Realizing the extremes: Estimation of tail-risk measures from a high-frequency perspective

Marco Bee[*], Debbie J. Dupuis[†] and Luca Trapin[‡]

May 17, 2015

Revised November 29, 2015

Abstract

This article applies realized volatility forecasting to Extreme Value Theory (EVT). We propose a two-step approach where returns are first pre-whitened with a high-frequency based volatility model, and then an EVT based model is fitted to the tails of the standardized residuals. This realized EVT approach is compared to the conditional EVT of McNeil & Frey (2000). We assess both approaches' ability to filter the dependence in the extremes and to produce stable out-of-sample VaR and ES estimates for one-day and ten-day time horizons. The main finding is that GARCH-type models perform well in filtering the dependence, while the realized EVT approach seems preferable in forecasting, especially at longer time horizons.

Keywords: Realized volatility; High-frequency data; Extreme Value Theory; Value-at-Risk; Expected Shortfall.

JEL codes: C4, C5, G1.

[*]Department of Economics and Management, University of Trento, via Inama, 5 - 38122 Trento, Italy Phone: +39-0461-282296, Fax: +39-0461-282241 E-mail: marco.bee@unitn.it

[†]Department of Decision Sciences, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal (Québec) H3T 2A7 Telephone: 514-340-6952 Fax: 514-340-5634 E-mail: debbie.dupuis@hec.ca

[‡]IMT Institute for Advanced Studies Lucca E-mail: luca.trapin@imtlucca.it

1. INTRODUCTION

Accurate assessment of the tail behavior of asset returns is of the utmost importance for financial market practitioners and regulators. Extreme Value Theory (EVT) is very useful as it provides probabilistic results which characterize the tail behavior of any distribution, without requiring knowledge of the main body of the distribution.

McNeil & Frey (2000) develop a two-step procedure to model the tails of the conditional returns distribution with EVT: first, the returns are pre-whitened with a GARCH-type model which explicitly accounts for the heteroskedasticity; then the tails of the standardized residuals from the GARCH model are fitted using the Peaks-over-Threshold (POT) method (Davison & Smith 1990). McNeil & Frey (2000) backtest this approach on different time series and provide evidence that it outperforms both the unconditional EVT model (Danielsson & de Vries 1997) and the GARCH models with normal and Student's t distributions. This two-step conditional EVT (C-EVT) approach is now considered standard in the financial community.

In a large simulation experiment, Jalal & Rockinger (2008) study the performance of C-EVT under different hypotheses regarding the underlying data generating process (DGP). They conclude that C-EVT performs fairly well in terms of one-day-ahead predictions of the conditional quantiles under misspecification of the conditional mean and variance dynamics.

In this paper, we develop a *realized* EVT (RV-EVT) approach which exploits high-frequency information to pre-whiten the returns in the first step, and uses the standardized residuals of the high-frequency based model in the second step. Recent work on realized volatility has emphasized how the use of high-frequency information can enhance the forecast of the conditional variance of the returns (Shephard & Sheppard 2010). We propose a class of high-frequency based volatility models that combines reduced form models for the realized volatility (Corsi 2009) with a *link function* relating the conditional return volatility with the prediction of the realized volatility. We consider three different *link functions* of increasing complexity and six reduced form models with both symmetric and asymmetric structures.

It is important to filter out the dependence in the first step before applying the POT approach in the second step. We investigate whether a high-frequency based volatility model produces standardized residuals closer to the ideal iid than those obtained under a GARCH-type model. We compare the degree of extremal dependence left in the standardized residuals of our models and the GARCH model for 17 time series of international stock indices from 2000 to 2014.

We then add to the simulation experiment of Jalal & Rockinger (2008) by examining the out-of-sample C-EVT forecasts of both Value-at-Risk (VaR) and Expected Shortfall (ES) for a 10-day horizon[1]. This period is relevant from the regulatory perspective as the risk capital of a bank must be sufficient to cover losses on the bank's trading portfolio over a 10-day holding period. We also consider an additional DGP where observations are generated according to the parametric model of Bandi & Renò (2015). The latter not only accommodates several stylized facts of the asset returns, but also allows us to draw intra-day observations and produce forecasts of the risk measures with the RV-EVT approach.

Finally, we compare C-EVT and RV-EVT for forecasting VaR and ES on the 17 international indices time series. The backtesting exercise is fully out-of-sample, with a training sample for the models (in-sample) of two different sizes, respectively 2000 and 500 observations. This part of the work is close to that of Giot & Laurent (2004), Clements, Galvão & Kim (2008), and Brownlees & Gallo (2009), in the sense that we assess the merit of using high-frequency data, but we do so within the context of EVT approaches.

The remainder of the paper is organized as follows: in Section 2, we present the C-EVT of McNeil & Frey (2000); in Section 3 we introduce the RV-EVT; in Section 4 we compare the two approaches, looking separately at the filtering and forecasting components; in Section 5 we perform robustness checks aimed at consolidating the evidence from the main analysis; in Section 6 we give concluding remarks. Some technical details appear in the appendix and

[1] Note that with ten-day-ahead risk measures, we mean the risk measures estimated on the conditional distribution of the sum of the next ten-day returns.

further data analyses can be found in the supplementary material.

2. THE CONDITIONAL EVT APPROACH

Let p_t be the logarithmic price at time t and define the conditional log-returns r_t as,

$$\begin{aligned} p_t - p_{t-1} = r_t &= \mu_t + \sigma_t \epsilon_t, \\ \mu_t &= f(\mathcal{F}_{t-1}), \\ \sigma_t^2 &= h(\mathcal{H}_{t-1}), \end{aligned} \tag{1}$$

where μ_t and σ_t^2, are respectively the conditional mean and variance, functions of the information sets $\mathcal{F}_{t-1}$ and $\mathcal{H}_{t-1}$, and ϵ_t an iid process with zero mean and unit variance. A notable amount of empirical research in financial markets shows time-variation and heaviness in the tails of the conditional returns distribution. To account for this evidence, Bollerslev (1986) proposes to model σ_t^2 as a function of its past values and the past values of ϵ_t, assuming ϵ_t to be normally distributed. This model is commonly referred to as GARCH.

McNeil & Frey (2000) propose to pre-whiten the returns using a standard GARCH and then model the tails of the estimated residuals by means of EVT. The theoretical justification is that Gaussian Quasi-Maximum Likelihood (QML) estimation of a GARCH model is consistent as long as $\mathbb{E}(\epsilon_t^4) < \infty$. Formally, they consider an AR(1)-GARCH(1,1),

$$r_t = \mu + \phi_1 r_{t-1} + \sigma_t \epsilon_t \tag{2}$$

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{3}$$

where $\epsilon_t \sim F$ with zero mean and unit variance, and $\alpha_1 + \beta_1 < 1$ to guarantee stationarity. Suppose[2] that F has upper endpoint $v_F := \sup \epsilon_t : F(\epsilon_t) < 1$. Given a high threshold u, $u < v_F$, Pickands (1975) shows that when $u \to v_F$, the distribution of the excesses $(\epsilon_t - u)_+$ converges to a Generalized Pareto (GP) distribution G with shape parameter ξ and scale

[2] The following argument equally applies to ϵ_t and the negated series $-\epsilon_t$. Throughout we will typically refer to the latter term or its distribution function that we call *loss distribution*.

parameter $\nu > 0$. That is, $\Pr(\epsilon_t - u \leq x | \epsilon_t > u)$ goes to

$$G(x;\xi,\nu) = \begin{cases} 1 - \{1 + \xi x/\nu\}^{-\frac{1}{\xi}} & \text{for} \quad \xi \neq 0 \\ 1 - \exp\{-x/\nu\} & \text{for} \quad \xi = 0 \end{cases} \tag{4}$$

as $u \to \upsilon_F$. When $\xi > 0$, F has Pareto-type upper tail with tail index $1/\xi$.

A model for the tail of the residuals $\widehat{\epsilon}$ can be defined following the POT method. The tail estimator of F is then

$$\widehat{\bar{F}}(\widehat{\epsilon}) = \frac{N_u}{T}\left(1 + \widehat{\xi}\frac{\widehat{\epsilon} - \widehat{u}}{\widehat{\nu}}\right)^{-\frac{1}{\widehat{\xi}}}, \tag{5}$$

where $\bar{F} = 1 - F$, $\widehat{u}$ is an appropriately chosen threshold, $\widehat{\xi}$ and $\widehat{\nu}$ are maximum likelihood (ML) estimates, and N_u is the number of observations exceeding the threshold $\widehat{u}$.

For $\alpha > 1 - N_u/T$, it is possible to use Equation 5 to obtain one-day-ahead prediction of the VaR and ES at level α,

$$\widehat{VaR}^{\alpha}_{t,t+1} = \widehat{\mu}_{t,t+1} + \widehat{\sigma}_{t,t+1}\left(\widehat{u} + \frac{\widehat{\nu}}{\widehat{\xi}}\left(\left(\frac{1-\alpha}{N_u/S}\right)^{-\widehat{\xi}} - 1\right)\right), \tag{6}$$

$$\widehat{ES}^{\alpha}_{t,t+1} = \widehat{\mu}_{t,t+1} + \widehat{\sigma}_{t,t+1}\widehat{\epsilon}_{\alpha}\left(\frac{1}{1+\widehat{\xi}} + \frac{\widehat{\nu} - \widehat{\xi}\widehat{u}}{(1-\widehat{\xi})\widehat{\epsilon}_{\alpha}}\right), \tag{7}$$

where $\widehat{\mu}_{t,t+1}$ and $\widehat{\sigma}_{t,t+1}$ are the forecast of the conditional mean and the conditional variance, and $\widehat{\epsilon}_{\alpha}$ is the $(1-\alpha)$quantile of the residuals at time t.

Financial decisions are often predicated on accurate multi-period-ahead forecasts of the risk measures, even though the risk management literature surprisingly focuses almost exclusively on the accuracy of one-period-ahead forecasts. The dominant long-horizon forecasting approach consists of scaling the one-period-ahead forecasts by $\sqrt{k}$ where k is the horizon of interest. Its popularity among practitioners stems mostly from its use in RiskMetrics. Alternatively, we follow the iterative approach of McNeil & Frey (2000) to obtain the regulatory relevant ten-day-ahead VaR and ES forecasts, as they were shown to outperform the $\sqrt{k}$ rule. We fit a GP distribution to both tails of the GARCH residuals and model the body by means of the empirical distribution function. Then, combining bootstrapping and GP simulation it is possible to obtain a semi-parametric estimate of the innovation distribution:

$$\widehat{F_\epsilon(\epsilon)} = \begin{cases} \frac{N_u^L}{S}\left(1+\widehat{\xi}^L \frac{|\epsilon-\widehat{u}^L|}{\nu^L}\right)^{-1/\widehat{\xi}^L} & \epsilon < \widehat{u}^L, \\ \frac{1}{S}\sum_{i=1}^{S} 1_{\widehat{\epsilon}_i \leq \epsilon} & \widehat{u}^L \leq \epsilon \leq \widehat{u}^H, \\ 1-\frac{N_u^H}{S}\left(1+\widehat{\xi}^H \frac{|\epsilon-\widehat{u}^H|}{\nu^H}\right)^{-1/\widehat{\xi}^H} & \epsilon > \widehat{u}^H, \end{cases} \tag{8}$$

where $\widehat{\epsilon}_i$ is the i-th residual, S is the sample size, N_u^k is the number of exceedances, $\widehat{\xi}^k$ and $\widehat{\nu}^k$ are GP parameter estimates, $\widehat{u}^k$ is a high threshold and $k = L, H$ denotes respectively the lower and upper tails. Through an iterative algorithm which consists of drawing observations from this distribution, see Appendix A, and updating the GARCH equation, it is possible to simulate several sample paths for the next ten days and obtain an estimate of the multi-period distribution. Throughout the analysis we generate 1000 sample paths. The ten-day-ahead VaR is obtained from the inversion of Equation 5 and ES as conditional expectation of exceedances over VaR.

3. THE REALIZED EVT APPROACH

Our proposal is to incorporate high frequency (intra-daily) information in the sets $\mathcal{F}_t$ and $\mathcal{H}_t$ of Equation (1), and use the POT approach to model the tails of the residuals. To the extent that high-frequency based volatility models provide better forecasts, they should produce residuals with a lower degree of extremal dependence and allow for a better estimation of the tail. As the POT approach is based on the assumption of iid observations, residuals $\widehat{\epsilon}_t$ that better proxy the innovations ϵ_t improve the inference on the GP parameters and the estimation of the probability of exceedances. At the same time, improved volatility predictions provide better forecasts of the conditional density and more accurate estimates of tail-related risk measures.

Assuming μ_t to be constant in Equation (1) (its effect is negligible), we specify a class of high-frequency based volatility models for the latent variable σ_t^2, using as a proxy the realized volatility (RV). The latter is a non-parametric estimator of the variation of the price path of an asset obtained as the sum of squares of intra-day returns recorded during the open hours of the stock exchange. Formally, let p_t denote the log-price of an asset at time t and $r_{t,\Delta} = p_t - p_{t-\Delta}$ be the discretely sampled Δ-period return. The RV on day t is then defined as

$$RV_t = \sum_{j=1}^{N} r^2_{t-1+(j\cdot\Delta)}. \tag{9}$$

If market microstructure noise is absent then, as $\Delta \to 0$, RV_t consistently estimates the quadratic variation of the price process on day t (Andersen, Bollerslev, Diebold & Labys 2001; Barndorff-Nielsen 2002). In practice, market microstructure noise plays an important role, and econometricians usually resort to 1- to 5-minute return data to mitigate the effect of the noise.

Let $\mathbb{E}(RV_t|\mathcal{H}_{t-1})$ be the conditional expectation of the RV given the information set $\mathcal{H}_{t-1}$ and assume that σ^2_t is defined as a function of $\mathbb{E}(RV_t|\mathcal{H}_{t-1})$, $\sigma^2_t = g(\mathbb{E}(RV_t|\mathcal{H}_{t-1}))$. We need to specify a model for the conditional expectation of the RV and a link function g that maps the conditional mean of the RV to the conditional variance of the returns. In what follows, we discuss these two issues separately and then show how to obtain forecasts of the risk measures.

3.1 The link function

Different approaches have been attempted to provide a connection between the conditional variance of the returns and RV but, to the best of our knowledge, a comparative study of these link functions has not been pursued so far. In this paper, we consider and discuss three different link functions, those of Giot & Laurent (2004), Clements et al. (2008) and Brownlees & Gallo (2009), and compare their performances.

The simplest approach is the one used in Clements et al. (2008), where the conditional variance of returns is defined by

$$\widehat{\sigma}^2_{t,t+h} = \exp\Big(\widehat{\log(RV_{t,t+h})}\Big). \tag{10}$$

We call this link function **type-I**. This specification bears two different sources of bias: one related to the log-transformation and the other one due to the imprecision of the volatility proxy. Specifically, given that returns are measured close-to-close while intra-daily observations neglect overnight information, this information mismatch may be a source of bias (Andersen, Bollerslev & Huang 2011). To circumvent this problem, Brownlees & Gallo (2009) assume the conditional variance σ_t^2 to be an affine function of the RV. This model is nested in the HEAVY class of Shephard & Sheppard (2010). Our logarithmic specification of the HAR function prohibits us from applying the idea directly, but we adapt it and call it **type-II**. We have

$$\widehat{\sigma}^2_{t,t+h} = c + d \exp\left(\widehat{\log(RV_{t,t+h})} \right), \tag{11}$$

where c and d are coefficients to be estimated. Finally, similarly to Giot & Laurent (2004), we incorporate a correction for the logarithmic transformation and call it **type-III**. We have

$$\widehat{\sigma}^2_{t,t+h} = c + d \exp\left(\widehat{\log(RV_{t,t+h})} + 0.5\widehat{\sigma}^2_{\widehat{\eta}} \right), \tag{12}$$

where $\widehat{\sigma}^2_{\widehat{\eta}}$ is the estimated variance of the logarithmic HAR regression residuals, see Sect. 3.2.

A word of caution is in order regarding models (10), (11) and (12), because they are reduced-form models that are misspecified in the presence of jumps and/or stochastic volatility, and this may have severe consequences on the estimation procedures (Nelson 1991; Tolver & Lange 2010).

3.2 Reduced form models

Several approaches have been proposed in the financial econometrics literature to model the dynamics of $\mathbb{E}(RV_t|\mathcal{H}_{t-1})$, see Andersen, Bollerslev, Diebold & Labys (2003), Bollerslev, Kretschmer, Pigorsch & Tauchen (2009), and Andersen et al. (2011). In this paper, we rely on the heterogeneous autoregressive (HAR) class of models, initially proposed by Corsi (2009), and now considered as the standard approach in the high-frequency literature. In particular, we use the logarithmic specification of these models for two specific reasons: constraints on the parameters to guarantee positive volatility are not necessary; the estimated residuals of the logarithmic specification are closer to normality, hence more amenable to standard time series procedures.

Let the multi-period normalized RV be denoted by

$$RV_{t,t+h} = h^{-1}[RV_{t+1} + \cdots + RV_{t+h}]. \tag{13}$$

Note that $RV_{t,t+1} \equiv RV_{t+1}$. The daily HAR model of Corsi (2009) can then be expressed as

$$\log(RV_{t,t+h}) = \beta_0 + \beta_D \log(RV_t) + \beta_W \log(RV_{t-5,t}) + \beta_M \log(RV_{t-22,t}) + \eta_{t,t+h}, \tag{14}$$

where $t = 1, \ldots, T$. This parsimonious specification exploits the information from the past 1-day, 5-day and 22-day average RV, reflecting the idea that heterogeneity in investor behavior creates different volatility components having different impacts on the future volatility.

The intuition that the disentanglement of the continuous and discontinuous sample paths could be a valuable source of information leads Andersen, Bollerslev & Diebold (2007) to extend the HAR model to include a jump component. This is important also in the light of the findings of Todorov & Tauchen (2011) and Bandi & Renò (2015) that volatility jumps are relevant. Adding the realized jump measure J_t, obtained as the difference between the realized variance and the bipower variation (Barndorff-Nielsen & Shephard 2004), as an explanatory variable, one obtains their HAR-J model

$$\log(RV_{t,t+h}) = \beta_0 + \beta_D \log(RV_t) + \beta_W \log(RV_{t-5,t}) + \beta_M \log(RV_{t-22,t}) + \beta_J \log(J_t + 1) + \eta_{t,t+h}. \tag{15}$$

According to Andersen et al. (2007) and Corsi, Pirino & Renò (2010), the jump component adds a significant contribution in terms of R^2.

The models in Equations (14)-(15) are both symmetric in the sense that they do not have a parameter accounting for the so called leverage effect, i.e. the generally negative correlation between an asset's returns and its changes of volatility (Nelson 1991). To generate the desired leverage effect, a first possibility is to use the signed jump variation obtained from the realized semi-variance (Patton & Sheppard 2015). By isolating the sign of the jump, we can differentiate its impact on RV. Letting RS_t^+ and RS_t^- be the positive and negative realized semi-variance respectively, the HAR-SJ model has specification

$$\begin{aligned}\log\left(RV_{t,t+h}\right) = \quad & \beta_0 + \beta_{J+}\log(\Delta J_t^{2+}+1) + \beta_{J-}\log(\Delta J_t^{2-}+1) \\ & +\beta_D\log(RV_t) + \beta_W\log(RV_{t-5,t}) + \beta_M\log(RV_{t-22,t}) + \eta_{t,t+h},\end{aligned} \tag{16}$$

with signed jump variations

$$\begin{aligned}\Delta J_t^{2+} &= (RS_t^+ - RS_t^-)I_{\{(RS_t^+-RS_t^-)>0\}},\\ \Delta J_t^{2-} &= (RS_t^+ - RS_t^-)I_{\{(RS_t^+-RS_t^-)<0\}}.\end{aligned}$$

Assuming that jumps are rare, these two quantities broadly capture the variation of positive and negative jumps. Nonetheless, if jumps of different signs occur on the same day, then these measures account only for a small part of the total jump variation. In this case, the HAR-SJ can miss valuable information coming from the actual size of the jumps. To overcome this issue, we propose a simple extension of this model and refer to it as the HAR-SJ*aug* model, see Appendix B for a discussion.

Another asymmetric model is the LHAR of Corsi & Renò (2012) that adds to the simple HAR specification three leverage components capturing the persistence of the leverage effect. In contrast to Patton & Sheppard (2015), they use the lagged negative returns to generate the desired effect,

$$\begin{aligned}\log(RV_{t,t+h}) \quad &= \beta_0 + \beta_D\log(RV_t) + \beta_W\log(RV_{t-5,t}) + \beta_M\log(RV_{t-22,t}) \\ & +\gamma_d r_t^- + +\gamma_w r_{t-5,t}^- + \gamma_m r_{t-22,t}^- + \eta_{t,t+h}\end{aligned} \tag{17}$$

where $r_{t-h,t} = \frac{1}{h}\sum_{j=1}^{h} r_{t-j}$, and $r_{t-h,t}^- = \min\left(r_{t-h,t}, 0\right)$.

3.3 Forecasting risk measures

Once we have a high-frequency based model for the conditional variance, we can filter the returns and model the tails of the residuals with the estimator in Equation (5). Whereas to obtain one-day-ahead predictions for C-EVT we rely on a closed-form approach (see Equations (6)-(7)), to obtain multiple-day-ahead predictions we have to use an iterative approach based on Equation (8) since the multiple-day conditional distribution of a daily GARCH is not available. We can proceed differently for RV-EVT. Considering that it is standard in the financial econometrics literature to fit the HAR model directly to the multi-period RV, we can compute explicitly the forecasts at both horizons for the RV-EVT. The h-day-ahead predictions of the VaR and ES at level α are then defined as

$$\widehat{VaR}^{\alpha}_{t,t+h} = \widehat{\mu}_{t,t+h} + \widehat{\sigma}_{t,t+h}\left(\widehat{u} + \frac{\widehat{\nu}}{\widehat{\xi}}\left(\left(\frac{1-\alpha}{N_u/T}\right)^{-\widehat{\xi}} - 1\right)\right) \tag{18}$$

$$\widehat{ES}^{\alpha}_{t,t+h} = \widehat{\mu}_{t,t+h} + \widehat{\sigma}_{t,t+h}\widehat{\epsilon}_{\alpha}\left(\frac{1}{1+\widehat{\xi}} + \frac{\widehat{\nu} - \widehat{\xi}\widehat{u}}{(1-\widehat{\xi})\widehat{\epsilon}_{\alpha}}\right) \tag{19}$$

where $\widehat{\mu}_{t,t+h}$ and $\widehat{\sigma}_{t,t+h}$ are respectively the h-period forecasts of the conditional mean and variance, and $\widehat{\epsilon}_{\alpha}$ is the $(1-\alpha)$ quantile of the residuals at time $t-1$.

4. CONDITIONAL EVT VS. REALIZED EVT

4.1 Data

The empirical analysis is based on the Oxford-Man Institute "Realised Library" version 0.2, (Heber, Lunde, Shephard & Sheppard 2009). We consider 17 different stock indices from the beginning of 2000 to the end of 2014, see Table 1. For each asset, the library currently records the daily returns and several daily realized measures. Among the latter, we consider the 5-min RV and use the 5-min Bipower Variation (BV) and the 5-min RSV to extract the jump components. The 5-min sampling frequency is standard in the literature as it mitigates the microstructure noise and assures good performance of the estimators (Liu, Patton & Sheppard 2012). To assess whether the microstructure noise is still a concern, we repeat the analysis with the *Second Best* estimator of Zhang, Mykland & Aït-Sahalia (2005) and the Realized Kernel of Barndorff-Nielsen, Hansen, Lunde & Shephard (2008) instead of the RV estimator, see Section 5.2.

Table 1: **Data description**. Time series of indices start January 2, 2000 and end December 31, 2014. Stock exchanges respect different holidays and the number of observations T subsequently differs.

Asset	Abbr.	T	Asset	Abbr.	T
Amsterdam Exchange Index	AEX	3816	IBEX35	IBX	3782
All Ordinaries Index	AOI	3743	IPC Mexico	IPC	3748
Bovespa Index	BVP	3664	Korea Composite Index	KCI	3690
CAC40	CAC	3817	Nasdaq 100	NSQ	3747
DAX30	DAX	3795	Nikkei 225	NK	3630
Dow Jones Industrial	DJ	3746	Russel 2000 Index	RUS	3745
Euro Stoxx 50	ESX	3794	SP500	SPX	3744
FTSE MIB	MIB	3778	Swiss Market Index	SMI	3749
FTSE100	FT	3764			

4.2 Methodology

Given the long series in Table 1 and the out-of-sample nature of our investigation, a rolling-window approach is used to obtain a time series of performance measures to compare the C-EVT and the RV-EVT. In particular, we use a fixed rolling-window length of $S = 2000$ days to train our models (in-sample) and daily updates.

We consider the exceedances over the threshold $\widehat{u}$ corresponding to the 95th quantile of the negated residuals $-\widehat{\epsilon}_t$. We perform Anderson-Darling and Cramér-von Mises goodness-of-fit tests at a level of significance of 0.05 to check the validity of the GP convergence (Choulakian & Stephens 2001). Recall from Pickands (1975) that convergence of the distribution of the excesses to the GP distribution occurs when the threshold u goes to the right endpoint of the support. Selecting a high empirical quantile as a threshold, we are using the results pre-asymptotically as an approximation only. One can thus expect the quality of the approximation to differ across the stock indices. When it is poor according to the goodness-of-fit tests, we raise the threshold u to the 98th quantile. These cases are flagged with the * symbol in the figures and the tables.

We assume the conditional mean of all models to be constant and equal to zero as in Brownlees & Gallo (2009). In the C-EVT, we use a GJR-GARCH(1,1) specification instead of the GARCH(1,1) to account for the evidence of a leverage effect in the financial returns.

We compare our EVT-based methods along two separate but related dimensions: the **filtering** and **forecasting** components. On the one hand, we assess how much of the extremal dependence inherent in the returns is captured by the different models. To this end, we rely on the concept of extremal index θ (Leadbetter 1983) which is a parameter measuring this dependence. As for forecasts, we produce one-day- and ten-day-ahead predictions of the VaR and ES at level $\alpha = 0.01$ and use standard performance measures to evaluate them, see Christoffersen (1998) and Engle & Manganelli (2004) for the VaR and McNeil & Frey (2000) for the ES.

4.3 The filtering component

Financial extremes tend to occur in clusters, and the degree of dependence in the extremes of a stationary series is measured by the extremal index θ. In particular, when $\theta < 1$ the extremes are dependent, conversely with $\theta = 1$ they are independent. The first step of C-EVT or RV-EVT should filter out this dependence so that the residuals used in the respective second steps are iid. If they are not, both the inference on the GP parameters and the probability of threshold exceedances will be affected. Given our finite sample size, we consider the standardized residuals of the models to be close to independent when $0.85 \leq \widehat{\theta} \leq 1$, where $\widehat{\theta}$ is the Ferro & Segers (2003) intervals estimate.

In what follows, we compare the performance of the GJR-GARCH(1,1) and the HAR models of Section 3.2 as filters. In particular, we estimate the extremal index $\widehat{\theta}$ of the model residuals obtained over each window. We repeat the analysis with the different link functions in Equations (10)-(12) to assess the merit of the complexity added to the high-frequency based model for the conditional volatility.

Figure 1 displays the estimates of θ for the type-I link function. The figure clearly shows that, in general, the residuals of each model exhibit a very low degree of extremal dependence and that the GJR-GARCH(1,1) model appears as the best performer. On a few occasions, the high-frequency based volatility models fail to capture the extremal dependence, particularly on the RUSSEL time series. Adding jumps and leverage does not always contribute positively.

In Figures 2-3, we report the estimates of θ respectively for the type-II and type-III link functions. The figures suggest that there is no real benefit from increasing the complexity of the conditional volatility model. The degree of extremal dependence left in the residuals by the high-frequency based models is in line with that observed in Figure 1 and GJR-GARCH(1,1) is still the best choice.

Although the results are compelling, they might seem counterintuitive. Indeed, to the extent that a model provides superior forecasts of the volatility, it should also provide a better proxy of the innovations. We argue that the observed difference in the estimated residuals might be explained by the different ability of the two approaches in generating extremal dependence.

Most of the evidence on the higher performance of the high-frequency based volatility models is based on comparisons of the R^2 statistic or other loss functions. But this does not say anything regarding the extremal dependence. The GARCH class of models is supported by theoretical arguments which show their ability to generate extremal dependence in the volatility process and consequently in the return process (Basrak, Davis & Mikosch 2002). The HAR class of models is a simple equation that fits nicely the stylized facts of the financial time series and is supported by economic arguments regarding the heterogeneous behavior of different investors. It is difficult (if possible) to derive a theoretical result on the extremal behavior of this model as it is for GARCH processes.

In order to get some insights on this issue, we can build upon the comparison between GARCH processes and stochastic volatility (SV) models from an extreme perspective. While GARCH processes exhibit extremal dependence, SV may or may not fail to account for this property depending on the specification (Davis & Mikosch 2009; Fasen, Klüppelberg &

Lindner 2006). As the HAR class of models can be considered in the domain of SV models (Corsi 2009), then it may either capture or not the extremal dependence in the returns.

To assess this possibility, we follow Liu & Tawn (2013) and rely on the conditional tail probability $c(x) = \mathbb{P}(X_{t+1} > q|X_t > q)$. Theoretically, $c(x) \to 0$ as $q \to \infty$ when the extremes are independent and $c(x) \to c > 0$ as $q \to \infty$ when they are dependent. As it is not possible to study the asymptotic behavior of $c(x)$ for $q \to \infty$, we approximate it considering a sequence of high quantiles q. Figure 4 displays the empirical conditional tail probability at different levels q for three sub-samples of the S&P500. Moreover, we report the same coefficients estimated on simulated data produced by the GJR-GARCH(1,1) and the HAR models fitted to these three samples. As expected, the tail probability implied by the GARCH model is positive at every quantile pointing toward the presence of extremal dependence. The conditional probability for the HAR model does not reach zero at the considered levels, but it is clear that its decay is much faster than the one of the GJR-GARCH model, and it fails to capture the empirical tail dependence.

Based on this evidence, it is worth investigating whether the time series for which the HAR provides a good filter exhibit a lower degree of extremal dependence. A positive response would be consistent with the idea that GARCH and HAR models imply two different degrees of dependence in the extremes. Figure 5 displays the estimates of θ obtained on the return series of the 17 stock indices in the different windows. The results are somewhat blurred: some cases are consistent with the findings in Figure 1 and some others are not. For instance, a high degree of dependence is shown by the AEX and the SPX and a low degree of dependence is shown by the BVP and the NK, which is in line with the dependence left in the residuals of the HAR models. But at the same time, IPC and RUS present lower dependence than other series, yet HAR models could not filter to iid.

4.4 The forecast component

Jalal & Rockinger (2008) carry out an extensive simulation experiment to investigate the performance of C-EVT when the DGP is not the standard GARCH(1,1) model of Bollerslev

Figure 1: **Extremal index estimates for the type-I link function**. Values of θ estimated on the standardized residuals obtained from the different windows of length $S = 2000$.

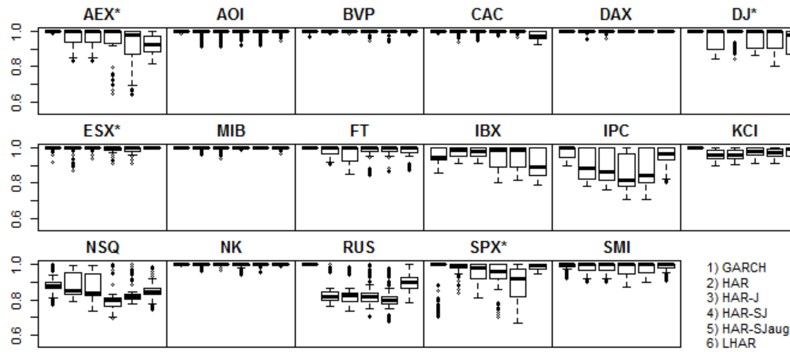


Figure 2: **Extremal index estimates for the type-II link function**. Values of θ estimated on the standardized residuals obtained from the different windows of length $S = 2000$.

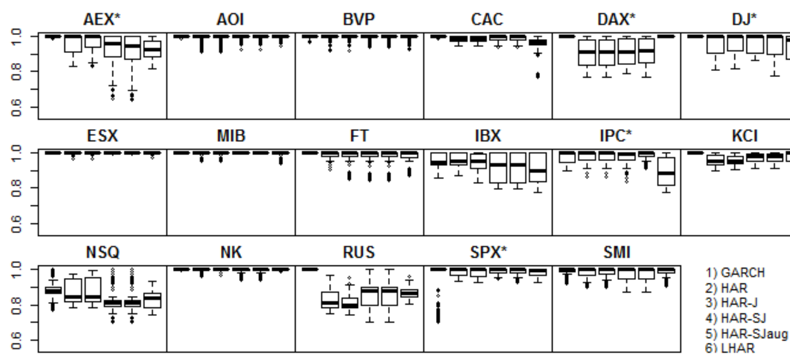


Figure 3: **Extremal index estimates for the type-III link function**. Values of θ estimated on the standardized residuals obtained from the different windows of length $S = 2000$.

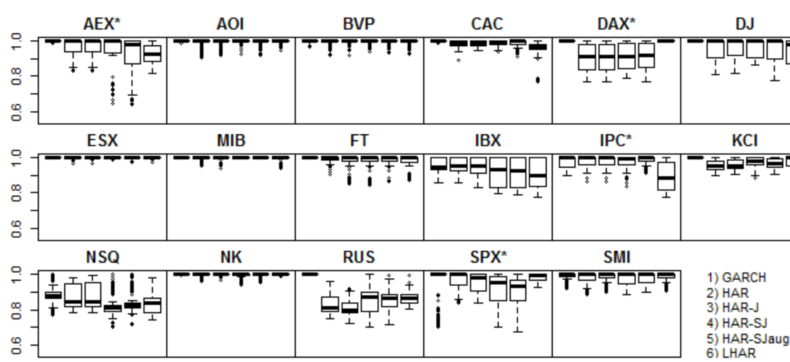


Figure 4: **Conditional tail probaility**. Estimates of $c(x)$ for the lower tail of three sub-samples of the S&P500 at 10 different quantiles: Empirical (*points*); simulation from the GJR-GARCH(1,1) model (*solid*) and tolerance levels (*dotted*); simulation from the HAR model (*dashed*) and tolerance levels (*dash-dotted*).

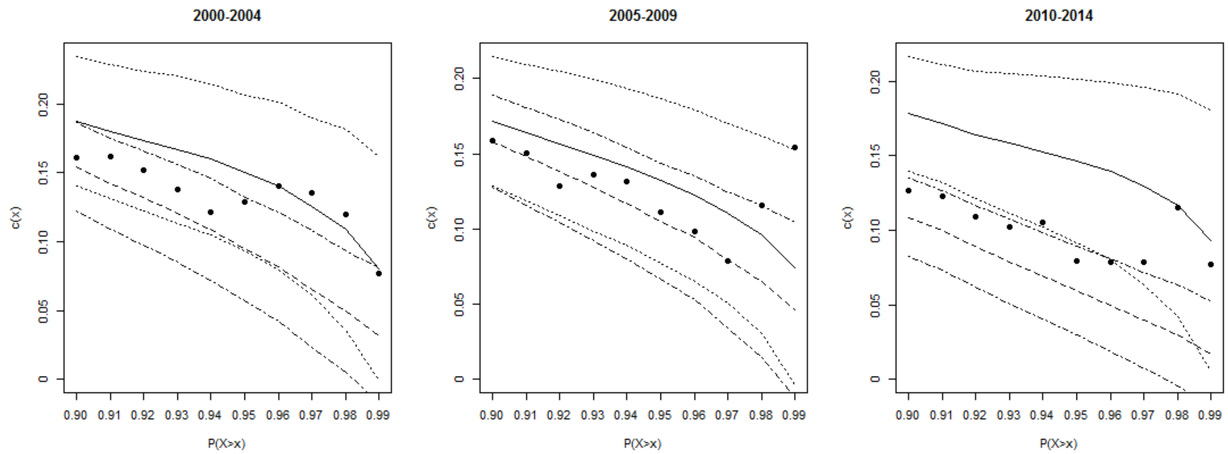
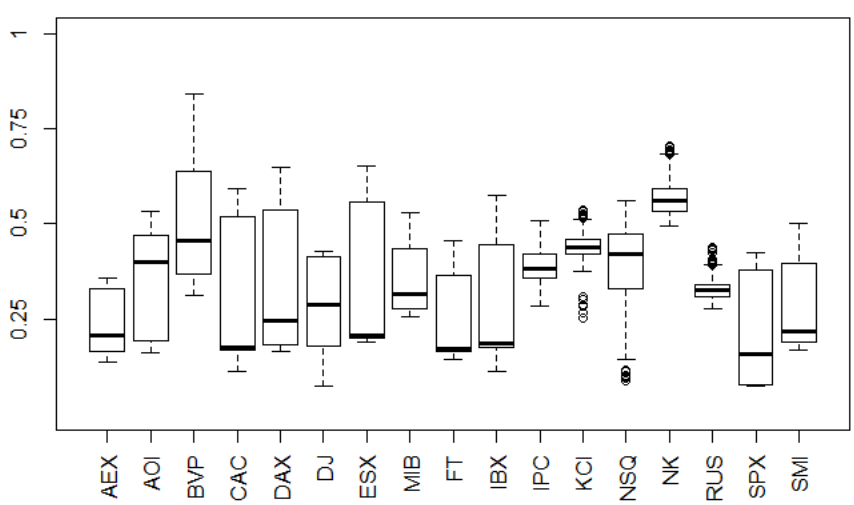


Figure 5: **Extremal index on the raw data**. Estimates of θ on the lower tail of the return series over moving windows of length $S = 2000$.



(1986). Their results emphasize how C-EVT provides accurate one-day-ahead forecasts of the VaR and ES regardless of the assumed DGP, thus strengthening the conclusions of McNeil & Frey (2000).

We suspect that the performance of the C-EVT deteriorates as estimation moves to longer time horizons. The quality of the estimation of the tail-risk measures depends on the quality of the volatility forecast. As it is shown that the volatility forecasting performance of methods based on daily returns deteriorates with increasing horizon length (Brownlees, Engle & Kelly 2011), similar deterioration should be seen in tail-risk forecasts.

We perform both a simulation experiment and an empirical data analysis. In both cases, we provide out-of-sample forecasts of the one-day- and ten-day-ahead VaR and ES at level $\alpha = 0.01$.

Simulation experiment. First, we reproduce the analysis of Jalal & Rockinger (2008). We consider six different DGP: the GARCH(1,1) with Gaussian and t innovations; the exponential EGARCH(1,1) of Nelson (1991); the regime-switching model of Schaller & Norden (1997); the discrete-time jump diffusion model of Pan (1997) and its pure jump version. For each process, we generate $B = 100$ samples of $T = 4000$ daily returns. Figure 6 shows the results for both the one-day- and ten-day-ahead forecasts. To evaluate the performance of the VaR, we count the violations in each window. When the length of the window is $S = 2000$, the expected number of violations is $(T - S) \times \alpha = 20$. As a performance measure for the ES, we compute its distance from the actual return when the VaR is exceeded, and expect this to be mean zero and symmetrically distributed.

Although we do not report any formal test, the results for the one-day-ahead predictions in Figure 6 are consistent with those of Jalal & Rockinger (2008). However, focussing particularly on the VaR, it is evident that the accuracy of the C-EVT is reduced at the longer time horizon. Indeed, for three out of six DGP the average number of violations moves further away from the expected number.

Can RV-EVT provide more accurate predictions? We set up another simulation experiment where the two approaches can be compared. We need to generate high-frequency observations, so we consider the stochastic volatility model of Bandi & Renò (2015). The latter involves 21 parameters and accommodates several attributes of financial time series. It accounts for time-variation in both the volatility and leverage, and allows for idiosyncratic and common jump discontinuities in the price process and the volatility, determined from a complex dependence structure.

We generate $B = 100$ samples of $T = 4000$ daily observations for both the returns and the 5-min realized measures. We use the Euler scheme to generate $N = 2520$ equally-spaced observations per day, assuming 7 hours of trading and changes in the price level occurring every ten seconds, leading to 6 ticks per minute ($6 \times 60 \times 7 = 2520$). This discretization method is standard in the literature (Huang & Tauchen 2005), while the choice of the frequency is mainly for computational reasons. The parameters of the model are set according to the GMM estimates obtained on the S&P500 futures from April 1982 to February 2009 in Bandi & Renò (2015).

In Figure 7, we report the one-day- and ten-day-ahead VaR and ES forecasts obtained over each window. We consider only a Type-I link function for RV-EVT. Consistent with the results in Figure 6, the C-EVT performs well for one-day-ahead forecasts but deteriorates as we extend the time horizon. Contrarily, the RV-EVT approach seems to be less affected by the increasing time horizon and performs well regardless of the high-frequency based volatility model for the RV. In particular, the LHAR model emerges as the best one since it presents less variability in terms of number of violations, and it is less upwardly biased in the estimation of the ES.

Empirical analysis. We now consider empirical data and assess whether the results suggested by the simulation experiment are confirmed.

In Table 2, we compare one-day-ahead VaR forecasts from the C-EVT and RV-EVT for the 17 indices. To evaluate the accuracy of the predictions, we use the standard tests of unconditional coverage (UC), independence assumption (IN) and conditional coverage (CC)

Figure 6: **Extending experiments in Jalal & Rockinger (2008)**. One- and ten-day-ahead predictions at level $\alpha = 1\%$. VaR violations (left panels) and actual returns minus ES when VaR is exceeded (right panels). Expected number of violations is 20. ES differences should be mean zero and symmetrically distributed. DGP are: Gaussian-GARCH(1,1) (a); Student's-GARCH(1,1) (b); EGARCH(1,1) (c); Switching-Markov (d); Jump-diffusion (e); Pure jumps (f).

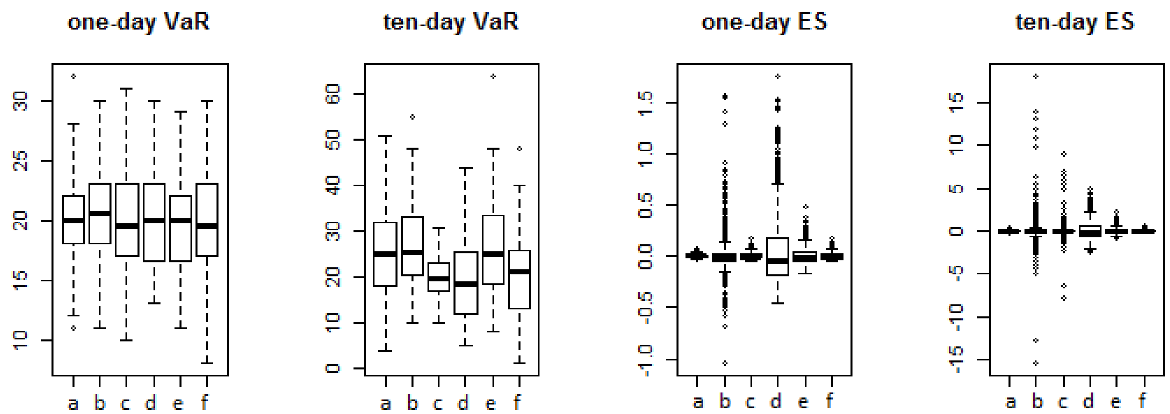
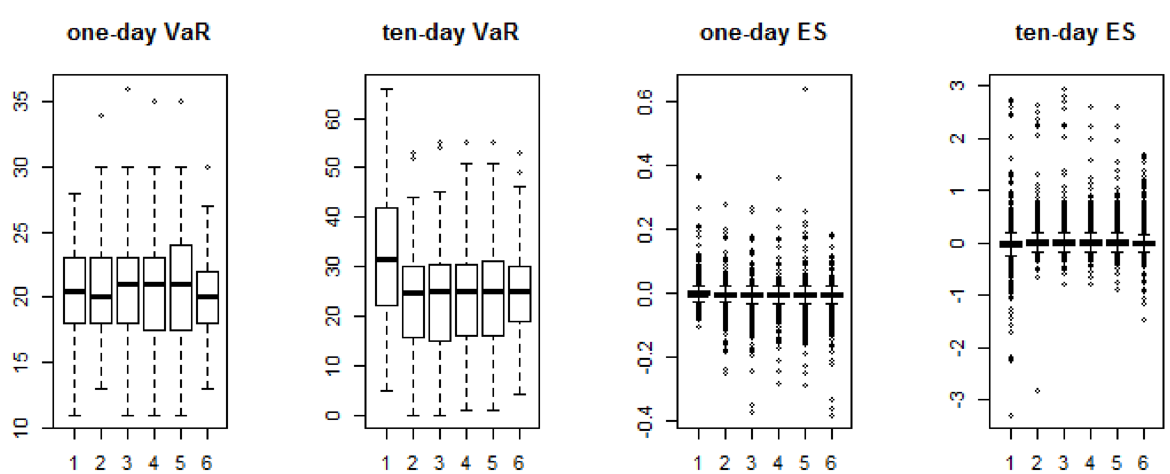


Figure 7: **Comparison of C-EVT and RV-EVT for simulated data following Bandi and Renó (2015) model**. One-day and ten-day-ahead predictions at level $\alpha = 1\%$. VaR violations (left panels) and actual returns minus ES when VaR is exceeded (right panels). Expected number of violations is 20. ES differences should be mean zero and symmetrically distributed. Models used in first step: (*C-EVT*) 1 - GARCH ; (*Symmetric RV-EVT*) 2 - HAR; 3 - HAR-J; (*Asymmetric RV-EVT*) 4 - HAR-SJ; 5 - HAR-SJaug; 6 - LHAR.



suggested by Christoffersen (1998), and the dynamic quantile test (DQ) of Engle & Manganelli (2004). These results suggest that the C-EVT and the RV-EVT perform equally well for one-day-ahead. In most of the cases, the number of violations for both approaches is very close to the expected, and the null hypothesis of independence is typically not rejected at the standard levels. Noteworthy is the DQ statistic, reflecting the predictability of a violation given the past information. The null hypothesis of independence between consecutive violations appears to be the one most frequently rejected. In sum, these results are in line with those obtained by Giot & Laurent (2004) and Brownlees & Gallo (2009) using non EVT-based modelling assumptions.

To compare the one-day-ahead ES predictions, we test the hypothesis that conditional upon exceeding the 99th quantile of the loss distribution, the difference between the actual return and the predicted ES has mean zero. We conduct a one-sided test with the alternative that the mean is greater than zero using a bootstrap that makes no assumption about the distribution of the differences, see Section 16.4 of Efron & Tibshirani (1994). The small number of p-values below 0.05 in Table 3 indicate that the good results for the VaR extend to ES.

Turning to the ten-day-ahead predictions of these tail-risk measures, hypothesis testing is complicated by the fact that the violations are inherently autocorrelated. Nonetheless, we provide two simple statistical tests for the VaR and the ES. First, let $I^{\alpha}_{t,t+10}$ take value one when the ten-day-ahead VaR prediction at level α is violated, and zero otherwise. $\{I_{t,t+10}, t = 1, 2, \dots\}$ is a sequence of correlated Bernoulli random variables. Letting $\Phi(\cdot)$ be a Probit link function, we assume that $I_{t,t+10} = \Phi(\gamma)$ and test the hypothesis of correct unconditional coverage $\mathcal{H}_0 : \gamma = \Phi^{-1}(\alpha)$ against the alternative $\mathcal{H}_1 =$ not $\mathcal{H}_0$. We perform a simple Wald test with robust standard errors computed with the Newey-West estimator to account for the autocorrelations. Table 4 shows the number of violations of each model along with the p-value of the test. The results suggest that at longer time horizons the RV-EVT approach is more accurate than C-EVT: in three cases the null of correct unconditional coverage is rejected for C-EVT, and most of the time the RV-EVT is more accurate in terms of number of violations.

To compare the ten-day-ahead ES predictions, we employ again the bootstrap test used for the one-day-ahead ES predictions, but we base it on a block bootstrap because of the autocorrelations in the violations. In Table 5 we report the p-values of the test. Contrasts are not as clear as in the case of the VaR as the number of rejections of the null hypothesis of mean zero difference between the ten-day-ahead ES forecasts and the actual ten-day returns is very similar in all cases.

5. FURTHER ANALYSIS

The results from the preceding sections highlight the ability of the GARCH model as a filter, and the merits of the high-frequency based volatility models for the forecast of the tail-risk measures. Now, as standard in finance, we verify that the results obtained are robust to the choices made.

5.1 An investigation during a period of turmoil

We evaluate the performance of C-EVT and RV-EVT when the training sample used to fit the models is limited to 500 days instead of 2000. Furthermore, we consider the time period from the beginning of 2007 to the end of 2011.

Given the reduced sample size, we consider the exceedances over the threshold $\widehat{u}$ corresponding to the 90th quantile of the negated residuals $-\widehat{\epsilon}_t$ of the considered model. Goodness-of-fit tests are carried out as described in Section 4.2. When the threshold seems to be still at the sub-asymptotic level, it is raised to the 92nd quantile. One-day- and ten-day-ahead forecasts of the VaR and the ES are produced at level $\alpha = 0.05$.

Results available in the supplementary material confirm the findings of Section 4, although the differences are less pronounced when looking only at the period of turmoil and the poorer performance of the GJR-GARCH(1,1) is clearer in the estimates of ES.

Table 2: **Performance measures for the one-day-ahead VaR forecast**. For each model (GARCH (a), HAR (b), HAR-J (c), HAR-SJ (d), HAR-SJaug (e), LHAR (f)), we report: the actual number of violations (NV); the p-values for the unconditional coverage (UC), the independence assumption (IN), the conditional coverage (CC), and the DQ test (DQ). Expected number of violations are in parentheses. Rejection at the level $\alpha = 5\%$ is in bold. The * denotes series for which a higher threshold was required for appropriate GPD behavior.

Index	PM	(a)	(b)	(c)	(d)	(e)	(f)	Index	PM	(a)	(b)	(c)	(d)	(e)	(f)
AEX*	NV	22	18	18	19	19	17	IBX	NV	26	19	19	20	21	18
(18)	UC	0.36	1	1	0.81	0.81	0.81	(18)	UC	0.07	0.81	0.81	0.64	0.49	1
	IN	0.46	0.55	0.55	0.52	0.52	0.57		IN	0.38	0.20	0.20	0.22	0.25	0.54
	CC	0.50	0.83	0.83	0.79	0.79	0.82		CC	0.14	0.42	0.42	0.42	0.40	0.83
	DQ	0.65	0.72	0.72	0.75	0.75	0.99		DQ	**0.003**	0.19	0.18	**0.001**	**0.002**	0.54
AOI	NV	16	13	15	16	17	11	IPC	NV	13	7	7	7	9	7
(17)	UC	0.80	0.30	0.62	0.81	1	0.12	(17)	UC	0.31	**0.005**	**0.005**	**0.005**	**0.03**	**0.005**
	IN	0.58	0.66	0.61	0.59	0.56	0.71		IN	0.66	0.81	0.81	0.81	0.76	0.81
	CC	0.82	0.54	0.77	0.83	0.84	0.27		CC	0.54	**0.02**	**0.02**	**0.02**	0.10	**0.02**
	DQ	0.99	0.34	0.51	0.61	0.63	0.12		DQ	0.27	0.46	0.49	0.49	0.75	0.47
BVP	NV	13	12	12	14	14	13	KCI	NV	13	20	20	19	21	17
(17)	UC	0.31	0.30	0.20	0.45	0.45	0.31	(17)	UC	0.31	0.48	0.48	0.63	0.35	1
	IN	0.65	0.66	0.68	0.63	0.63	0.65		IN	0.65	0.23	0.23	0.21	0.26	0.56
	CC	0.53	0.54	0.40	0.66	0.66	0.53		CC	0.53	0.38	0.38	0.40	0.34	0.83
	DQ	0.25	0.34	0.27	0.06	0.07	0.31		DQ	0.97	0.36	0.34	0.16	0.27	0.42
CAC	NV	29	26	27	26	27	22	NSQ	NV	20	22	22	20	20	17
(18)	UC	**0.01**	0.08	**0.05**	0.08	**0.05**	0.36	(17)	UC	0.48	0.24	0.24	0.48	0.48	1
	IN	0.33	0.38	0.37	0.38	0.37	0.46		IN	0.50	0.45	0.45	0.50	0.49	0.56
	CC	**0.03**	0.14	0.09	0.14	0.09	0.49		CC	0.61	0.38	0.38	0.61	0.61	0.84
	DQ	**0.02**	**0.05**	**0.04**	**0.05**	**0.04**	0.09		DQ	0.0008	0.13	0.13	0.18	0.40	0.16
DAX	NV	19	18	17	16	16	15	NK	NV	15	16	16	15	16	15
(18)	UC	0.81	1	0.81	0.63	0.63	0.46	(16)	UC	0.80	1	1	0.80	1	0.80
	IN	0.52	0.54	0.57	0.59	0.59	0.61		IN	0.13	0.15	0.15	0.13	0.15	0.13
	CC	0.78	0.82	0.81	0.76	0.76	0.66		CC	0.30	0.34	0.34	0.30	0.35	0.29
	DQ	0.16	**0.04**	0.14	0.11	0.42	0.29		DQ	0.50	**0.007**	**0.008**	**0.01**	**0.02**	**0.007**
DJ*	NV	21	22	21	20	21	17	RUS	NV	23	25	25	23	24	20
(17)	UC	0.35	0.25	0.35	0.48	0.34	1	(17)	UC	0.16	0.07	0.07	0.16	0.11	0.48
	IN	0.47	0.28	0.48	0.50	0.47	0.56		IN	0.43	0.39	0.39	0.43	0.41	0.49
	CC	0.49	0.28	0.49	0.61	0.49	0.83		CC	0.28	0.13	0.13	0.28	0.19	0.61
	DQ	0.52	**0.008**	**0.01**	0.95	0.69	0.99		DQ	**0.005**	**0.001**	**0.001**	**0.001**	**0.02**	0.42
ESX*	NV	31	18	18	19	19	15	SPX*	NV	27	25	28	25	25	23
(18)	UC	**0.01**	1	1	0.82	0.81	0.46	(17)	UC	**0.02**	0.07	**0.01**	0.07	0.07	0.16
	IN	0.29	0.54	0.55	0.53	0.53	0.61		IN	0.36	0.40	0.34	0.40	0.40	0.43
	CC	**0.01**	0.82	0.82	0.79	0.79	0.66		CC	0.05	0.13	0.03	0.13	0.13	0.27
	DQ	**0.001**	0.37	0.37	0.13	0.12	0.41		DQ	0.11	0.40	0.003	0.20	0.28	0.73
MIB	NV	31	23	23	24	25	23	SMI	NV	17	19	19	18	18	15
(18)	UC	**0.01**	0.26	0.26	0.17	0.12	0.26	(17)	UC	1	0.63	0.63	0.81	0.81	0.62
	IN	0.29	0.44	0.44	0.42	0.40	0.43		IN	0.56	0.20	0.20	0.54	0.54	0.62
	CC	**0.01**	0.38	0.38	0.29	0.20	0.38		CC	0.84	0.39	0.39	0.79	0.79	0.77
	DQ	**0.003**	**0.02**	**0.02**	**0.04**	0.22	**0.04**		DQ	0.68	0.13	0.13	0.17	0.17	0.93
FT	NV	20	16	16	15	15	13								
(18)	UC	0.64	0.63	0.63	0.46	0.46	0.26								
	IN	0.50	0.59	0.59	0.61	0.61	0.43								
	CC	0.70	0.76	0.76	0.67	0.67	0.38								
	DQ	0.71	0.96	0.97	0.98	0.98	0.34								

Table 3: **Performance of one-day-ahead ES forecast**. P-values from the one-sided upper-tail bootstrap test for the null of mean zero difference between actual returns and ES when VaR is exceeded. Rejection at level $\alpha = 5\%$ is in bold. The * denotes series for which a higher threshold was required for appropriate GPD behavior.

	GARCH	HAR	HAR-J	HAR-SJ	HAR-SJaug	LHAR
AEX*	**0.020**	0.090	0.113	0.097	**0.048**	0.135
AOI	0.920	0.577	0.480	0.787	0.630	0.96
BVP	0.414	0.988	0.982	1.000	1.000	1.000
CAC	0.103	0.599	0.457	0.425	0.148	0.655
DAX	0.262	0.315	0.307	0.303	0.376	0.235
DJ*	0.998	1.000	1.000	0.993	1.000	1.000
ESX*	0.737	0.406	0.431	0.555	0.567	0.491
MIB	0.821	0.807	0.840	0.857	0.844	0.906
FT	0.410	0.647	0.719	0.791	0.693	0.719
IBX	0.328	0.435	0.317	0.257	0.213	0682
IPC	0.827	0.991	0.989	0.999	0.956	0.953
KCI	0.063	0.254	0.135	0.177	0.076	0.289
NSQ	0.490	0.900	0.910	0.797	0.755	0.678
NK	**0.023**	**0.034**	**0.023**	**0.017**	**0.008**	**0.023**
RUS	0.960	0.781	0.671	0.523	0.500	0.539
SPX*	0.999	0.995	0.993	0.999	1.000	1.000
SMI	0.198	0.860	0.875	0.776	0.799	0.638

Table 4: **Performance measures for the ten-day-ahead VaR forecast**. Actual and expected number of violations of the VaR. Cases in which the null hypothesis of correct unconditional coverage is rejected at level $\alpha = 10\%$ are in bold. The best performer is in *italics*. The * indicates series for which a higher threshold was required for appropriate GPD behavior.

Index	Expected	GARCH	HAR	HAR-J	HAR-SJ	HAR-SJ-AUG	LHAR
AEX	18	**34 (0.08)**	18 (1.00)	*16 (0.81)*	19 (0.89)	20 (0.79)	18 (0.99)
AOI*	17	25 (0.43)	25 (0.43)	25 (0.43)	23 (0.60)	22 (0.68)	*21 (0.63)*
BVP	17	12 (0.49)	22 (0.54)	22 (0.54)	21 (0.60)	22 (0.54)	*19 (0.77)*
CAC	18	29 (0.23)	*27 (0.36)*	*27 (0.36)*	*27 (0.34)*	*27 (0.34)*	*27 (0.30)*
DAX	18	27 (0.33)	22 (0.66)	*20 (0.82)*	21 (0.73)	21 (0.73)	22 (0.66)
DJ	17	12 (0.51)	20 (0.71)	20 (0.71)	*18 (0.92)*	*18 (0.92)*	19 (0.83)
ESX	18	32 (0.17)	21 (0.72)	21 (0.72)	21 (0.72)	22 (0.64)	*20 (0.81)*
MIB	18	**50 (0.02)**	*25 (0.36)*	*25 (0.36)*	*25 (0.36)*	*25 (0.36)*	*25 (0.34)*
FT	18	33 (0.14)	19 (0.95)	*18 (0.95)*	*18 (0.95)*	19 (0.95)	24 (0.41)
IBX	18	**39 (0.07)**	*24 (0.87)*	25 (0.78)	26 (0.69)	26 (0.69)	25 (0.35)
IPC	17	*14 (0.68)*	7 (0.16)	7 (0.16)	7 (0.16)	7 (0.16)	9 (0.30)
KCI*	17	*11 (0.45)*	26 (0.21)	27 (0.25)	27 (0.25)	28 (0.23)	26 (0.21)
NSQ	17	20 (0.73)	*18 (0.92)*	18 (0.92)	15 (0.75)	15 (0.75)	14 (0.62)
NK	16	*24 (0.57)*	25 (0.47)	25 (0.47)	26 (0.39)	26 (0.39)	25 (0.48)
RUS	17	18 (0.94)	18 (0.93)	*17 (0.98)*	*17 (0.98)*	16 (0.88)	18 (0.94)
SPX	17	10 (0.38)	*17 (0.98)*	*17 (0.97)*	16 (0.88)	16 (0.88)	19 (0.83)
SMI	17	24 (0.41)	*11 (0.35)*	*11 (0.35)*	*11 (0.35)*	*11 (0.35)*	*11 (0.35)*

Table 5: **Performance of ten-day-ahead ES forecast**. P-values from the one-sided upper-tail block-bootstrap test for the null of mean zero difference between actual returns and ES when VaR is exceeded. Rejection at level $\alpha = 5\%$ is in bold. The * denotes series for which a higher threshold was required for appropriate GPD behavior.

	GARCH	HAR	HAR-J	HAR-SJ	HAR-SJ-AUG	LHAR
AEX	0.114	0.124	0.071	0.123	0.333	0.294
AOI*	0.946	0.244	0.249	0.243	0.345	0.122
BVP	0.673	0.128	0.138	0.118	**0.035**	0.073
CAC	0.283	0.534	0.562	0.414	0.391	0.601
DAX	**0.045**	0.101	**0.035**	0.060	0.059	0.056
DJ	0.345	0.288	0.246	0.232	0.217	0.211
ESX	0.153	0.170	0.203	0.153	0.057	**0.025**
MIB	0.288	0.743	0.763	0.776	0.787	0.816
FT	0.119	0.451	0.396	0.312	0.379	0.622
IBX	0.738	0.157	0.196	0.291	0.238	0.183
IPC	0.220	0.767	0.741	0.908	0.868	1.000
KCI*	**0.027**	0.123	0.140	0.251	0.281	0.319
NSQ	0.062	**0.028**	**0.022**	**0.000**	**0.000**	**0.003**
NK	**0.007**	**0.021**	**0.030**	**0.019**	**0.024**	**0.035**
RUS	0.089	0.156	0.085	0.134	0.052	0.178
SPX	0.282	**0.047**	**0.043**	**0.038**	**0.038**	0.092
SMI	0.479	0.511	0.483	0.493	0.484	0.341

5.2 Microstructure noise

To investigate whether the microstructure noise has an impact on the analysis of Section 4, we repeat the analysis on both the full and restricted samples using two estimators of the quadratic variation that mitigate the effect of the microstructure noise: the *Second-Best Approach: Sub-sampling and Averaging* estimator of Zhang et al. (2005), which is biased in the presence of microstructure noise but presents a lower variance than the RV estimator; and the Realized Kernel (RK) of Barndorff-Nielsen, Hansen, Lunde & Shephard (2008).

We assess both the filtering and forecasting component of the RV-EVT using first the sub-sampled measures and then substituting the sub-sampled RV with the RK. Results reported in the supplementary material lead to the same conclusions of Section 4.

6. CONCLUDING REMARKS

This article questions whether combining the recent advances in the high-frequency financial econometrics with results from the Extreme Value Theory can improve the fit of the tails of the conditional returns distribution.

We propose an RV-EVT approach where returns are pre-whitened with a high-frequency based volatility model and the POT approach is applied to the tails of the standardized residuals. We use three different functions to link the predictions of the realized volatility to the conditional variance of returns, and employ six different econometric specifications to model the realized volatility.

This approach is compared to the standard C-EVT technique both through simulation and with an extensive empirical analysis on 17 international indices. We assess both approaches' ability of filtering the dependence in the extremes and of producing stable out-of-sample VaR and ES predictions for one- and ten-day time horizons.

Summarizing the results, it seems that a GARCH-type filter performs slightly better than a high-frequency based filter, even though they both tend to produce estimated residuals which are close to independent. From a risk management perspective however, the RV-EVT approach seems preferable, especially at the longer time horizons which are of interest for regulatory purposes.

Acknowledgments. The second author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada. The authors wish to thank two anonymous referees for comments that helped improve an earlier version of the manuscript.

REFERENCES

Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2007), "Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility," *The Review of Economics and Statistics*, 89(4), 701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001), "The distribution of realized exchange rate volatility," *Journal of the American Statistical Association*, 96(453), 42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003), "Modeling and forecasting realized volatility," *Econometrica*, 71(2), 579–625.

Andersen, T. G., Bollerslev, T., & Huang, X. (2011), "A reduced-form framework for modeling volatility of speculative prices based on realized variation measures," *Journal of Econometrics*, 160(1), 176–189.

Baillie, R. T., & Bollerslev, T. (2002), "The message in daily exchange rates: a conditional-variance tale," *Journal of Business & Economic Statistics*, 20(1), 60–68.

Bandi, F. M. and Renò, R. (2015), "Price and volatility co-jumps," *Journal of Financial Economics* (Forthcoming).

Barndorff-Nielsen, O. E. (2002), "Econometric analysis of realized volatility and its use in estimating stochastic volatility models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2), 253–280.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008), "Designing realized kernels to measure the ex post variation of equity prices in the presence of noise," *Econometrica*, 76(6), 1481–1536.

Barndorff-Nielsen, O. E., Kinnebrock, S., & Shephard, N. (2008), "Measuring downside risk-realised semivariance," CREATES Research Paper, 2008-42.

Barndorff-Nielsen, O. E., & Shephard, N. (2004), "Power and bipower variation with stochastic volatility and jumps," *Journal of Financial Econometrics*, 2(1), 1–37.

Basrak, B., Davis, R. A., & Mikosch, T. (2002), "Regular variation of GARCH processes," *Stochastic Processes and their Applications*, 99(1), 95–115.

Bollerslev, T. (1986), "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, 31(3), 307–327.

Bollerslev, T., Kretschmer, U., Pigorsch, C., & Tauchen, G. (2009), "A discrete-time model for daily S&P500 returns and realized variations: Jumps and leverage effects," *Journal of Econometrics*, 150(2), 151–166.

Brownlees, C. T., Engle, R. F., & Kelly, B. T. (2011), "A practical guide to volatility forecasting through calm and storm," Available at SSRN 1502915.

Brownlees, C. T., & Gallo, G. M. (2009), "Comparison of volatility measures: a risk management perspective," *Journal of Financial Econometrics*, 8(1), 29–56.

Choulakian, V., & Stephens, M. (2001), "Goodness-of-fit tests for the generalized Pareto distribution," *Technometrics*, 43(4), 478–484.

Christoffersen, P. F. (1998), "Evaluating interval forecasts," *International Economic Review*, 39(4), 841–862.

Clements, M. P., Galvão, A. B., & Kim, J. H. (2008), "Quantile forecasts of daily exchange rate returns from forecasts of realized volatility," *Journal of Empirical Finance*, 15(4), 729–750.

Corsi, F. (2009), "A simple approximate long-memory model of realized volatility," *Journal of Financial Econometrics*, 7(2), 174–196.

Corsi, F. & Renò, R. (2012), "Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling," *Journal of Business & Economic Statistics*, 30(3), 368–380.

Corsi, F., Pirino, D., & Renò, R. (2010), "Threshold bipower variation and the impact of jumps on volatility forecasting," *Journal of Econometrics*, 159(2), 276–288.

Danielsson, J., & de Vries, C. G. (1997), "Tail index and quantile estimation with very high frequency data," *Journal of Empirical Finance*, 4(2), 241–257.

Davis, R. A., & Mikosch, T. (2009), "Extremes of stochastic volatility models," in Handbook of Financial Time Series, Springer, 355–364.

Davison, A. C., & Smith, R. L. (1990), "Models for exceedances over high thresholds," *Journal of the Royal Statistical Society (Series B)*, 52(3), 393-442.

Diebold, F. X., Schuermann, T., & Stroughair, J. D. (1998), Pitfalls and Opportunities in the Use of Extreme Value Theory in Risk Management, Springer.

Efron, B., & Tibshirani, R. J. (1994), An Introduction to the Bootstrap, CRC press.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997), Modelling Extremal Events: for Insurance and Finance, Springer.

Engle, R. F., & Gallo, G. M. (2006), "A multiple indicators model for volatility using intra-daily data," *Journal of Econometrics*, 131(1), 3–27.

Engle, R. F., & Manganelli, S. (2004), "CAViaR: Conditional autoregressive Value at Risk by regression quantiles," *Journal of Business & Economic Statistics*, 22(4), 367–381.

Fasen, V., Klüppelberg, C., & Lindner, A. (2006), "Extremal behavior of stochastic volatility models," *Stochastic Finance*, Springer, 107–155.

Ferro, C. A., & Segers, J. (2003), "Inference for clusters of extreme values," *Journal of the Royal Statistical Society (Series B)*, 65(2), 545–556.

Giot, P., & Laurent, S. (2004), "Modelling daily Value-at-Risk using realized volatility and ARCH type models," *Journal of Empirical Finance*, 11(3), 379–398.

Hansen, P. R., & Lunde, A. (2006), "Realized variance and market microstructure noise," *Journal of Business & Economic Statistics*, 24(2), 127–161.

Heber, G., Lunde, A., Shephard, N., & Sheppard, K. (2009), "Oxford-Man Institute realized library, version 0.2,", Oxford-Man Institute, University of Oxford.

Huang, X., & Tauchen, G. (2005), "The relative contribution of jumps to total price variance," *Journal of Financial Econometrics*, 3(4), 456–499.

Jalal, A., & Rockinger, M. (2008), "Predicting tail-related risk measures: The consequences of using GARCH filters for non-GARCH data," *Journal of Empirical Finance*, 15(5), 868–877.

Kuester, K., Mittnik, S., & Paolella, M. S. (2006), "Value-at-risk prediction: A comparison of alternative strategies," *Journal of Financial Econometrics*, 4(1), 53–89.

Laurini, F., & Tawn, J. A. (2008), "Regular variation and extremal dependence of GARCH residuals with application to market risk measures," *Econometric Reviews*, 28(1-3), 146–169.

Leadbetter, M. (1983), "Extremes and local dependence in stationary sequences," *Probability Theory and Related Fields*, 65(2), 291–306.

Liu, L. Y., Patton, A. J., & Sheppard, K. (2012), "Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes," Workimg paper, December 2012.

Liu, Y., & Tawn, J. A. (2013), "Volatility model selection for extremes of financial time series," *Journal of Statistical Planning and Inference*, (143), 520–530.

Longin, F. M. (2000), "From Value at Risk to stress testing: The extreme value approach," *Journal of Banking & Finance*, 24(7), 1097–1130.

Mancini, L., & Trojani, F. (2011), "Robust Value at Risk prediction," *Journal of Financial Econometrics*, 9(2), 281–313.

McNeil, A. J., & Frey, R. (2000), "Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach," *Journal of Empirical Finance*, 7(3), 271–300.

Nelson, D. B. (1991), "Conditional heteroskedasticity in asset returns: A new approach," *Econometrica*, 59(2), 347–370.

Pan, J. (1997), "Stochastic volatility with reset at jumps," Available at SSRN 135048.

Patton, A. J., & Sheppard, K. (2015), "Good volatility, bad volatility: Signed jumps and the persistence of volatility," *Review of Economics and Statistics* (Forthcoming).

Pickands, J. (1975), "Statistical inference using extreme order statistics," *Annals of Statistics*, 3(1), 119–131.

Schaller, H., & Norden, S. V. (1997), "Regime switching in stock market returns," *Applied Financial Economics*, 7(2), 177–191.

Shephard, N., & Sheppard, K. (2010), "Realising the future: forecasting with high-frequency-based volatility (HEAVY) models," *Journal of Applied Econometrics*, 25(2), 197–231.

Todorov, V., & Tauchen, G. (2011), "Volatility jumps," *Journal of Business & Economic Statistics*, 29(3), 356–371.

Tolver, J. A. & Lange, T. (2010), "On Convergence of the QMLE for Misspecified GARCH Models," *Journal of Time Series Econometrics*, 2(1), 1–31.

Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005), "A tale of two time scales," *Journal of the American Statistical Association*, 100(472), 1394–1411.

A. SAMPLING FROM THE SEMI-PARAMETRIC DISTRIBUTION OF THE RESIDUALS

1. Randomly select with replacement a residual from the sample of S residuals;

2. If the residual exceeds an upper threshold $\widehat{u}^H$ sample a GP$(\widehat{\xi}^H, \widehat{\eta}^H)$ distributed observation y^H from the right tail and return $\widehat{u}^H + y^H$;

3. If the residual is less than a lower threshold $\widehat{u}^L$ sample a GP$(\widehat{\xi}^L, \widehat{\eta}^L)$ distributed observation y^L from the left tail and return $\widehat{u}^L + y^L$;

4. Otherwise return the residual itself;

5. Repeat.

B. AUGMENTED HAR-RV MODEL WITH SIGNED JUMPS

Patton & Sheppard (2015) propose to extend the HAR-J model of Andersen et al. (2007), isolating the information coming from the sign of the jumps. They use a measure called signed jump variation $\Delta J_t^2 \equiv RS_t^+ - RS_t^-$ which is positive when a day is dominated by an upward jump and negative when a day is dominated by a downward jump. They explicitly decompose this measure into two components

$$\begin{aligned} \Delta J_t^{2+} &= (RS_t^+ - RS_t^-) I_{\{(RS_t^+ - RS_t^-)>0\}}, \\ \Delta J_t^{2-} &= (RS_t^+ - RS_t^-) I_{\{(RS_t^+ - RS_t^-)<0\}}. \end{aligned} \tag{A.1}$$

If jumps are rare then these measures should respectively correspond to the jump variation when either a positive or a negative jump occurs. However, if jumps tend to cluster, there might be more than one jump in a day and the measures in Equation A.1 will fail to capture the whole jump contribution. To verify whether the actual size of the jumps could still generate valuable information, we extend the model of Patton & Sheppard (2015) to obtain the following HAR-SJ*aug* model,

$$\begin{aligned} log(RV_{t,t+h}) &= \beta_0 + \beta_J \log(1+J) + \beta_{J+} \log(\Delta J_t^{2+} + 1) + \beta_{J-} \log(\Delta J_t^{2-} + 1) \\ &+ \beta_{JJ+} \log(1+J) * \log(\Delta J_t^{2+} + 1) + \beta_{JJ-} \log(1+J) \log(\Delta J_t^{2-} + 1) \\ &+ \beta_C \log(RV_t) + \beta_W \log(RV_{t-5,t}) + \beta_M \log(RV_{t-22,t}) + \eta_{t,t+h}. \end{aligned}$$