Replication of *Time Series Analysis for Financial Market Meltdowns* (Kim et al., 2011)

1 Summary of *Time Series Analysis for Financial Market Meltdowns* (Kim et al., 2011)

Introduction and Motivation

In the context of the paper, there were recent financial crises that exposed an issue regarding risk models, where they were not well equipped enough to capture extreme market behaviour. The authors Kim, Rachev, Bianchi, Mitov, and Fabozzi (2011) address this issue in their paper *Time Series Analysis for Financial Market Meltdowns*. Their primary argument is that the heavy reliance on assumptions of Normality in volatility models has led to a systemic underestimation of risk, particularly during periods of crisis. Normality cannot explain the "25-standard-deviation events" that practitioners like David Viniar, Goldman Sachs' CFO referenced (Kim et al., 2011). As a solution, the authors propose using ARIMA-GARCH models with non-normal innovations, specifically α-stable and tempered stable distributions, combined with risk measures that go beyond Value-at-Risk (VaR) to include Average Value-at-Risk (AVaR), a more reliable risk assessment during financial crises (Kim et al., 2011). This corresponds with the idea that extreme events are not rare outliers, rather they are regular features of financial markets when heavy-tailed distributions are used.

The concept of volatility clustering is highly important, where it is a phenomenon well modeled by ARCH and GARCH processes, but emphasizes that volatility dynamics are not enough if the innovations are mis-specified. By incorporating non-normal, heavy-tailed innovations into time series models, the authors demonstrate that the probability of market crash occurrences and tail risk can be forecast more realistically. They also propose that AVaR is a stronger measure than VaR mathematically and empirically, as evaluated on the S&P 500 across historical crash periods (Kim et al., 2011).

ARMA-GARCH model with α-stable and tempered stable innovations

The authors define the following ARMA(1,1)-GARCH(1,1) model:

$$y_t = \alpha y_{t-1} + \beta r_{t-1}\varepsilon_{t-1} + \sigma_t\varepsilon_t + \gamma, \quad \sigma_t^2 = \omega + \alpha_1\sigma_{t-1}^2\varepsilon_{t-1}^2 + \beta_1\sigma_{t-1}^2,$$

The innovations ε_t follow different distributions depending on the time series model in question. Four types of time series models are considered (Kim et al., 2011).

1. *Normal time series models*: CV, GARCH, ARIMA-GARCH, all with normal innovations

2. *Student-t time series models*: CV, GARCH, ARIMA-GARCH, all with t-distribution innovations

3. *Stable time series models*: CV, GARCH, ARIMA-GARCH, all with stable innovations

4. *Tempered stable time series models*: CV, GARCH(1,1), ARIMA(1,1)-GARCH(1,1), with CTS, MTS, and RDTS innovations.

Tempered stable distributions differ from α-stable distributions, in that α-stable distributions capture skewness and heavy tails, but can sometimes imply infinite variance. Tempered stable distributions retain their flexibility while ensuring finite moments by exponentially tempering the tails (Kim et al., 2011).

Expanding on the three tempered stable distributions, there are

- Classical tempered stable (CTS) - Modified tempered stable (MTS) - Rapidly decreasing tempered stable (RDTS)

Parameter Estimation and Forecasting Return Distributions

The method of estimating model parameters and testing on the historical S&P 500 data is outlined in this section.

The study focuses on a select five days of market crashes to represent the daily returns of the index.

- October 19, 1987 (Black Monday)
- October 27, 1997 (Asian Turmoil)
- August 31, 1998 (Russian Default)
- April 14, 2000 (Dotcom Collapse)
- September 29, 2008 (US financial crisis)

Parameters are estimated using ten years of historical data from the closest trading day before the day of the crash (Kim et al., 2011).

Models and estimation windows. The authors estimate parameters for three time–series specifications: the constant–volatility (CV) model, the GARCH(1,1) model, and the ARMA(1,1)–GARCH(1,1) model. All models are applied to S&P 500 daily returns. The fitted models are then used to analyze crash probabilities.

Estimation for normal models. For the normal–CV, normal–GARCH, and normal–ARMA–GARCH specifications, parameters are obtained by maximum likelihood estimation (MLE).

Estimation for non-normal innovations (stable and tempered stable). For the remaining models (i.e., models with α-stable or tempered stable innovations such as CTS, MTS, and RDTS), estimation takes three steps (Kim et al., 2011):

1. **Time–series parameters via Student-t MLE.** Estimate the structural parameters $\alpha_0, \alpha_1, \beta_1, a, b, c$ assuming *Student-t* innovation by MLE. For the CV models, all parameters except α_0 are set to zero; for the GARCH(1,1) model, the ARMA coefficients a and b are set to zero.
2. **Residual extraction.** Compute standardized residuals using the estimates from step 1.
3. **Innovation law fit by MLE.** Fit the parameters of the target innovation distribution to these residuals by MLE — specifically, the α-stable family or one of the tempered stable families (CTS, MTS, RDTS).

This procedure ensures comparable volatility dynamics across models while allowing the tail behavior to be governed by the appropriate heavy–tailed distribution (Kim et al., 2011). With these estimates, the paper evaluates crash probabilities and, in later sections, conducts VaR/AVaR forecasting and backtesting.

VaR and AVaR for ARMA-GARCH Model

VaR and Backtesting

The authors compute 1% daily VaR for the S&P 500 from December 14, 2004 to December 31, 2008 using five models:

- Exponentially weighted moving average (EWMA)
- normal-ARMA-GARCH
- t-ARMA-GARCH
- stable-ARMA-GARCH
- CTS-ARMA-GARCH

Goodness of fit results

The Kolmogorov-Smirnov (KS) statistics demonstrate clear results. The EWMA and normal-ARCH-GARCH models are always rejected at the 1% significance level, confirming that normal innovations do not fit the actual distribution of returns. The t-ARMA-GARCH and stable-ARMA-GARCH models perform better but still occasionally fail, as seen during the high volatility 2007-2008 period. The CTS-ARMA-GARCH is consistently the best performer of all the models, having KS statistics well below the rejection threshold. This indicates that tempered stable innovations provide the most realistic representation of return distributions (Kim et al., 2011).

Backtesting Performance

As another measure of performance, the authors conduct backtesting using two approaches:

- Christoffersen Likelihood Ratio (CLR) Test: assesses if the frequency of VaR violations matches expectations. Consists of three parts:
    1. Test of unconditional coverage
    2. Test of independence
    3. Joint test of coverage and independence
- Berkowitz Likelihood Ratio (BLR) Test
    1. Test of independence
    2. Test of tail distribution

The time periods considered for these tests are:

- 1-year test
    - 2005: December 14, 2004 - December 15, 2005
    - 2006: December 16, 2005 - December 20, 2006
    - 2007: December 21, 2006 - December 27, 2007
    - 2008: December 28, 2007 - December 31, 2008
- 2-year test
    - 2005-2006: December 14, 2004 - December 20, 2006
    - 2007-2008: December 21, 2006 - December 31, 2008
- 4-year test
    - 2005-2008: December 14, 2004 - December 31, 2008

Findings from Backtesting: CLR tests

Unconditional Coverage: For the 1-year tests in 2005 and 2006, and the 2-year test from 2005-2006, none of the five models are rejected at the 1% significance level. This demonstrates that none of the models underestimated risk.

For the 1-year tests in 2007 and 2008, and the 2-year test from 2007-2008, the three non-normal models are not rejected, but the two normal models are rejected at the 1% significance level. This indicates that the normal models systemically underestimate the probability of extreme losses.

For the 4-year test for 2005-2008, the same applies, where the non-normal models are not rejected, while the normal models are rejected at the 1% significance level. *******

The filtered historical simulation (FHS) method is rejected for the 1-year test in 2007 and the 2-year test from 2007-2008 at the 1% significance level.

Thus, it can be concluded that normal-based models cannot correctly predict the frequency of losses during periods of crisis, whereas models with non-normal innovations can. Therefore, VaR estimated under normal assumptions underestimates in periods of crisis (Kim et al., 2011).

Independence: None of the five models or the FHS method are rejected at the 1% significance level, with an exception being the 1-year test in 2005. Regarding the three non-normal models, there are no violations at all in some periods , meaning the independence test cannot be performed. This suggests that the non-normal models may yield slightly conservative VaR thresholds, potentially overestimating risk, or indicating that the required capital charge is higher than necessary (Kim et al., 2011).

Joint Test of Coverage and Independence: For the 1-year test in 2005, the test again cannot be performed for the three non-normal models due to the lack of VaR violations. For the 1-year test in 2006 and the 2-year test from 2005-2006, none of the models are rejected at the 1% significance level. For the 1-year tests in 2007 and 2008, and the 2-year test from 2007-2008, the non-normal models are not rejected, but the normal models are rejected at the 1% significance level. For the 4-year test from 2005-2008, it is the same. The FHS method is again rejected for the 2-year test from 2007-2008 (Kim et al., 2011).

Findings from Backtesting: BLR tests

Test of independence

The three non-normal models are not rejected at the 1% significance level for any time period. This indicates that they are able to consistently measure the independence of violations. The EWMA model is rejected for the 1-year test in 2008 and the 2-year test for 2007-2008, indicating the presence of violation clustering likely due to the 2008 financial crisis. The normal-ARMA-GARCH model is also rejected for the 2-year test from 2007-2008 at the 1% significance level (Kim et al., 2011).

Test of tail distribution

The three non-normal models are not rejected at the 1% significance level across all time periods. The normal models are not rejected in the 1-year tests for 2005 and 2006, as well as for the 2-year tests from 2005-2006. However, they are rejected for the 1-year tests for 2007 and 2008, the 2-year test from 2007-2008, and the 4-year test from 2005-2008. This is consistent with assumptions of normality being insufficient in times of financial crisis (Kim et al., 2011).

VaR Dynamics

It is found that all non-normal models typically have higher VaR values than normal models. The authors used the average of the relative difference (ARD) between a normal model and a non-normal model to properly quantify this divergence. The ARD formula is

$$\text{ARD} = \frac{\text{VaR}_{\text{model}} - \text{VaR}_{\text{normal}}}{\text{VaR}_{\text{normal}}} \times 100\%.$$

Practically, the ARD informs the cost of risk management, relating to the excess capital charge required when using one model as opposed to another. Smaller ARD values are more economically efficient (Kim et al., 2011).

Empirical results (Table 6, p. 1887) show that in calm periods such as 2005–2006, non-normal VaR estimates exceed those from normal models by approximately 10–30%. For example, the Student-*t* ARMA–GARCH model yields VaR values about 30% higher than those from the EWMA model, while the stable and CTS–ARMA–GARCH models are approximately 21% and 23% higher in 2005 and 2006, respectively. In comparison to the normal-ARMA–GARCH model for the same years, the t-ARMA–GARCH model yields VaR estimates approximately 21% and 20% larger, the stable-ARMA–GARCH about 12% larger for both periods, and the CTS–ARMA–GARCH about 11% and 10% larger.

Transitioning to the 2007–2008 financial crisis period, this difference narrows but is still significant. For 2007, the t-ARMA–GARCH, stable-ARMA–GARCH, and CTS–ARMA–GARCH models yield approximately 25%, 18%, and 15% larger VaR values than the EWMA model, and 21%, 13%, and 10% larger VaR values, respectively, than the normal-ARMA–GARCH benchmark. For 2008, these models still yield higher VaR estimates that are about 18%, 11%, and 13% larger than the EWMA model, and 23%, 16%, and 17% larger than the normal-ARMA–GARCH model.

Overall, the authors find that even in periods of crisis, models with heavy-tailed innovations yield significantly higher VaR estimates than normal models. This is consistent with the theory that normally distributed VaR measures tend to underestimate. It is also seen that the tempered-stable models provide more realistic VaR estimates that remain consistent across changing market regimes.

As can be seen in Table 6, the FHS method has the smallest ARD values, followed by the stable-ARMA–GARCH and CTS–ARMA–GARCH models, and the t-ARMA–GARCH model has the largest. From these results, the authors conclude that the CTS–ARMA–GARCH model performs best among the non-normal models for three reasons. First, the CTS–ARMA–GARCH model consistently produces lower Kolmogorov–Smirnov (KS) statistics than the t- and stable-ARMA–GARCH models across all four years. Second, the model is not rejected by either the CLR or BLR tests in either non-volatile periods (2005–2006) and volatile periods (2007–2008). Third, the model exhibits lower ARD values than the other two non-normal models in every year except 2008 relative to both normal benchmarks, where the stable-ARMA–GARCH model has slightly smaller ARD values. While the FHS method yields the smallest ARD values overall, it is rejected by the CLR test in 2007, and thus is likely unreliable.

Hence, the CTS–ARMA–GARCH model offers a balanced performance, where its VaR estimates are less extreme than those from the t-ARMA–GARCH model in non-volatile periods (2005–2006) while also remaining statistically significant in both CLR and BLR tests during volatile periods (2007–2008).

Average Value-at-Risk (AVaR)

In this section, the authors introduce and derive the Average Value-at-Risk (AVaR) for the ARMA–GARCH model with classical tempered stable (CTS) innovations and provide an empirical application to S&P 500 returns. AVaR represents the expected loss conditional on losses exceeding the VaR threshold, offering a more comprehensive measure of left-tailed risk than conventional VaR.

AVaR at significance level η is defined as

$$\mathrm{AVaR}_\eta(X) = \frac{1}{\eta}\int_0^\eta \mathrm{VaR}_\epsilon(X)\, d\epsilon,$$

where $\mathrm{VaR}_\epsilon(X)$ denotes the VaR of the random variable X at level ϵ. For continuous distributions, this can be expressed as

$$\mathrm{AVaR}_\eta(X) = -E[X \mid X < -\mathrm{VaR}_\eta(X)].$$

The authors extend this definition to the conditional AVaR for returns of the next time step in the ARMA–GARCH model. Let y_{t+1} be the forecasted return conditional on information available up to time t. Then, with significance level η, the conditional AVaR is written as

$$\mathrm{AVaR}_{t,\eta}(y_{t+1}) = -E_t\big[y_{t+1}\,\big|\,y_{t+1} < -\mathrm{VaR}_{t,\eta}(y_{t+1})\big],$$

where $\mathrm{VaR}_{t,\eta}(y_{t+1})$ is the conditional VaR at level η and $E_t[\cdot]$ denotes the conditional expectation given information at time t.

By substituting the ARMA–GARCH dynamics defined earlier, the authors derive that

$$\mathrm{AVaR}_{t,\eta}(y_{t+1}) = -(c + ay_t + b\sigma_t\varepsilon_t) + \sigma_{t+1}\mathrm{AVaR}_\eta(\varepsilon_{t+1}),$$

where ε_{t+1} represents the standardized innovation term. This expression demonstrates that conditional AVaR can be decomposed into a predictable component, depending on past returns and volatility, and a distributional component determined by the AVaR of the innovation process.

The next step involves deriving a closed-form expression for $\mathrm{AVaR}_\eta(\varepsilon_{t+1})$ when ε_{t+1} follows a tempered stable distribution. If ε_{t+1} is tempered stable, Kim *et al.* (2010b) show that AVaR can be expressed through the following proposition:

Proposition. Let Y be a continuous, infinitely divisible random variable with characteristic function $\phi_Y(u)$. If there exists $\rho > 0$ such that $|\phi_Y(-u + i\rho)| < \infty$ for all $u \in R$, then

$$\mathrm{AVaR}_\eta(Y) = \mathrm{VaR}_\eta(Y) - \frac{e^{-\mathrm{VaR}_\eta(Y)\rho}}{\pi\eta}\Re\left(\int_0^\infty e^{-iu\mathrm{VaR}_\eta(Y)}\frac{\phi_Y(-u+i\rho)}{(-u+i\rho)^2}\,du\right),$$

where $\Re(\cdot)$ denotes the real part of a complex number. This analytic form allows efficient computation of AVaR directly from the characteristic function of the tempered stable distribution, providing a closed-form solution within the CTS–ARMA–GARCH framework.

The authors then empirically evaluate the model using S&P 500 daily returns over the period 14 December 2004 to 31 December 2008. They calculate 1% daily AVaR values for both the normal-ARMA–GARCH and CTS–ARMA–GARCH models, as well as the daily differences between them, defined by

$$d_{\mathrm{CTS\&Normal}}(t) = \mathrm{AVaR}^{\mathrm{CTS}}_{t,0.01}(y_{t+1}) - \mathrm{VaR}^{\mathrm{Normal}}_{t,0.01}(y_{t+1}),$$

which measures how much more conservative the CTS–ARMA–GARCH AVaR is relative to the normal-ARMA–GARCH VaR.

The time series of $d_{\mathrm{CTS\&Normal}}(t)$ (Fig. 3) shows that AVaR is consistently more conservative under stressed conditions. From July 2007 to September 2008, these differences are significantly larger than in earlier periods, indicating that AVaR appropriately responds to increased market volatility. The index reaches its peak (in the study) in around September-October 2008, coinciding with the market collapse, demonstrating that AVaR can be considered a strong indicator of financial stress.

A similar result holds when comparing AVaRs between the two models:

$$D_{\mathrm{CTS\&Normal}}(t) = \mathrm{AVaR}^{\mathrm{CTS}}_{t,0.01}(y_{t+1}) - \mathrm{AVaR}^{\mathrm{Normal}}_{t,0.01}(y_{t+1}),$$

with the time series (Fig. 4) showing parallel behavior. Both $d_{\mathrm{CTS\&Normal}}(t)$ and $D_{\mathrm{CTS\&Normal}}(t)$ exhibit clear upward trends beginning in mid-2007 and reach their maximum levels near the end of 2008. This means that these indicators can be viewed as early warning signals of a market downturn, as they start rising roughly one year before the 2008 crash, providing advance evidence of systemic stress.

Conclusion

This paper concludes that time series models which assume innovations are normal fail to accurately forecast return distributions. The authors found that models with stable and tempered stable innovations - the CTS-ARMA-GARCH model in particular - are stronger predictors of market risk. Empirically, non-normal models are not rejected by backtesting tests (CLR and BLR) at the 1% significance level, whereas normal models are. The CTS-ARMA-GARCH model is the highest performer, as well as being practical.

The authors derived a closed form expression for AVaR under tempered stable innovations and observed that the spread between AVaR from the CTS and normal models began to widen around a year prior to the 2008 market crash. This indicates that such financial crises are potentially predictable. The comparisons of distributions also led to practical implications, where the CTS model provided comparable risk estimates to one based on the Student-t distribution, however required a lower capital charge than the t or α-stable models. It was also more conservative than the normal models.

Model Replication

Note

I encountered some issues with using the entire 20 years of data, so many of the results use less than the full period.

CV Model

This section replicates the constant volatility (CV) market model in Kim et al. (2011). Setting $a = 0$, $b = 0$, $\alpha_1 = 0$, and $\beta_1 = 0$ in their general specification yields a constant conditional standard deviation, i.e.

$$\sigma_t \equiv \sigma_0 = \sqrt{\omega_0} \quad \text{for all } t \geq 0,$$

so that returns follow

$$r_t = \mu + \sigma_0 z_t, \qquad z_t \sim D(0,1),$$

where $D(0,1)$ denotes either the standard Normal or the standardized Student–t distribution. We consider two CV variants:

$$\begin{aligned} \text{Normal–CV:} \quad & z_t \sim \mathcal{N}(0,1), \\ t\text{–CV:} \quad & z_t \sim t_\nu(0,1), \quad \nu > 2. \end{aligned}$$

The parameter vector θ equals (μ, σ_0) for Normal–CV and (μ, σ_0, ν) for t–CV, and is estimated by maximum likelihood.

Given $\hat{\theta}$, the fitted one–period predictive distribution is

$$F(r \mid \hat{\theta}) = \begin{cases} \Phi\left(\dfrac{r - \hat{\mu}}{\hat{\sigma}_0}\right), & \text{Normal–CV,} \\ T_\nu\left(\dfrac{r - \hat{\mu}}{\hat{\sigma}_0}\right), & t\text{–CV,} \end{cases}$$

where Φ is the standard Normal CDF and T_ν is the Student–t CDF with ν degrees of freedom. Probability integral transforms (PITs) are computed as

$$u_t \; = \; F(r_t \mid \hat{\theta}),$$

which should be i.i.d. Uniform$(0, 1)$ under correct specification.

To assess distributional adequacy we apply two tests: (i) the Kolmogorov–Smirnov statistic,

$$D \; = \; \sup_x \big| \hat{F}_n(x) - x \big|,$$

and (ii) the Anderson–Darling statistic, which emphasizes tail deviations,

$$A^2 \; = \; -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \Big[\ln u_{(i)} + \ln\big(1 - u_{(n+1-i)}\big) \Big],$$

with $u_{(i)}$ the ith order statistic of the PITs. Rejection indicates that the assumed innovation distribution (Normal vs. Student–t) is inadequate for the CV benchmark.

Results

Table 1: Goodness-of-fit test results for the CV model under Normal and Student-t innovations.

Model	Parameters	KS Statistic	KS 1% Critical Value	KS Test Result	AD Statistic
CV–Normal	$\mu = 0.00016$, $\sigma = 0.01190$	0.0943	0.02298	Reject	9884.119
CV–Student-t	$\mu = 0.00059$, $\sigma = 0.00684$, $\nu = 2.57$	0.0205	0.02298	Fail to Reject	10059.4683

Table 1 reports the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) test results for the conditional volatility (CV) model under both Normal and Student–t innovation assumptions. The KS test statistic for the Normal model (0.0943) exceeds the 1% critical value of 0.02298, leading to null hypothesis of uniformity being rejected. In contrast, the Student–t specification yields a significantly smaller KS statistic (0.0205), failing to reject uniformity at the 1% level.

This outcome indicates that the conditional return distribution with Student–t innovations provides a closer fit to the empirical data, particularly in capturing the heavy-tailed nature of returns. This is consistent with known theory and empirical findings from other studies. Although both models exhibit large AD statistics (both are blown out of proportion, but I am examining their difference, not magnitude)—reflecting some residual tail deviations—the Student–t model is a more adequate distribution, consistent with the findings of Kim et al. (2011). In summary, the results suggest that the heavier-tailed Student–t distribution significantly improves the fit of the CV model relative to the Normal benchmark.

GARCH(1,1) Model

Volatility in financial returns is time-varying and exhibits clustering, where large movements are followed by large movements and small by small. The GARCH(1,1) model (Bollerslev, 1986) captures this behaviour by modelling conditional variance as a function of past shocks and volatility. However, assuming normally distributed innovations often underestimates extreme returns during crises. To better capture the heavy tails observed in financial data, the Student-t GARCH model introduces a degrees-of-freedom parameter controlling tail thickness. Following Kim et al. (2011), both models are used to assess volatility dynamics and tail risk under normal and fat-tailed innovations.

Let y_t denote the log-return:

$$y_t = \mu + \varepsilon_t, \qquad \varepsilon_t = \sigma_t z_t,$$

where μ is the conditional mean, σ_t^2 the conditional variance, and z_t an i.i.d. innovation with zero mean and unit variance.

The GARCH(1,1) process is

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \quad \omega > 0,\ \alpha, \beta \geq 0,\ \alpha + \beta < 1.$$

Normal GARCH(1,1). Assuming $z_t \sim \mathcal{N}(0,1)$, parameters $\theta = (\mu, \omega, \alpha, \beta)$ are estimated via maximum likelihood:

$$\ell(\theta) = -\tfrac{1}{2}\sum_{t=1}^{T}\left[\log(2\pi) + \log(\sigma_t^2) + \frac{(y_t - \mu)^2}{\sigma_t^2}\right].$$

Student-*t* GARCH(1,1). If $z_t \sim t_\nu(0,1)$ with $\nu > 2$, the likelihood becomes

$$\ell(\theta) = \sum_{t=1}^{T}\left[\log\Gamma\!\left(\tfrac{\nu+1}{2}\right) - \log\Gamma\!\left(\tfrac{\nu}{2}\right) - \tfrac{1}{2}\log[(\nu-2)\pi] - \log\sigma_t - \tfrac{\nu+1}{2}\log\left(1 + \frac{(y_t-\mu)^2}{(\nu-2)\sigma_t^2}\right)\right].$$

As $\nu \to \infty$, the Student-t GARCH converges to the normal case. Smaller ν values produce heavier tails, allowing more realistic modelling of extreme returns and improved risk estimates (VaR, AVaR) during market stress.

Model parameters are estimated via Maximum Likelihood Estimation (MLE) using ten years of daily S&P 500 returns preceding each market crash, as in Kim et al. (2011). For the Normal-GARCH model, $(\mu, \omega, \alpha, \beta)$ are obtained by maximizing the Gaussian log-likelihood, while the Student-t GARCH additionally estimates the degrees of freedom ν controlling tail thickness.

Goodness-of-Fit Tests

For each fitted model, we test standardized residuals against the assumed innovation law. The Kolmogorov–Smirnov (KS) test compares the empirical CDF to the model CDF via the sup–norm distance; rejection occurs when the statistic exceeds the critical value (here, the 1% threshold). The Anderson–Darling (AD) statistic upweights tail deviations, making it more sensitive to extreme-return misfit—a key feature in crisis periods. This KS/AD residual-based workflow follows Kim et al. (2011).

Table 2: Residual-based goodness-of-fit for replicated GARCH(1,1) models. KS critical value at 1% level = 0.022983.

Model	**Parameters (subset)**	**KS Stat.**	**KS 1% Crit.**	**KS Result**	**AD Stat.**
GARCH–Normal	$\mu = 0.20243,\ \omega = 3.17781\times10^{-6},\ \alpha = 0.10022,\ \beta = 0.88003$	0.808396	0.022983	Reject	1.0734
GARCH–Student's t	$\mu = 0.00029,\ \omega = 1.36998\times10^{-3},\ \alpha = 0.99987,\ \beta = 0.00010\ldots$	0.290623	0.022983	Reject	0.1961

Both models are rejected by the KS test at the 1% significance level, indicating significant deviations from the assumed innovation distributions. However, the Student-t GARCH model displays a smaller KS and AD statistic compared to the Normal-GARCH model, indicating that it is a better fit to the empirical residuals. As highlighted by Kim et al. (2011), the AD statistic is most relevant in evaluating tail behaviour, where lower AD values indicate a better description of extreme returns. The Normal-GARCH model has a higher AD statistic (1.0734), demonstrating that it is less capable in capturing heavy tails, whereas the Student-t model (AD = 0.1961) better represents tail risk. These results reinforce the conclusion that heavy-tailed innovations are essential for realistic volatility and risk estimation during crisis periods.

ARMA(1,1)–GARCH(1,1) Models with Normal and Student-t Innovations

While the GARCH(1,1) model captures time-varying volatility, it assumes a constant conditional mean. In reality, financial return series often exhibit short-term autocorrelation arising from market microstructure effects or delayed information flow. To address this, an autoregressive moving average (ARMA) component can be added to the conditional mean equation, forming the ARMA(1,1)–GARCH(1,1) model. This allows both the mean and variance to evolve dynamically, improving short-term predictive accuracy.

The model is fitted under both normal and Student-t innovations. The normal version is a baseline for comparison, while the Student-t variant accommodates heavy-tailed return distributions that are often observed during turbulent market periods.

Model Specification

Let y_t denote the log-return series and define

$$y_t = c + \phi y_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t, \qquad \varepsilon_t = \sigma_t z_t,$$

where c is a constant, ϕ and θ are the AR and MA coefficients, and σ_t^2 follows the GARCH(1,1) process:

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \quad \omega > 0,\ \alpha, \beta \geq 0,\ \alpha + \beta < 1.$$

Here, z_t are i.i.d. standardized innovations with zero mean and unit variance.

Normal ARMA(1,1)–GARCH(1,1). Under the normal specification,

$$z_t \sim \mathcal{N}(0,1),$$

and parameters $\Theta = (c, \phi, \theta, \omega, \alpha, \beta)$ are estimated by maximizing the Gaussian log-likelihood:

$$\ell(\Theta) = -\tfrac{1}{2} \sum_{t=1}^{T} \left[\log(2\pi) + \log(\sigma_t^2) + \frac{\varepsilon_t^2}{\sigma_t^2} \right].$$

The normal model provides a convenient benchmark but tends to underestimate the probability of large shocks, particularly in crisis periods.

Student-t ARMA(1,1)–GARCH(1,1). For the Student-t variant,

$$z_t \sim t_\nu(0,1),$$

with $\nu > 2$ degrees of freedom. The log-likelihood becomes

$$\ell(\Theta) = \sum_{t=1}^{T} \left[\log\Gamma\left(\tfrac{\nu+1}{2}\right) - \log\Gamma\left(\tfrac{\nu}{2}\right) - \tfrac{1}{2}\log[(\nu-2)\pi] - \log\sigma_t - \tfrac{\nu+1}{2}\log\left(1 + \frac{\varepsilon_t^2}{(\nu-2)\sigma_t^2}\right)\right].$$

As $\nu \to \infty$, the Student-t model converges to the Gaussian case. Smaller ν values capture heavier tails, allowing the model to better reproduce the empirical distribution of returns during extreme market movements.

Parameter Estimation and Goodness-of-Fit

Parameters for the ARMA(1,1)–GARCH(1,1) models were estimated using maximum likelihood, with residual-based goodness-of-fit evaluated through the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) tests. The KS test measures the maximum distance between the empirical and theoretical cumulative distributions, while the AD statistic emphasizes discrepancies in the tails, offering greater sensitivity to extreme values. Both tests assess whether the standardized residuals conform to the assumed innovation distribution.

Table 3: Residual-based goodness-of-fit for ARMA(1,1)–GARCH(1,1) models.

Model	**Parameters (subset)**	**KS Stat.**	**KS Result**	**AD Stat.**
ARMA–GARCH (Normal)	$\phi = -0.09062,\ \mu = -0.03850,\ \omega = 3.74067{\times}10^{-6},\ \alpha = 0.100\ldots$	0.659139	Reject	0.8229
ARMA–GARCH (Student's t)	$\phi = 0.27065,\ \mu = -0.00009,\ \omega = 1.31252{\times}10^{-3},\ \alpha = 0.311\ldots$	0.344951	Reject	0.2581

Both models are rejected by the KS test, implying that neither innovation distribution fully captures the observed residual behaviour. However, the Student-t ARMA–GARCH yields considerably lower KS and AD statistics, indicating a better fit. As observed in Kim et al. (2011), the AD statistic is a stronger indicator of tail adequacy, and its reduction from 0.8229 to 0.2581 demonstrates that the heavy-tailed model captures large changes in the market more effectively than the normal case. Overall, the KS test rejections suggest that even with ARMA dynamics and Student-t innovations, residuals retain heavier tails than those implied by the model, highlighting the persistent challenge of modelling crisis-period returns.

Christoffersen Likelihood Ratio (CLR) Tests

To assess the reliability of Value-at-Risk (VaR) forecasts, the Christoffersen Likelihood Ratio (CLR) tests examine both the correct unconditional coverage and the independence of VaR violations. Following the approach in Kim et al. (2011), these tests determine whether the observed exceedances of the predicted VaR occur with the correct frequency and do not exhibit temporal dependance.

Let $I_t = 1\{L_t > \widehat{\mathrm{VaR}}_{\alpha,t}\}$ be the indicator variable of a violation at time t, where L_t is the portfolio loss and α is the confidence level. Define $T_1 = \sum_{t=1}^{T} I_t$ as the total number of violations and $T_0 = T - T_1$ as the number of non-violations.

1. Unconditional Coverage Test (CLR$_{uc}$). The null hypothesis $H_0\colon \pi = 1-\alpha$ tests whether the frequency of violations equals the expected rate. The likelihood ratio statistic is

$$\text{CLR}_{uc} = -2[T_0 \ln(1-\alpha) + T_1 \ln(\alpha) - T_0 \ln(1-\widehat{\pi}) - T_1 \ln(\widehat{\pi})]\,, \qquad \widehat{\pi} = \frac{T_1}{T},$$

which asymptotically follows $\chi^2(1)$ under H_0.

2. Independence Test (CLR$_{ind}$). This test evaluates whether violations occur independently over time. Let T_{ij} be the number of transitions from state i at $t-1$ to state j at t, for $i, j \in \{0, 1\}$. Then define

$$\widehat{\pi}_{01} = \frac{T_{01}}{T_{00}+T_{01}}, \qquad \widehat{\pi}_{11} = \frac{T_{11}}{T_{10}+T_{11}},$$

and compute

$$\text{CLR}_{ind} = -2\ln \frac{(1-\widehat{\pi})^{T_{00}+T_{10}}\widehat{\pi}^{T_{01}+T_{11}}}{(1-\widehat{\pi}_{01})^{T_{00}}\widehat{\pi}_{01}^{T_{01}}(1-\widehat{\pi}_{11})^{T_{10}}\widehat{\pi}_{11}^{T_{11}}},$$

which also follows $\chi^2(1)$ under the null hypothesis of independence.

3. Conditional Coverage Test (CLR$_{cc}$). The joint test of both coverage and independence is performed by summing the two components:

$$\text{CLR}_{cc} = \text{CLR}_{uc} + \text{CLR}_{ind},$$

and follows $\chi^2(2)$ under the null. Low p-values indicate either incorrect coverage, violation clustering, or both.

These tests are performed in this replications, and the results are below.

CLR Test Results

Table 4: Christoffersen Likelihood Ratio (CLR) test results for ARMA(1,1)–GARCH(1,1) models across horizons.

Horizon	**Model**	N	CLR$_{uc}$ (p)	CLR$_{ind}$ (p)	CLR$_{cc}$ (p)
1-year	Normal–ARMA–GARCH	252	5.42 (0.0199)	0.40 (0.5262)	5.83 (0.0543)
1-year	Student-*t*–ARMA–GARCH	252	0.75 (0.3880)	0.13 (0.7189)	0.87 (0.6458)
2-year	Normal–ARMA–GARCH	504	10.85 (0.0010)	0.72 (0.3959)	11.57 (0.0031)
2-year	Student-*t*–ARMA–GARCH	504	5.32 (0.0211)	1.44 (0.2299)	6.76 (0.0340)
4-year	Normal–ARMA–GARCH	1008	7.67 (0.0056)	0.68 (0.4100)	8.34 (0.0154)
4-year	Student-*t*–ARMA–GARCH	1008	2.11 (0.1464)	1.53 (0.2164)	3.64 (0.1623)
10-year	Normal–ARMA–GARCH	2510	35.27 (0.0000)	3.30 (0.0692)	38.57 (0.0000)
10-year	Student-*t*–ARMA–GARCH	2510	14.11 (0.0002)	3.51 (0.0609)	17.62 (0.0001)

Across all time periods, the Student-t ARMA–GARCH generally delivers lower CLR statistics and higher p-values than the normal specification, especially for unconditional coverage. At the 1-year horizon, both models pass the joint conditional coverage test (Normal: CLR$_{cc}$ = 5.83, $p = 0.054$; Student-t: CLR$_{cc}$ = 0.87, $p = 0.646$). At 2 years, the normal model is rejected (CLR$_{cc}$ = 11.57,

$p = 0.003$) and the Student-t model is borderline but still rejected at the 5% level ($\mathrm{CLR}_{cc} = 6.76$, $p = 0.034$). At 4 years, the normal model is rejected ($\mathrm{CLR}_{cc} = 8.34$, $p = 0.015$) while the Student-t model is not ($\mathrm{CLR}_{cc} = 3.64$, $p = 0.162$). Over long horizons (10–20 years), both models fail (Normal: $\mathrm{CLR}_{cc} \approx 38.6$, $p < 0.001$; Student-t: $\mathrm{CLR}_{cc} \approx 17.6$, $p < 0.001$). These rejections are primarily due to coverage (CLR_{uc}) rather than independence (CLR_{ind} $p \in [0.06, 0.72]$), which demonstrates mild evidence of violation clustering at long horizons. These patterns are consistent with Kim et al. (2011)in that heavy-tailed innovations improve short- to medium-horizon VaR calibration, but neither specification maintains correct coverage over longer-period samples.

Berkowitz Likelihood Ratio (BLR) Tests

While the Christoffersen tests assess the frequency and independence of VaR exceedances, they do not evaluate the full predictive distribution of losses. The Berkowitz Likelihood Ratio (BLR) test provides a stricter validation by testing whether the Probability Integral Transform (PIT) values derived from the model are consistent with a standard normal distribution. In this study, following Kim et al. (2011), the test is applied to forecasts of both Value-at-Risk (VaR) and Average Value-at-Risk (AVaR), as AVaR incorporates information from the tail beyond the VaR threshold and provides a smoother loss distribution for evaluation.

Mathematical Framework. Let p_t denote the PIT values obtained from the model, defined as

$$p_t = F_t(L_t),$$

where $F_t(\cdot)$ is the model-implied cumulative loss distribution. Under a correctly specified model, $p_t \sim \mathrm{Uniform}(0, 1)$. Transforming to the standard normal scale gives

$$z_t = \Phi^{-1}(p_t),$$

where Φ^{-1} is the inverse standard normal CDF. If the model is correctly calibrated, then $z_t \sim \mathcal{N}(0, 1)$ and should exhibit no serial dependence.

The test estimates an AR(1) model for the transformed series:

$$z_t = \mu + \rho z_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

The null hypothesis $H_0 : (\mu, \rho, \sigma^2) = (0, 0, 1)$ implies correct distributional calibration. The likelihood ratio statistic is computed as

$$\mathrm{BLR} = -2(\ln L_0 - \ln L_1),$$

where L_0 and L_1 are the restricted and unrestricted log-likelihoods respectively. Under H_0, $\mathrm{BLR} \sim \chi^2(3)$.

BLR Test Results

Table 5: Berkowitz Likelihood Ratio (BLR) test results for ARMA(1,1)–GARCH(1,1) models. BLR_{ind} tests full PIT independence; BLR_{tail} tests tail calibration for PITs $\leq 1\%$.

Horizon	**Model**	BLR_{ind} LR3	BLR_{ind} p	BLR_{tail} LR3	BLR_{tail} p	Tail n
1-year	Normal–ARMA–GARCH	4.6998	0.1951	5.0379	0.1690	7
1-year	Student-t–ARMA–GARCH	3.4033	0.3335	–	–	4
2-year	Normal–ARMA–GARCH	8.3121	0.0400	32.6492	0.0000	14
2-year	Student-t–ARMA–GARCH	8.0055	0.0459	2.4254	0.4889	11
4-year	Normal–ARMA–GARCH	13.2712	0.0041	65.0591	0.0000	20
4-year	Student-t–ARMA–GARCH	12.5067	0.0058	6.6306	0.0847	15
10-year	Normal–ARMA–GARCH	14.4016	0.0024	67.9269	0.0000	60
10-year	Student-t–ARMA–GARCH	14.0842	0.0028	7.9910	0.0462	46

The BLR results provide further insight into how well each model captures both the dynamics and tail behaviour of the loss distribution. For the independence test (BLR_{ind}), the Normal–ARMA–GARCH model exhibits serial dependence from the 2-year horizon ($p = 0.04$), while the Student-t model does so beyond 4 years. Overall, the Student-t specification produces smaller test statistics and higher p-values, indicating a more consistent calibration of conditional losses.

The tail test (BLR_{tail}), which focuses on the lowest 1% of PIT values, highlights clear differences in tail calibration. The Normal model is strongly rejected at nearly all horizons, consistent with its known tendency to underestimate the severity of extreme losses. The Student-t model performs better, as its heavier tails align PITs more closely with the empirical loss distribution. However, at a long horizon (10 years) it also fails the tail test at the 5% level ($p \approx 0.046$). Hence, while both models lose accuracy over longer-period samples, the Student-t ARMA–GARCH remains the better-calibrated and more realistic representation of tail risk, consistent with the findings of Kim et al. (2011).

AVaR Comparison

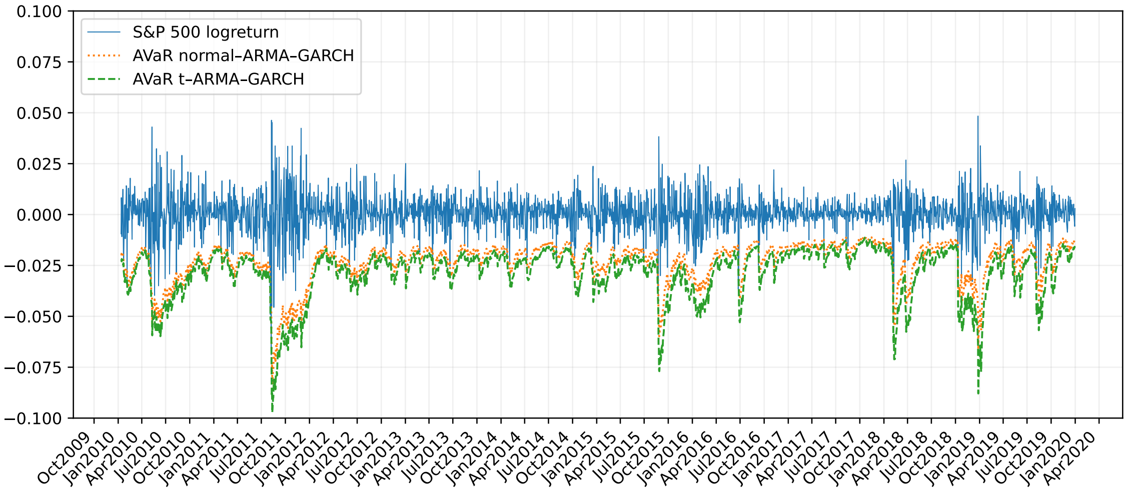

Figure 1: The 1% Average Value-at-Risk (AVaR) series, $-\mathrm{AVaR}_{t,0.01}(y_{t+1})$, for the S&P 500 under the normal–ARMA–GARCH and Student–t–ARMA–GARCH models.

Figure 1 displays the 1% Average Value-at-Risk (AVaR) series from the normal–ARMA–GARCH and Student–t–ARMA–GARCH models, plotted against the S&P 500 daily log returns. Both models demonstrate strong volatility clustering, which is consistent with real-world financial returns, with the t-ARMA–GARCH model producing deeper and more variable tail estimates due to its heavier-tailed innovation distribution. This is consistent with the findings in Kim *et al.* (2011).

Absolute Relative Difference (ARD)

The Absolute Relative Difference (ARD) quantifies the magnitude of difference between the Student-t and Normal ARMA–GARCH Value-at-Risk (VaR) forecasts, as in Kim et al. (2011).

It is a scale-free measure of how much the VaR estimates diverge between the two distributional assumptions across different forecast horizons.

Mathematical Definition. Let $\mathrm{VaR}_{t,\alpha}^{(t)}$ and $\mathrm{VaR}_{t,\alpha}^{(n)}$ denote the VaR estimates at confidence level α from the Student-t and Normal models respectively. The ARD at each horizon is defined as:

$$\mathrm{ARD} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\mathrm{VaR}_{t,\alpha}^{(t)} - \mathrm{VaR}_{t,\alpha}^{(n)}}{\mathrm{VaR}_{t,\alpha}^{(n)}}\right|.$$

A smaller ARD value indicates that the two models produce similar VaR forecasts, while larger values imply that the Student-t model deviates substantially from the Normal model—typically reflecting periods of heightened volatility or heavier tails.

Results

Table 6: Absolute Relative Difference (ARD) between Student-t and Normal ARMA–GARCH VaR forecasts across horizons.

Horizon	**ARD (t vs Normal)**
1-year	0.140649
2-year	0.132810
4-year	0.100105
10-year	0.090967

The ARD values decrease with horizon duration, falling from approximately 14% at the 1-year horizon to 9% beyond 10 years. This indicates that differences between the Student-t and Normal VaR estimates are larger in shorter-term forecasts, where volatility and tail sensitivity have a stronger effect. As the horizon extends, the forecasts seem to converge, suggesting that longer-horizon VaR estimates are less sensitive to the innovation distribution. This mirrors the findings of Kim et al. (2011), where the Student-t model produced higher short-term VaR levels, better capturing extreme downside risk, while the two models aligned more closely in the long run.

Conclusion

Overall, the results from this replication closely align with the findings of Kim et al. (2011). Across all model specifications, the Student-t versions consistently outperform the Normal models, particularly in capturing the heavy tails and volatility clustering characteristic of financial crisis periods. The GARCH and ARMA–GARCH frameworks both show that while volatility persistence is well modelled, normal innovations systematically underestimate extreme losses, leading to poor tail calibration. In contrast, the Student-t models deliver more accurate VaR and AVaR forecasts and produce lower KS, AD, CLR, and BLR statistics, indicating a better overall fit. The ARD results further confirm that the gap between the two distributions narrows over longer horizons, as the effect of tail thickness decreases. In general, these outcomes reinforce the paper's central conclusion that incorporating heavy-tailed innovations is essential for realistic risk measurement and improved predictive performance during periods of financial stress.

References

[Kim et al.(2011)Kim, Rachev, Bianchi, Mitov, and Fabozzi] Kim, Y. S., S. T. Rachev, M. L. Bianchi, I. Mitov, and F. J. Fabozzi. 2011. Time series analysis for financial market meltdowns. Journal of Banking & Finance 35:1879–1891.