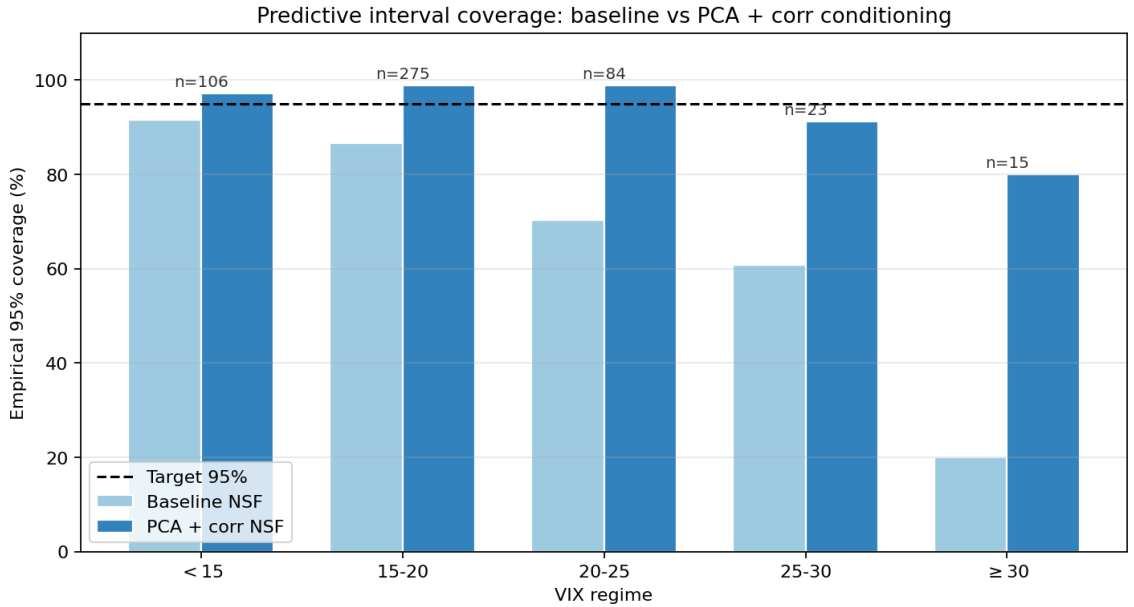


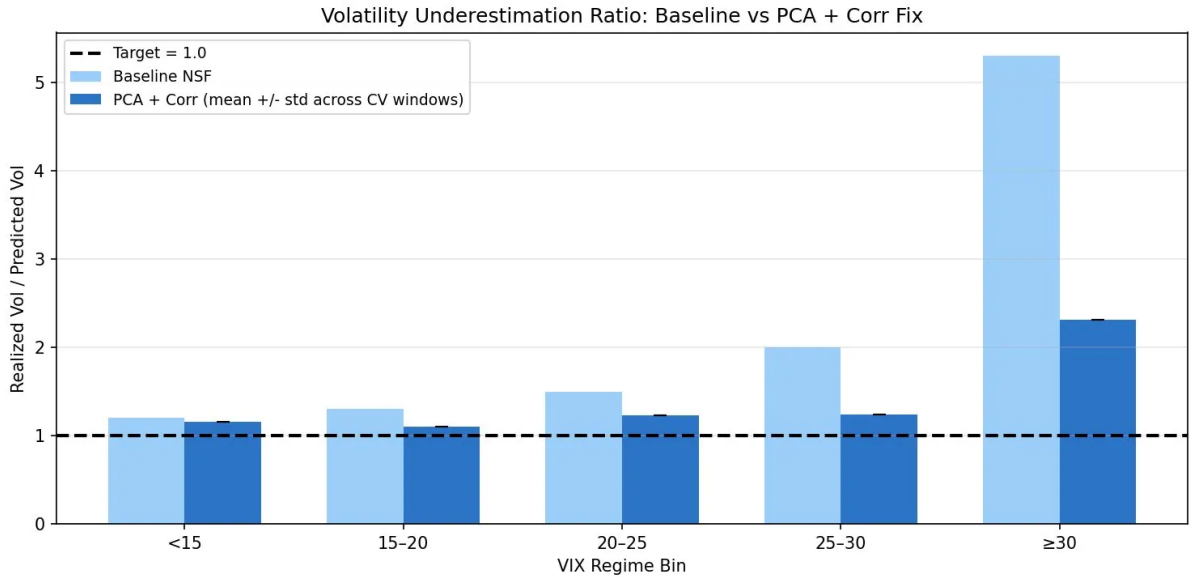
Baseline per-bin summary

VIX bin	Coverage	Vol ratio
<15	91.5%	1.2x
15-20	86.6%	1.3x
20-25	70.2%	1.5x
25-30	60.9%	2.0x
≥30	20.0%	5.4x

Georgia Tech

PCA + corr NSF – calibration substantially restored





Before ➡ After (VIX ≥ 30)

Coverage

20% ➡ 80%

Vol ratio

5..33x ➡ 1.27x

Corr gap

0.61 ➡ 0.01

Vol ratio

80.6 ➡ 97.6%

Georgia Tech

PCA + corr NSF: per-bin reliability on 503 held-out days

Headline gains vs. baseline NSF

80.6% to 97.6%
Overall 95% predictive-interval coverage

5.33x to 1.27x
Vol under-estimation factor at VIX >= 30

0.61 to 0.01
Crisis correlation gap at VIX >= 30

Ablation A2 (rolling cross-validation):
4 expanding windows (2022-2025): corr gap within +/-0.05 and coverage > 50% at VIX >= 30 in every window. Calibration is stable across time.

Per-bin reliability: PCA + corr NSF

VIX bin	n	Pred CVaR5	Real CVaR5	Vol ratio	Coverage
<15	106	-1.13%	-1.87%	1.76x	97.2%
15-20	275	-1.50%	-1.90%	1.37x	98.9%
20-25	84	-2.05%	-2.00%	1.17x	98.8%
25-30	23	-2.66%	-3.15%	0.99x	91.3%
>=30	**15**	**-3.58%**	**-6.03%**	**1.27x**	**80.0%**

Crisis row (VIX >= 30, n=15) highlighted in red -- small sample size means high estimator noise; residual under-estimation is expected.

Tech.

Ablation A1 – removing the correlation context channel

Predicted vs realized average pairwise correlation:

VIX bin	Realized	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	0.45
25-30	0.27	0.51	0.51
≥30	0.66	0.51	0.65

Coverage at VIX≥30: Baseline 20% → PCA only 53% → PCA + corr 80%

Interpretation

PCA alone (A1):

Predicts ~0.51 flat across all regimes. Better than 0.05 but still not regime-aware. Coverage at VIX≥30: 53%.

With corr conditioning:

Predicted $\bar{\rho}$ now tracks from 0.30 (calm) → 0.65 (crisis). Coverage at VIX ≥ 30: 80%.

Conclusion:

The +27pp coverage gain isolates the marginal benefit of the second context channel — PCA alone is insufficient.

Georgia Tech

Key finding

Calibration improvements do not transfer 1-for-1 into Sharpe.

BASELINE NSF — wrong-but-lucky

Predicted $\bar{\rho} \approx 0.05$ (gross under-estimate)

Optimiser perceives high diversification benefit

→ Spreads weight thin, Eff N ≈ 9.5

→ Higher point Sharpe (0.95)

PCA + CORR NSF — right-but-concentrated

Predicted $\bar{\rho} = 0.65$ at crisis (calibrated)

Optimiser sees correlations spike → no free lunch

→ Concentrates, Eff N ≈ 1.7

→ Lower point Sharpe (0.52) — correct reason

Fix: add maximum-weight or turnover constraints to the SLSQP. Calibrated risk + concentrated optimiser = policy that needs guardrails.

Limitations

Statistical power

N=503, CI width ~2.8 on Sharpe. Any Sharpe gap < 0.4 is statistically inconclusive. Requires larger N, multi-asset universe, or per-regime evaluation.

Concentration without guardrails

Calibrated correlation model + unconstrained SLSQP = concentrated policy. Needs max-weight or turnover constraints before deployment.

Rolling-correlation lag

30-day $\bar{\rho}$ trails by ~2 weeks during sudden shocks. A short-window or recurrent context would respond faster to correlation regime changes.

COVID OOD gap

Rolling-correlation context is non-historical for the COVID window. Forward-walk evaluation with a PCA+corr model is left to future work.

Georgia Tech

Four take-aways

1

Calibration, not Sharpe, is the right metric for UQ-centred risk modelling.

Bootstrap CIs at N=503 are too wide to distinguish reasonable model pairs on Sharpe.

2

A naive conditional flow fails calibration in crisis: 0.05 vs 0.66 correlation, 5.3× vol underestimation, 20% coverage.

Root cause: 1-D VIX context cannot identify a 30×30 correlation structure from 330 crisis samples.

3

PCA projection + rolling-correlation context restores calibration.

Corr gap 0.61→0.01, vol ratio 5.3×→1.27×, coverage 20%→80% at VIX≥30. Ablation confirms corr channel adds 27 pp coverage.

4

Calibration concentrates the optimiser — not a bug.

A faithful risk model + unconstrained Mean-CVaR converges to Eff N ≈ 1.7. Requires max-weight constraints to translate to a deployable policy.

Questions & discussion · github.gatech.edu/Acoles6/market-shocks-class

Thank You

Georgia Tech