VRIJE UNIVERSITEIT

Group Assignment 2: Model Validation

Quantitative Financial Risk Management

Authors: A 15

Jip de Boer j.a4.de.boer@student.vu.nl (2710065)

Yunji Eo y.eo@student.vu.nl (2735445)

Lars de Graaff l.de.graaff@student.vu.nl (2790679)

Xuan Zhou x.zhou9@student.vu.nl (2653511)

Supervisor:

Dr. C.S. Bos

May 1, 2025

1 Documentation

The model documentation of the report from group 14 demonstrates a thorough analysis accompanied by a systematic structure. It comprehensively addresses the most important components necessary for a robust financial risk assessment. It systematically outlines the data foundation, modeling choices and validation procedures where the assumptions are clearly justified. The data analysis could benefit from more comprehensive tabular statistics as most analysis is now only based on graphs. The interpretations coming forward from their analysis are insightful, specifically where results are linked to past crisis events.

Despite the many topics discussed, the report remains largely concise and is understandable to a reader with a background in quantitative finance. It avoids unnecessary derivations or standard formulas. Furthermore, the report avoids redundancy as it clearly moves from model description to result visualization to backtesting and finally to stress testing. However, a notable drawback could be that readers who are not familiar with statistical backtesting might find parts like the Weibull likelihood ratio test and the t-tests on the K-residuals a hard to understand without a more extensive explanation of such methods.

The graphs and tables are highly effective in supporting and illustrating the report its findings. Graphical outputs, such as time series plots of VaR/ES estimates and Q-Q plots, are integrated correctly and relevant to the discussion that they are tied to. Additionally, the tables which summarize the VaR/ES outcomes, backtesting results and stress testing impact are organized, clear, and make it easy to interpret complex information. Furthermore, the distinction between the full sample, crisis and non-crisis periods is clearly defined, which enhances the clarity and depth of the analysis; nevertheless, they do not specify the dates, which poses potential deviation for reimplementation.

Overall, the documentation of the report is of a good standard and in the realm what you would expect from a professional financial risk analyst.

2 Theoretical Assessment

The report begins by establishing a solid data foundation. First, it details the securities incorporated into the portfolio, which include three market indices alongside a loan component derived from the EURIBOR rate plus a credit spread. The data sourcing and input methodology is explained clearly and handled accurately. Specifically, the closing prices of the market indices were obtained via the yfinance Python package, which is from Yahoo Finance, while the 1-week EURIBOR rates were sourced from the Deutsche Bundesbank. Both sources are widely recognized for their reliability and suitability for financial data and analysis. Missing values within the time series were addressed using time-based interpolation. This is regarded as a standard and widely accepted technique for handling gaps in financial data. Furthermore, loan returns were calculated appropriately as they applied the correct actual/360 day-count convention to the interest rate data.

Nevertheless, certain methodological decisions can be questioned. Specifically, the report adjusts all asset returns to a common EUR base as they apply exchange rate transformations which is a step that seems somewhat unnecessary given that portfolio construction is performed on a percentage return basis. This additional step may introduce unneeded complexity and the potential for inaccuracies.

Furthermore, several important details are missing or presented unclearly for the reader. First, the rationale behind the specific asset weights is not discussed. Why for example are GSPC and Nikkei each given a 25% weight while IBEX-35 is only 20%? Also, the choice of a 1.5% arbitrary credit spread in the loan component is also unexplained. This could undermine the credibility of the synthetic asset. Moreover, the description of the loan component oversimplifies the complexities involved in modeling credit risk. To give an example, there is no mention of the credit rating, default probability nor of the duration. Even though these choices might have been arbitrary, a trustworthy financial risk advisor would have come up with a reason to choose these specific portfolio settings.

Continuing from the data section, the report presents a highly structured and detailed selection of methodologies for modeling financial risk which is in line with industry standards. The authors employ a variety of models to estimate portfolio Value-at-Risk (VaR) and Expected Shortfall (ES). This includes the Variance-Covariance (VC) method under both normal and Student's t-distributional assumptions, Historical Simulation (HS), GARCH(1,1) with Constant Conditional Correlation (CCC) and Filtered Historical Simulation (FHS-EWMA).

For the VC model under the normality assumption where portfolio returns are presumed to be jointly normally distributed, the authors assess its validity using a Q-Q plot. The Q-Q plot reveals significant deviations from the expected linear pattern, especially in the tails. This indicates the presence of fat tails. This on its turn suggests that the normal distribution underestimates the probability of extreme losses as the authors suggest.

To explore alternative distributional assumptions, the authors also assess the goodness-of-fit of the Student's t-distribution for the VC model with varying degrees of freedom (df). To do so they again deploy Q-Q plots. As expected, higher degrees of freedom lead the t-distribution to approximate the normal distribution more closely. Among the tested values (df = 3, 4, 5, and 6), the distribution with df = 3 demonstrates the best fit to the empirical data which is shown in the graphical analysis of the report.

The formulas used for calculating the VaR and ES under both the normal and Student's t assumptions using the VC method are correctly specified and also consistent with the lecture materials. Notably, the transformation from the t-distribution scale to standard deviation is handled appropriately through the use of the correct value for the degrees of freedom (df = 3).

In addition to parametric approaches the authors implement the Historical Simulation (HS) method which differs as it makes no assumptions about the return distribution. Instead, it directly uses empirical quantiles derived from historical portfolio returns to estimate risk. The Historical Simulation steps are documented correctly.

The report also includes a GARCH(1,1) model within the Constant Conditional Correlation (CCC) framework. This multivariate GARCH-CCC model assumes constant correlations between assets while allowing time-varying volatility. While the CCC framework holds correlations fixed, more advanced models such as the Dynamic Conditional Correlation (DCC-GARCH) permit correlations to evolve over time. This offers greater flexibility in capturing market dynamics. However, the GARCH(1,1)-CCC steps are documented correctly and are in line with the lecture slides.

Lastly, the report discusses the Filtered Historical Simulation (FHS) method which includes aspects of both historical and parametric approaches by applying a volatility filter: specifically, an Exponentially Weighted Moving Average (EWMA). Unlike the other modeling approaches, which are described in considerable detail, the documentation of the FHS method is notably brief. The authors provide only a short reference to McNeil et al. (2015) without a clear step-by-step explanation of the methodology. This lack of elaboration places a heavy burden on the reader who because of this is expected to possess some high level prior knowledge to fully understand the implementation. It would have been preferred to explain the FHS procedure in a structured manner similar to the presentation of the GARCH(1,1)-CCC model. Moreover, the assumptions of the FHS approach are not sufficiently addressed, in example the EWMA λ which leaves an important gap in the methodological calibration.

The report demonstrates a thoughtful and technically competent approach to calibration. For both the normal and Student's t specifications in the VC method a rolling estimation window of 261 trading days is employed to dynamically recalibrate mean return vectors and covariance matrices. This indicates a correct recalibration process. Furthermore, as already mentioned the Student's t-distribution is calibrated manually through the inspection of Q-Q plots for the degrees of freedom (choosing df = 3). The GARCH(1,1) model with Constant Conditional Correlation (CCC) was calibrated via Maximum Likelihood Estimation (MLE) for each individual asset. The time-varying conditional variances were derived recursively while the correlation matrix was estimated from standardized residuals and assumed constant. Similarly, the

FHS model was calibrated using an EWMA filter, although, as already mentioned, the decay factor (λ) is not reported which makes it difficult to assess the sensitivity of the estimates to this parameter.

Unfortunately, the report does not specify how initial parameter values were chosen for any of the models. This is a notable omission as the initialization strategy can influence convergence and final estimates in models like GARCH. Additionally, the distinction between in-sample and out-of-sample performance is not addressed. This may leave the reader uncertain about how the models perform in predictive contexts, despite the ample dataset. A more nuanced discussion of parameter estimates, including their economic interpretation and relevance in different market regimes, would enhance the report's depth. For instance, an illustration of how parameters relate to specific economic conditions would help the reader better understand the practical implications of the modeling choices.

3 Quantitative Assessment

In the report, from the perspective of the estimation, backtesting and stress testing results, the recommended modeling approaches are *GARCH-CCC(1,1)* and *FHS-EWMA*, and it is further concluded that these two approaches can be the most efficient to capture portfolio risks in both normal and stress periods. Thus, for the quantitative assessment part, we will focus on validation towards these two models with the components of implementation accuracy, model performance, calibration and assumption tests, modeling techniques, and reporting.

Implementation

The report includes a detailed calculation procedure and VaR and ES results for five methods and demonstrates it in the form of the Q-Q plot, the time series plot and tables. For the estimation, the authors deploy the rolling window approach with a duration of 261 days, which is the mean of sample trading days, and compute the 1 day VaR and ES at the levels 97. 5% and 99%. In addition, the authors also show the risk measure estimates of the full sample, the crisis periods, and the non-crisis periods to consider the impacts coming from shocks and stresses. Note that for the crisis & non-crisis periods, we follow the definition of stress periods in the tested report, although the definition does not specify dates, we define the periods graphically based on Figure 1 of the validated report. (from 2015-06-01 to 2016-06-01, 2020-02-01 to 2020-05-01, and 2021-01-01 to 2023-01-01).

Figure 1 and Figure 2 show the Q-Q plot with respect to 97.5% and 99% VaR violation spacing with exponential distribution of the GARCH-CCC and FHS-EWMA methods. 97.5% VaR level of GARCH-CCC shows better fitting compared to 99% level one. The plot shows that this method is capable to capture

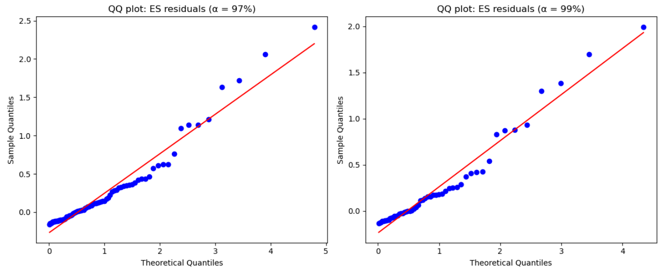


Figure 1: Q-Q plot for VaR of the spacings between violations of the **97.5% VaR** and **99% VaR** against the corresponding quintiles of the exponential distribution of GARCH-CCC

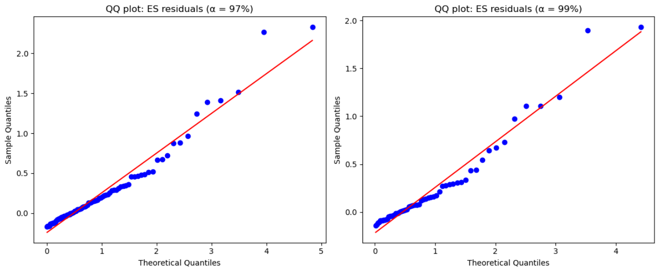


Figure 2: Q-Q plot for VaR of the spacings between violations of estimates of the **97.5% VaR** and **99% VaR** against the corresponding quintiles of the exponential distribution of FHS-EWMA

the most of risks while may underestimate risks of extreme cases. And for the FHS-EWMA case, the 99% VaR violation spacings are closer to the red line, underscoring that this method can capture tail risks. The conclusion is aligned with that of the validated report Figure A5.

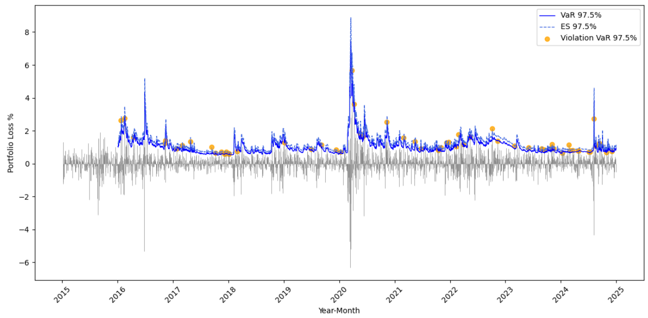


Figure 3: 97.5% level of VaR, VaR violation and ES of GARCH-CCC

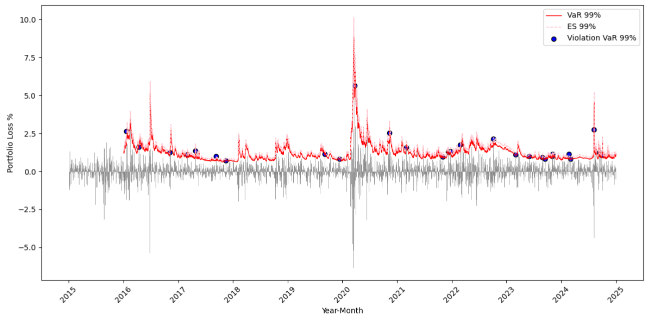


Figure 4: 99% level of VaR, VaR violation and ES of GARCH-CCC

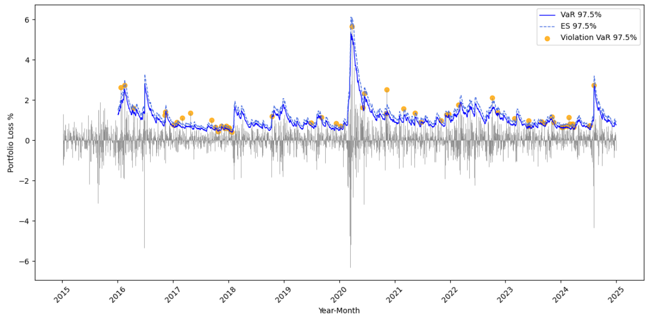


Figure 5: 97.5% level of VaR, VaR violation and ES of FHS-EWMA

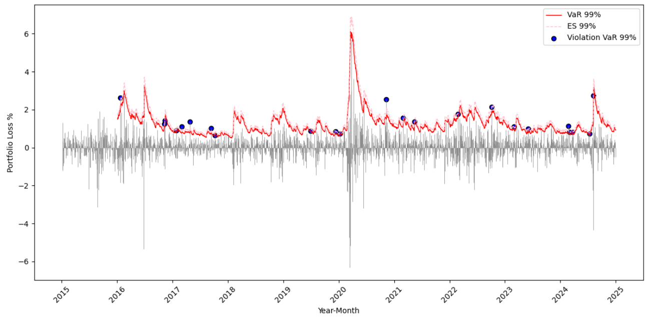


Figure 6: 99% level of VaR, VaR violation and ES of FHS-EWMA

We follow the innovative rolling window VaR and ES estimation as the one indicated in the validated report, the resulting plots are shown as Figure 3 and Figure 4 for GARCH-CCC and Figure 5 and Figure 6 for FHS-EWMA for the level of level of 97.5% and 99%, respectively. For the case of GARCH-CCC, the shape and magnitude of VaR and ES for 97.5% and 99% are replicated similarly, and the clustered violations around 2018 and 2020 are successfully captured by both implementations, although the original report shows more violation points for VaR 99% than what we reproduced, for example, around 2020. For the case of FHS-EWMA, the replicated VaR and ES are lower than the ones in the validated report, and clustered violation around 2020 does not show for the re-implementation results, while the high peak around 2020 especially for 99% level is captured by both estimations.

Table 1: VaR and ES estimates of GARCH-CCC and FHS-EWMA methods under different periods

Method	Level	Full Sample	Crisis Period	Non-Crisis Period
GARCH-CCC	VaR 97.5%	0.935%	1.159%	0.870%
	VaR 99%	1.114%	1.373%	1.041%
	ES 97.5%	1.120%	1.380%	1.046%
	ES 99%	1.280%	1.570%	1.199%
FHS-EWMA	VaR 97.5%	0.864%	1.102%	0.833%
	VaR 99%	1.030%	1.305%	0.998%
	ES 97.5%	1.036%	1.312%	1.003%
	ES 99%	1.184%	1.493%	1.150%

Furthermore, we calculate the risk measures for the entire sample, and divide it into crisis and non-crisis periods to compare Tables 4 and 5 in the validated report with the results indicated in Table 1. For both methods, the replicated VaR and ES for both levels are slightly lower than the original reporting results, no matter for the full sample, crisis periods, or non-crisis periods. Another difference is that the estimates from FHS-EWMA are lower than those from GARCH-CCC. While the observation that the risk measures in crisis periods are higher than non-crisis periods is aligned with the validated report. The reasons for different results of estimates can be several, which can include, different initial values in estimation, different portfolio weights / mean update approach, for example, in our estimation case, we apply the fixed portfolio mean vector as the mean value, and the covariance matrix is from the last day, the different optimizer for GARCH-CCC model. In addition, for the GARCH-CCC method, we exclude the cash from the portfolio and start the estimation, which may also contribute to the implementation differences.

Table 2: Weibull LR test for VaR and t-test on K-residuals for ES

			VaR		ES	
Method	Level	n	LR-stat	LR p-value	t-stat	t p-value
GARCH-CCC	97.5%	83	5.0297	0.0249	4.3506	0.0000
	99%	53	9.1837	0.0024	3.8635	0.0003
FHS-EWMA	97.5%	87	8.0372	0.0046	4.7339	0.0000
	99%	57	9.3238	0.0023	4.1233	0.0001

For the backtesting part, the Weibull LR test for VaR and the K residual t-test for ES are also performed; the results are indicated in Table 2. Weibull LR test can be treated as an alternative to the Q-Q plot of spacing data against quintiles of exponential distribution. As the validated report indicates, the GARCH-CCC cannot reject the null hypothesis that the exceedance follows standard exponential distributions at the significance level $\alpha = 0.05$, and FHS-EWMA can pass if the significance level changes to $\alpha = 0.01$. And for the replicated results, GARCH-CCC at 97.5% level can support the null hypothesis in the 0.01 case, while the rest cannot. For ES, the violation-based t-test is performed, since empirical violation residuals are expected to be like the realizations of i.i.d. variables from the distributions with the mean of zero, the null hypothesis is that the violation residuals have the same characteristics. The replicated violation numbers are higher than the original results for both the α level, and the p-values indicate that the two selected models fail the violation residual test and can reject the null hypothesis that the non-zero violation residuals have a mean of zero.

Model performance - Sensitivity Analysis

Table 3: Sensitivity Analysis of FHS-EWMA Model for Different λ Values

λ	Confidence Level	VaR (%)	ES (%)
0.94	97.5%	0.9481	1.1498
0.94	99.0%	1.1153	1.3397
0.97	97.5%	1.1061	1.3772
0.97	99.0%	1.3013	1.6119
0.99	97.5%	1.3123	1.6944
0.99	99.0%	1.6808	2.0522

In the original report, sensitivity analysis was not performed. However, we considered it important to assess the impact of key parameters, particularly λ in the FHS-EWMA model, as this model demonstrated strong performance. Therefore, we selected the FHS-EWMA model for additional sensitivity analysis. As shown in Table 3, the model exhibits significant variation depending on the chosen value of λ, indicating high parameter sensitivity. For instance, increasing λ from 0.94 to 0.99 raises the 99% VaR from 1.1153% to 1.6808%, a substantial difference of approximately 50.7%.

Model performance - Stress Testing

The report provides a structured stress testing framework, applying shocks to equities, FX rates, and interest rates. The scenarios are clearly defined and applied consistently across different risk models. However, no benchmark was initially used to evaluate whether the resulting changes in Value-at-Risk (VaR) are realistic. Hence, we compare the VaR values between the crisis period and the non-crisis period and calculate the percentage change based on our Implementation part above. These actual changes provide a benchmark to assess whether the VaR values changes in the original report are reasonable.

Table 4: Crisis vs. Non-crisis VaR Percentage Change

Method	VaR97.5_Crisis Change(%)	VaR99_Crisis Change(%)
GARCH-CCC	23.96	23.25
FHS-EWMA	27.55	26.70

Table 4 indicates that during the crisis period, the VaR values increased by approximately 24–27%. In the original report, under downward shock scenarios, the GARCH and FHS models showed around 20% value changes, except in the case of the equity –20% scenario that has 4% value changes. Overall, the observed VaR changes under both crisis and stress testing scenarios can be considered realistic while it should be noted that the models may still underestimate risks under extreme scenarios.

Calibration test

For the calibration approach test, since the models do not involve parameter calibration explicitly, but it re-estimates 1 day VaR and ES based on the rolling window of 261 days, which is the average yearly number of trading days of the selected sample. Thus, the calibration test focuses on validating the stability of parameter evolution, rolling window size selection, and the consistency of the underlying market dynamics.

Figure 7 and Figure 8 compare the estimated volatility of the portfolio and the VaR and ES levels of 97.5% and 99% for the GARCH-CCC and FHS-EWMA method. It can be observed that both models can capture the volatility of the market and computed VaR and ES changes along it. GARCH-CCC generates the risk

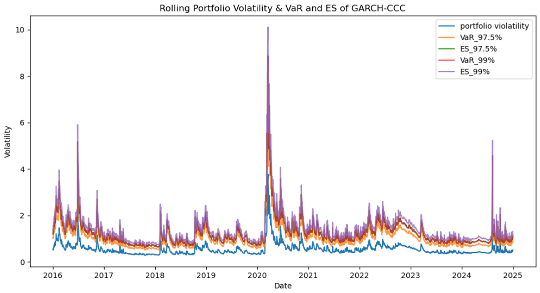


Figure 7: Comparison of volatility and **97.5%** and **99%** VaR and ES of GARCH-CCC

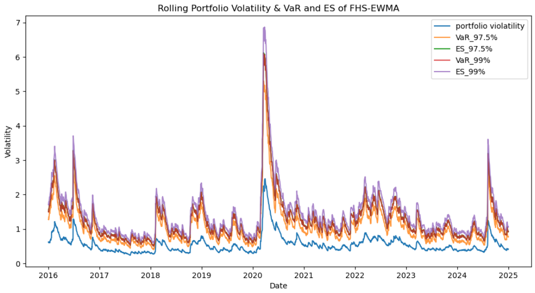


Figure 8: Comparison of volatility and **97.5%** and **99%** VaR and ES of FHS-EWMA

measures with the significant higher magnitudes and FHS-EWMA creates smoother estimates but still captures the shocks such as the covid-pandemic around 2020, this observation stands for the takeaways that the parameter evolution matches with statistical logic of risk measures and features of market condition.

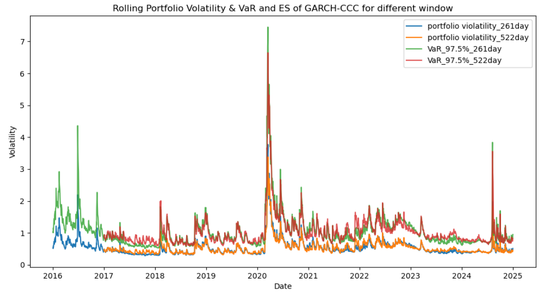


Figure 9: Comparison of volatility and **97.5%** VaR and ES of GARCH-CCC for 1-year and 2-year window

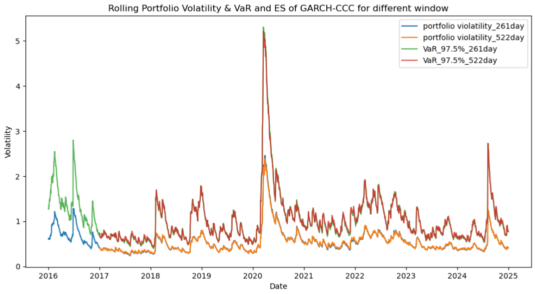


Figure 10: Comparison of volatility and **97.5%** VaR and ES of FHS-EWMA for 1-year and 2-year window

Regarding the perspective of rolling window selection calibration, in Figure 9 and Figure 10, we further compare the estimates of the window of 261 and 522 days for both methods, approximately accounting for 1-year and 2-year in-sample periods. We observe that the GARCH-CCC model reacts sharply to shocks, but if the horizon is widened, the estimates tend to be relatively smoother. It can be concluded that the selection of the current window may be the robust option from the point of view of risk analysis. For the FHS-EWMA method, the reaction is less sensitive and the peak magnitude is lower than that for the GARCH-CCC case. The selection of different estimation horizons in FHS-EWMA is less distinctly since the generated VaR and ES has little differences graphically.

Assumption test

In this subsection, in order to test for assumptions lying within the validated report, we carried out the testing on the selection of the GARCH-CCC model distribution, autocorrelation, and whether the standardized residuals are drawn from the same distribution, also including the discussion of constant conditional correlation within the model setting for both methods.

For the implementation of GARCH-CCC, the standardized residuals in the GARCH part modeling are assumed to follow a normal distribution. We also estimate the one with the student's t distribution to compare the results based on this assumption. The results are shown in Figure 11, which demonstrates the finding that both two distributions generate quite close estimates in the regular periods, as the lines mostly intersect. However, during turmoil periods, like periods around the first half of 2020, the student's t distribution generates quite high peak compared to the normal distribution, showing the feature of heavy-tailed and capability to capture extreme risks.

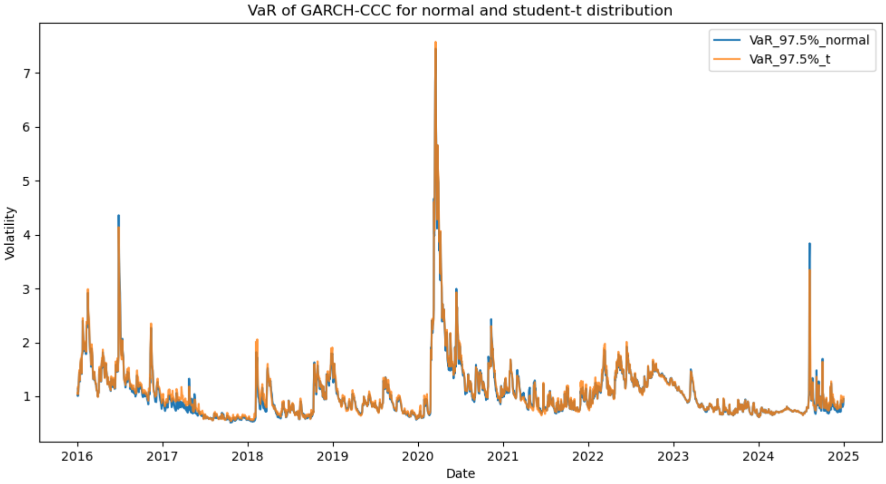


Figure 11: Comparison between normal and student's t distribution estimates of the **97.5% VaR** of GARCH-CCC

Considering the fact that the standardized residuals of estimation form the bases to estimate emprical VaR and ES, we also conducted the Ljung-Box (LB) test and Kolmogorov-Smirnov (KS) towards the standardized residuals of estimation, where the former is used to check the model assumption that the standardized residuals are i.i.d., i.e., there is no serial autocorrelation, and the Kolmogorov-Smirnov (KS) test is also included to test if the standardized residuals from the sub-samples are drawn from the same distribution to check the distributional invariance. The packages of `acorr_ljungbox` and `ks_2samp` are applied for this purpose. For the LB test, the lags of 10 and 20 are selected, and for the KS test, we selected two scenarios for testing, the 50% cut means we split the sample into half, one from 2016-01-04 to 2020-07-01 and another from 2020-07-02 to 2024-12-31. For the crisis & non-crisis indicator, we follow the definition stated in the beginning of this section. The results are listed in Table 5.

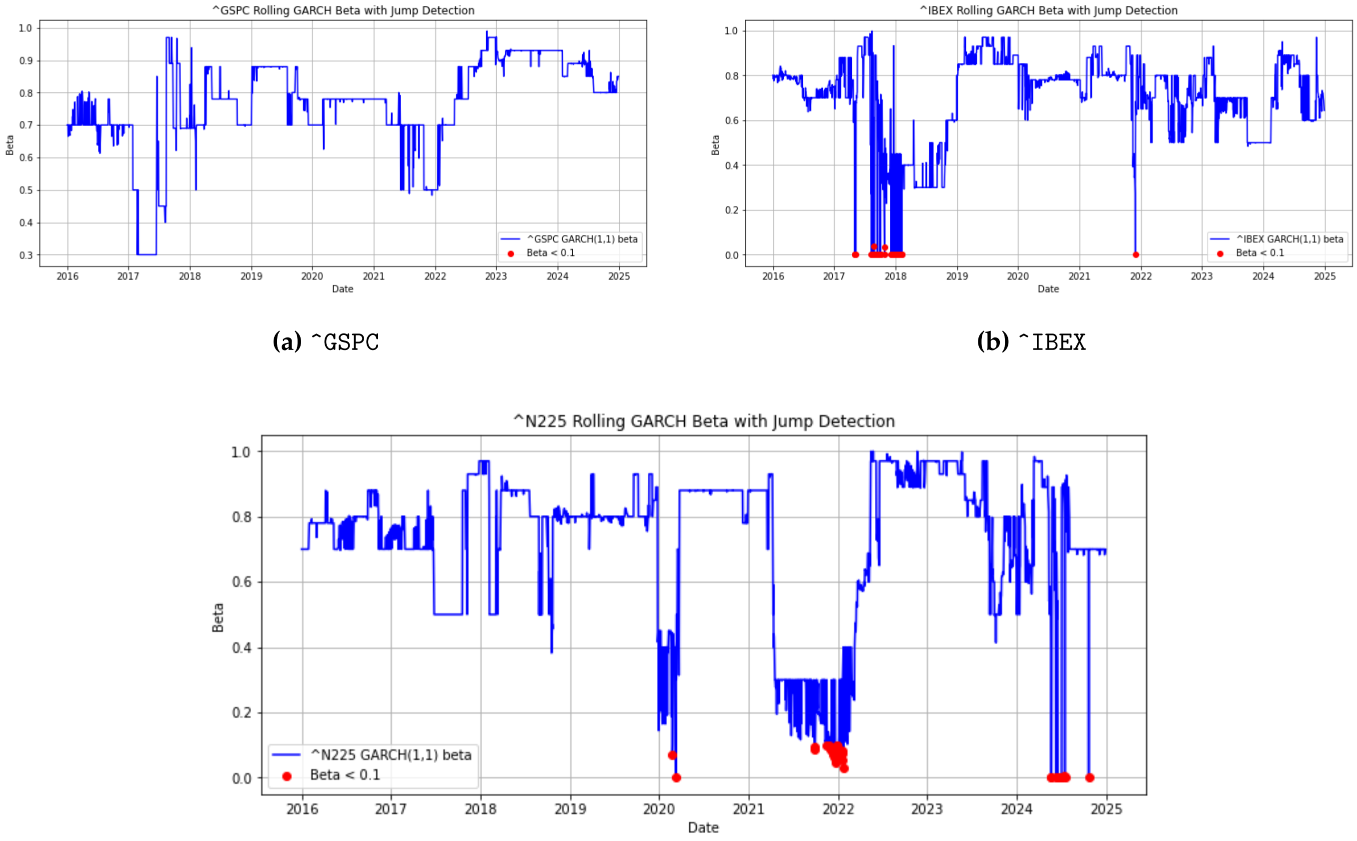
Table 5: Ljung-Box (LB) test and Kolmogorov-Smirnov (KS) test results for GARCH-CCC and FHS-EWMA

Model	Lag/Period	LB or KS Stat	p-value
GARCH-CCC	Lag 10	131.62	0.000
	Lag 20	141.99	0.000
FHS-EWMA	Lag 10	411.81	0.000
	Lag 20	459.68	0.000
GARCH-CCC	50% Cut	0.0226	0.915
	Crisis & Non-crisis	0.0772	0.006
FHS-EWMA	50% Cut	0.0858	0.000
	Crisis & Non-crisis	0.0686	0.019

For the LB test results, for two methods, the hypothesis of i.i.d. standardized residuals is rejected, meaning there exists autocorrelation and the models need to be adjusted or re-considered. For the results of the KS test, it shows that for the GARCH-CCC model, the 50% cut scenario can reject the hypothesis and conclude that the data of the standardized residuals are drawn from the same distribution, while the rest cannot reject this hypothesis. We interpret this as the shocks or stresses within the sample bring challenges towards the method of historical simulation, which relies on quintiles of empirical portfolio returns and losses, and the method of separating different market condition regimes can somehow deal with the drawbacks of this modeling approach.

Also, it should be noted that for both GARCH-CCC and FHS-EWMA, one of the assumptions is that the conditional correlation matrix is assumed to be constant for all t, which has the benefit of easy implementation, while the characteristic of constant conditional correlation is considered unrealistic, the performance of the estimation could be influenced during times of stress, and the impacts of news on financial markets require models that can capture the dynamic evolution of conditional correlation and volatility. The possible innovation point is to introduce the dynamics of conditional correlation, for example, to deploy the dynamic conditional correlation (DCC) approach, which is developed on the basis of CCC and allows conditional correlations to change dynamically.

Modeling techniques - Convergence and Stability Test



(a) `^GSPC`

(b) `^IBEX`

(c) `^N225`

Figure 12: Rolling GARCH(1,1) β for Individual Indices with Jump Detection

There is no indication in the original report of convergence or stability tests for the numerical methods. In our re-implementation, the rolling-window VaR and ES estimates evolve smoothly over time, with noticeable increases primarily during known crisis periods such as COVID-19 and the 2015–2016 market stress. This initially suggested no clear convergence or stability issues.

However, to further examine the robustness of the model, we investigated the behavior of estimated parameters—specifically the GARCH(1,1) β coefficient—under rolling window estimation. Figure 12 shows that while the β values for GSPC, IBEX, and N225 mostly remain in the 0.7–0.9 range, indicating general stability, there are periods of abrupt changes. For example, N225 shows a sharp drop in β during the COVID-19 crisis, which aligns with a spike in VaR. Yet, in other periods such as 2022, we observe β fluctuations without a corresponding VaR shift. This discrepancy suggests that parameter instability may exist even when the VaR measure appear smooth, suggesting there might be potential limitations of the GARCH-CCC model's ability to fully reflect underlying dynamics.

To check if the results of the underlying numerical method are stable, alternatively, we also compare the one-day risk measures estimated with a fixed window based on the full sample, the first window, and the last window, where the first window stands for the first 261 trading days and the last window represents the last 261 trading days of the full sample. The results are listed in Table 6.

Table 6: VaR and ES estimates of full sample, first window and last window for GARCH-CCC and FHS-EWMA

Model	Measures	Full sample	First window	Last window
GARCH-CCC	*97.5% level result*			
	VaR	0.935%	1.033%	0.863%
	ES	1.120%	1.234%	1.040%
	99% level result			
	VaR	1.114%	1.228%	1.034%
	ES	1.280%	1.407%	1.19%
FHS-EWMA	*97.5% level result*			
	VaR	0.864%	1.206%	0.761%
	ES	1.036%	1.441%	0.919%
	99% level result			
	VaR	1.030%	1.434%	0.915%
	ES	1.184%	1.645%	1.057%

Clearly, we observe that the risk measures estimated based on the first window sample are the highest, then followed with the full sample, the last ranked is the ones from the last window, this finding holds for both VaR and ES, and for both methods, which indicates that the dynamic changes across the time and the simple estimation methods with fixed window may be less efficient to tackle with this fact, the approach indicated in validated report of rolling window fits better for this dynamics.

Reporting

The risk reporting in this document is generally well-structured, transparent and consistent accross multiple models. The authors clearly define the confidence levels (97.5% and 99%) for both Value-at-Risk (VaR) and Expected Shortfall (ES) and present results not only for the full sample but also broken down by crisis and non-crisis periods. This helps contextualize the numbers and gives the reader a more granular view of risk under different regimes. The VaR and ES values are visualized over time and violations are clearly marked which enhances the interpretability. Additionally, the authors go a step further by explaining deviations from expected results, such as the t-distribution sometimes yielding higher VaR than the normal case and they justify these findings with a derivation in section 3.2.

However, some elements can be improved for clarity and completeness. First, while the formulas for each model are provided, the parameter estimation methods (e.g., how degrees of freedom for the t-distribution were chosen beyond the Q-Q visual fit) can be discussed in more depth. Also, while they mention backtesting and providing p-values, they do not always explain how to interpret them from a practitioner's perspective. To give an example, what it means for model performance if the null hypothesis is rejected in the LR test. Adding more narrative around practical interpretation would make risk reporting more robust for readers who are not necessarily educated in this financial risk domain.

4 Limitation

Theoretical assessment

The report provides a diverse range of models and applies them correctly. That being said, there remain some limitations in the theoretical modeling choices and assumptions. Most notably, the Constant Conditional Correlation (CCC) assumption in the GARCH framework is overly restrictive. This is particularly the case during the crisis periods when correlations between assets shift significantly. A more flexible approach such as Dynamic Conditional Correlation (DCC) would better capture these evolving dependencies of different asset prices during such turmoil periods. Similarly the Filtered Historical Simulation (FHS) method is not documented in sufficient detail. This is specifically the case with regards to the volatility filtering step which limits the reader its ability to evaluate the methodology. Moreover, the report omits distributional tests for model residuals or simulation validity checks for the empirical quantile-based methods which are important for assessing the model fit. Also, the justification for the selected degrees of freedom in the Student-t distributions relies solely on visual inspection of Q-Q plots without any formal goodness of fit test. Lastly, the assumptions underlying the synthetic loan component, which includes the fixed credit spread and lack of credit risk modeling, reduce the realism of the portfolio under stress scenarios.

Quantitative assessment

- The report does not include sensitivity analyses for important model parameters, such as the decay factor λ in the EWMA filter, the rolling window length, or the initial values in the GARCH estimation.
- The Q-Q plot of Variance-Covariance methods with student's t distribution shows the better fit from the one with degree of freedom of 4 while the authors selected df=3 instead.
- Backtesting results are reported, but there is limited discussion on their economic or operational implications. For example, the consequences of rejecting the null hypothesis in the LR or t-tests are

not translated into practical takeaways. Besides, the authors provide the graphical and statistical test about if violations are distributed by an exponential distribution by Q-Q plot and Weibull LR test, while neglecting the VaR binomial test which quantify the discrepancy between observed and expected violation counts.

- The analysis would benefit from performance comparisons across different periods beyond crisis and non-crisis periods. For example, pre-COVID, COVID and post-COVID would be an interesting extension given that these periods likely yield different dynamics among the selected portfolio assets possibly leading to new valuable insights.
- Finally, no out-of-sample forecast evaluation is conducted. Including an out-of-sample validation would strengthen the credibility of the proposed risk estimates.

5 Final Validation Result

Based on the detailed review and validation results discussed in this report based on the report of Group 14, we assign the final verdict to be yellow.

The report demonstrates strong command for financial risk modeling and includes sufficient documented implementations of various VaR and ES models. In particular, the use of multiple models (such as Historical Simulation, GARCH(1,1)-CCC and Filtered Historical Simulation) offers a comprehensive risk assessment that align with both academic and semi-professional standards. Nevertheless, several areas require improvement before full endorsement can be given. From a theoretical standpoint, the GARCH-CCC model assumes constant conditional correlation, which is a restrictive simplification during crisis periods. The FHS methodology, while promising, lacks methodological transparency, especially regarding its EWMA filtering process and the decay parameter λ. Although the report includes thorough backtesting and sensitivity analysis, it omits some crucial robustness checks, such as the interpretation of test outcomes and tests for autocorrelation in residuals. In addition, areas like initial parameter specification and in-sample versus out-of-sample performance are not addressed.

To conclude, while the report is well structured an includes insightful analysis, the model assumptions and methodological clarity must be further substantiated. These issues do not disqualify the report from practical use, but they should be addressed to ensure robustness and reliability for the client for whom the report is written.

References

McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative risk management : Concepts, techniques and tools*. Princeton University Press.