Model Risk of Risk Models*

Jon Danielsson
Systemic Risk Centre
London School of Economics

Kevin James
Systemic Risk Centre
London School of Economics

Marcela Valenzuela
University of Chile, DII

Ilknur Zer
Federal Reserve Board

April 2014
FEDS Working Paper No. 2014–34

Abstract

This paper evaluates the model risk of models used for forecasting systemic and market risk. Model risk, which is the potential for different models to provide inconsistent outcomes, is shown to be increasing with and caused by market uncertainty. During calm periods, the underlying risk forecast models produce similar risk readings, hence, model risk is typically negligible. However, the disagreement between the various candidate models increases significantly during market distress, with a no obvious way to identify which method is the best. Finally, we discuss the main problems in risk forecasting for macro prudential purposes and propose an evaluation criteria for such models.

Keywords: Value–at–Risk, expected shortfall, systemic risk, model risk, CoVaR, MES, financial stability, risk management, Basel III

*Corresponding author Ilknur Zer, `ilknur.zerboudet@frb.gov`. The views in this paper are solely the responsibility of the author and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System. The early version of this paper is circulated under the title "Model Risk of Systemic Risk Models". We thank the Economic and Social Research Council (UK) [grant number: ES/K002309/1], the AXA Research Fund for its financial support provided via the LSE Financial Market Group's research programme on risk management and regulation of financial institutions. We also thank Kezhou (Spencer) Xiao for excellent research assistance. Finally we thank to Seth Pruitt, Kyle Moore, participants at various seminars and conferences where earlier versions of this paper were presented. All errors are ours. Updated versions of this paper can be found on `www.RiskResearch.org` and the Webappendix for the paper is at `www.ModelsandRisk.org/modelrisk`.

1 Introduction

Following the 2008 crisis, risk forecasting has emerged as a key public concern, with policy makers under considerable pressure to find new and better ways to accurately identify and forecast risk. This has led to rapid developments in macro prudential motivated statistical methods of systemic risk and market risk. This means in practice that statistical risk measures are set to play a much more fundamental role in policymaking and decision making within financial institutions, than before the crisis. Considering that the output of those risk measures has a real economic impact, an understanding of the model risk of risk forecast models, that is, the potential for different underlying risk forecast models to provide inconsistent outcomes, is of considerable interest to both policymakers and practitioners. The study of such model risk constitutes the main motivation in this paper. We first propose a classification system for systemic risk models, after which we measure the model risk of both systemic risk and regulatory risk forecast models when estimated by the most commonly used statistical techniques. Finally, we propose criteria for evaluating methods for macro prudentially motivated risk identification and forecasting.

Market risk regulations have been based on daily 99% Value–at–Risk (VaR) ever since the 1996 amendment to Basel I. After the crisis started in 2007, the extant market risk regulatory models (MRRMs) became to be seen as lacking in robustness, especially when it comes to tail risk and risk measure manipulation. In response, the Basel Committee proposed three major changes to the existing regulatory regime in 2013, to be incorporated into Basel III: The replacement of 99% VaR with 97.5% expected shortfall (ES), the use of overlapping estimation windows, and the calibration of a risk forecast to the historically worst outcome.

Parallel to these developments, and often undertaken by the same authorities, the literature on systemic risk identification and forecast methods has now emerged as a key priority for policymakers. A wide variety of systemic risk measures have been proposed, see Bisias et al. (2012) for a survey. Perhaps the most common way to construct a systemic risk model (SRM) is to adopt existing market risk regulation methodologies to the systemic risk problem, an approach we term *market data based methods*.[1] Those measures generally take a leaf from the Basel II market risk regulations and use price

[1]Besides the market data based methods, other approaches exist to construct SRMs, such as those based on credit risk techniques, market implied losses, and macroeconomic conditions. See for instance Segoviano and Goodhart (2009), Huang et al. (2009), Alessi and Detken (2009), Borio and Drehmann (2009), Tarashev et al. (2010), Drehmann and Tarashev (2013), Gray and Jobst (2011), Huang et al. (2012), Suh (2012), and Gray and Jobst (2013). However, given the preeminence of market data based methods amongst SRMs, that is where we focus our attention.

data to forecast VaR as a first step in the calculation of the SRM, perhaps along with ES as an intermediate step.

In other words, while intended for different purposes, both the market data based systemic risk methods and the market risk regulation techniques are closely related, sharing a methodological common root — VaR. Therefore, any model risk analysis of VaR will apply to both most SRMs and MRRMs.

In our first contribution, we propose a general setup for the classification of SRMs. Starting with the joint distributions of individual financial institutions and the entire financial system; one can get the conditional densities — an institution given the system or system given an institution. With this classification system, both the existing and proposed Basel market risk measures are then obtained from the marginal densities of individual financial institutions. This general setup provides the lens through which to analyze the various SRMs. The prominent MES (Acharya et al., 2010), CoVaR (Adrian and Brunnermeier, 2011), SRISK (Brownlees and Engle, 2012; Acharya et al., 2012), Co–Risk (IMF, 2009), and BIS's Shapley value method (Tarashev et al., 2010) all fall under this classification setup. Each and every one of these SRMs, as many others, is elementally founded on VaR as a fundamental building block, suggesting that the study of the model risk of VaR is a logical starting point for analyzing the model risk of market data based SRMs.

It has been known, from the very first days of financial risk forecasting, that different models can produce vastly different outcomes, where it can be difficult or impossible to identify the best model, as noted for example by Hendricks (1996), Berkowitz and O'Brien (2002), and Daníelsson (2002). This problem arises because financial risk cannot be directly measured and instead has to be forecasted by a statistical model. Since the ultimate use of these risk models is decision making, it is of key importance that the reliability of the underlying model is verifiable. In spite of this, very little formal model risk analysis has been done on VaR, with a few exceptions, such as Kuester et al. (2006) and Boucher et al. (2013). This paper contributes to this literature by studying the model risk of the most common market risk measures — VaR and ES — along with the most frequently used statistical models for implementing these risk measures in practice. We focus in particular on the risk measuring methodology proposed in Basel III. In addition, we provide the first empirical evidence in the literature that VaR–ES model risk passes through towards the market data based systemic risk measures.

The main avenue for assessing the veracity of market risk forecasts is backtesting, a somewhat informal way to evaluate model risk. While straightforward to implement, backtesting is not a good substitute for formal model risk analysis for several reasons. First, backtesting is based on strong distributional assumptions, second, it is focused on a particular criteria, like frequency of violations, while operationally any number of other criteria, such as clustering, magnitude or volatility of risk forecasts might be more relevant. Third, the paucity of observations of financial crises can make it difficult to obtain reliable tests, especially when we are concerned with tail events. Furthermore, it can be hard to design backtests in the special case of SRMs, which are generally based on conditional distributions. Finally, because the underlying risk forecast models are generally non–nested, backtesting does not enable formal model comparison. Taken together, these issues imply that other techniques for ascertaining model risk become necessary.

In order to assess the model risk, that is, the disagreement between the common risk forecast methodologies, we propose a new method we term *risk ratios*. This entails applying a range of common risk forecast methodologies to a particular asset on a given day, and then calculating the ratio of the maximum to the minimum risk forecasts. This provides a succinct way of capturing model risk because as long as the underlying models have passed some model evaluation criteria by the authorities and financial institutions, they can all be considered as a reputable candidate for forecasting risk. Supposing that a true number representing the latent level of risk exists, if this risk is forecasted by a number of equally good models, the risk ratio should be close to 1. If the risk ratio is very different from 1, it therefore captures the degree to which different models disagree, and hence, provides a measure of model risk.

While there is a large number of candidate methods for forecasting risk, the following six techniques in our experience are by far the most common in practical use: historical simulation, moving average, exponentially weighted moving average, normal GARCH, student–t GARCH, and extreme value theory. For that reason, we focus our attention on those six. It is straightforward to expand the universe of methods if another prominent candidate emerges.

Our risk ratio method is agnostic as to the specific risk measure chosen, however since the most commonly used risk measure is VaR, and VaR is usually the first elemental step in both the implementation of market data based SRMs and other market risk measures, such as ES, we opted to focus most of our attention on VaR and ES. We investigate if the VaR–ES results carry through to any VaR–ES based SRM. In the interest of brevity, we focus our risk ratio SRM analysis on MES and CoVaR.

The data set consists of large financial institutions traded on the NYSE, AMEX, and NASDAQ exchanges from the banking, insurance, real estate, and trading sectors over a sample period spanning January 1970 to December 2012. Considering the equities and 99% VaR, the mean model risk across all stocks and observations is 2.26, whilst it is 1.76 and 1.84 for S&P-500 and

Fama-French financial portfolios, respectively. If we consider the maximum model risk for each stock across time, the median across the whole sample is 7.62, and for 95% of companies it is below 23.14. In the most extreme case it is 71.43. Not surprisingly, the lowest risk forecasts tend to come from the conditionally normal methods, like MA, EWMA and GARCH, whilst the highest forecasts resulted from the fat tailed and (semi)– nonparametric approaches (t–GARCH, HS, and EVT). The average VaR across all assets is the lowest for EWMA at 4.65 and highest for t–GARCH at 7.70. The least volatile forecast method is MA with the standard deviation of VaR at 2.4, whilst it is highest for t–GARCH at 9.93.

By further segmenting the sample into calm and turmoil periods, we find that model risk is much higher when market risk is high, especially during financial crises. We investigate this in more detail by studying the relationship between the risk ratios and the Chicago Board Options Exchange Market Volatility Index (VIX), finding that model risk is positively and significantly correlated with market volatility, where the VIX Granger causes model risk, whilst the opposite causality is insignificant. In other words, market volatility provides statistically significant information about future values of the risk readings' disagreement. Finally, we develop a procedure to obtain the distribution of the risk ratios. The results reveal that the model risk during crisis periods is significantly higher than in the immediate preceding period.

When we apply the risk ratio analysis to the overlapping 97.5% ES approach proposed by the Basel Committee, instead of the current regulations with non–overlapping 99% VaR, we find that model risk increases by a factor of three, on average, with the bulk of the deteriorating performance of the risk models due to the use of overlapping estimation windows.

In the case of the SRMs considered, we find quite similar results as for VaR; the systemic risk forecasts of MES and CoVaR highly depend on the chosen model, especially during crisis periods. This supports our contention that any VaR based SRM is subject to the same fundamental model risk as VaR. A further analysis of CoVaR reveals that both theoretically and empirically, the time series correlation of ΔCoVaR and VaR is almost 1, with quite a high estimation uncertainty, implying that the use of CoVaR might not have much of an advantage over just using VaR.

We suspect the problem of poor risk model performance arises for two reasons. The first is the low frequency of actual financial crises. Developing a model to capture risk during crises is quite challenging, since the actual events of interest has never, or almost never, happened during the observation period. Such modeling requires strong assumptions about the stochastic processes governing market prices, assumptions that are likely to fail when the economy transits from a calm period to a crisis. Second, each and every statistical model in common use is founded on risk being exogenous, in other words, the assumption that extreme events arrive to the markets from the outside, like an asteroid would, where the behavior of market participants has nothing to do with the crisis event. However, as argued by Daníelsson and Shin (2003), risk is really endogenous, created by the interaction between market participants, and their desire to bypass risk control systems. As both risk takers and regulators learn over time, we can also expect the price dynamics to change, further frustrating statistical modeling.

Overall, the empirical results are a cause for concern, considering that the output of the risk forecast models is used as an input into expensive decisions, be they portfolio allocations or the amount of capital. From a market risk and systemic risk point of view, the risk forecast models are most important in identifying risk levels during periods of market stress and crises. The absence of model risk during calm times might provide a false of confidence in the risk forecasts. From a macro prudential point of view, this is worrying, since the models are most needed during crisis, but that this when they perform the worst. This applies directly to the SRMs but also increasingly to MRRMs.

The observation that the risk forecast models perform well most of the time, but tend to fail during periods of turmoil and crisis, is not necessarily all that important for the models original intended use: market risk management. because in that case the financial institution is concerned with managing day–to–day risk, rather than tail risk or systemic risk. However, given the ultimate objective of an SRM and MRRM, the cost of a type I or type II error is significant. For that reason, the minimum acceptable criteria for a risk model should not be to weakly beat noise, instead the bar should be much higher, as discussed in Daníelsson, James, Valenzuela and Zer (2014a). To this end, we finally propose a set of evaluation criteria for macro prudentially motivated risk forecast models. First, point forecasts are not sufficient: confidence intervals, backtesting, and robustness analysis should be required. Second, models should not only rely on observed market prices, instead, they ought to aim at capturing the pre–crisis built–up of risk as well. Finally, the probabilities assumed in the modeling should correspond to actual crisis probabilities.

Ultimately, we conclude that one should be careful in applying successful market risk methodologies, originally designed for the day–to–day management the market risk and financial institutions, to the more demanding job of systemic risk identification and tail risk.

The outline of the rest of the paper is as follows: in the next section, we provide a classification system for market risk and systemic risk methodologies, especially those with an empirical bent. Section 3 introduces the our main tool of analysis, risk ratios, as well as the data and risk forecast methodologies used in the paper. In Section 4 we present the empirical findings.

This is followed by Section 5 analyzing the results and proposing criteria for systemic risk measures. Finally, Section 6 concludes.

2 Classification of systemic risk measures

The various market data based methods systemic risk measures (SRMs) that have been proposed, generally fall into one of three categories: the risk of an institution given the system, the risk of the system given the institution or the risk of the system or institution by itself. In order to facilitate the comparison of the various SRMs, it is of benefit to develop a formal classification scheme.

The joint distribution of the financial system and the individual financial institutions sits at the top of the classification system. By the application of Bayes' theorem we obtain the risk of the system given an individual bank or alternatively the system given the bank.

Let R_i be the risky outcome of a financial institution i on which the risk measures are calculated. This could be for example, daily return risk of such an institution. Similarly, we denote the risky outcome of the entire financial system by R_S. We can then define the joint density of an institution and the system by

$$f\left(R_i, R_S\right).$$

The marginal density of the institution is then $f(R_i)$, and the two conditional densities are $f\left(R_i|R_S\right)$ and $f\left(R_S|R_i\right)$. If we then consider the marginal density of the system as a normalizing constant, we get the risk of the institution conditional on the system by Bayes' theorem:

$$f\left(R_i|R_S\right) \propto f\left(R_S|R_i\right) f(R_i). \tag{1}$$

The risk of the system conditional on the institution is similarly defined;

$$f\left(R_S|R_i\right) \propto f\left(R_i|R_S\right) f(R_S). \tag{2}$$

Suppose we use VaR as a risk measure. Defining Q as an event such that:

$$\mathrm{pr}[R \leq Q] = p,$$

where Q is some extreme negative quantile and p the probability. Then VaR equals to $-Q$. Expected shortfall (ES) is similarly defined;

$$\mathrm{ES} = \mathrm{E}[R|R \leq Q].$$

CoVaR_i is then obtained from (1) with VaR being the risk measure;[2]

$$\mathrm{CoVaR}_i = \mathrm{pr}[R_S \leq Q_S|R_i \leq Q_i] = p$$

[2] Adrian and Brunnermeier (2011) identify an institution being under distress if its return is *exactly* at its VaR level rather than *at most* at its VaR.

and if instead we use (2) and ES as a risk measure, we get MES;

$$\mathrm{MES}_i = \mathrm{E}[R_i | R_S \leq Q_S]. \tag{3}$$

We could just as easily have defined MVaR as

$$\mathrm{MVaR}_i = \mathrm{pr}[R_i \leq Q_i | R_S \leq Q_S] = p$$

and CoES as

$$\mathrm{CoES}_i = \mathrm{E}[R_S | R_i \leq Q_i].$$

To summarize, see Table 1;

Table 1: Classifying systemic risk measures

Marginal risk measure	Condition on system	Condition on institution
VaR	MVaR $\mathrm{pr}[R_i \leq Q_i \vert R_S \leq Q_S] = p$	CoVaR $\mathrm{pr}[R_S \leq Q_S \vert R_i \leq Q_i] = p$
ES	MES $\mathrm{E}[R_i \vert R_S \leq Q_S]$	CoES $\mathrm{E}[R_S \vert R_i \leq Q_i]$

The Shapley value (SV) methodology falls under this classification scheme, by adding a characteristic function, which maps any subgroup of institutions into a measure of risk. The SV of an institution i is a function of a characteristic function θ and the system S. If we choose θ as VaR, then

$$SV_i = g(S, \theta) = g(S, \mathrm{VaR}).$$

If the characteristic function is chosen as the expected loss of a subsystem given that the entire system is in a tail event, we end up the same definition as MES. Similarly, the Co–Risk measure of (IMF, 2009) and systemic expected shortfall (SRISK) of Brownlees and Engle (2012); Acharya et al. (2012) also fall under this general classification system. SRISK is a function of MES, leverage, and firm size, where MES is calculated as in (3) with a DCC and TARCH model to estimate volatility. On the other hand, Co-Risk is similar in structure to CoVaR, except that it focuses the co-dependence between two financial institutions, rather than the co-dependence of an institution and the overall financial system. In other words, it depends on the conditional density of institution i given institution j and can be estimated via quantile regressions with market prices, specifically the CDS mid-prices, being the input.

Ultimately, regardless of the risk measure or conditioning, the empirical performance of the market based systemic risk measures fundamentally depends on VaR. This applies equally whether the risk measure is directly based on VaR like CoVaR or indirectly like MES. Empirical analysis of VaR will therefore provide useful guidance on how we can expect the systemic risk measures to perform.

3 Model risk analysis

Broadly speaking, model risk relates to the uncertainty created by not knowing the data generating process. That high level definition does not provide guidance on how to assess model risk, and any test for model risk will be context dependent. Within the finance literature, some authors have defined model risk as the uncertainty about the risk factor distribution (e.g., Gibson, 2000), misspecified underlying model (e.g., Green and Figlewski, 1999; Cont, 2006), the discrepancy relative to a benchmark model (e.g., Hull and Suo, 2002; Alexander and Sarabia, 2012), and inaccuracy in risk forecasting that arises from estimation error and the use of an incorrect model (e.g., Hendricks, 1996; Boucher et al., 2013). In this paper, we are primarily interested in a particular aspect of model risk, how the use of different models can lead to widely different risk forecasts. Assessing that aspect of model risk is the main motivation of proposing our risk ratio approach.

That leaves the question of why we implement risk ratio analysis instead of just doing backtesting. After all, backtesting is a common and very useful methodology to see how a particular risk model performs, based on the subjective criteria set by the model designer. For our purpose, backtesting is not as useful for four important reasons. First, in backtesting, any systematic occurrence of violations quickly shows up in the back test results. However, there are a number of different criteria for judging risk models, be they violation ratios, clustering, magnitudes or volatility of risk forecasts, each of which can be assessed by a number of different, and often conflicting, statistical procedures. A particular model may pass one set of criteria with flying colors and fail on different criteria.

Second, we are particularly interested in model risk during periods of financial turmoil and the applicability of backtesting to model risk is not as clear–cut during such periods. There are several reasons for this; the underlying assumption behind most backtesting methodologies is that violations are i.i.d. Bernoulli distributed, however, the embedded stationary assumption is violated when the economy transits from a calm period to a turmoil period. This might for example show up in the clustering of violations during market turmoil, something very difficult to test without making stringent assumptions. Moreover financial crisis or systemic events, for which SRMs are designed to analyze, are by definition very infrequent. The paucity of data on during such time periods makes it difficult, if not impossible, to formally test for violations and to obtain robust backtest results.

Third, because the underlying risk forecast models are generally non–nested, backtesting does not enable formal model comparison, except for forecast violations. Finally, in the special case of the SRMs which are based on conditional distributions, as discussed in Table 1, backtesting in practice is difficult since they would need much larger sample sizes than available. Taken together, this suggests that a more general model risk approach, such as the risk ratio method proposed here, is necessary for ascertaining model risk.

3.1 Evaluating model risk: Risk ratios

With a range of plausible risk forecast models, one obtains a range of risk readings. Given our objective, we propose a new method, the ratio of the highest to the lowest risk forecasts, *risk ratios,* across the range of these candidate models. This provides a clear unit free way to compare the degree of divergence, as long as the underlying models are in daily use by the regulated financial institutions and have passed muster by the authorities. The baseline risk ratio estimate is 1, and if a true number representing the latent level of risk exists, and we forecast the risk by a number of equally good models, the risk ratio should be close to 1, a small deviance can be explained by estimation risk. If the risk ratio is very different from 1, it therefore captures the degree to which different models disagree.

We further propose a procedure to evaluate the statistical significances of risk ratios during different market conditions. We assume that an investor holds the 100 biggest financial institutions in her portfolio. The stocks in the portfolio are allowed to change at the beginning of each year and portfolio weights are random. We calculate the highest to the lowest VaR estimates for the random portfolios employing the six VaR approaches. The following algorithm illustrates the main steps:

1. Select the biggest 100 institutions in terms of market capitalization at the beginning of each year and obtain the daily holding period return for each stock.

2. For a given year, select a random portfolio of positions for the stocks selected in step (1) by drawing the portfolio weights from a unit-simplex. Hence, get the daily return of the random portfolio for the sample period.

3. Calculate the daily 99% VaR by employing each of the six candidate risk models for the random portfolio chosen in step (2) with an estimation window size of 1,000.

4. For a given day calculate the ratio of the highest to the lowest VaR readings (VaR risk ratios) across all methods.

5. Repeat the steps two through four 1,000 times. This gives a matrix of risk ratios with a dimension of number of days $\times$ number of trials.

6. Identify the crisis and pre-crisis periods. For a given episode, we consider the previous 12 months as a pre-crisis period. For instance, for the 2008 global financial crisis, which has peak on December 2007 and trough on June 2009, the pre-crises period covers from December 2006 to November 2007.

7. For each trial, obtain the time-series averages of risk ratios over the crisis and pre-crisis periods and calculate the confidence intervals.

3.2 Data and models

We focus our attention on the six most common risk forecast models used by industry: historical simulation (HS), moving average (MA), exponentially weighted moving average (EWMA), normal GARCH (G), student$-t$ GARCH (tG), and extreme value theory (EVT) and compare the risk forecasts produced by those models. The models are explained in detail in Appendix A. We estimate daily 99% VaR values for each method, where the portfolio value is set to be $100 and the estimation window is 1,000 days. Then, we calculate the ratio of the highest to the lowest VaR readings (risk ratios) across all methods. If there is no model risk, one would expect the VaR readings to be similar across the models employed, i.e., the ratio to be close to 1.[3]

Since our focus is on systemic risk, it is natural to consider a sample of financial institutions. In order to keep the estimation manageable and avoid

[3]It was not possible to obtain VaR forecasts for every estimation method and institution each day. In some cases, the nonlinear optimization methods would not converge, usually for tGARCH. In other cases, the optimizer did converge but the estimated degrees of freedom parameter of the tGARCH model was unusually low, just over two, making the tails of the condition of distribution quite fat, pushing up the VaR numbers. Generally, risk forecast methods that aim to capture fat tails, are estimated with more uncertainty than those who don't, and the particular combination of data and estimation method is what caused these apparent anomalous results. While one might be tempted to use different optimizer, our investigation showed that the optimization failed because the model was badly misspecified given some of the extreme outcomes. In particular, the models were unable to simultaneously find parameter combinations that work for market outcomes when a company is not traded for consecutive days. While investors are not subject to risk on those days, many consecutive zeros adversely affect some of the risk forecast methods, biasing the results. For this reason, we do not use any part of a stock's sample that contains more than one week worth of zero returns; that is we truncated the sample instead of just removing the zeros. Increasing or decreasing that number did not materially alter the results.

problems of holidays and time zones, we focus on the largest financial market in the world, the US. We start with all NYSE, AMEX, and NASDAQ–traded financial institutions from the banking, insurance, real estate, and trading sectors with SIC codes from 6000 to 6799. We collect daily prices, holding period returns, and number of shares outstanding from CRSP 1925 US Stock Database for the period January 1970 to December 2012. We then keep a company in the sample if (1) it has more than 1,010 return observations, (2) it has less than 30 days of consecutively missing return data, and (3) it is one of the largest 100 institutions in terms of market capitalization at the beginning of each year. This yields a sample of 439 institutions.

Below we present the results from a small number of stocks for illustrative purposes, with the full results relegated to the Webappendix, `www.ModelsandRisk.org/modelrisk`. We consider the biggest depository–JP Morgan (JPM), non-depository–American Express (AXP), insurance – American International Group (AIG), and broker-dealer–Goldman Sachs (GS) in the sample.[4] Besides the individual stock risk ratios, in order to study the model risk of the overall system, we employ the daily returns of the S&P-500 index and the Fama-French value-weighted financial industry portfolio (FF). In addition, we create a financial equity portfolio, Fin100, by assuming that an investor holds the 100 biggest financial institutions in her portfolio. The portfolio is rebalanced annually and the weights are calculated based on the market capitalization of each stock at the beginning of the year.

4 Empirical findings

In our empirical application, we apply the six most common risk forecast methods discussed above to our extensive financial data set, evaluating model risk by risk ratio. More specifically, we both address the model risk of MRRMs and SRMs, focusing on the latest regulatory developments and the most popular systemic risk models.

The Basel Committee, motivated by the poor performance of risk forecast models prior to the 2008 crisis, has proposed significant changes to the market risk regulatory regime, aiming to both better capture tail risk and also reduce the potential for model manipulation. To this end, the Committee made three key proposals in 2013: First, changing the core measure of probability from 99% VaR to 97.5% ES. Second, estimating the model with n–day overlapping time intervals, where n depends on the type of asset. In practice, this means that one would use the returns from day 1 to n as the

[4]Metlife and Prudential are the first and the second biggest insurance companies in our sample in terms of asset size, respectively. However, we present the results for the American International Group (AIG), which is the third biggest insurance company in the sample because both Metlife and Prudential have available observations only after 2000.

first observation, day 2 to $n+1$ for the second, and so forth. Finally, the ES risk forecast is to be calibrated to the historically worst outcome. Below, we analyze all three aspects of the proposed regulations from the point of view of model risk, by means of the risk ratio approach.

In addition to the market risk models, we also consider two of the most popular systemic risk models, MES and CoVaR. Both measures are quite related, as shown in Table 1, and are elementally based on VaR. One could apply the risk ratio approach to other market data based SRMs, but given their common ancestry, we expect the results to be fundamentally the same, and in the interest of brevity we focus on the two SRMs.

4.1 VaR and ES model risk

We start the model risk analysis by examining the model risk of VaR. In Section 4.1.1 we focus the VaR risk forecasts of JP Morgan to visualize the model risk. In Section 4.1.2 we study the model risk of market risk models specifically focusing on the current and proposed Basel III regulations. Finally, in Section 4.1.3 we assess the model risk based on market conditions in detail.

4.1.1 Case study: JP Morgan

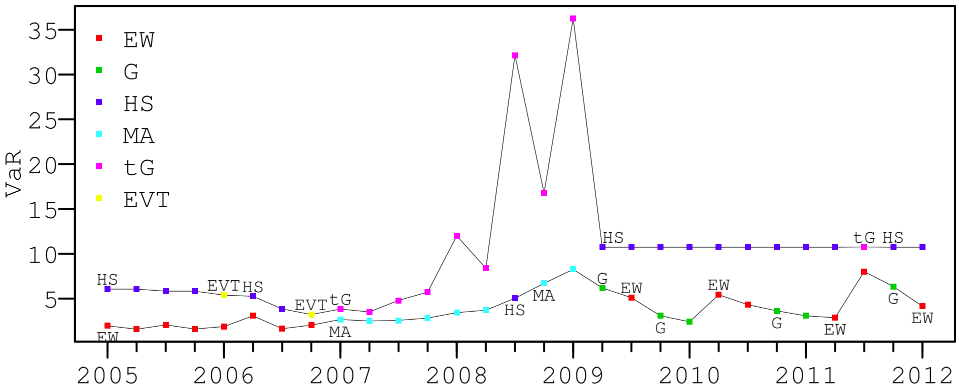
To visualize the model risk embedded in risk forecast methods, we present detailed results for the biggest stock in our sample in terms of asset size; JP Morgan. Results for the other stocks give the similar material results as can be seen from the web appendix. Consistent with the existing market risk regulations in Basel I and Basel II, dating back to the 1996 amendment, we start our analysis with the risk ratios calculated based on daily VaR at a 99% level.[5]

The results are illustrated in Figure 1, which shows end of quarter the highest and the lowest VaR forecasts, along with the method generating the highest and the lowest readings. As expected, the fatter methods; historical simulation, student-t GARCH, and extreme value theory produce the highest risk forecasts, whereas the thinner tailed methods; EWMA, moving average, and GARCH produce the lowest risk forecasts. The figure clearly shows the degree of model disagreement, and hence, model risk, across the sample period. Prior to the 2008 crisis, the models mostly agree, they sharply move apart during the crisis and have only partially come together since then.

[5]The rules stipulate a 10 day holding period, but then allow for the square of the time calculation, almost always used in practice, so the 10 day VaR is just 1 day VaR times a constant.

Figure 1: Model risk for JP Morgan

The highest and the lowest 99% daily VaR forecasts for JP Morgan based on six different methods; historical simulation (HS), moving average (MA), exponentially weighted moving average (EW), normal GARCH (G), student-t GARCH (tG), and extreme value theory (EVT). Estimation window is 1,000. To minimize clutter end of quarter results are plotted. Every time the VaR method changes, the label changes. Portfolio value is $100. Data is obtained from CRSP 1925 US Stock Database.



4.1.2 Model risk under the current and the proposed Basel regulations

Table 2 presents the maximum daily risk ratios across the NBER recession dates,[6] the stock market crashes of 1977 and 1987, and the 1998 LTCM/Russian crisis. It makes use of the three equity indices and four stocks introduced in Section 3.2. Panel 2(a) shows the results where the risk is calculated via daily 99% VaR using non-overlapping estimation windows, in line with the current market risk regulations. In Panel 2(b) we calculate the risk ratios via 97.5% daily ES with 10–day overlapping estimation windows, hence, we consider the recent Basel III proposal.

The VaR results in Panel 2(a) show that the average risk ratio, across the entire time period, ranges from 1.76 to 1.88 for the portfolios, and from 1.88 to 2.19 for the individual stocks, suggesting that model risk is generally quite moderate throughout the sample period. A clearer picture emerges by examining the maximum risk ratios across the various subsamples. Model risk remains quite temperate during economic recessions, but increases substantially during periods of financial turmoils, exceeding 9 during the 1987 crash or 5 during the 2008 global crisis for the market portfolio.

On the other hand, Panel 2(b) focuses on the proposed changes to the Basel Accords; with 97.5% ES 10–day overlapping estimation windows. We see that the model risk increases sharply, with the risk ratios during turmoil periods, on average, almost double for S&500, triple for Fama–French finan-

[6] www.nber.org/cycles.html

Table 2: Daily risk ratios: non-overlapping 99% VaR and overlapping 97.5% ES

This table reports the maximum of the ratio of the highest to the lowest daily VaR and ES forecasts (risk ratios) for the period from January 1974 to December 2012 for the S&P-500, Fama-French financial sector portfolio (FF), the value-weighted portfolio of the biggest 100 stocks in our sample (Fin100), JP Morgan (JPM), American Express (AXP), American International Group (AIG), and Goldman Sachs (GS). Panel 2(a) presents the risk ratio estimates where the risk calculated via daily 99% VaR. In Panel 2(b) we calculate the risk ratios via 97.5% daily ES with 10–day overlapping estimation windows. Six different methods; historical simulation, moving average, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory are employed to calculate the VaR and ES estimates. Estimation window size is 1,000. Finally, the last row of each panel reports the average risk ratio for the whole sample period.

(a) Basel II requirements: VaR, $p = 99\%$, non-overlapping

Event	Peak	Trough	SP-500	FF	Fin100	JPM	AXP	AIG	GS
1977 crash	1977-05	1977-10	2.64	3.16	3.20	3.39	4.30	13.02	
1980 recession	1980-01	1980-07	2.05	2.65	2.36	2.49	2.00	3.06	
1981 recession	1981-07	1982-11	2.23	2.41	2.46	2.97	2.88	3.57	
1987 crash	1987-10	1988-01	9.52	10.00	9.76	10.38	6.07	3.72	
1990 recession	1990-07	1991-03	2.06	2.77	2.50	3.82	2.30	1.97	
LTCM crisis	1998-08	1998-11	4.73	4.01	3.53	3.33	5.13	3.27	
2001 recession	2001-03	2001-11	2.02	2.48	2.45	2.31	2.28	2.80	
2008 recession	2007-12	2009-06	6.74	5.69	7.07	6.90	6.76	13.89	6.15
Full sample (ave.)	1974-01	2012-12	1.76	1.84	1.88	1.88	1.87	2.15	2.19

(b) Basel III proposals: ES, $p = 97.5\%$, 10–day overlapping

Event	Peak	Trough	SP-500	FF	Fin100	JPM	AXP	AIG	GS
1977 crash	1977-05	1977-10	4.23	11.00	4.67	5.93	6.69	3.57	
1980 recession	1980-01	1980-07	5.42	23.74	12.32	15.61	5.26	7.62	
1981 recession	1981-07	1982-11	7.40	12.28	17.68	21.64	4.82	8.44	
1987 crash	1987-10	1988-01	11.84	13.68	62.81	7.37	7.27	5.19	
1990 recession	1990-07	1991-03	9.80	18.72	27.02	7.05	4.44	18.69	
LTCM crisis	1998-08	1998-11	4.94	5.13	5.38	8.13	5.76	7.06	
2001 recession	2001-03	2001-11	5.08	4.18	3.99	3.20	4.47	3.18	
2008 recession	2007-12	2009-06	9.46	10.78	8.48	25.15	12.26	41.85	7.44
Full sample (ave.)	1974-01	2012-12	2.63	2.77	2.73	2.41	2.42	2.70	2.63

cial sector portfolio (FF), quadruple for the value-weighted portfolio of the biggest 100 stocks in our sample (Fin100).

To understand whether the shift to ES instead of VaR, or to using overlapping windows increases model risk, in Table B.1 in the appendix we report the risk ratios calculated based on 99% VaR 10–day overlapping and 97.5% ES non–overlapping estimation windows. Further investigation shows that the main causal factor behind the deterioration in model performance is due to the overlapping estimation windows, whilst the contribution of the move to 97.5% ES to the increase in model risk is positive but quite moderate.

We suspect the reason for the impact of the overlapping estimation windows on model risk is because of how observations are repeated. Not only it will introduce dependence in the underlying time series, which then may bias the estimation, but also that anomalous events will be repeated in sample for n times, giving them artificial prominence, which in turn also biases the estimation. Since different estimation methods react differently to these artifacts introduced by the overlapping estimation windows, it is not surprising that model risk increases so sharply.

The third innovation by the Basel Committee to the market risk accords is the calibration of the forecast to the historically worst outcome. These results show that because historically worst outcomes are subject to the highest degree of model risk, the proposed methodology will carry model risk forward, arbitrarily introducing it in time periods when it otherwise would be low.

4.1.3 Model risk and market conditions

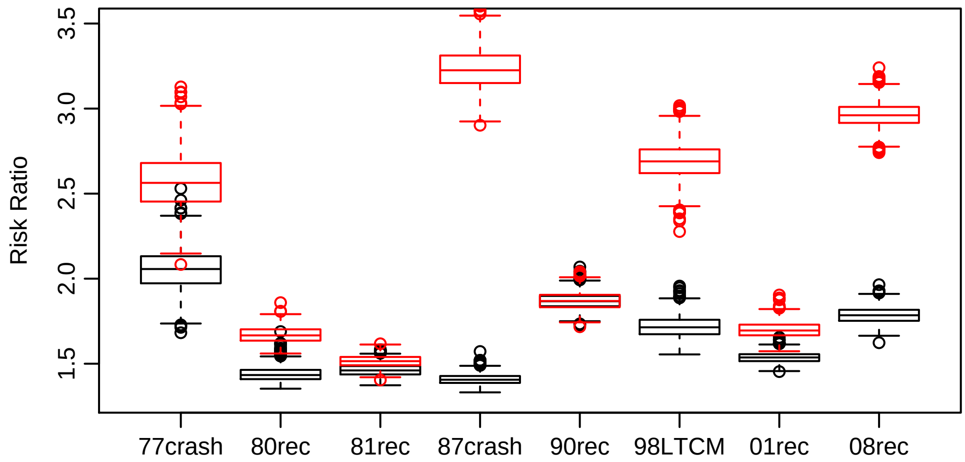
Table 2 reveals that while modeling risk is typically quite moderate, it sharply increases when overall market risk increases. To investigate this further, we compare the model risk with the Chicago Board Options Exchange Market Volatility Index (VIX). As expected, the VIX is significantly highly correlated with the 99% VaR risk ratios of S&P500 at 19.2%. In addition, we formally test for causality between the VIX and the model risk of S&P500 by a Granger causality test. We find that model risk does not significantly cause VIX, but the converse is not true. The VIX does cause model risk significantly at the 95% level.

Given that three of the six risk measures we use are based on conditional volatilities, estimated by past data, a part of the explanation is mechanical; whenever the volatility increases, a conditional historical volatility method, such as GARCH, will produce higher risk readings. More fundamentally, however, the results indicate that not the VaR readings, but the disagreement between those readings increases. All of the risk forecast models employed can be considered industry–standard, even if different users might hold strong views on their relative merits. Given that the models have entered the canon based on their performance during non–crisis times, it is not surprising that they broadly agree at such periods, otherwise any model that sharply disagreed, might have been dismissed. However the models all treat history and shocks quite differently and therefore can be expected to differ when faced with a change in statistical regimes. Given that none of the methods produce systematically highest or the lowest VaR estimates throughout the sample period, we surmise that this is what we are picking up in our analysis.

Finally, we assess whether the model risk is significantly higher during crisis compared to immediate pre-crisis periods by employing the procedure outlined in Section 3.1. Figure 2 plots the first and the third quartiles of risk ratios for each of the episodes separately. The intervals for the crisis periods are plotted in red, whereas the pre-crisis periods are in black. For all crisis periods, except the 1990 recession, we find that the risk ratios are higher during the crises compared to calm periods. Moreover, the difference is statistically significant for the 1987 crash, 1998 LTCM crisis, and the 2008 global financial crisis. In other words, systemic risk forecast models perform the worst when needed the most.

Figure 2: Model risk–confidence intervals

The plot displays the first and the third quartiles of risk ratios for the crises and non-crisis periods separately between January 1974 and December 2012. The intervals for the crisis periods are plotted in red, whereas the pre-crisis periods are identified as black. The risk ratio is the ratio of the highest to the lowest VaR estimates of the simulated portfolio outlined in Section 3.1. Estimation window size is 1,000 and VaR estimates are calculated at a 99% probability level based on six different methods; historical simulation, moving average, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory. Data is obtained from CRSP 1925 US Stock Database.



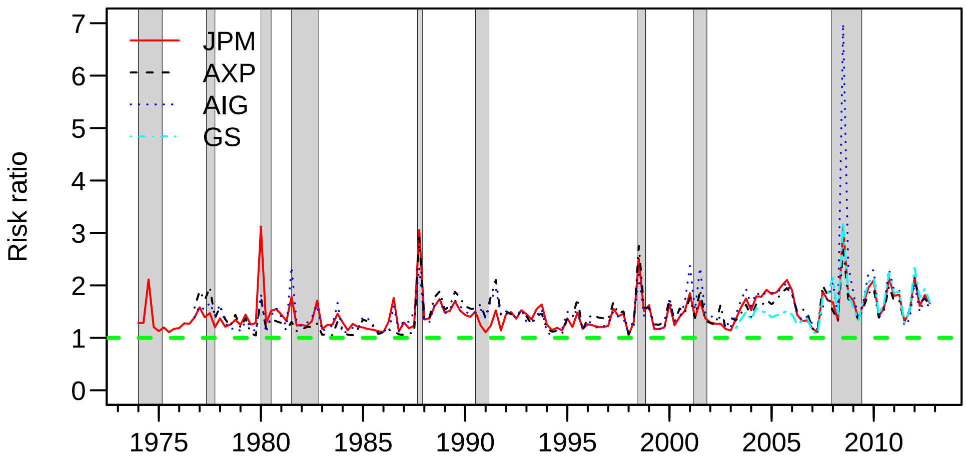
4.2 MES

As noted in Section 2, the first step in most common market data based systemic risk measures (SRMs) is the calculation of VaR, hence, we expect the risk ratio analysis hold for them as well. In this section we illustrate this by investigating the model risk in a popular SRM, MES, defined as an institution's expected equity loss given that the system is in a tail event. Hence, it is an expected shortfall (ES) estimate modified to use a threshold from the overall system rather than the returns of the institution itself and the first step requires the calculation of VaR of the market portfolio. Following Acharya et al. (2010) we use a 95% probability level with S&P500 as the market portfolio. This procedure results in six MES forecasts for each day, one for each of the six risk forecast methods. We then finally calculate the risk ratios across the risk readings.

Figure 3 illustrates end of the quarter risk ratios for the same four companies as above. The NBER recession dates, the stock market crashes of 1977 and 1987, and the 1998 LTCM/Russian crisis are marked with gray shades to visualize the trends in model risk during the turmoil times. The results are in line to those for VaR, as presented in Table 2. Model risk remains low most of the time, but spikes up during periods of market turmoil.

Figure 3: MES model risk

Ratio of the highest to the lowest daily 95% MES estimates for JP Morgan (JPM), American Express (AXP), American International Group (AIG), and Goldman Sachs (GS). S&P500 index is used as market portfolio. Six different methods; historical simulation, moving average, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory are employed to calculate the system–VaR estimates. Estimation window size is 1,000. To minimize clutter, end of quarter results are plotted. Data is obtained from CRSP 1925 US Stock Database. The NBER recession dates, the stock market crashes of 1977 and 1987, and the LTCM/Russian crisis are marked with gray shades.



Note that, in general, MES risk ratios presented in Figure 3 are closer to 1 than the VaR ratios presented in Table 2. It is because one gets much more accurate risk forecasts in the center of the distribution compared to the tails, and therefore 95% risk forecasts are more accurate than 99% risk forecasts. The downside is that a 95% daily probability is an event that happens more than once a month. This highlights a common conclusion, it is easier to forecast risk for non–extreme events than extreme events and the less extreme the probability is, the better the forecast. That does not mean that one should therefore make use of a non–extreme probability, because the probability needs to be tailored to the ultimate objective for the risk forecast.

4.3 CoVaR and ΔCoVaR

The other market based systemic risk measure we study in detail is CoVaR (Adrian and Brunnermeier, 2011). CoVaR of an institution is defined as the VaR of the financial system given that the institution is under financial distress whilst, ΔCoVaR captures the marginal contribution of a particular institution to the systemic risk.

While Adrian and Brunnermeier (2011) estimate the CoVaR model by means of quantile regression methods (see Appendix C for details), one can estimate the model with five of the six industry–standard methods considered here. The one exception is historical simulation, which is quite easy to implement, but requires at least $1/0.01^2 = 10,000$ observations at the 99% level. For this reason, the risk ratio results for CoVaR will inevitably be biased towards one.

If one defines an institution being under distress as its return being *at most* at its VaR, rather than being *exactly* at its VaR, then CoVaR is defined as:[7]

$$\mathrm{pr}[R_S \leq \mathrm{CoVaR}_{S|i} \, | R_i \leq \mathrm{VaR}_i] = p.$$

It is then straightforward to show that:

$$\int_{-\infty}^{\mathrm{CoVaR}_{S|i}} \int_{-\infty}^{\mathrm{VaR}_i} f(x,y) dx dy = p^2. \tag{4}$$

Hence, one can estimate CoVaR under any distributional assumptions by solving (4). Girardi and Ergun (2013) estimate CoVaR under normal GARCH and Hansen's (1994) skewed$-t$ distribution. We further extend this analysis to bivariate moving average (MA), exponentially weighted moving average (EWMA), student$-t$ GARCH (tG), and extreme value theory (EVT) and compare the risk forecasts produced by these models. We model the correlation structure with Engle's (2002) DCC model and obtain CoVaR by numerically solving for CoVaR by applying (4) to the conditional density. The EVT application was based on using EVT for the tails and an extreme value copula for the dependence structure.

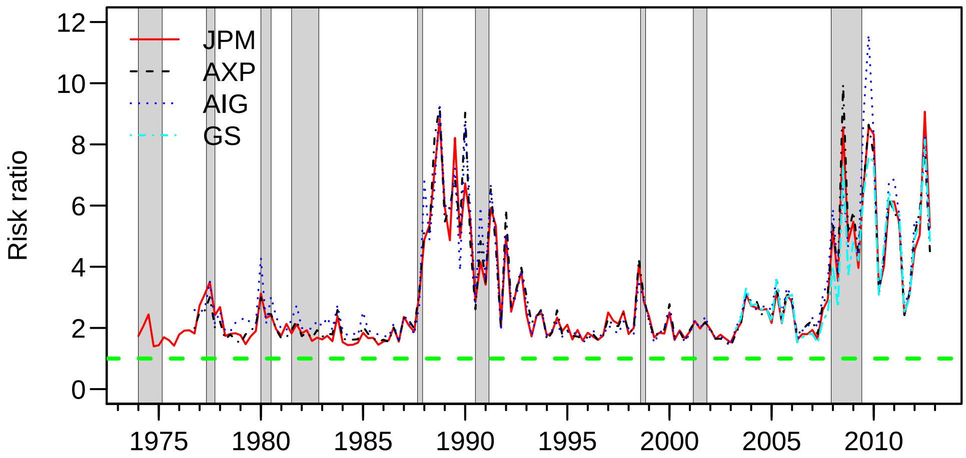
Figure 4 illustrates the end of quarter risk ratios for the same four companies. The NBER recession dates, the stock market crashes of 1977 and 1987, and the 1998 LTCM/Russian crisis are marked with gray shades to visualize the

[7] Mainik and Schaanning (2012) and Girardi and Ergun (2013) estimate the dynamics of CoVaR under the conditioning event $R_i \leq \mathrm{VaR}_i$. Their results show that the resulting CoVaR does not significantly different from the original CoVaR analysis proposed by Adrian and Brunnermeier (2011) conditioned on $R_i = \mathrm{VaR}_i$. This suggests that without loss of generality one can condition the CoVaR measure on $R_i \leq \mathrm{VaR}_i$ rather than on $R_i = \mathrm{VaR}_i$, and yet it allows us to estimate the CoVaR under different distributional assumptions.

trends in model risk during the turmoil times. We find that the model risk of CoVaR is higher on average compared to the model risk of VaR and MES, especially after the 2008 period. In line with the other results reported, it increases sharply with market turmoil.

Figure 4: CoVaR model risk

Ratio of the highest to the lowest daily 99% CoVaR estimates for JP Morgan (JPM), American Express (AXP), American International Group (AIG), and Goldman Sachs (GS). The Fama-French value-weighted financial industry portfolio index is used as market portfolio. Five different methods; moving average, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory are employed to calculate the individual stock VaR estimates and CoVaR is estimated by numerically integrating (4). Estimation window size is 1,000. To minimize clutter, end of quarter results are plotted. Data is obtained from CRSP 1925 US Stock Database. The NBER recession dates, the stock market crashes of 1977 and 1987, and the LTCM/Russian crisis are marked with gray shades.



We also investigate the statistical properties of the CoVaR measure, as well as the ΔCoVaR measure, estimated by the quantile regression methods of Adrian and Brunnermeier (2011). The results are reported in Appendix C. First we find that the unconditional correlation between VaR and ΔCoVaR mostly exceeds 99%, suggesting that the scaled signal provided by ΔCoVaR is very similar to the signal provided by VaR. Second, we show that when the estimation noise in the quantile regression is carried through to the ΔCoVaR estimates, it is hard to significantly discriminate between different financial institutions based on ΔCoVaR.

5 Analysis

Our findings indicate significant levels of model risk in the most common risk forecast methods, affecting both applications of market risk regulatory models (MRRMs) and systemic risk measures (SRMs). Unfortunately, the results are somewhat negative, casting a doubt on appositeness of market data based SRMs and MRRMs to macro prudential policy making. However, one needs to interpret the results with care, considering where they apply and especially how best to make use of them. We surmise that the reason for the rather poor performance of the risk forecast models, especially during times of market turmoil, is because of the nature of financial risk, as discussed below. Ultimately, this has particular implications for the implementation of risk forecast models in policy and in particular the quality criteria that should be applied to them.

5.1 Relevance of data

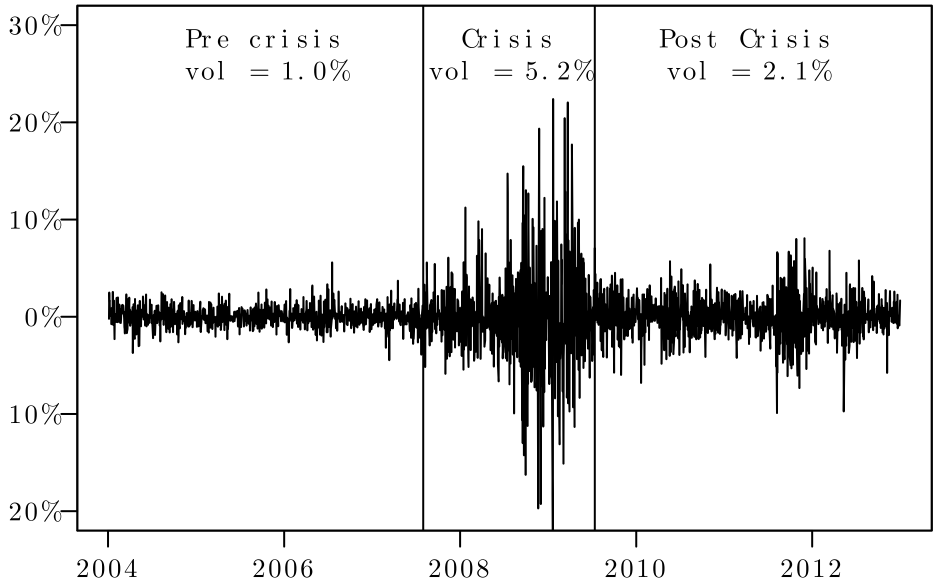
The high degree of model risk, as documented above, does not inspire confidence. We suspect there are two main reasons for this rather negative result: The low frequency of financial crises and the presence of endogenous risk. Perhaps the main problem in systemic risk forecasting/identifying is the low frequency of financial crises. While fortunate from a social point of view, it causes significant difficulties for any empirical analysis. For OECD countries the unconditional probability of a financial crisis is 2.3% a year, or alternatively, a typical country will suffer a financial crisis once every 43 years (Daníelsson, Valenzuela and Zer, 2014b). Therefore, the empirical analyst has to make use of data from non–crisis periods to impute statistical inference on the behavior of financial markets during crises.

The challenge in building an empirical systemic risk model is therefore capturing the risk of an event that has almost never happened using market variables during times when not much is going on. In order to do so, one needs to make stronger assumptions about the stochastic process governing market prices, assumptions that may not hold as the economy transits from a calm period to a turmoil period. At the very least, this implies that a reliable method would need to consider the transition from one state of the world to another. It requires a leap of faith to believe that price dynamics during calm time have much to say about price dynamics during crisis, especially when there is no real crisis to compare the forecast to. Ultimately this implies that from a statistical point of view, the financial system may transit between distinct stochastic processes, frustrating modeling.

We illustrate the issues that arise by a time series of JP Morgan returns, as seen in Figure 5. Visual identification shows the presence of three distinct regimes, where the volatility and extreme outcomes before the crisis do not seem to indicate the potential for future crisis events, and similarly, data during the crisis would lead to the conclusion that risk is too high after the crisis.

In other words, if one were to estimate a model that does not allow for

Figure 5: Daily JP Morgan returns before, during and after the last crisis, along with daily volatility. Data is obtained from CRSP 1925 US Stock Database.



structural breaks, one is likely to get it wrong in all states of the world; risk assessments would be too low before the crisis and too high after the crisis.

The second reason for the rather high levels of model risk witnessed here is how risk arises in practice. Almost all risk models assume risk is exogenous, in other words that adverse events arise from the outside. However, in the language of Daníelsson and Shin (2003), risk is endogenous, created by the interaction between market participants. Because market participants have an incentive to undermine any extant rules aimed at controlling risk–taking and hence, take risk in the least visible way possible, risk–taking is not visible until the actual event is realized. In the words of the former head of the BIS, Andrew Crockett (2000):

> "The received wisdom is that risk increases in recessions and falls in booms. In contrast, it may be more helpful to think of risk as increasing you upswings, as financial imbalances build up, and materializing in recessions."

5.2 Quality control for risk measures

Given the fundamental importance of risk models in macro prudential policy, the poor model performance is a cause for concern. After all, policymakers would like to use their outputs for important purposes; perhaps to determine capital for systematically important institutions, or in the design of financial regulations, where the costs of type I and type II errors are significant. This requires the statistical results from an SRM or MRRM being precise. However, our analysis shows that the risk readings depend on the model employed, so it is not possible to accurately conclude which institution is (systemically) riskier than the other.

For this reason, any risk forecast methodology that is to be embedded into macro prudential policy should be evaluated by the following four criteria:

1. Point forecasts are not sufficient, confidence intervals incorporating the uncertainty from a particular model should be provided along with any point forecast. Methods need to be analyzed for robustness and model risk;

2. Data should be predictive and not reactive. A large number of time series coincide in signaling market turmoil. However, most of these show market turmoil only after it has occurred. It is therefore important that a systemic risk forecast be properly predictive and an indicator be real–time, and not lagging;

3. The proposed statistical models should incorporate backtesting. Simply proposing a methodology without demonstrating how it has performed over historical periods of market turmoil is not sufficient. In particular, demonstrating that a particular method predicted or indicated the crisis of 2008 is not sufficient, after all, it is only one observation from which to draw important inference;

4. Event probabilities need to correspond with the probability of actual market turmoil. If financial crisis happen only once every 40 years, daily 99% event probabilities, which happen 2.5 times a year, are of little relevance. The existing and the proposed MRRMs focus on daily 99% and 97.5%, respectively and most SRMs are based on 99% or lower daily probabilities. Hence, neither the SRMs nor MRRMs seem to have much to say about crises.

6 Conclusion

Risk forecasting is a central element in macro prudential policy, both when addressing systemic risk and in the regulation of market risk. The fundamental problem of model risk in any risk model such as VaR arises because risk cannot be measured, but has to be estimated by the means of a statistical model. Many different candidate statistical models have been proposed where one cannot robustly discriminate between them. Therefore, with a range of different plausible models one obtains a range of risk readings, and their disagreement provides a succinct measure of model risk.

We propose a method, termed risk ratios, for the estimation of model risk in macro prudential motivated risk forecasting, be it for systemic risk or regulatory purposes. Our results indicate that, predominately during times of no financial distress, model risk is low. In other words, the various candidate statistical models are roughly equally informative. However, model risk is significantly correlated with and caused by market uncertainty, proxied by the VIX index. Macro prudential motivated risk forecast models are subject to a significant model risk during financial distress periods, unfortunately those times when they are most needed. Hence, policymakers should interpret the risk readings with caution since this may lead to costly decision mistakes.

A Statistical methods

We employ the six VaR forecast methodologies most commonly used by industry: historical simulation (HS), moving average (MA), exponentially weighted moving average (EWMA), normal GARCH (G), student$-t$ GARCH (tG), and extreme value theory (EVT).

Historical simulation is the simplest non-parametric method to forecast risk. It employs the p^{th} quantile of historic return data as the VaR estimate. The method does not require an assumption regarding the underlying distribution on asset returns. However, it relies on the assumption that returns are independent and identically distributed. Moreover, it gives the same importance to all returns, ignoring structural breaks and clustering in volatility.

All the other five methods we consider in this study are parametric methods. For the first four, the VaR is calculated as follows:

$$\mathrm{VaR}(p)_{t+1} = -\sigma_t F_R^{-1}(\vec{\theta})\vartheta, \tag{5}$$

where σ_t is the time-dependent return volatility at time t, $F_R(\cdot)$ is the distribution of standardized simple returns with a set of parameters $\vec{\theta}$, and ϑ is the portfolio value. Hence, these approaches require a volatility estimate and distributional assumptions on asset returns.

One of the simplest ways to estimate the time–varying volatility is the moving average models. Under the assumption that the returns are conditionally normally distributed, $F_R(\cdot)$ represents the standard normal cumulative distribution $\Phi(\cdot)$. The volatility is calculated as:

$$\hat{\sigma}^2_{\mathrm{MA},t+1} = \frac{1}{E_W}\sum_{i=1}^{E_W} y^2_{t-i+1}, \tag{6}$$

where E_W is the estimation window size and equals to 1,000 in our analysis. The moving average model gives the same weight $1/E_W$ to each return in the sample. On the other hand, the exponentially weighted moving average (EWMA) model modifies the MA model by applying exponentially decaying weights into the past:

$$\hat{\sigma}^2_{\mathrm{EWMA},t+1} = (1-\lambda)y_t^2 + \lambda\hat{\sigma}^2_{\mathrm{EWMA},t},$$

where λ is the decay factor and set to 0.94 as suggested by J.P. Morgan for daily returns (J.P. Morgan, 1995).

In addition, we estimate the volatility by employing a standard GARCH(1,1) model both under the assumption that returns are normally and student-t distributed. We denote the former model as normal GARCH (G) and the latter one as the student-t distribution GARCH (tG).

$$\hat{\sigma}^2_{G,t+1} = \omega + \alpha y_t^2 + \beta\sigma^2_{G,t}.$$

The degrees of freedom parameter for the student-t distribution GARCH (tG) is estimated through a maximum-likelihood estimation.

Finally, we use of Extreme Value Theory (EVT) which is based on the fact that for any fat tailed distribution, as applies to all asset returns, the tails are asymptotically

Pareto distributed.

$$F(x) \approx 1 - Ax^{-\iota}$$

where A is a scaling constant whose value is not needed for VaR and ι the tail index, estimated by maximum likelihood (Hill, 1975):

$$\frac{1}{\hat{\iota}} = \frac{1}{q}\sum_{i=1}^{q} \log \frac{x_{(i)}}{x_{(q-1)}},$$

where q is the number of observations in the tail. The notation $x_{(i)}$ indicates sorted data. We follow the VaR derivation in Daníelsson and de Vries (1997):

$$\mathrm{VaR}(p) = x_{(q-1)} \left(\frac{q/T}{p}\right)^{1/\hat{\iota}}.$$

ES is then:

$$\mathrm{ES}(p) = \mathrm{VaR}\,\frac{\hat{\iota}}{\hat{\iota}-1}.$$

B Daily risk ratios

Table B.1: Daily risk ratios: overlapping 99% VaR and non-overlapping 97.5% ES
This table reports the maximum of the ratio of the highest to the lowest daily VaR and ES forecasts (risk ratios) for the period from January 1974 to December 2012 for the S&P-500, Fama-French financial sector portfolio (FF), the value-weighted portfolio of the biggest 100 stocks in our sample (Fin100), JP Morgan (JPM), American Express (AXP), American International Group (AIG), and Goldman Sachs (GS). Panels 1(a) and 1(b) present the risk ratio estimates where the risk is calculated via daily 99% VaR 10–day overlapping and 97.5% ES with non-overlapping estimation windows, respectively. Six different methods; historical simulation, moving average, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory are employed to calculate the VaR and ES estimates. Estimation window size is 1,000. Finally, the last row of each panel reports the average risk ratio for the whole sample period.

(a) VaR, $p = 99\%$, 10–day overlapping

Event	Peak	Trough	SP-500	FF	Fin100	JPM	AXP	AIG	GS
1977 crash	1977-05	1977-10	4.29	9.16	4.80	5.48	6.67	3.26	
1980 recession	1980-01	1980-07	5.42	19.81	10.97	13.16	5.05	6.34	
1981 recession	1981-07	1982-11	7.01	11.40	14.87	18.09	4.82	7.05	
1987 crash	1987-10	1988-01	11.82	13.68	52.36	7.37	7.27	5.22	
1990 recession	1990-07	1991-03	8.43	15.72	22.53	7.05	4.44	15.52	
LTCM crisis	1998-08	1998-11	4.93	5.13	5.32	8.04	5.76	6.95	
2001 recession	2001-03	2001-11	4.76	4.13	3.88	3.18	4.45	3.18	
2008 recession	2007-12	2009-06	9.45	10.24	8.44	21.60	11.04	35.31	7.10
Full sample (ave.)	1974-01	2012-12	2.40	2.59	2.59	2.36	2.34	2.65	2.66

(b) ES, $p = 97.5\%$, non-overlapping

Event	Peak	Trough	SP-500	FF	Fin100	JPM	AXP	AIG	GS
1977 crash	1977-05	1977-10	2.56	3.21	3.30	3.38	4.26	16.23	
1980 recession	1980-01	1980-07	2.08	2.71	2.41	2.58	2.06	3.36	
1981 recession	1981-07	1982-11	2.28	2.48	2.52	3.11	2.95	3.91	
1987 crash	1987-10	1988-01	8.55	9.84	8.94	9.92	5.52	3.83	
1990 recession	1990-07	1991-03	2.65	2.92	2.64	4.00	2.38	2.16	
LTCM crisis	1998-08	1998-11	4.91	4.12	3.62	3.48	5.26	3.36	
2001 recession	2001-03	2001-11	2.06	2.59	2.57	2.28	2.31	2.90	
2008 recession	2007-12	2009-06	6.94	5.82	7.41	7.09	7.14	14.77	6.36
Full sample (ave.)	1974-01	2012-12	1.84	1.91	1.95	1.96	1.93	2.29	2.20

C CoVaR

Following Adrian and Brunnermeier (2011) we estimate the time-varying CoVaR via quantile regressions. For stock i and the system S:

$$\begin{aligned} R_{t,i} &= \alpha_i + \gamma_i M_{t-1} + \varepsilon_{t,i} \\ R_{t,S} &= \alpha_{S|i} + \beta_{S|i} R_{t,i} + \gamma_{S|i} M_{t-1} + \varepsilon_{t,S|i}, \end{aligned}$$

where R is defined as the growth rate of marked-valued total assets. The overall financial system portfolio $R_{t,S}$ is the weighted average of individual stock $R_{t,i}$s, where the lagged market value of assets is used as weights. Finally, M denotes the set of state variables that are listed in detail below.

By definition, VaR and CoVaR are obtained by the predicted values of the quantile regressions:

$$\begin{aligned} \mathrm{VaR}_{t,i} &= \widehat{\alpha}_i + \widehat{\gamma}_i M_{t-1} \\ \mathrm{CoVaR}_{t,i} &= \widehat{\alpha}_{S|i} + \widehat{\beta}_{S|i}\,\mathrm{VaR}_{t,i} + \widehat{\gamma}_{S|i} M_{t-1}. \end{aligned} \tag{7}$$

The marginal contribution of an institution, ΔCoVaR, is defined as:

$$\Delta\mathrm{CoVaR}_{t,i}(p) = \widehat{\beta}_{S|i}\left[\mathrm{VaR}_{t,i}(p) - \mathrm{VaR}_{t,i}(50\%)\right]. \tag{8}$$

In order to calculate CoVaR estimates, we collapse daily market value data to a weekly frequency and merged it with quarterly balance sheet data from the CRSP/Compustat Merged Database. Following Adrian and Brunnermeier (2011), the quarterly data are filtered to remove leverage and book-to-market ratios less than zero and greater than 100, respectively.

We start our analysis by considering the time series relationship between ΔCoVaR and VaR. ΔCoVaR is defined as the difference between the CoVaR conditional on the institution is under distress and CoVaR calculated in the median state of the same institution. Given that the financial returns are (almost) symmetrically distributed, VaR calculated at 50% is almost equal to zero. Our empirical investigation confirms this theoretic observation; we find that the unconditional correlation between VaR and ΔCoVaR mostly exceeds 99%. This suggests that the scaled signal provided by ΔCoVaR is very similar to the signal provided by VaR.

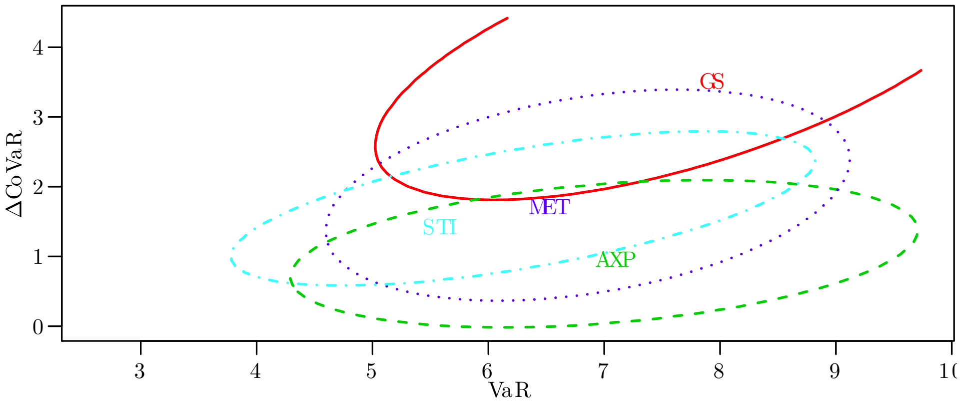
On the other hand, in a cross sectional setting, in what is perhaps their key result, Adrian and Brunnermeier (2011) find that even if the VaR of two institutions is similar, their ΔCoVaR can be significantly different, implying that the policy maker should consider this while forming policy regarding institutions' risk.

In order to get the idea of the model risk embedded in this estimation, we employ a bootstrapping exercise. For each of the stocks we re-run the quantile regressions 1,000 times by reshuffling the error terms and estimate VaR, CoVaR, and ΔCoVaR for each trial. Figure 6 shows 99% confidence intervals of the bootstrapped estimates along with the point estimates. An institution's ΔCoVaR is plotted on the y-axis and its VaR on the x-axis, estimated as of 2006Q4 at a 1% probability level. For the ease of presentation, we present the confidence intervals for Goldman Sachs (GS), American Express (AXP), Metlife (MET), and Suntrust Banks (STI). The point estimates show that there is a considerable difference between VaR and ΔCoVaR cross–sectionally, confirming the results of Figure 1 in Adrian and Brunnermeier (2011). For instance, although the VaR estimate of Goldman Sachs (GS) is comparable to its peers, its contribution to systemic risk, ΔCoVaR, is the highest. However concluding that Goldman Sachs (GS) is the systemically riskiest requires substantially caution since the confidence intervals overlap in quite wide ranges.

The following set of state variables (M) are included in the time–varying CoVaR analysis:

Figure 6: 99% confidence intervals

99% confidence intervals of the 1,000 bootstrapped quantile regressions outlined in (7). VaR is the 1% quantile of firm asset returns, and ∆CoVaR is the marginal contribution of an institution to the systemic risk. The confidence intervals of Goldman Sachs (GS), Metlife (MET), Suntrust Banks (STI), and American Express (AXP) are presented. Portfolio value is equal to \$100. Stock data is obtained from CRSP 1925 US Stock Database and CRSP/Compustat Merged Database.



1. *Chicago Board Options Exchange Market Volatility Index (VIX):* Captures the implied volatility in the stock market. Index is available on the Chicago Board Options Exchange's website since 1990.

2. *Short–term liquidity spread:* Calculated as the difference between three-months US repo rate and three-months US Treasury bill rate. The former is available in Bloomberg since 1991 whereas bill rate is from the Federal Reserve Board's H.15 release.

3. The change in the three-months Treasury bill rate.

4. *Credit spread change:* Difference between BAA-rated corporate bonds from Moody's and 10-years treasury rate, from H.15 release.

5. *The change in the slope of the yield curve:* The change in difference of the yield spread between the 10-years Treasury rate and the three-months bill rate.

6. S&P500 returns as a proxy for market return.

7. Real estate industry portfolio obtained from Kenneth French's website.

References

Acharya, V. V., R. Engle and M. Richardson, "Capital Shortfall: A New Approach to Ranking and Regulating Systemic Risk," *American Economic Review* 102 (2012), 59–64.

Acharya, V. V., L. H. Pedersen, T. Philippon and M. Richardson, "Measuring Systemic Risk," (May 2010), Working Paper.

Adrian, T. and M. K. Brunnermeier, "CoVaR," (2011), Working Paper, NBER–17454.

Alessi, L. and C. Detken, "Real Time Early Warning Indicators for Costly Asset Price Boom/Bust Cycles," ECB Working Paper Series 1039, European Central Bank, 2009.

Alexander, C. and M. Sarabia, J., "Quantile Uncertainty and Value-at-Risk Model Risk," *Risk Analysis* 32 (2012), 1293–1308.

Basel Committee, *Amendment to the Capital Accord to Incorporate Market Risks*, Basel Committee on Banking Supervision (1996), `http://www.bis.org/publ/bcbs24.pdf`.

Basel Committee on Banking Supervision, "Fundamental Review of the Trading Book: A Revised Market Risk Framework," Technical Report, Basel Committee on Banking Supervision, 2013.

Berkowitz, J. and J. O'Brien, "How Accurate Are Value-at-Risk Models at Commercial Banks?," *Journal of Finance* 57 (2002), 977–987.

Bisias, D., M. Flood, A. Lo and S. Valavanis, "A Survey of Systemic Risk Analytics," Technical Report, U.S. Department of Treasury, January 2012.

Borio, C. and M. Drehmann, "Assessing the Risk of Banking Crises–Revisited," BIS Quarterly Review, pp. 29–46, Bank of International Settlements, 2009.

Boucher, M., C., J. Daníelsson, S. Kouontchou, P. and B. Maillet, B., "Risk Models–at–Risk," *forthcoming, Journal of Banking and Finance* (2013).

Brownlees, C. T. and R. Engle, "Volatility, Correlation and Tails for Systemic Risk Measurement," (July 2012), Working Paper, NYU Stern School of Business.

Cont, R., "Model uncertanity and its impact on the pricing of derivative instruments," *Mathematical Finance* 16 (2006), 519–547.

Crockett, A., "Marrying the micro- and macro-prudential dimensions of financial stability," (2000), the General Manager of the Bank for International Settlements; `http://www.bis.org/review/rr000921b.pdf`.

Daníelsson, J., "The Emperor has no Clothes: Limits to Risk Modelling," *Journal of Banking and Finance* 26 (2002), 1273–1296.

Daníelsson, J. and C. G. de Vries, "Tail index and quantile estimation with very high frequency data," *Journal of Empirical Finance* 4 (1997), 241–257.

DANÍELSSON, J., K. JAMES, M. VALENZUELA AND I. ZER, "Can We Prove a Bank Guilty of Creating Systemic Risk? A Minority Report," (2014a), London School of Economics Working Paper.

DANÍELSSON, J. AND H. S. SHIN, "Endogenous Risk," in *Modern Risk Management — A History* (Risk Books, 2003), `http://www.RiskResearch.org`.

DANÍELSSON, J., M. VALENZUELA AND I. ZER, "Learning from History: What Statistical Analysis of Crises Passed Tells Us About Crises Future," (2014b), London School of Economics Working Paper.

DREHMANN, M. AND N. TARASHEV, "Measuring the Systemic Importance of Interconnected Banks," *Journal of Financial Intermediation* 22 (2013), 586–607.

ENGLE, R., "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business and Economic Statistics* 20 (2002), 339 – 350.

GIBSON, R., ed., *Model Risk: Concepts, Calibration and Pricing* (Risk Books, 2000).

GIRARDI, G. AND T. ERGUN, A., "Systemic risk measurement: Multivariate GARCH estimation of CoVaR," *Journal of Banking and Finance* 37 (2013), 3169–3180.

GRAY, D. AND A. JOBST, *Systemic Contingent Claims Analysis–A Model Approach to Systemic Risk* (London: Edward Elgar, 2011), 143–185.

———, "Systemic Contingent Claims Analysis–Estimating Market-Implied Systemic Risk," Technical Report, IMF, 2013, iMF, WP/13/54.

GREEN, T., C. AND S. FIGLEWSKI, "Market Risk and Model Risk for a Financial Institution Writing Options," *Journal of Finance* 54 (1999), 1465–1499.

HANSEN, B. E., "Autoregressive Conditional Density Estimation," *International Economic Review* 35, (3) (1994), 705–30.

HENDRICKS, D., "Evaluation of Value-at-Risk Models Using Historical Data," Technical Report, FRBNY Economic Policy Review, April 1996.

HILL, B. M., "A simple general approach to inference about the tail of a distribution," *Ann. Statist.* 35 (1975), 1163–1173.

HUANG, X., H. ZHOU AND H. ZHU, "A Framework for Assessing the Systemic Risk of Major Financial Institutions," *Journal of Banking and Finance* 33 (2009), 2036–2049.

———, "Assessing the Systemic Risk of a Heterogeneous Portfolio of Banks during the Recent Financial Crisis," *Journal of Financial Stability* 8 (2012), 193–205.

HULL, J. AND W. SUO, "A methodology for assessing model risk and its application to the implied volatility function model," *Journal of Financial and Quantitative Analysis* 37 (2002), 297–318.

IMF, "Assessing the Systemic Implications of Financial Linkages," Technical Report, International Monetary Fund, April 2009.

J.P. Morgan, *RiskMetrics-technical manual*, third edition (1995).

Kuester, K., S. Mittnik and S. Paolella, M., "Value-at-Risk Prediction: A Comparison of Alternative Strategies," *Journal of Financial Econometrics* 4 (2006), 53–89.

Mainik, G. and E. Schaanning, "On dependence consistency of CoVaR and some other systemic risk measures," (August 2012), Working Paper.

Segoviano, A., M. and C. Goodhart, "Banking Stability Measures," Technical Report, IMF, WP/09/4, 2009.

Suh, S., "Measuring systemic risk: A factor-augmented correlated default approach," *Journal of Financial Intermediation* 21 (2012), 341–358.

Tarashev, N., C. Borio and K. Tsatsaronis, "Attributing systemic risk to individual institutions," Technical Report, BIS, May 2010.