Managing extreme risks in tranquil and volatile markets using conditional extreme value theory

Hans N.E. Byström *,1

School of Finance and Economics, University of Technology Sydney, P.O. Box 123, Broadway NSW 2007, Australia

Accepted 13 February 2004

Abstract

Financial risk management typically deals with low-probability events in the tails of asset price distributions. To capture the behavior of these tails, one should therefore rely on models that explicitly focus on the tails. Extreme value theory (EVT)-based models do exactly that, and in this paper, we apply both unconditional and conditional EVT models to the management of extreme market risks in stock markets. We find conditional EVT models to give particularly accurate Value-at-Risk (VaR) measures, and a comparison with traditional (Generalized ARCH (GARCH)) approaches to calculate VaR demonstrates EVT as being the superior approach both for standard and more extreme VaR quantiles.

© 2004 Elsevier Inc. All rights reserved.

JEL classification: C22; C53; G19
Keywords: Value-at-Risk; Conditional extreme value theory; GARCH; Backtesting

1. Introduction

Financial risk management typically deals with low-probability events in the tails of asset return distributions. The further out in the tails we go, the smaller the probability of an event becomes at the same time as its consequences will be larger. It is therefore important to be able to model these, the most extreme price movements, accurately. Traditional risk management models, however, might fail to give us accurate estimates and
forecasts of the tails because they usually focus on the whole distribution, of which the tails are just small portions. Extreme value theory (EVT)-based risk management, on the other hand, focuses directly on the tails and could therefore potentially give us better estimates and forecasts of risk.

The modelling of extreme events is the central issue in EVT, and the main purpose of the theory is to provide asymptotic models with which we can model the tails of a distribution. EVT has been around for some time, from the pioneering work on block maxima of Fischer and Tippett (1928) and Gnedenko (1943) to the exposé by Gumbel (1958). More recently, Balkema and de Haan (1974) and Pickands (1975) have presented results for threshold-based extreme value methods. Applications of the theory have since appeared in different areas over the years, for instance, in hydrology and wind engineering, and lately we have seen an increased interest in EVT applied to finance and insurance. Presentations emphasizing these applications can be found in Embrechts, Klueppelberg, and Mikosch (1997), and in Reiss and Thomas (1997).

In studies of today’s volatile financial markets, it is common to make a distinction between conditional and unconditional asset return distributions. In risk management, the unconditional distribution is interesting when we look at long-term investment decisions and the occurrence of very rare (stress) events while the conditional distribution is more appropriate when we deal with day-to-day risks and short-term risk management. In this paper, we follow McNeil and Frey (2000) by focusing on day-to-day stock market risk and Value-at-Risk (VaR) calculations, and to forecast VaR measures, we consider both conditional and unconditional EVT models. McNeil and Frey calculate conditional VaR$_p$ measures by filtering return series with a Generalized ARCH (GARCH) model and then apply threshold-based EVT tools to the (near) independently identically distributed (IID) residuals. We extend their approach to the block maxima method and create conditional VaR$_p$ forecasts based on the block maxima method. We also compare the performance of the block maxima approach with the threshold approach in general, and find the two models to perform very similarly. We continue by emphasizing the relation between extreme returns and the level of volatility in the market, and evaluate unconditional and conditional VaR$_p$ forecasts in both tranquil and volatile periods. Finally, the particular period at and around the Asian financial crisis and the behavior of the most extreme returns is studied in more detail.

Although both the negative and the positive tails of stock return distributions are interesting from a risk management perspective, most studies of extreme stock returns focus on losses (as opposed to gains), and large crashes are generally considered more interesting than large booms. In this paper, we look at two countries’ aggregated stock markets and we expect losses to be of more general interest than gains. To save space, we therefore chose to focus solely on the negative tail of the distributions of two different stock return series; the Swedish “Affärsvärlden’s Generalindex” (AFF) and the U.S. Dow Jones Industrial Average (DOW) from 1980 to 1999.

Section 2 introduces different unconditional and conditional models used to describe both “normal” and extreme stock market returns, Section 3 gives a brief introduction to the data, the different stock markets, and the behavior of the most extreme returns, and Section 4 presents the results from both in-sample and out-of-sample evaluations of VaR$_p$ measures in tranquil and volatile periods. Section 5 concludes the paper.
2. Modelling the tails of stock return distributions

In the following subsections, we present the different stock return models that we use to estimate and forecast risk measures (VaR, quantiles) in stock markets. The main focus is on EVT-based models and the fundamental theory underlying such models. One can divide traditional (unconditional) EVT into at least two groups; either one looks at those events in the data that exceed a high threshold (% peaks over threshold (POT) methods), or one divides the data into consecutive blocks and focuses on the series of maxima (or minima) in these blocks (block maxima methods).

2.1. The block maxima method

The theory behind the block maxima method can be found in work by various authors (for instance, in Embrechts et al., 1997; Kellezi & Gilli, 2000; McNeil, 1998) and here, we only give a brief presentation of the method. We let \( R_i \) be the daily returns in our data series and initially we assume these returns to be independently identically distributed (IID). From these returns, which are drawn from the (unknown) distribution \( F_R \), we draw maxima \( X_i \) from consecutive nonoverlapping blocks of length \( n \); in total \( m \) maxima. These maxima are then normalized to fit our specific distribution as \( Y_i = (X_i - \beta) / \alpha \), where \( \alpha \) and \( \beta \) are constant location and scale parameters. The limit distribution of these normalized maxima, when \( n \rightarrow \infty \), was then proved by Fischer and Tippett (1928) and Gnedenko (1943) to be either the Gumbel, the Fréchet, or the Weibull distribution. Later, Jenkinson (1955) suggested the generalized extreme value (GEV) distribution that contains the three types distinguished by Gnedenko:

\[
H_Y(y) = e^{-(1+\xi)y} - 1/\xi, \quad \text{if } \xi \neq 0,
\]

\[
H_Y(y) = e^{-e^{-y}}, \quad \text{if } \xi = 0. \tag{1}
\]

The parameter \( \xi \) is called the tail index and is related to the shape of the underlying distribution \( F_R \). So-called thin-tailed distributions, like the normal distribution, lead to the Gumbel case, \( \xi = 0 \), while fat-tailed distributions, like Student’s \( t \) distribution and the Pareto distributions, lead to the Fréchet case, \( \xi > 0 \). For financial time series, which are known to be fat-tailed, one therefore normally gets positive values for \( \xi \) when data is fitted to the GEV distribution.

The tail index, \( \xi \), as well as the constants \( \alpha \) and \( \beta \), are estimated by fitting the GEV distribution to the actual data, and in this paper, we use the maximum likelihood method:

\[
\max L_H(\xi, \alpha, \beta, x) = \max \sum_i \ln(h_{\xi,\alpha,\beta}(x_i)), \tag{2}
\]

where

\[
h_{\xi,\alpha,\beta}(x) = \frac{1}{\alpha} \left( 1 + \xi \left( \frac{x - \beta}{\alpha} \right) \right)^{-1} - 1 \left( 1 + \xi \left( \frac{x - \beta}{\alpha} \right) \right)^{-1} e^{-\left( 1 + \xi \left( \frac{x - \beta}{\alpha} \right) \right)^{-1}}, \tag{3}
\]
is the density function of the GEV distribution if $\xi \neq 0$ and $1+\xi \left(\frac{x-a}{\alpha}\right) > 0$. To get confidence intervals for the parameters, we use the standard bootstrap method.

Up until now, we have only considered the IID case. Under the additional assumption of “weak long-range dependency” (which is supposed to hold in most financial markets), we can fit a (slightly modified) GEV distribution to stationary series that show the kind of clustering behavior seen for instance in the stock market (Embrechts et al., 1997; McNeil, 1998). What we need to do is to define a new parameter called the extremal index, $\theta$. To define $\theta$, we observe that for $n$ IID returns, $R_i$, drawn from a distribution $F_R$, we can easily calculate the distribution of the sample maxima as

$$F_X = \left[F_R\right]^n. \quad (4)$$

$\theta$ is then defined, asymptotically, by the corresponding expression for non-IID stationary series with clusters

$$F_X = \left[F_R\right]^{n\theta}. \quad (5)$$

This means that the asymptotic distribution of (normalized) maxima for the non-IID series is in fact a GEV distribution; more specifically, it is the GEV distribution $H_y(y)$, which we would have had if we had an IID series, but raised to the power $\theta$. As McNeil (1998) points out, the extremal index $\theta$ can then be interpreted as the reciprocal of the mean cluster size.

The extremal index must be introduced if one wants to modify one’s calculations to incorporate clustering when going from the distribution of maxima and its quantiles to the underlying (unknown) return distribution and its quantiles. For instance, if VaR$_p$ quantiles of the return distribution, $F_R$, is to be calculated, as in this paper, one needs an estimate of $\theta$, the extremal index. Using the so-called % blocks method (Embrechts et al., 1997), $\theta$ can be estimated asymptotically as

$$\theta = \frac{1}{n} \frac{\ln \left(1 - \frac{K_u}{m}\right)}{\ln \left(1 - \frac{N_u}{mn}\right)} \quad (6)$$

where $N_u$ is the number of returns exceeding a certain high threshold, $K%_u$ is the number of blocks in which this threshold is exceeded, and $m$ and $n$, as before, are, respectively, the number of blocks and the length of these blocks.

---

2 The descriptive study of our data, see below, indicates that we have the Fréchet case in our two stock markets.

3 A discussion on the distributional mixing condition describing the sort of weak long-range dependencies we are talking about can be found in chapter 4.4 of Embrechts et al. (1997).

4 Throughout this paper, we make the (rather weak) assumption that the stock return series actually have extremal indices.

5 The GEV distributions for IID and non-IID series have the same tail index, because raising $H_y(y)$ to the power $\theta$ only affects the scale and location parameters $\alpha$ and $\beta$. 

---
Finally, by inverting Eq. (5), we can go from the asymptotic GEV distribution of maxima to the distribution of the observations themselves, and by using the expression for the distribution of the maxima in Eq. (1) as well as inserting the estimates of \( \theta, \xi, \alpha, \) and \( \beta, \)
we may express the (unconditional) VaR \( p \) measures (tail quantiles) for our return series at a certain probability, \( p \), as

\[
\text{VaR}_p = \beta + \frac{\alpha}{\xi} [n\theta \ln(p)]^{-\xi} - 1].
\] (7)

2.2. The POT method

As an alternative to looking at blocks and block maxima, one can collect those returns in the series that exceed a certain high threshold, \( u \), and model these returns separately from the rest of the distribution. This is the POT method.

We start by calling a daily return in our data series \( R \) and assume that it comes from a distribution \( F_R \). The returns above the threshold \( u \) then follow the excess distribution \( F_u(y) \) that is given by

\[
F_u(y) = P(R - u \leq y \mid R > u) = \frac{F_R(u + y) - F_R(u)}{1 - F_R(u)}, \quad 0 \leq y \leq R_F - u
\] (8)

where \( y \) is the excess over \( u \), and \( R_F \) is the right endpoint of \( F_R \). If the threshold, \( u \), is high enough, Balkema and de Haan (1974) and Pickands (1975) show that for a large class of distributions, \( F_R \), the excess distribution, \( F_u(y) \), can be approximated by the so-called generalized Pareto distribution (GPD)

\[
G_{\xi, \alpha}(y) = \left[ 1 - \left( 1 + \frac{\xi}{\alpha} y \right) \right]^{\frac{1}{\alpha}}, \quad \text{if} \ \xi \neq 0
\]

\[
G_{\xi, \alpha}(y) = 1 - e^{\frac{-y}{\alpha}}, \quad \text{if} \ \xi = 0
\] (9)

for \( 0 \leq y \leq R_F - u \). \( \xi \) is the tail index and for the fat-tailed distributions found in finance, one can expect a positive \( \xi \). \( \alpha \) is just a positive scaling parameter. A more detailed discussion of the GPD and the POT method can be found in Embrechts et al. (1997).

The tail index, \( \xi \), as well as the scaling parameter, \( \alpha \), have to be determined by fitting the GPD to the actual data and just as for the GEV distribution above, we estimate the parameters with the maximum likelihood method:

\[
\max L_G(\xi, \alpha; y) = \max \sum \ln(g_{\xi, \alpha}(y_i)),
\] (10)

where

\[
g_{\xi, \alpha}(y) = \frac{1}{\alpha} \left( 1 + \frac{\xi}{\alpha} y \right)^{(1+\xi)}
\] (11)
is the density function of the GPD distribution if $\xi \neq 0$ and $1 + \xi/\alpha > 0$. To get confidence intervals for the parameters, we again apply the standard bootstrap method.\(^6\)

When the GPD distribution and its parameters are estimated, we continue by calculating VaR\(_p\) quantiles of the underlying return distribution $F_R$ which can be written as

$$F_R(u + y) = (1 - F_R(u))F_u(y) + F_R(u).$$

(12)

Acknowledging that $F_R(u)$ can be written as $(n - N_u)/n$ where $n$ is the total number of returns and $N_u$ is the number of returns above the threshold $u$, and that $F_u(y)$ can be replaced by $G_{\xi,\alpha}(y)$ (as well as rewriting $u + y$ as $x$), this expression can be simplified to

$$F_R(x) = 1 - \frac{N_u}{n} \left(1 + \frac{\xi}{\alpha}(x - u)\right)^{-\frac{1}{\xi}}.$$  

(13)

By inverting this expression, we get an expression for (unconditional) VaR\(_p\) quantiles associated with certain probabilities $p$:

$$\text{VaR}_p = u + \frac{\alpha}{\xi} \left(\left(\frac{n}{N_u}p\right)^{-\frac{1}{\xi}} - 1\right).$$

(14)

This expression can be compared to Eq. (7) that is based on the block maxima method.

2.3. GARCH models

In the literature, there are several different approaches to the modelling of financial asset returns. Among the most successful is the ARCH family and its extensions (Engle, 1982; Bollerslev, 1986). GARCH, in particular, has become the workhorse of many financial economists; not only within the academic world, but also in the industry. This popularity is also the motivation behind our choice of GARCH as representing a typical parametric model for stock returns. As opposed to the EVT-based models described above, GARCH models do not focus directly on the returns in the tails. Instead, by acknowledging the tendency of financial return volatilities to be time dependent, GARCH models explicitly model the conditional volatility as a function of past conditional volatilities and returns.

In this paper, we use the simplest possible AR-GARCH model with the mean return modelled as an AR(1) process and the conditional variance of the return as a GARCH(1,1) model:

$$r_t = \alpha_0 + \alpha_1 r_{t-1} + \varepsilon_t$$

$$\varepsilon_t^2 = \phi_0 + \phi_1 \varepsilon_{t-1}^2 + \phi_2 \varepsilon_{t-1}^2$$

(15)

with $\varepsilon_t | \Omega_{t-1} \sim N(0, \sigma_t^2)$ or $\sim$ Student’s $t$ distribution with mean $= 0$, variance $= \sigma_t^2$ and degree of freedom parameter, $v$, and where $\Omega_t$ is the information set of all information through time $t$.

---

\(^6\) The descriptive study of our data, see below, indicates that we have fat-tailed distributions.

\(^7\) We assume that the returns above the threshold are essentially IID. If this is not the case, one can instead apply the moving block bootstrap.
The reason for including the $t$ distribution is that empirical evidence strongly rejects the idea that financial returns are normally distributed. Among others, Baillie and DeGennaro (1990) as well as Poon and Taylor (1992) clearly demonstrate that the $t$ distribution gives a better fit to financial return series.

When the AR-GARCH model in Eq. (15) has been fitted to data by maximization of the likelihood function, one can estimate (or forecast) dynamic VaR$_p$ measures by assuming either the normal distribution or the $t$ distribution, multiplying one’s estimates (or forecasts) of $\sigma_t$ with the standard quantiles of each distribution, and finally adding the conditional mean. Compared to the unconditional EVT-based methods described earlier, the AR-GARCH models have the advantage of producing time varying VaR$_p$ measures.

2.4. Conditional EVT

We have now described two different approaches to calculating risk measures like VaR: the unconditional EVT approach that focuses directly on the tail but does not acknowledge the fact that financial asset returns are non-IID, and the conditional AR-GARCH approach that does model the conditional return distribution as time varying, but focuses on the whole return distribution and not only on the part we are primarily interested in, the tail.

The obvious step is to combine these two approaches as suggested by McNeil and Frey (2000) and Diebold, Schuermann, and Stroughair (2000). McNeil and Frey first filters different financial time series with a GARCH model, and then applies the EVT tool kit to the residuals coming out of such a model. In this paper, we follow McNeil and Frey in combining the extreme value approach with GARCH models, but in addition to applying the POT method to our residual series, we also apply the block maxima method.

The first step is to fit the AR-GARCH model in Eq. (15) to the return data. The residuals, $\eta_t$, from this model are hopefully close to IID (the better the fit of the GARCH model, the closer the residuals are to being IID) which means that it is straightforward to use either the POT method or the block maxima method to model the tails of these residual distributions.8 At the very least, the residuals should be closer to IID than the original return series. Finally, if we call the (unconditional) residual distribution quantiles, $x_p$, we can easily calculate the forecasted VaR$_p$ quantiles of our return distribution tomorrow as

$$\text{VaR}_{t+1,p} = z_0 + z_1 r_t + z_{t+1} x_p$$  \hspace{1cm} (16)

where $z_0 + z_1 r_t$ is the conditional mean and $\sigma_{t+1}$ is the GARCH forecast of tomorrow’s conditional volatility.

What we do is essentially to scale our unconditional EVT-based tail estimates with the expected return and volatility. In this way, we get forecasts of tail risks that are conditional on the actual market conditions. Another major advantage with first filtering the returns is that we get essentially IID series on which it is straightforward to apply the EVT machinery. It should perhaps be mentioned that it is common in the finance literature to apply the EVT machinery to financial return series that are known to be non-IID. Whether

---

8 No estimate of $\theta$ is needed in the block maxima method because we expect the GARCH filtered series to be close to IID.
this deviation from the assumptions of the underlying theory is critical for the validity of
the results is an open issue. We do believe that a filtering, such as ours, minimizes this
problem, though.

As a final model, in addition to the more elaborate model above, we also include
the rather naive unconditional normal model of stock returns where we assume that
the returns come from the normal distribution (IID) with historically estimated mean
and variance.

3. Swedish and U.S. stock market characteristics

To compare the performance of our different risk models, we apply the models to two
different stock indices: AFF and DOW. These indices are among the most important in
each country, and both have a long history; DOW was created in 1896 and AFF was
created in 1937. The two indices have some differences; while AFF is a very broad index
containing around 300 of the largest stocks on the Stockholm Stock Exchange, DOW
contains (and has done so since 1928) only 30 stocks, while AFF is relatively heavy in
high-tech stocks, DOW contains mostly traditional stocks, and while AFF is capitalization
weighted, DOW is unweighted. The value of AFF has also grown much faster than that of
DOW over the last 20 years (the time period we look at); while the price level of DOW
today is around 13 times the level we saw 20 years ago, the price level of AFF has
increased around 48 times.

We look at daily prices (trading days) and the time periods are January 2, 1980 to
November 3, 1999 for AFF and January 2, 1980 to September 8, 1999 for DOW. By
choosing the time periods in this way, we get two equally long series containing 4961
daily observations each.9,10 From these prices, we calculate 4960 returns (log differences)
and focus on the extreme lower quantiles of these distributions. Table 1 reports some
statistics on the return series and we can see how the unconditional distributions of the
returns are nonnormal, as evidenced by high skewness, high excess kurtosis, and highly
significant Jarque–Bera statistics.11 In Table 1, we are also presenting Phillips–Perron
statistics that tell us that the return series are stationary. Ljung–Box statistics for the
returns themselves and for the squared returns are also presented, and these statistics tell us
that we have both autocorrelation and GARCH effects in our return series. The existence
of (G)ARCH effects is further demonstrated by the Engle LM Test.12 The GARCH effects
are particularly significant for the AFF series.

9 The data was downloaded from the EcoWin database.
10 The reason for the slightly different endpoints for the two indices is that there are different numbers of
holidays in the two countries; in Sweden, there are about 11 days a year when the stock exchange is closed due to
holidays and in the United States, there are on average, 9 days a year. In the context of our paper, it is important to
know that no extreme returns (< −4% or >4%) occur at, or immediately before or after, these holidays.
11 The Jarque–Bera (J–B) statistic is \( \chi^2 \)-distributed under the null of normality. The statistic is
\( n \frac{\text{skewness}^2}{6} + \frac{\text{excess kurtosis}}{24} \), where \( n \) is the sample size.
12 We include the results for the LM test using 6 lags. The results for 1 to 12 lags are all also highly
significant.
In addition to the statistics above on the entire return distribution, it is also interesting to compare the sizes and dependencies of the most extreme returns (losses) in the negative tails of the two distributions. For AFF, the top three negative returns are $-9.15\%$, $-7.76\%$, and $-7.34\%$ and 28 losses are larger than $-4\%$. Finally, 499 losses (almost exactly 10% of the sample) are larger than one standard deviation away from the mean. For DOW, the top three negative returns are $-25.6\%$, $-8.38\%$ and $-7.45\%$, 12 losses are larger than $-4\%$, and 510 losses are larger than one standard deviation away from the mean.

When it comes to dependencies in the tail, we have tried to capture some of that by looking at clustering of extreme losses. In Table 2, we present the distribution of the time distances between the 50 most extreme losses (above $-3.0\%$ for AFF and above $-2.4\%$ for DOW) over the 20-year period. We see that for both indices, about half of the extremes come closer after each other than 1 month (21 days) and that two thirds are closer to each other than 3 months. If the extreme losses were independent, they should on average come with 100-day intervals. In other words, there is evidence of clustering within the months as well as within the quarters. While this clustering in the medium to long range is very similar for the two indices, the short-range situation is quite different. For the Swedish index, about one fourth of the extreme losses are closer to each other than 3 days while for the DOW, there is much less of these short-term extreme dependencies.

There is of course a connection between volatility and extremes and between volatility clustering and clustering of extremes. The short-term extreme dependencies and clustering that we find in AFF but not in DOW also corresponds well with the behavior for more normal returns; the $Q^2(6)$ statistic for the squared returns in Table 1 is much larger for AFF than for DOW and the variance parameter $\phi_1$ in Table 4 is larger for AFF than for DOW. The value of the Engle LM Test for (G)ARCH effects is also much higher for AFF than for DOW. In other words, AFF has more of both volatility clustering (GARCH) and extreme loss clustering.

As we have seen, there are some differences between the two indices when it comes to the size and frequency of extreme returns; while DOW has one very extreme negative

<table>
<thead>
<tr>
<th></th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>J–B</th>
<th>P–P</th>
<th>Q(6)</th>
<th>Q(12)</th>
<th>Q(24)</th>
<th>$Q^2(6)$</th>
<th>$Q^2(12)$</th>
<th>$Q^2(24)$</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFF</td>
<td>20.0</td>
<td>17.7</td>
<td>−0.31</td>
<td>8.34</td>
<td>14,450</td>
<td>−4186</td>
<td>122.9</td>
<td>140.8</td>
<td>190.04</td>
<td>1536.8</td>
<td>2459.8</td>
<td>2878.4</td>
<td>1217.4</td>
</tr>
<tr>
<td>DOW</td>
<td>12.8</td>
<td>16.4</td>
<td>−3.20</td>
<td>80.68</td>
<td>1,354,000</td>
<td>−4666</td>
<td>25.7</td>
<td>31.3</td>
<td>46.11</td>
<td>227.8</td>
<td>241.9</td>
<td>261.9</td>
<td>249.4</td>
</tr>
</tbody>
</table>

Mean and S.D. are annualized and in percent. J–B is the Jarque–Bera test for normality and the 99% critical value is 9.21. P–P is the Phillips–Perron test (without trend) and the 99% critical value is $-3.43$. Q(.) are the Ljung–Box tests for the returns and the squared returns and the 99% critical values are 16.8, 26.2, and 43.0. LM refers to the Engle (1982) LM test for presence of ARCH at lag 6. The 99% critical value is 16.8.
return observation (−25% during the October crash 1987), AFF overall has a larger number of extreme negative returns. Both skewness and kurtosis are also much higher for DOW than for AFF, but on the other hand, there is more autocorrelation in AFF than in DOW, both for the returns themselves and for the squared returns.

4. VaR calculations

We divide the empirical study into two parts: an in-sample study to examine the fit of the models to our data, and an out-of-sample study to assess how efficient each of the models is in forecasting both typical and extreme \( \text{VaR}_p \) quantiles. The out-of-sample experiment is comparable to the “backtesting” procedure recommended by the Basle Committee of Banking Supervision. All through this section, we look at losses; that is, we have chosen to look at extremes in the negative part of the original return distributions.

4.1. In-sample evidence

In this subsection, we fit different models to the full samples of our stock index returns. Starting with the block maxima method, we fit the GEV distribution in Eq. (1) to maxima (a maxima corresponds to the maximal loss in a block) from the two series. To do so, one first has to choose a proper block length. Considering both the asymptotic properties of the theory and the fairly short sample, we have chosen a block length of approximately one quarter \((n = 62)\) which gives us 80 blocks \((m = 80)\). Maximum likelihood estimates of \( \xi, \alpha, \) and \( \beta \) together with 95% confidence intervals (calculated by bootstrapping) are found in the left part of Table 3. The tail index estimates, 0.28 for AFF and 0.38 for DOW, together with the confidence intervals, indicate the Fréchet case for both stock indices.

To capture some of the clustering in the stock returns, we continue by estimating the extremal index, \( \theta \). We use the estimator in Eq. (6) for a number of different high thresholds between \(-1.5\%\) and \(-5\%\) (in steps of 0.5%) and the mean \( \theta \) value is estimated to be

<table>
<thead>
<tr>
<th>Time between extremes</th>
<th>AFF</th>
<th>DOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 day</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>≤3 days</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>≤1 week</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>≤2 weeks</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>≤1 month</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>≤1 quarter</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>≤1/2 year</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>≤1 year</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>≤2 years</td>
<td>49</td>
<td>49</td>
</tr>
</tbody>
</table>
0.56 for AFF and 0.66 for DOW. This means that we have an average clustering size of extremes equal to 1.8 days for AFF and equal to 1.5 days for DOW.

Finally, inserting our estimates of $h$, $n$, and $a$ in Eq. (7) gives us (1-day) unconditional VaR quantiles (in percent) for the negative tail of the marginal distributions. By comparing these VaR quantiles with the actual returns (losses) over the whole sample, and count the number of losses that are larger than the estimated unconditional VaR measure, we get information on the accuracy of these measures. For instance, the theoretical number of exceedences of a VaR 99% measure over a time period of 4960 days is 49.6 ($0.01\times 4960$). In Table 5, we see how the block maxima approach (GEV) to estimate extreme VaR measures performs well and that we have a close correspondence between the actual and the theoretical number of exceedences.

We continue with the POT method and fit the GPD to excess returns above a certain high threshold. Just as the block maxima method leaves us with a choice of an optimal block length, the POT method relies on a reasonable choice of threshold; too low threshold value and the asymptotic theory breaks down, too high threshold value and one does not have enough data points to estimate the parameters in the excess distribution. To have enough data points for estimation, but still only use observations in the extreme tails, we have chosen to fix the number of excess returns to 300 for each index. With this number, we also follow the recommendations (5–10% of the sample) from the simulation study in McNeil and Frey (2000). We get different threshold values, $u$, for the two indices: for AFF, the threshold is approximately $-1.5\%$, and for DOW, it is about $-1.3\%$ (Table 3).

Next, the tail index, $\xi$, as well as the scaling parameter, $\alpha$, have to be determined by fitting the GPD to the actual data. Just as for the GEV distribution above, we estimate the parameters with the maximum likelihood method and $\xi$, the tail parameter, is indeed positive and estimated to be 0.27 for AFF and 0.29 for DOW. Confidence intervals are given in Table 3.

Finally, we insert our parameter estimates in Eq. (14) and just as for the block maxima method, we calculate VaR quantiles for the unconditional distributions. The results are presented in Table 5 and are similar to those of the block maxima method.

15 The two parameters $N_u$ and $K_u$ in Eq. (6) are direct functions of the threshold value. We have chosen thresholds that give reasonable values for both $N_u$ and $K_u$. One can equally well go the other way around and choose $N_u$ and $K_u$ directly.
For the two AR-GARCH models with normally distributed and with $t$-distributed errors, maximum likelihood estimates as well as some statistics on the standardized residuals are presented in Table 4. For both models and both indices, we get significant parameter estimates and $\phi$ parameters that are positive and whose sum is less than 1. Statistics on the standardized residuals indicate a reasonably good fit of our models to data. While most of the dependencies are removed [$Q(6)$ and $Q(12)$ are much smaller for the residuals than for the original return series] some skewness and kurtosis remains. This is common in GARCH modelling of financial time series, particularly for stock return series when the 1987 crash is included in the data, and means that not all the fat-tailedness is captured by the time-varying volatility. Moreover, it should be understood that the major objective for us is to get residuals that are as close to IID as possible so that EVT models can be applied. We think we have succeeded in that. However, we do not expect the residuals to be normally distributed, and that is the reason why we chose to apply EVT-based models in the first place.

Daily VaR$_p$ quantiles are easily calculated by multiplying the AR-GARCH volatilities with quantiles from the standard normal and the $t$ distribution and finally adding the conditional mean. In Table 5, we present the number of exceedences of the AR-GARCH-based VaR$_p$ measures, and it is clear how the VaR$_p$ quantiles from the two AR-GARCH models differ from the EVT-based VaR$_p$ quantiles; the normal AR-GARCH model seriously underestimates more extreme VaR$_p$ quantiles while AR-GARCH $t$ instead overestimates less extreme VaR$_p$ quantiles. Obviously, the distribution of the error term has an important impact on the AR-GARCH-based VaR$_p$ measures.

Compared to the static EVT-based methods described earlier, the AR-GARCH model has the advantage of giving us time-varying VaR$_p$ measures; when we enter periods with

<table>
<thead>
<tr>
<th>Table 4</th>
<th>AR(1)–GARCH(1,1) parameters, as well as statistics on the standardized residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AFF</td>
</tr>
<tr>
<td>Normal</td>
<td></td>
</tr>
<tr>
<td>$\alpha_0 \times 10^4$</td>
<td>$-7.21_{1.31}$</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>$0.205_{0.0147}$</td>
</tr>
<tr>
<td>$\phi_0 \times 10^6$</td>
<td>$4.77_{2.12}$</td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>$0.128_{0.00743}$</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>$0.833_{0.00916}$</td>
</tr>
<tr>
<td>$\nu$</td>
<td></td>
</tr>
</tbody>
</table>

| Residuals |                 |                |                 |                 |
| Mean      | $0.0176$         | $0.0189$       | $0.0297$        | $0.0129$        |
| Standard deviation | $0.999$ | $0.999$ | $1.028$ | $1.022$ |
| Skewness  | $0.906$          | $0.756$        | $1.207$         | $1.014$         |
| Excess kurtosis | $8.45$ | $8.66$ | $13.02$ | $13.15$ |
| $Q(6)$    | $4.72$           | $9.97$         | $6.59$          | $11.84$         |
| $Q(12)$   | $14.18$          | $13.95$        | $17.07$         | $15.90$         |

Small figures below estimates are standard errors. $Q(.)$ are the Ljung–Box tests and the 99% critical values are 16.8 and 26.2.
high (low) volatility, the VaR\(_p\) measure is also increasing (decreasing). The EVT-based risk estimators, on the other hand, have the advantage of treating the tails more efficiently. Therefore, the next step is to combine these two approaches into a conditional extreme value model as described above.

Looking at the figures in Table 5, it is found that compared to the unconditional counterparts, the improvement is only marginal. A comparison of the results for the two conditional EVT models in Table 5 further shows that at more extreme VaR\(_p\) levels, the conditional block maxima approach gives slightly better estimates than the conditional POT model. At the 95% level, the situation is reversed.

Finally, the simple unconditional normal approximation, where we simply fit the normal distribution to the unconditional return distributions, obviously does a very bad job in estimating the more extreme VaR\(_p\) measures. As expected, all VaR\(_p\) measures, except for the 95% level, are seriously underestimated.

The general conclusion from the in-sample investigation has to be that VaR\(_p\)-estimates based on EVT dominate VaR\(_p\) estimates based on simple historical averages and GARCH models. The unconditional EVT models do a good job already and AR-GARCH scaling of EVT estimates do not significantly improve the results. In addition, whether we use the block maxima method or the POT method does not seem to be of importance.

A comparison of the results for AFF and DOW shows no major difference in performance between the series despite their slightly different characteristics. Possibly, the larger amount of autocorrelation and volatility clustering in the AFF case is the cause of the larger improvement in performance of the AR-GARCH model (compared to the unconditional normal model) on the Swedish index than on the U.S. index.

### Table 5

<table>
<thead>
<tr>
<th></th>
<th>Expected</th>
<th>Normal</th>
<th>AR-GARCH</th>
<th>AR-GARCH-t</th>
<th>GEV</th>
<th>GPD</th>
<th>Cond. GEV</th>
<th>Cond. GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFF VaR(_{95%})</td>
<td>248</td>
<td>209</td>
<td>227</td>
<td>165</td>
<td>256</td>
<td>229</td>
<td>226</td>
<td>253</td>
</tr>
<tr>
<td>VaR(_{99%})</td>
<td>49.6</td>
<td>83</td>
<td>68</td>
<td>31</td>
<td>42</td>
<td>42</td>
<td>43</td>
<td>42</td>
</tr>
<tr>
<td>VaR(_{99.5%})</td>
<td>24.8</td>
<td>57</td>
<td>46</td>
<td>19</td>
<td>27</td>
<td>26</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>VaR(_{99.9%})</td>
<td>4.96</td>
<td>38</td>
<td>24</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>VaR(_{99.95%})</td>
<td>2.48</td>
<td>34</td>
<td>22</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>DOW VaR(_{95%})</td>
<td>248</td>
<td>175</td>
<td>218</td>
<td>144</td>
<td>314</td>
<td>241</td>
<td>202</td>
<td>244</td>
</tr>
<tr>
<td>VaR(_{99%})</td>
<td>49.6</td>
<td>58</td>
<td>66</td>
<td>25</td>
<td>42</td>
<td>35</td>
<td>44</td>
<td>42</td>
</tr>
<tr>
<td>VaR(_{99.5%})</td>
<td>24.8</td>
<td>36</td>
<td>43</td>
<td>13</td>
<td>21</td>
<td>19</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>VaR(_{99.9%})</td>
<td>4.96</td>
<td>23</td>
<td>24</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>VaR(_{99.95%})</td>
<td>2.48</td>
<td>19</td>
<td>20</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

The numbers are empirically observed number of VaR exceedences. They should be as close as possible to the theoretically expected number in the first column.

### 4.2. Out-of-sample evidence

Up until now, we have only looked in-sample, essentially investigating the fit of the models to (extreme) data. In practice however, a risk manager is probably more interested in how well he can predict future extreme movements than in accurately modelling the
past. To assess the predictive performance of the different risk models treated in this paper, we use backtesting. We look both at the entire sample and at individual years. Particular emphasis is put on the period around the Asian financial crisis.

In applying backtesting, we reestimate the models each day, something that immediately reveals possible stability problems of a model. GARCH models are well known to behave well in maximum likelihood estimations and it is interesting to note that the same holds for the EVT models in this study. The EVT-based models are actually estimated faster than the GARCH models (in GAUSS using the BHHH algorithm) and they are also more parsimonious (they rely on fewer parameters). This might turn out to be important in practical applications of the models.

### 4.2.1. Backtesting

Starting with the 1000th observation (January 11, 1984 for AFF and December 14, 1983 for DOW) in our samples of 4960 returns, each day we reestimate the different models using the last 1000 days’ returns. Using each of the models, we produce (1-day) $\text{VaR}_p$ forecasts for the following day. These $\text{VaR}_p$ forecasts are then compared to the actual return that particular day, and the number of days when the actual loss is larger than the $\text{VaR}_p$ forecast is counted. We call this the number of exceedences and the results are presented in Table 6.

A look at the numbers in Table 6 reveals a general underestimation of risk, whatever model we use. This is not too surprising; however, one must remember that the evaluation period contains two “stress” periods with very high volatility, the October crash 1987 and the Asian financial crisis 1997–1998. When it comes to the relative model performance, the results from the in-sample study remain essentially unchanged. The general observation would be that for the $\text{VaR}_{0.95}$ measure, the EVT-based models and the traditional models produce equally good $\text{VaR}_p$ estimates while for higher $\text{VaR}_p$ measures ($\text{VaR}_{0.99}$ and above), the EVT-based models do a better job. Whether one uses conditional or

---

Table 6

Out-of-sample evaluation of forecasted 1-day VaR quantiles (number of exceedences)

<table>
<thead>
<tr>
<th>VaR$_p$</th>
<th>Expected</th>
<th>Normal</th>
<th>AR-GARCH</th>
<th>AR-GARCH-(t)</th>
<th>GEV</th>
<th>GPD</th>
<th>Cond. GEV</th>
<th>Cond. GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{VaR}_{0.95}$</td>
<td>198</td>
<td>184</td>
<td>198</td>
<td>145</td>
<td>211</td>
<td>215</td>
<td>255</td>
<td>228</td>
</tr>
<tr>
<td>$\text{VaR}_{0.99}$</td>
<td>39.6</td>
<td>77</td>
<td>77</td>
<td>29</td>
<td>44</td>
<td>48</td>
<td>45</td>
<td>43</td>
</tr>
<tr>
<td>$\text{VaR}_{0.99,5}$</td>
<td>19.8</td>
<td>65</td>
<td>45</td>
<td>17</td>
<td>23</td>
<td>30</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>$\text{VaR}_{0.99,9}$</td>
<td>3.96</td>
<td>42</td>
<td>23</td>
<td>8</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>DOW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{VaR}_{0.95}$</td>
<td>198</td>
<td>174</td>
<td>188</td>
<td>126</td>
<td>197</td>
<td>209</td>
<td>232</td>
<td>212</td>
</tr>
<tr>
<td>$\text{VaR}_{0.99}$</td>
<td>39.6</td>
<td>70</td>
<td>68</td>
<td>28</td>
<td>45</td>
<td>54</td>
<td>45</td>
<td>43</td>
</tr>
<tr>
<td>$\text{VaR}_{0.99,5}$</td>
<td>19.8</td>
<td>56</td>
<td>48</td>
<td>17</td>
<td>26</td>
<td>32</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>$\text{VaR}_{0.99,9}$</td>
<td>3.96</td>
<td>34</td>
<td>29</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

The numbers are empirically observed number of VaR exceedences. They should be as close as possible to the theoretically expected number in the first column.

---

16 The Basle Committee of Banking Supervision suggests backtesting to be performed on a yearly basis.
17 We have tried to (partly) avoid data snooping as well as to facilitate comparisons by simply using the same estimation window, 1000 days, as used by McNeil and Frey (2000). For the same reason, we follow McNeil and Frey in the choice of threshold, $u$, in the GPD estimations; we have chosen to keep the number of excess returns constant, equal to 100, over the entire sample. In the GEV estimation, we have chosen $n=25$ and $m=40$. 

unconditional models do not seem to be of great importance, however. As an alternative to the EVT models, one could also rely on a nonnormal AR-GARCH model; the AR-GARCH \( t \) model performs better than the normal AR-GARCH model, at least at more extreme VaR \( p \) levels. Finally, comparing the different EVT models, we see that the block maxima models and the POT model present very similar forecasts, particularly when prefiltered by AR-GARCH.

The number of exceedences obviously differ to some extent from the theoretical numbers but then one must remember that it is a very difficult task to estimate as high quantiles as we are trying to. For example, out of the 3960 returns in the evaluation sample, we expect as few as 4 returns to exceed the VaR\( 99\% \) forecast. In the light of this, the observed numbers are not that bad.

4.2.2. Backtesting year by year

To study when the different models underestimate VaR\( p \), we divide the evaluation period into sixteen 1-year long subperiods, and for each of these subperiods, we count the number of exceedences above the VaR\( 99\% \) measure.\(^{18}\) In this way, we can tell something about the stability of the earlier results. In evaluating the different VaR\( p \) forecasts in Table 7, it is important to remember that even if underprediction of VaR\( p \) measures is the most important sign of weak performance in a risk management context, overprediction is not good either; an accurate VaR\( 99\% \) measure is expected to be exceeded by exactly 1% of the returns over a certain (long enough) period, not more, not less.

Table 7
Number of exceedences of forecasted 1-day 99% VaR quantiles each year for AFF and DOW (within parenthesis)

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>AR-GARCH</th>
<th>AR-GARCH-( t )</th>
<th>GEV</th>
<th>GPD</th>
<th>Cond. GEV</th>
<th>Cond. GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1984</td>
<td>4 (0)</td>
<td>5 (2)</td>
<td>2 (0)</td>
<td>1 (0)</td>
<td>1 (0)</td>
<td>3 (2)</td>
<td>3 (2)</td>
</tr>
<tr>
<td>1985</td>
<td>2 (0)</td>
<td>3 (1)</td>
<td>2 (0)</td>
<td>1 (0)</td>
<td>1 (0)</td>
<td>2 (1)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>1986</td>
<td>8 (6)</td>
<td>4 (6)</td>
<td>2 (3)</td>
<td>3 (7)</td>
<td>5 (7)</td>
<td>3 (7)</td>
<td>3 (7)</td>
</tr>
<tr>
<td>1987</td>
<td>12 (17)</td>
<td>5 (6)</td>
<td>2 (4)</td>
<td>11 (12)</td>
<td>11 (15)</td>
<td>2 (5)</td>
<td>2 (5)</td>
</tr>
<tr>
<td>1988</td>
<td>2 (2)</td>
<td>7 (6)</td>
<td>3 (2)</td>
<td>2 (2)</td>
<td>1 (2)</td>
<td>3 (2)</td>
<td>3 (3)</td>
</tr>
<tr>
<td>1989</td>
<td>1 (1)</td>
<td>2 (3)</td>
<td>2 (1)</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>2 (1)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>1990</td>
<td>9 (1)</td>
<td>8 (5)</td>
<td>2 (1)</td>
<td>2 (0)</td>
<td>1 (0)</td>
<td>3 (2)</td>
<td>3 (0)</td>
</tr>
<tr>
<td>1991</td>
<td>1 (1)</td>
<td>6 (3)</td>
<td>4 (1)</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>2 (2)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>1992</td>
<td>8 (0)</td>
<td>4 (0)</td>
<td>2 (0)</td>
<td>2 (0)</td>
<td>5 (0)</td>
<td>2 (0)</td>
<td>2 (0)</td>
</tr>
<tr>
<td>1993</td>
<td>0 (1)</td>
<td>2 (2)</td>
<td>0 (2)</td>
<td>0 (0)</td>
<td>0 (1)</td>
<td>1 (2)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>1994</td>
<td>1 (3)</td>
<td>3 (6)</td>
<td>0 (1)</td>
<td>0 (1)</td>
<td>0 (1)</td>
<td>1 (4)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>1995</td>
<td>0 (2)</td>
<td>1 (2)</td>
<td>0 (2)</td>
<td>0 (1)</td>
<td>0 (2)</td>
<td>1 (2)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>1996</td>
<td>0 (7)</td>
<td>4 (6)</td>
<td>0 (5)</td>
<td>0 (5)</td>
<td>0 (5)</td>
<td>3 (5)</td>
<td>3 (5)</td>
</tr>
<tr>
<td>1997</td>
<td>9 (16)</td>
<td>8 (10)</td>
<td>4 (3)</td>
<td>7 (9)</td>
<td>8 (11)</td>
<td>8 (5)</td>
<td>8 (5)</td>
</tr>
<tr>
<td>1998</td>
<td>20 (11)</td>
<td>12 (7)</td>
<td>4 (3)</td>
<td>13 (6)</td>
<td>13 (8)</td>
<td>9 (4)</td>
<td>7 (4)</td>
</tr>
<tr>
<td>1999</td>
<td>0 (2)</td>
<td>3 (3)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (1)</td>
<td>0 (1)</td>
</tr>
</tbody>
</table>

The expected number of yearly exceedences is 2.5.

\(^{18}\) The results for the other quantiles are similar to those of the VaR\( 99\% \) quantile, and the choice of the 99% quantile is due to the weak performance of the traditional models at higher measures and the difficulties in evaluating the most extreme tail quantiles. The 99% measure is also widely used in practice (it is, for instance, the level suggested by The Basle Committee of Banking Supervision).
Looking at Table 7, we can see how all forecasting models underestimate VaR during the Asian crisis 1997–1998. Some models underestimate much more than others though, and the conditional models generally capture more of the current market conditions than the unconditional models. The latter (not surprisingly) seriously underestimate the risk in stress periods, and an example of this is the difference between the unconditional and the conditional block maxima methods during the turbulent 1987; in this year, we have 11 (12) exceedences of the VaR99% forecast for AFF (DOW) using the unconditional model and only 2 (5) exceedences using the conditional model.

Comparing the results for AFF and DOW (whose numbers are within parentheses), we can see how the relative performance of the different models essentially is the same for the two indices. Generally, however, there are some interesting differences between the U.S. and the Swedish stock indices. These differences are related to different macroeconomic situations in the two countries over the investigated time period. As we saw earlier, DOW was more turbulent than AFF during the October crash 1987, and all models generally underestimate more for the U.S. index than for the Swedish index for that particular year. In 1998, the situation is reversed; during the Asian (and Russian) crisis 1998, the Swedish index was more volatile than the U.S. index. This is also the reason for the larger number of underpredictions for AFF than for DOW in 1998. Finally, the Swedish currency crisis 1992 is also shining through; while several models underpredict the risk for AFF for the year 1992, no model gives even a single exceedence for DOW for that year.

To summarize, the division of the evaluation period into subperiods reveals a difference between the conditional and the unconditional models that we did not observe when the whole evaluation period was used. The performance of the AR-GARCH filtered models is more stable over time and in that sense, they must be considered superior to the unconditional ones when we deal with day-by-day forecasts of VaRp.

4.2.3. Backtesting, extremes, and the Asian financial crisis

To further assess the improvement in performance from using conditional models when forecasting VaRp measures, we look more carefully at the period 1996 to 1998 that covers the Asian financial crisis 1997 and 1998. In Fig. 1, we plot the negative returns of the DOW together with VaR99% forecasts, and we observe a quick reaction of the conditional VaRp forecasts to a changing market volatility.19 The unconditional models produce VaRp forecasts that react much slower to changing market conditions.

At first, one might draw the conclusion from studying Table 7 and Fig. 1 that it is solely in stress situations like the Asian crisis 1997 and 1998 that the use of conditional models is warranted. However, even if conditional models are particularly important in these situations, they also improve the accuracy of risk estimates in tranquil periods. In Fig. 1, we have plotted conditional VaRp forecasts that not only increase with increasing volatility but also decrease with decreasing volatility, and overall, these conditional VaRp measures correspond more closely to the actual returns than the unconditional measures. The important observation to make is that while unconditional models (using the 1000 last

19 The two conditional (as well as the unconditional) EVT-based VaR forecasts are essentially identical to each other and cannot be separated in Fig. 1. Most of the time, the conditional forecasts lie between the normal-distribution-based and the t-distribution-based conditional forecasts. The picture for the AFF is very similar.
observations) underestimate risks during, for instance, 1997 and 1998, they instead tend to be too conservative during more tranquil periods.

Even if the conditional models produce VaR forecasts that react to changing market conditions, not even these are able to capture the most extreme losses; they are simply too slow to react to these sudden and unexpected extreme returns. This is the reason behind the general weak performance of the models during, for instance, the Asian crisis; even if conditioning on current market conditions reduces the number of exceedences significantly, there is no model that does not underestimate the risk in this period.

To study the importance of the most extreme returns in the evaluation of the different models, we now focus only on the 50 most extreme returns over the sample (approximately 1% of the sample). Not surprisingly, a large part of these returns appear during the Asian crisis 1997 and 1998; in the Swedish market we have 15 extreme returns (out of 50) in those 2 years, and in the U.S. market, we have 11 extreme returns in the same period. For the unconditional models, most of the VaR exceedences of course come from these extreme returns, and the interesting question is whether they are also responsible for the exceedences of the conditional models’ VaR forecasts.

In Table 8, we present the percentage of the 50 most extreme returns as well as the percentage of all out-of-sample returns that actually exceed the forecasted VaR quantiles. Comparing the normal unconditional model with the normal AR-GARCH conditional model, we can see that both capture around 98% (instead of the expected 99%) of the total number of returns. However, focusing on the extreme returns, more than a third of these returns are captured by the AR-GARCH VaR quantiles while only a few extreme returns are captured by the normal unconditional model.

Fig. 1. DOW; negative returns and VaR forecasts.

20 We focus only on the commonly used VaR quantile. As expected, for the most extreme VaR quantiles, essentially all the exceedences are due to the extreme returns.
returns are captured by the unconditional model. The AR-GARCH-\(t\) model’s conservative VaR\(_p\) forecasts are of course capturing an even larger share of both the extreme returns and the rest of the returns than the normal AR-GARCH model.

Similarly, comparing the unconditional and the conditional EVT models, we observe that all EVT models capture close to 99% of the returns. However, while the conditional EVT models capture close to half of the 50 most extreme losses, the unconditional models only capture about a quarter of the extremes.

In other words, for the conditional models, not all the most extreme returns contribute to the number of exceedences; a large share of the extreme losses are actually predicted by the models and the good performance of the two conditional EVT models is evidently partly associated with the models’ ability of capturing also the most extreme losses and to spread out the exceedences over the entire sample. Just as the volatility pattern is captured by the conditional and not by the unconditional models in Table 7, the most extreme returns are captured essentially only by the conditional models in Table 8.

5. Conclusions

The purpose of this paper has been to evaluate how well EVT-based models perform in modelling the tails of distributions and in estimating and forecasting VaR\(_p\) measures. We have found that the combination of either the block maxima method or the POT method with traditional time series modelling, into what we call conditional EVT, does the best job both in tranquil and volatile periods.

Two different stock indices, the Swedish AFF, and the DOW, have been investigated, and some differences between the indices have been pointed at. While a study of the in-sample performance of the different models tells us that EVT-based risk models outperform both the naive unconditional normal model and more realistic conditional AR-GARCH models, the real test of a risk management model is of course in forecasting risks. In an out-of-sample evaluation, we have therefore tried to evaluate the practical use of including EVT in both unconditional and conditional VaR\(_p\) forecasts. An overall

<table>
<thead>
<tr>
<th>Table 8</th>
<th>Percentage of the 50 most extreme losses and of all losses that exceed 1-day 99% VaR quantiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
</tr>
<tr>
<td>AFF</td>
<td>Percentage of extreme returns that exceed</td>
</tr>
<tr>
<td></td>
<td>Percentage of all returns that exceed</td>
</tr>
<tr>
<td>DOW</td>
<td>Percentage of extreme returns that exceed</td>
</tr>
<tr>
<td></td>
<td>Percentage of all returns that exceed</td>
</tr>
</tbody>
</table>
conclusion would be that for the VaR$_{95\%}$ measure, there is no great difference in performance between EVT-based models and traditional models while for higher VaR$_p$ levels (99% and above), the EVT-based models do a better job. The results for the Swedish and the U.S. indices are overall very similar.

The next question is whether conditional EVT models improve upon traditional unconditional EVT models. When we look at the average performance of the models over the entire sample using out-of-sample backtesting methods, conditional EVT first seems less of an improvement upon unconditional EVT. However, when we perform a more realistic year-by-year backtesting (as suggested by the Basle Committee of Banking Supervision), conditional EVT appears to be the best approach. This is one of the major results in the paper. Although extreme market conditions occur at slightly different years in the United States and in Sweden, it is clear that for either of the indices, only the conditional EVT models succeed in capturing the behavior of the indices during volatile years. At the same time, the conditional EVT models are also producing the most accurate (neither too conservative nor too low) VaR$_p$ forecasts in tranquil periods. In addition to being able to capture the swings of volatility, the conditional EVT models also capture close to half of the (1%) most extreme returns with its VaR$_{99\%}$ forecasts.

To sum up, the best VaR$_p$ forecasts are produced by AR-GARCH filtered EVT-based models and we would advocate the use of EVT when managing market risks in the equity market, particularly more extreme risks. The EVT techniques are not difficult to implement, and compared to ordinary time series models like GARCH, they are both faster and easier to estimate as well as more efficient in its use of parameters. In addition, the EVT machinery is by no means limited to 1-day VaR$_p$ calculations but it can also easily be applied to multiperiod forecasting as well as to pricing, hedging, and risk management of derivatives. A drawback with EVT is that the theory is based on asymptotic arguments and that in particular, the block maxima method requires quite long histories for estimation. Still, the block maxima method is found, in this paper, to be at least as good as the POT method in estimating and forecasting both unconditional and conditional VaR.

Acknowledgements

The author is particularly grateful for helpful comments received from participants at the 8th APFA Annual Conference in Bangkok, Thailand. Financial support from Jan Wallanders och Tom Hedelius Stiftelse and Crafoordska stiftelsen is also gratefully acknowledged.

References


