2T-POT Hawkes model for left- and right-tail conditional quantile forecasts of financial log returns: Out-of-sample comparison of conditional EVT models

Matthew F. Tomlinson [a,b,*], David Greenwood [c], Marcin Mucha-Kruczyński [a,d]

[a] *Department of Physics, University of Bath, Bath BA2 7AY, United Kingdom*
[b] *Centre for Networks and Collective Behaviour, University of Bath, Bath BA2 7AY, United Kingdom*
[c] *CheckRisk LLP, 4 Miles's Buildings, George Street, Bath BA1 2QS, United Kingdom*
[d] *Centre for Nanoscience and Nanotechnology, University of Bath, Bath BA2 7AY, United Kingdom*

ARTICLE INFO

ABSTRACT

Conditional extreme value theory (EVT) methods promise enhanced forecasting of the extreme tail events that often dominate systemic risk. We present an improved two-tailed peaks-over-threshold (2T-POT) Hawkes model that is adapted for conditional quantile forecasting in both the left and right tails of a univariate time series. This is applied to the daily log returns of six large-cap indices. We also take the unique step of fitting the model at multiple exceedance thresholds (from the 1.25% to 25.00% mirrored quantiles). Quantitatively similar asymmetries in Hawkes parameters are found across all six indices, adding further empirical support to a temporal leverage effect in financial price time series in which the impact of losses is not only larger but also more immediate. Out-of-sample backtests find that our 2T-POT Hawkes model is more reliably accurate than the GARCH-EVT model when forecasting (mirrored) value at risk and expected shortfall at the 5% coverage level and below. This suggests that asymmetric Hawkes-type arrival dynamics are a better approximation of the true generating process for extreme daily log returns than GARCH-type conditional volatility. Our 2T-POT Hawkes model therefore presents a better-performing alternative for financial risk modelling.

1. Introduction

A notable feature of many complex systems is that outcomes are often influenced more by rare extreme events than by more typical fluctuations (Sornette, 2006). As a result, these extreme tail events often dominate the associated systemic risk, which makes accurately forecasting them a vital objective in many disciplines. Extreme value theory (EVT) presents an approach to this problem in which asymptotic tail behaviour is modelled independently from the typical 'bulk' fluctuations, under the justification that the two are often generated by distinct mechanisms (Coles, 2001; de Haan & Ferreira, 2006). In practical terms, this means that values beyond a defined threshold are classified as extremes and are described by an exceedance distribution that is fitted independently from the sub-threshold (i.e. non-extreme or *bulk*) distribution. The simplest application is made when the generating process for extreme events is stationary. In this case, not only is the distribution of exceedances itself unconditional, but the intensity (i.e. arrival rate in time) of events from this distribution is constant. Together, these two conditions mean that arrivals of extreme events beyond any arbitrary threshold within the tail distribution

---

* Corresponding author at: Department of Physics, University of Bath, Bath BA2 7AY, United Kingdom.

E-mail address: mft28@bath.ac.uk (M.F. Tomlinson).

occur in time according to a homogeneous Poisson point process (as defined by a constant intensity). However, in some systems of interest, such as financial markets, complex dynamics and feedbacks give rise to non-stationary behaviour that means that the conditions of constant intensity and of an unconditional distribution of exceedance magnitudes cannot be assumed to hold. In such cases, extreme events tend to cluster together in time, and their arrival intensity is often correlated with their magnitude. Descriptions of these systems demand the development of conditional EVT methods that can account for these effects.

This raises a question about whether these dynamics should also be treated independently from the bulk behaviour or whether they should be related to the conditional moments of the full distribution. In the context of finance, both of these approaches have been used to describe extreme price changes, as measured by extreme log returns. A seminal example of the latter approach is the GARCH-EVT model created by McNeil and Frey (2000). This is an extension to the family of generalised autoregressive conditional heteroscedasticity (GARCH) processes – a ubiquitous class of reduced-form models in financial analysis in which log returns are described by independent and identically distributed (i.i.d.) innovations (i.e. white noise) scaled by the conditional heteroscedasticity, more often called the *volatility* (Ruppert & Matteson, 2015). The GARCH-EVT model appends a generalised Pareto (GP) distribution to the innovation distribution of the GARCH process, in order to account for the residual heavy tails that are typically observed in log returns when fitted with a standard (i.e. non-EVT) GARCH process (Cont, 2001). It follows that the intensity and magnitude of extreme events are then simply functions of the conditional volatility. Conversely, a pure conditional EVT model in which the dynamics of threshold-exceeding extremes are entirely self-contained is presented by the peaks-over-threshold (POT) Hawkes model. This combines an exceedance distribution with a self-exciting Hawkes point process to describe the time-inhomogeneous arrivals of tail events. The Hawkes point process, which first emerged as a stochastic model for the self-reflexive pattern of foreshocks and aftershocks that decorate major seismic activity (Adamopoulos, 1976; Hawkes, 1971a, 1971b; Shcherbakov et al., 2019), is defined by past events causing a time-decaying increase in the intensity of future events (Hawkes, 2018; Reinhart, 2018). This approach has found broad application to many systems characterised by activity bursts, including neural networks (Pernice et al., 2012; Tannenbaum & Burak, 2017), crime (Mohler, 2014; Mohler et al., 2018), conflict (Johnson et al., 2018; Short et al., 2014), epidemics (Chiang et al., 2022), social media (Fujita et al., 2018), and financial markets (Bacry et al., 2012, 2015; Filimonov & Sornette, 2012; Hardiman et al., 2013; Hardiman & Bouchaud, 2014; Hawkes, 2018, 2020).

Chavez-Demoulin et al. (2005) were the first to apply a POT Hawkes model to the extreme log returns of a financial price time series, which they defined as those less than the 10% in-sample quantile. In their approach, a Hawkes point process describes the arrival intensity of threshold-exceeding log returns, while the excess magnitudes of these events are described by an unconditional GP distribution. Subsequent literature has developed this work, by proposing alternative parameterisations (Bień-Barkowska, 2020; Gresnigt et al., 2015), incorporating exogenous signals (Rambaldi et al., 2015), and through multivariate extensions that describe contagious shocks between different price series (Aït-Sahalia et al., 2015; Embrechts et al., 2011; Grothe, 2014). In this literature, the POT Hawkes model has almost always been used to describe the left tail (i.e. extreme losses) exclusively, but this neglects that log returns also experience right-tail extremes (extreme gains).[1] This reflects a similarly exclusive focus on left-tail risk in the broader financial risk literature. Indeed, the two most commonly used tail-risk measures – value at risk (VaR), which is a conditional quantile for the return of an investment over a given holding period; and the expected shortfall (ES), which is the expected value of the return given that it is less than the VaR – are named and defined for the left tail only. However, this focus disguises the importance of right-tail extremes, which are highly correlated with their left-tail counterparts (Embrechts et al., 2011), and can either quickly mitigate the impact of their mirror opposites or present an equivalent risk (or opportunity) under certain investment strategies. This is especially apparent in the wake of the Covid-19 pandemic and the impact its outbreak had on global financial markets. On 12th March 2020, the S&P 500 index suffered its worst daily loss since the infamous 1987 Black Monday crash, declining by -9.5%. Five of the next 12 trading days saw losses of -12.0%, -5.2%, -4.3%, -2.9%, and -3.4%. Crucially, however, the same period also saw daily gains of $+9.3\%$, $+6.0\%$, $+9.4\%$, $+6.2\%$, and $+3.4\%$. Without these strong upswings, the index would have lost a staggering 31.1% of its value over this period rather than the 4.2% aggregate drop that it did experience.

Tomlinson et al. (2021) developed the two-tail peaks-over-threshold (2T-POT) Hawkes model to investigate the interaction of and asymmetries between left- and right-tail extremes in financial log returns. In this model, left- and right-tail threshold exceedances are described by an asymmetric self- and cross-exciting Hawkes-type arrivals process combined with asymmetric conditional GP tails. This can account for the time clustering of threshold-exceeding extremes from both tails, the correlation between the arrival intensity and the magnitude of exceedances, the heavy tailed distributions of those excess magnitudes, and the propensity for all these features to exhibit left-versus-right tail asymmetry – all of which are observed in the fluctuations of financial asset

---

[1] In the few cases where right-tail extremes have been considered in the POT Hawkes literature, assumptions of left–right symmetry have been made. Embrechts et al. (2011) treat left- and right-tail extreme log returns as distinct point processes in a bivariate Hawkes model, but they assume that the excitement produced by past events in either tail decays at the same rate (i.e. they assume symmetry in the decay vector $\boldsymbol{\beta}$, such that $\beta^{\leftarrow} = \beta^{\rightarrow}$). Gresnigt et al. (2015) apply their univariate Hawkes model to threshold exceedances in the series of absolute log returns $|X_t|$; this approach assumes that the two tails are symmetric in all respects.

prices as measured by daily log returns (Chicheportiche & Chakraborti, 2014; Cont, 2001; Davies & Krämer, 2016; Ruppert & Matteson, 2015; Tsay, 2010). Tomlinson et al. (2021) applied their model to the daily log returns of the S&P 500 (SPX), with left- and right-tail extremes defined by thresholds set at the 2.5% and 97.5% quantiles, respectively. They found that extreme daily losses and gains shared a common conditional intensity: losses were found to contribute 2.2 times as much to this intensity overall, and this contribution was found to decay 4.6 times as quickly. These two asymmetries are connected to the leverage effect – a well-known stylised fact of financial markets, which states that volatility is negatively correlated with the sign of past log returns (Cont, 2001). However, while the first of these asymmetries can be explained by conventional models of this effect, the latter cannot.[2] If this temporal aspect of the leverage effect is demonstrated to be a common property of this and other classes of financial data, it would prove a novel insight into one of the lowest-order non-zero correlations in the price signal (Chen et al., 2013). This could provide an enhanced understanding of the generating mechanism for extreme log returns from both tails, which, if exploited, could then improve forecasts of these most consequential events.

In this paper, we directly compare the 2T-POT model with a set of GARCH models, including GARCH-EVT, by testing the accuracy of one-step-ahead forecasts of conditional quantile-based risk measures for all the approaches. To achieve this, we first build on the work of Tomlinson et al. (2021) and reparameterise the 2T-POT Hawkes model so that the expected average intensity replaces the exogenous background intensity as a fitting parameter. This halves the optimisation time and enables a constraint that achieves a dimension reduction of 1 at a negligible cost to the goodness of fit. The resulting improvements to the speed and reliability of the optimization procedure enable us to fit the 2T-POT Hawkes model to the daily log returns of six international large-cap equity indices over the in-sample period, 1975-01-01 to 2015-01-01, and, in a novel step in the POT Hawkes literature, repeat this over a wide range of exceedance thresholds, from the 1.25% to 25.00% mirrored in-sample quantiles. This allows us to determine the sensitivity of our results to threshold selection. Comparable asymmetries to those reported in Tomlinson et al. (2021) are found in the fitted parameters of all six indices across a wide range of thresholds; this provides evidence that the temporal leverage effect is a universal property of this class of assets.

We further expand the 2T-POT model by extending its support to the full distribution of log returns via a subordinate bulk (i.e. intra-threshold) distribution that is conditional on the Hawkes exceedance process. This guarantees that forecasts of conditional quantile-based risk measures are always defined at all coverage levels.

We compute next-step-ahead VaR and ES forecasts for both the left and right tails and assess their accuracy and serial independence in the out-of-sample period, 2015-01-01 to 2022-09-10, through backtesting methods. This greatly expands upon similar analysis in previous literature (Bień-Barkowska, 2020; Echaust & Just, 2020; Jalal & Rockinger, 2008; Taylor, 2020), both by extending the analysis to the right tail and by evaluating forecasts over a much wider and finer range of coverage levels, from 0.25% to 15.00%. We find that our asymmetric 2T-POT Hawkes model produces the most reliably accurate forecasts in both the far left tail (5% quantile or less) and the far right tail (95% quantile or greater). This is not only of practical use, but also suggests that the asymmetric Hawkes-type arrival dynamics are a better approximation of the true generating process for extreme daily log returns than GARCH-type variance dynamics.

We present the reparameterised 2T-POT model with the subordinate bulk distribution in Section 2, where we also specify the GARCH-EVT model. The improved 2T-POT Hawkes model is fitted to the in-sample data in Section 3, where we then examine the estimated parameters across different thresholds. In Section 4, we validate and compare the accuracy and serial independence of forecasts produced by the different models using backtesting methods in the out-of-sample data. In Appendix A, we derive the reparameterisation of the 2T-POT Hawkes model in terms of the expected intensity. In turn, Appendix B defines the likelihood function and optimisation procedure for the model and shows the results of the likelihood ratio tests used for model selection.

2. Model specification

2.1. Conditional quantile-based risk measures

We first explicitly define the conditional quantile-based risk measures that are subject to validation in Section 4 so that they can be explicitly defined for each model specified later in this section. When applied to the left tail, these measures are known in the financial risk literature as value at risk and expected shortfall. However, because we generalise these to include equivalent measures for the right tail, we adopt the more generic names: *conditional quantile* and *conditional violation expectation*.

2.1.1. Conditional quantile
If a discrete stochastic time series X_t is generated according to the conditional cumulative distribution function (cdf) $F_{X,t}$, then the left-tail conditional quantile (i.e. the value at risk) at the coverage level a_q over the holding period from $t-1$ to t is

$$Q_{a_q,t}^{\leftarrow} = F_{X,t}^{-1}(a_q), \tag{1}$$

such that a fraction a_q of X_t are less than $Q_{a_q,t}^{\leftarrow}$. Violations of the left-tail conditional quantile are indicated by the left-tail violation series

$$I_{a_q,t}^{\leftarrow} = \mathcal{I}\left[-\left(X_t - Q_{a_q,t}^{\leftarrow}\right)\right], \tag{2}$$

---

[2] When the 2T-POT Hawkes model was fitted to simulated data generated by a GJR-GARCH process (defined in Section 2.3.1) the impact asymmetry found in the branching vector $\gamma^{\leftrightarrow}$ was reproduced but the time asymmetry found in the decay vector β was not (Tomlinson et al., 2021).

where

$$\mathcal{I}[x] = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0, \end{cases} \tag{3}$$

is the Heaviside step function. The right-tail conditional quantile is similarly defined:

$$Q_{a_q,t}^{\rightarrow} = F_{X,t}^{-1}(1 - a_q), \tag{4}$$

such that a fraction a_q of X_t are right-tail violations, as identified by the series $I_{a_q,t}^{\rightarrow} = \mathcal{I}[+(X_t - Q_{a_q,t}^{\rightarrow})]$.

Note that the superscripts $\leftarrow$ and $\rightarrow$ are used to denote the left and right tails, respectively. Henceforth, the superscript $\leftrightharpoons$ is used to represent either tail (i.e. either $\leftarrow$ or $\rightarrow$) in generic expressions, and the superscript $\leftrightarrow$ denotes a union of the left and right tails.

2.1.2. Conditional violation expectation

The conditional quantile has the limitation of providing no information on the distribution beyond itself. In the context of finance, this has drawn the interest of many practitioners to expected shortfall (known as left-tail conditional violation expectation, under our nomenclature) as an alternative risk measure. This has also been recognised at the level of regulation, where expected shortfall is now recommended as a risk measure by the Basel Committee on Banking Supervision (Basel Committee, 2016). The left- and right-tail conditional violation expectations are defined as the conditional expectation of X_t given a left- or right-tail violation, respectively:

$$E_{a_q,t}^{\leftrightharpoons} = \mathbb{E}\left[X_t \big| I_{a_q}^{\leftrightharpoons} = 1\right] = \pm \frac{1}{a_q} \int_{\mp\infty}^{Q_{a_q,t}^{\leftrightharpoons}} X' f_{X,t}(X') dX', \tag{5}$$

where $\mathbb{E}[.]$ is the expectation operator, and $f_{X,t} = dF_{X,t}/dX$ is the conditional probability density function (pdf) of X_t. Note that throughout this paper, F and f are used to denote cumulative distribution functions and probability density functions, respectively. These are accompanied with subscripts to denote the variable or model associated with the distribution. Conditional distributions are also indexed in discrete time with the subscript t.

2.2. 2T-POT Hawkes model

2.2.1. Hawkes exceedance model

The two-tailed peaks-over-threshold (2T-POT) Hawkes model developed in Tomlinson et al. (2021) defines two sets of exceedance events in X_t. Left-tail exceedance events are defined as the values of X_t that are less than the left-tail exceedance threshold $u^{\leftarrow}$. This series of events, which is indexed with k, is fully described by the series of left-tail excess magnitudes $\{M_k^{\leftarrow}\} = \{-(X_t-u^{\leftarrow})|-(X_t-u^{\leftarrow}) > 0\}$ and arrival times $\{t_k^{\leftarrow}\} = \{t| - (X_t - u^{\leftarrow}) > 0\}$. Similarly, right-tail exceedance events are defined with respect to the right-tail exceedance threshold $u^{\rightarrow}$, and are fully specified by the series $\{M_k^{\rightarrow}\} = \{+(X_t - u^{\rightarrow})| + (X_t - u^{\rightarrow}) > 0\}$ and $\{t_k^{\rightarrow}\} = \{t| + (X_t - u^{\rightarrow}) > 0\}$. The thresholds $\mathbf{u} = [u^{\leftarrow}, u^{\rightarrow}]^{\mathrm{T}}$ are defined symmetrically so that number of exceedance events from each tail in the sample $X_{0:T}$ is asymptotically equal, as the length of the sample T tends to infinity. This is simply achieved

by setting the thresholds equal to the value of mirrored in-sample quantiles, which is the typical approach to threshold selection in the financial POT Hawkes literature (Aït-Sahalia et al., 2015; Bień-Barkowska, 2020; Chavez-Demoulin et al., 2005; Embrechts et al., 2011; Gresnigt et al., 2015; Grothe et al., 2014). Here, these quantiles are specified by the threshold level $a_u \in [0, 0.5]$, such that the left-tail threshold is equal to the in-sample a_u-quantile, $u^{\leftarrow} = \hat{Q}_{a_u}(X_{0:T_{\text{in}}})$, and the right-tail threshold is equal to its mirrored in-sample quantile, $u^{\rightarrow} = \hat{Q}_{1-a_u}(X_{0:T_{\text{in}}})$.

In the most general description, the arrivals of left- and right-tail exceedance events are counted by two distinct point processes, which can be viewed as the components of the bivariate point process $\mathbf{N}(t) = [N^{\leftarrow}(t), N^{\rightarrow}(t)]^{\mathrm{T}}$, such that

$$dN^{\leftrightharpoons}(t) = \sum_k \delta\left(t - t_k^{\leftrightharpoons}\right), \tag{6}$$

where $\delta(t')$ is the Dirac delta function. The arrival rate at time t of events in either point process is the conditional intensity for that process:

$$\lambda^{\leftrightharpoons}\left(t\big|\mathcal{M}_t\right) = \mathbb{E}\left[\frac{dN^{\leftrightharpoons}(t)}{dt}\bigg|\mathcal{M}_t\right]. \tag{7}$$

The explicit time-dependence of $\lambda^{\leftrightharpoons}(t|\mathcal{M}_t)$ specifies $N^{\leftrightharpoons}(t)$ to be an *inhomogeneous* point processes; Hawkes-type behaviour is specified by the conditional dependence on the event history up to the present time t, $\mathcal{M}_t = \mathcal{M}_{0:t} = \{(t_k^{\leftrightharpoons}, M_k^{\leftrightharpoons}) : t_k^{\leftrightharpoons} < t\}$. If $N^{\leftarrow}$ and $N^{\rightarrow}$ are treated as distinct point processes, the conditional probability of a left-tail (right-tail) exceedance event occurring at time t is

$$p_t^{\leftrightharpoons} = \Pr\{\mathcal{I}[\mp(X_t - u^{\leftrightharpoons})] = 1|\mathcal{M}_t\}$$
$$= \Pr\{dN^{\leftrightharpoons}(t) = 1|\mathcal{M}_t\}$$
$$= 1 - \exp\left[-\int_{t-1}^t \lambda^{\leftrightharpoons}(t'|\mathcal{M}_t)dt'\right]. \tag{8}$$

Note, however, that this is incompatible with the fact that the arrivals of these events in X_t are mutually exclusive. Moreover, this treatment cannot natively forbid the case where the probability of an exceedance from either tail occurring at time t is found to be greater than 1, i.e. $p_t^{\leftrightarrow} = p_t^{\leftarrow} + p_t^{\rightarrow} > 1$. It is established in Tomlinson et al. (2021) that, in the context of financial log returns, these two requirements can be enforced at an insignificant cost to the goodness of fit by using the common intensity 2T-POT Hawkes model in which both types of extremes are counted in the same common point process $N^{\leftrightarrow}(t)$, whose arrival rate is given by the one-dimensional common conditional intensity $\lambda^{\leftrightarrow}$. Each exceedance event in $N^{\leftrightarrow}(t)$ is then stochastically drawn from either tail upon arrival. Since the left- and right-tail thresholds are selected so that the average expected intensities of events from either tail, $a_{\lambda}^{\leftrightharpoons} \equiv \mathbb{E}[\lambda^{\leftrightharpoons}]$, are approximately equal ($a_{\lambda}^{\leftarrow} = a_{\lambda}^{\rightarrow} = a_{\lambda}^{\leftrightarrow}/2$), the high correlation of left- and right-tail extremes (Embrechts et al., 2011; Tomlinson et al., 2021) means that it can be assumed that each event in $N^{\leftrightarrow}(t)$ is drawn from either tail with equal probability. Under
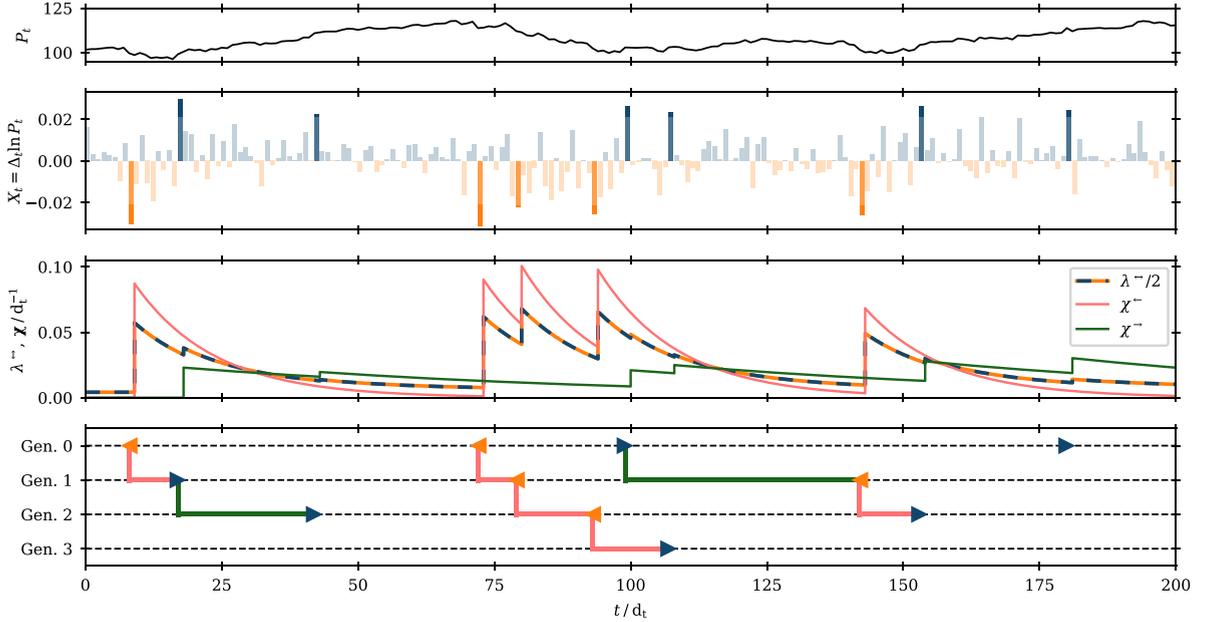
Fig. 1. Illustration of the self-exciting 2T-POT Hawkes exceedance model as a branching process. The geometric random walk discrete time series P_t (top panel) is transformed into the series of log differences, $X_t = \Delta_t \ln P_t = \ln P_t - \ln P_{t-1} = \ln P_t / P_{t-1}$ (second panel from top), with threshold-exceeding 'extreme' values marked in bolder shading. The 2T-POT Hawkes exceedance model (third panel) describes the common conditional intensity $\lambda^{\leftrightarrow}$ (light orange and dark blue dashed) of left- and right-tail exceedance events as a linear sum of the endogenous excitements χ generated by the arrival of past left- (light red) and right-tail (dark green) exceedances. This may be understood as a branching process (fourth panel) in which the daughter events in generation $n+1$ are spawned from the endogenous intensity produced by mother events in generation n.

this assumption, the conditional probability of a left-tail (right-tail) exceedance event occurring at time t is

$$p_t^{\rightleftharpoons} = \frac{1}{2} \Pr\{dN^{\leftrightarrow}(t) = 1 | \mathcal{M}_t\}$$
$$= \frac{1}{2} \left\{ 1 - \exp\left[-\int_{t-1}^{t} \lambda^{\leftrightarrow}(t'|\mathcal{M}_t) dt' \right] \right\}. \quad (9)$$

The Hawkes-type arrival dynamics of the common intensity process are constructed through a constrained bivariate model[3] that still allows for asymmetric self- and cross-excitement between asymmetric tails. This takes the form

$$\lambda^{\leftrightarrow}(t|\boldsymbol{\theta}_u; \mathcal{M}_t) = \mu^{\leftrightarrow} + \boldsymbol{\gamma}^{\leftrightarrow\mathrm{T}} \boldsymbol{\chi}(t|\boldsymbol{\theta}_u; \mathcal{M}_t), \quad (10)$$

where $\boldsymbol{\theta}_u$ is the parameter vector for the Hawkes exceedance model, $\mu^{\leftrightarrow}$ is the constant exogenous background intensity for the common arrival process, $\boldsymbol{\gamma}^{\leftrightarrow} \equiv [\gamma^{\leftrightarrow\leftarrow}, \gamma^{\leftrightarrow\rightarrow}]^{\mathrm{T}}$ is the branching vector, and $\boldsymbol{\chi}(t|\boldsymbol{\theta}_u; \mathcal{M}_t) \equiv [\chi^{\leftarrow}, \chi^{\rightarrow}]^{\mathrm{T}}$ is the vector of endogenous excitements generated by the arrivals of left- and right-tail exceedance events.

As is illustrated in Fig. 1, the self-exciting dynamics of the Hawkes process can be understood as a branching

process, in which *daughter* events are triggered by the additional endogenous intensity produced by the arrival of prior *mother* events. $\boldsymbol{\gamma}^{\leftrightarrow}$ is called the branching vector, because $\gamma^{\leftrightarrow\rightleftharpoons}$ is the mean number of daughter events in $N^{\leftrightarrow}$ that are triggered by a mother event in $N^{\rightleftharpoons}$. This is because the endogenous excitement $\chi^{\rightleftharpoons}$ is normalised such that the expected lifetime contribution of each event in $N^{\rightleftharpoons}$ to $\chi^{\rightleftharpoons}$ is 1. This normalisation also guarantees that the model is uniquely fitted. The process is sub-critical (i.e. non-explosive) provided the aggregate branching ratio $(\gamma^{\leftrightarrow\leftarrow} + \gamma^{\leftrightarrow\rightarrow})/2$ is less than 1 (Wheatley et al., 2019).

The components $\chi^{\rightleftharpoons}$ of the endogenous excitement vector $\boldsymbol{\chi}$ are the sums of contributions from all past events in each tail:

$$\chi^{\rightleftharpoons}(t|\boldsymbol{\theta}_u; \mathcal{M}_t) = \sum_{k:t_k^{\rightleftharpoons} < t} \phi^{\rightleftharpoons}(t - t_k^{\rightleftharpoons}) \kappa^{\rightleftharpoons}(M_k^{\rightleftharpoons}|t_k^{\rightleftharpoons}), \quad (11)$$

where for each component the decay kernel $\phi^{\rightleftharpoons}$ is a monotonically decreasing function of the time between the arrival of the past event, $t_k^{\rightleftharpoons}$, and the present, t. The 2T-POT Hawkes model was conceived as a parametric model: for the sake of parsimony when describing multiple tails independently, the decay kernels (along with the other functions on which the endogenous excitement depends) are described with parametric functions adapted from previous financial POT literature (Gresnigt et al., 2015; Grothe et al., 2014). The decay kernel has previously been taken as either an exponential or power law decay (Gresnigt et al., 2015). Tomlinson et al. (2021) used an exponential decay, which is scaled differently for the

---

[3] When the thresholds are selected so that $a_\lambda^{\leftarrow} = a_\lambda^{\rightarrow}$, an unconstrained bivariate model of the form $\boldsymbol{\lambda} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\chi}$ will tend to have equal rows in $\boldsymbol{\mu}$ and the branching matrix $\boldsymbol{\Gamma}$, such that $\lambda^{\leftarrow} \approx \lambda^{\rightarrow}$. Table B.4 shows that likelihood ratio test almost never finds a significant difference in goodness of fit between the full bivariate model $[H_{2,\mathrm{bi}}(a_u)]$ and the common intensity model $[H_2(a_u)]$, even in the out-of-sample period where the $a_\lambda^{\leftarrow} = a_\lambda^{\rightarrow}$ condition is not strictly guaranteed by the thresholds (which are set to mirrored in-sample quantiles).

left and right tails according to the decay vector $\boldsymbol{\beta} = [\beta^{\leftarrow}, \beta^{\rightarrow}]^{\mathsf{T}}$, such that

$$\phi^{\rightleftharpoons}(t') = \beta^{\rightleftharpoons}e^{-\beta^{\rightleftharpoons}t'}. \tag{12}$$

The advantages of this choice are that Eq. (11) can be recast in Markov form,

$$d\chi^{\rightleftharpoons} = \beta^{\rightleftharpoons}\left[-\chi^{\rightleftharpoons}dt + \kappa^{\rightleftharpoons}(M_k^{\rightleftharpoons}|t_k^{\rightleftharpoons})dN^{\rightleftharpoons}\right], \tag{13}$$

and the decay vector $\boldsymbol{\beta}$ can be used to infer the characteristic timescales over which daughter events are triggered by mother events from either tail.

The conditional impact function $\kappa^{\rightleftharpoons}(M|t)$ is a monotonically increasing function of the excess magnitude M. Following the approach of Gresnigt et al. (2015), Grothe et al. (2014), this is defined so that the intensity jump from the exceedance event arriving at time $t_k^{\rightleftharpoons}$ is determined by the contemporaneous value of the conditional cumulative distribution function of excess magnitudes for that tail:

$$\kappa^{\rightleftharpoons}(M|t) = \frac{1 - \alpha^{\rightleftharpoons}\ln\left[1 - F_{M,t}^{\mathcal{P},\rightleftharpoons}(M)\right]}{1 + \alpha^{\rightleftharpoons}}, \tag{14}$$

where the mark parameter $\alpha^{\rightleftharpoons} \geq 0$. When $\alpha^{\rightleftharpoons} \neq 0$, larger-magnitude events produce greater jumps in the excitement. This reduces sensitivity to the choice of threshold value $u^{\rightleftharpoons}$, since $\kappa^{\rightleftharpoons}(M|t) \rightarrow (1+\alpha^{\rightleftharpoons})^{-1}$ as $M \rightarrow 0$. Conversely, when $\alpha^{\rightleftharpoons} = 0$, $\kappa_t^{\rightleftharpoons}$ becomes unity and an *unmarked* Hawkes process, in which $\chi^{\rightleftharpoons}$ is independent of the magnitudes of past events, is recovered. Eq. (14) is specified so that $\mathbb{E}[\kappa^{\rightleftharpoons}(M|t)] \equiv 1$ for all values of $\alpha^{\rightleftharpoons}$ and t. Note that throughout this paper, superscript calligraphic letters symbolise parametric probability distributions: $\mathcal{N}$ for the normal distribution, $\mathcal{S}$ for the Student-t distribution, and $\mathcal{P}$ for the generalised Pareto distribution. Hence, $F_{M,t}^{\mathcal{P},\rightleftharpoons}$ in Eq. (14) denotes that the conditional cdf of excess magnitudes from either tail is described by the GP distribution.

In EVT methods, the tail is almost always described by the GP distribution. This follows from the Gnedenko–Pickands–Balkema–De Haan (GPBH) theorem, which states that the GP distribution is the limiting distribution for threshold excesses (Balkema & de Haan, 1974; Pickands, 1975; Sornette, 2006). Accordingly, we assume that the excess magnitudes are distributed according to a conditional GP distribution, as specified by the cdf:

$$F_{M,t}^{\mathcal{P},\rightleftharpoons}(M) = \begin{cases} 1 - \left[1 + \xi^{\rightleftharpoons}M/\sigma_t^{\rightleftharpoons}\right]^{-1/\xi^{\rightleftharpoons}}, & \xi^{\rightleftharpoons} \neq 0, \\ 1 - \exp\left[-M/\sigma_t^{\rightleftharpoons}\right], & \xi^{\rightleftharpoons} = 0, \end{cases} \tag{15}$$

where the shape parameter $\xi^{\rightleftharpoons}$ can specify a range of tail heaviness over three distinct phases: from the finite decay of the Weibull distribution ($\xi^{\rightleftharpoons} < 0$), through the exponential decay of the Gumbel distribution ($\xi^{\rightleftharpoons} = 0$), to the increasingly leptokurtic power-law decay of the Fréchet distribution ($\xi^{\rightleftharpoons} > 0$) (Grothe et al., 2014; Sornette, 2006). Conditional dependence on the endogenous (i.e. non-background) intensity of the Hawkes process is introduced via the conditional scale parameter:

$$\sigma_t^{\rightleftharpoons} = \varsigma^{\rightleftharpoons} + \eta^{\rightleftharpoons}\left[\lambda^{\leftrightarrow}(t) - \mu^{\leftrightarrow}\right]/2. \tag{16}$$

Thus, when $\eta^{\rightleftharpoons} > 0$, larger magnitude events become more likely when the conditional intensity of exceedances

is high, as is generally observed in financial returns (Cont, 2001). Conversely, when $\eta^{\rightleftharpoons} = 0$, the excess magnitudes are drawn from an unconditional GP distribution with a fixed scale parameter $\sigma_t^{\rightleftharpoons} = \varsigma^{\rightleftharpoons}$.

In this paper, we greatly expand the application of the 2T-POT Hawkes model to historic market data: in Tomlinson et al. (2021), the model was fitted to a single data sample (SPX) at a single threshold level, $a_u = 0.025$; in Section 3 of this paper, we not only expand this to six data samples, but, uniquely within the financial POT Hawkes literature, we also fit the model to each sample at 20 different values of the threshold level, $a_u = 0.0125k_u$, where $k_u \in \mathbb{Z} \cup [1, 20]$. Given that the computational cost of calibrating the exceedance model scales with a_u, the 120 independent calibrations in this paper are approximately 630 times more expensive in total than the single calibration performed in Tomlinson et al. (2021). This greatly expanded application of the 2T-POT Hawkes model required improvements to the speed and reliability of the optimisation procedure. We achieved this by reparameterising the exceedance model so that expected average intensity $a_\lambda^{\leftrightarrow} = \mathbb{E}[\lambda^{\leftrightarrow}]$ replaced the background intensity $\mu^{\leftrightarrow}$ as a fitting parameter, with the latter then calculated as

$$\mu^{\leftrightarrow} = [2 - (\gamma^{\leftrightarrow\leftarrow} + \gamma^{\leftrightarrow\rightarrow})]a_\lambda^{\leftrightarrow}, \tag{17}$$

as is derived in Appendix A. It is more efficient to use $a_\lambda^{\leftrightarrow}$ as a fitting parameter instead of $\mu^{\leftrightarrow}$, since the former is orthogonal to $\boldsymbol{\gamma}^{\leftrightarrow}$. When this updated model was used to reproduce the single calibration from Tomlinson et al. (2021), we found that the maximum likelihood (ML) optimisation time under the SLSQP method in SciPy (Nocedal & Wright, 2006) was reduced by 53%. More significantly, this modification also greatly reduced the number of failed optimisations in the expanded set of applications performed in this paper. Indeed, we emphasise that the large-scale analysis in Sections 3 and 4 was only made feasible because of this reparameterisation. Even greater efficiency was achieved by noting the relationship between the expected average intensity $a_\lambda^{\leftrightarrow}$ and the threshold level a_u. Because the thresholds $\mathbf{u}$ are set equal to the $(a_u, 1 - a_u)$ in-sample quantiles, it follows that the average intensity $a_\lambda^{\leftrightarrow}$ in the in-sample period is asymptotically equal to $2a_u d_t$, where d_t is the unit measuring one step in discrete time (in the case of daily log returns, d_t denotes trading days), as the size of the in-sample period T_{in} tends to infinity. Thus, $a_\lambda^{\leftrightarrow} = 2a_u d_t$ can be used either as an initial value in estimation or used as a fixed constraint. It is shown by likelihood ratio tests in Appendix B that the constraint achieves a parameter reduction of 1 at negligible cost to the goodness of fit. Accordingly, this constraint is enforced for all calibrations in this paper. We also note that the constraint reduced the total optimisation time for this paper by 12%. Having demonstrated these significant gains in optimisation speed and reliability, we strongly recommend that our reparameterisation be applied to all other Hawkes models that use the background intensity as a fitting parameter.

The common intensity 2T-POT Hawkes exceedance model is fully specified by the set of parameters, $\boldsymbol{\theta}_u = \{\boldsymbol{\gamma}^{\leftrightarrow}, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\varsigma}, \boldsymbol{\eta}, \boldsymbol{\alpha}; a_u\}$, where vector quantities are of the

form $\boldsymbol{\beta} \equiv [\beta^{\leftarrow}, \beta^{\rightarrow}]^{\mathrm{T}}$. This model is hereafter denoted as $H_2(a_u)$. Note that, if all parameters in $\boldsymbol{\theta}_u$ are constrained to be symmetric (so that the left- and right-tail components of all vector parameters are equal), then the common intensity 2T-POT model is equivalent to the classical single-tailed peaks-over-threshold Hawkes model applied to the absolute values of a copy of the original time series that is centred on the midpoint between the thresholds. That is, the set of absolute exceedances $\{M_k^{\leftrightarrow}\} = \{|X_t - (u_{\rightarrow} + u_{\leftarrow})/2| - u^{\leftrightarrow} > 0\}$, where $u^{\leftrightarrow} = (u^{\rightarrow} - u^{\leftarrow})/2$ is a union of $\{M_k^{\leftarrow}\}$ and $\{M_k^{\rightarrow}\}$, and a univariate Hawkes model applied to this exceedance series describes equal self- and cross-excitations between left and right tails that are symmetric in all properties. Hereafter, this is referred to as the symmetric 2T-POT Hawkes exceedance model and is denoted as $H_1(a_u)$.

2.2.2. Subordinate bulk distribution

It is simple to calculate the left-tail (right-tail) conditional quantile at the coverage level a_q over the holding period $t-1$ to t (i.e. to calculate $Q_{a_q,t}^{=}$) using the 2T-POT Hawkes exceedance model, provided that the conditional probability of a left-tail (right-tail) exceedance estimated by the model is greater than or equal to the coverage level, i.e. $p_t^{=} \geq a_q$. In this case, $Q_{a_q,t}^{=}$ will lie within the GP tail distribution, $F_{M,t}^{\mathcal{P},=}$, specified by the exceedance model. Thus,

$$
\begin{aligned}
a_q &= \Pr\{\mathcal{I}[\mp(X_t - u^{=})] = 1 | \mathcal{M}_t\} \\
&\quad \times \Pr\left\{\mathcal{I}\left[\mp\left(X_t - Q_{a_q}^{=}\right)\right] = 1 \Big| \mathcal{I}[\mp(X_t - u^{=})] = 1; \mathcal{M}_t\right\} \\
&= p_t^{=}\left[1 \mp \xi^{=} \frac{Q_{a_q,t}^{=} - u^{=}}{\sigma_t^{=}}\right]^{-1/\xi^{=}}.
\end{aligned} \tag{18}
$$

Hence,

$$
\left(Q_{a_q,t}^{=} \Big| p_t^{=} \geq a_q\right) = u^{=} \mp \left[\left(\frac{a_q}{p_t^{=}}\right)^{-\xi^{=}} - 1\right]\frac{\sigma_t^{=}}{\xi^{=}}. \tag{19}
$$

The conditional violation expectation over the same holding period can then be calculated as

$$
\left(E_{a_q,t}^{=} \Big| p_t^{=} \geq a_q\right) = Q_{a_q,t}^{=} \pm \frac{\xi^{=}(Q_{a_q,t}^{=} - u^{=}) - \sigma_t^{=}}{1 - \xi^{=}}. \tag{20}
$$

If, however, the left-tail (right-tail) exceedance probability is less than the coverage level (i.e. $p^{=,t} < a_q$), then the hypothetical conditional quantile would lie between the two thresholds – within the bulk of log returns not supported by either the left- or right-tail exceedance distributions of the 2T-POT Hawkes model. It is therefore desirable to extend the support of the 2T-POT Hawkes model to the full distribution of X_t. We do this by incorporating it as the jump process in a jump-diffusion-type model where the intra-threshold bulk is supported by a supplementary conditional distribution that describes diffusion. Even when solely concerned with forecasting extremes, the distribution of the intra-threshold bulk becomes increasingly pertinent as the coverage level a_q approaches the threshold level a_u from below, and even more so if forecasts are to be made over more than a single time step. This extension was essential for this

paper because the backtesting methods used in Section 4 require that $Q_{a_q,t}^{=}$ and $E_{a_q,t}^{=}$ are defined at all values of t, and this can only be guaranteed when the full distribution of X_t is supported.

Because we are interested in the Hawkes process as a pure conditional EVT model and how this compares to the conditional volatility mechanism described by the GARCH-EVT model, we take the unusual step of subordinating the diffusion model (responsible for the bulk) to the jump model (generating the extreme events). Specifically, we introduce an intra-threshold bulk distribution that is transformed by parameters that are conditional on the intensity of the Hawkes exceedance process. This subordinate bulk distribution is described by the conditional cdf $F_{B,t}^{\mathcal{D}}$, where the superscript $\mathcal{D}$ is replaced by the symbol of the specified parametric distribution ($\mathcal{N}$ for the normal distribution, and $\mathcal{S}$ for the Student-t distribution). $F_{B,t}^{\mathcal{D}}$ is transformed by the conditional location and scale parameters, m_t and s_t, so that its value at the thresholds **u** matches the probability of an exceedance event at time t, as determined by the Hawkes arrival process. These two constraints are expressed by the following equations:

$$
F_{B,t}^{\mathcal{D}}\left(u^{\leftarrow} | m_t, s_t^2; \boldsymbol{\theta}_B^D\right) = p_t^{\leftarrow}, \tag{21a}
$$
$$
1 - F_{B,t}^{\mathcal{D}}\left(u^{\rightarrow} | m_t, s_t^2; \boldsymbol{\theta}_B^D\right) = p_t^{\rightarrow}. \tag{21b}
$$

where $\boldsymbol{\theta}_B^D$ is a vector containing the additional unconstrained parameters of the chosen parametric distribution $\mathcal{D}$. Here, it becomes essential that the condition, $p_t^{\leftarrow} + p_t^{\rightarrow} \leq 1$, is strictly enforced by the use of a common conditional intensity, as described in Eq. (9). In contrast, the bivariate model described in Eq. (8) does not enforce this condition, and so it will attribute negative probability mass to the intra-threshold bulk when the components of the bivariate conditional intensity $\lambda^{=}$ are sufficiently high.

The conditional pdf for X_t at time t is then a weighted piecewise union of the bulk and tail distributions:

$$
f_{H,t}^{\mathcal{D}}(X) = \begin{cases} p_t^{\leftarrow} f_{M,t}^{\mathcal{P},\leftarrow}[-(X - u^{\leftarrow})], & X < u^{\leftarrow}, \\ f_{B,t}^{\mathcal{D}}(X | m_t, s_t), & u^{\leftarrow} \leq X \leq u^{\rightarrow}, \\ p_t^{\rightarrow} f_{M,t}^{\mathcal{P},\rightarrow}[+(X - u^{\rightarrow})], & X > u^{\rightarrow}. \end{cases} \tag{22}
$$

The fully supported 2T-POT Hawkes model at threshold level a_u is denoted as $H_2^{\mathcal{D}}(a_u)$. The equivalent fully supported symmetric 2T-POT Hawkes model is denoted as $H_1^{\mathcal{D}}(a_u)$.

2.3. GARCH-EVT model

2.3.1. GARCH

In this paper, the 2T-POT Hawkes model is compared with the family of generalised autoregressive conditional heteroscedasticity (GARCH) models. These are the standard reduced form models for log returns in industry and they are a ubiquitous baseline in the financial risk literature (Ruppert & Matteson, 2015). GARCH models have earned this status due to their parsimonious description of two key stylised facts of financial returns: volatility clustering, which states that the standard deviation of log returns (known as the volatility) is non-constant and exhibits significant positive autocorrelation; and the leverage effect, which states that volatility increases when

returns become more negative (Cont, 2001; Ruppert & Matteson, 2015). GARCH models describe a conditional volatility process in which a real univariate discrete time series is generated as

$$X_t = \mu + \sigma_t \epsilon_t^{\mathcal{D}}, \qquad (23)$$

where μ is the unconditional mean, σ_t is the conditional volatility, and $\epsilon_t^{\mathcal{D}}$ are i.i.d. random innovations drawn from a parametric distribution $\mathcal{D}$, typically with zero mean and unit variance. The GARCH(p, r, q) model describes the conditional variance σ_t^2 with an autoregressive moving average (ARMA) process:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i (\sigma_{t-i} \epsilon_{t-i})^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$
$$+ \sum_{k=1}^{r} \gamma_k (\sigma_{t-k} \epsilon_{t-k})^2 \mathcal{I}[-\epsilon_{t-k}], \qquad (24)$$

where ω is the minimum conditional variance and $\{\alpha_i, \beta_j, \gamma_k\}$ are the ARCH coefficients. When $r > 0$, Eq. (24) is a GJR-GARCH model (Glosten et al., 1993) that accounts for the leverage effect – interpreted as an asymmetric impact of negative log returns on volatility – through the Heaviside step function $\mathcal{I}$.

In this paper, we consider the GARCH($p = 1, r, q = 1$) model with ($r = 1$) and without ($r = 0$) the GJR leverage effect. For the innovation distribution, we consider the unit normal distribution ($\mathcal{D} = \mathcal{N}_{0,1}$) and the unit Student-$t$ distribution with ν degrees of freedom ($\mathcal{D} = \mathcal{S}_{0,1,\nu}$). Hereafter, we label these specifications of the GARCH model as $G_r^{\mathcal{D}}(0)$.

2.3.2. GARCH-EVT

While the standard GARCH models specified in Section 2.3.1 provide the typical benchmark for forecasting log returns, we are primarily interested in comparing the 2T-POT Hawkes model with the GARCH-EVT model (Echaust & Just, 2020; McNeil & Frey, 2000). This model is constructed by appending GP tails to the distribution of innovations, so that the (unconditional) pdf for ϵ_t now takes the form

$$f_\epsilon^{\mathcal{D}}(\epsilon) = \begin{cases} a_u f_{M_\epsilon}^{\mathcal{P}, \leftarrow}\left(-\left\{\epsilon - u_\epsilon^\leftarrow\right\}\right), & \epsilon_t < u_\epsilon^\leftarrow, \\ f_{B_\epsilon}^{\mathcal{D}}(\epsilon), & u_\epsilon^\leftarrow \leq \epsilon_t \leq u_\epsilon^\rightarrow, \\ a_u f_{M_\epsilon}^{\mathcal{P}, \rightarrow}\left(+\left\{\epsilon - u_\epsilon^\rightarrow\right\}\right), & \epsilon_t > u_\epsilon^\rightarrow, \end{cases} \quad (25)$$

where $f_{B_\epsilon}^D$ is the pdf of a continuous parametric distribution of zero mean and unit variance,[4] and the GP tail distributions $f_{M_\epsilon}^{\mathcal{P}, \rightleftharpoons}$ are parameterised equivalently to Eq. (15), except that the scale parameter remains constant and is therefore denoted as $\varsigma_\epsilon^\rightleftharpoons$. Thus,

$$f_{M_\epsilon}^{\mathcal{P}, \rightleftharpoons}(M_\epsilon) = \begin{cases} 1 - \left[1 + \xi_\epsilon^\rightleftharpoons M_\epsilon / \varsigma_\epsilon^\rightleftharpoons\right]^{-1/\xi_\epsilon^\rightleftharpoons}, & \xi_\epsilon^\rightleftharpoons \neq 0, \\ 1 - \exp\left[-M_\epsilon / \varsigma_\epsilon^\rightleftharpoons\right], & \xi_\epsilon^\rightleftharpoons = 0. \end{cases} \quad (26)$$

The innovation thresholds are set according to the threshold level a_u, such that $u_\epsilon^\leftarrow \equiv F_{B_\epsilon}^{D\ -1}(a_u)$ and $u_\epsilon^\rightarrow \equiv F_{B_\epsilon}^{D\ -1}(a_u)$. Thus, when $a_u = 0$, the standard (i.e. non-EVT)

GARCH models are recovered. Accordingly, the GARCH-EVT model is labelled as $G_r^{\mathcal{D}}(a_u)$. Estimates for the innovation tail parameters, $\hat{\xi}_\epsilon$ and $\hat{\varsigma}_\epsilon$, are obtained by ML estimation over the estimated innovations $\hat{\epsilon}_t = (X_t - \hat{\mu})/\hat{\sigma}_t$ (Mc-Neil & Frey, 2000). Note that we use hat accents to denote values that are estimated or are derived from estimates.

As well as being one of the best performing univariate risk models in the financial risk literature, the GARCH-EVT model represents an alternative conditional EVT approach compared with the 2T-POT Hawkes model. In the latter, the dynamics are solely influenced by the threshold-exceeding events. Conversely, the dynamic properties of the former are functions of the conditional volatility σ_t, which, per Eq. (24), is influenced by each innovation in ϵ_t in proportion to their magnitude. Thus, while the 2T-POT Hawkes model faithfully extends the foundational principle of EVT – that 'extreme events should speak for themselves' – to the conditional case, the GARCH-EVT model represents a compromise. The comparison of the forecasting accuracy of these two models in Section 4 therefore acts as comparison between these two distinct views of how extreme events are generated.

3. Data and empirical study

3.1. Data description

The models specified in Section 2 are applied to the daily log returns of six international large-cap equity indices: the S&P 500 (SPX) and Dow Jones Industrial Average (DJI) from the U.S.A., the DAX 30 (DAX) of Germany, the CAC 40 (CAC) of France, the Nikkei 225 (NKX) of Japan, and the Hang Seng index (HSI) of Hong Kong. These series are widely investigated financial benchmarks that are often perceived as proxies for the broader equity market in the world's three major financial centres: North America, Western Europe, and East Asia. We specifically use the daily log returns over the period beginning 1975-01-01 and ending 2022-09-10.[5] This is divided into an in-sample training period (1975-01-01 to 2015-01-01) used to calibrate the models in this section, and an out-of-sample period (2015-01-01 to 2022-09-10) used to backtest the next-step-ahead forecasts of $\hat{Q}_{a_q, t}^\rightleftharpoons$ and $\hat{E}_{a_q, t}^\rightleftharpoons$ in Section 4. The in-sample period is a 40-year span that includes among other notable episodes the 1987 Black Monday crash, the 1997–1998 Asian Financial Crisis, the 2000–2002 Tech Bubble crash, and the 2007–2008 Global Financial Crisis. The out-of-sample period encompasses more than 92 months of financial data and includes the severe fluctuations caused by the outbreak of the Covid-19 pandemic in March 2020 and by the Russian invasion of Ukraine in February 2022. The data were sourced from the stooq.com online database (Stooq, 2021); summary statistics are provided in Table 1. We note that, for both the in- and out-of-sample periods, the six series are not of the exact same length T, due to idiosyncratic holidays and suspensions.

---

[4] Note that $f_\epsilon^{\mathcal{D}}$ will have a different variance compared with $f_{B_\epsilon}^{\mathcal{D}}$.

[5] The official base date for both the DAX 30 and CAC 40 indices is 1987-12-31. However, both of these series can be extended backwards using their direct predecessors: the former DAX index and the Insee de la Bourse de Paris, respectively.

Table 1

Summary statistics for the series of daily log returns of the six large-cap international equity indices. For each sample of log returns X_t, the number of observations T, mean $\bar{X}$, standard deviation σ_X, median $\hat{Q}_{0.5}(X)$, and median absolute deviation MAD_X are given.

Stat.	In-sample (1975-01-01 – 2015-01-01)						Out-of-sample (2015-01-01 – 2022-09-10)					
	SPX	DJI	DAX	CAC	NKX	HSI	SPX	DJI	DAX	CAC	NKX	HSI
T	10092	10092	10025	9988	9461	9870	1936	1936	1944	1962	1814	1895
$\bar{X}$	3.37E−04	3.33E−04	3.09E−04	2.96E−04	1.18E−04	4.97E−04	3.52E−04	3.05E−04	1.37E−04	1.76E−04	1.50E−04	−1.05E−04
σ_X	1.09E−02	1.08E−02	1.30E−02	1.31E−02	1.33E−02	1.77E−02	1.17E−02	1.18E−02	1.30E−02	1.26E−02	1.28E−02	1.27E−02
$\hat{Q}_{0.5}(X)$	5.33E−04	4.76E−04	5.91E−04	4.23E−04	4.18E−04	6.63E−04	6.33E−04	5.68E−04	6.91E−04	7.12E−04	6.12E−04	4.85E−04
MAD_X	5.15E−03	5.25E−03	6.37E−03	6.52E−03	6.12E−03	8.00E−03	4.57E−03	4.70E−03	6.07E−03	5.76E−03	6.35E−03	6.84E−03

3.2. Threshold selection and in-sample calibration

As is discussed in Section 2.2.1, the exceedance thresholds **u** of the POT Hawkes models are determined by the choice of threshold level a_u, which specifies the quantiles at which the thresholds are set and is regarded as a tuneable parameter. In unconditional EVT applications, where an unconditional GP distribution is fitted to a stationary tail, the only concern with the threshold is that it determines the sample of excess magnitudes $M_k^{\leftrightharpoons}$. Threshold selection is therefore treated as a classic bias–variance trade-off: a less extreme threshold (i.e. higher threshold level a_u) ensures a greater number of observations in the tail, reducing noise and improving the stability of parameter estimates. However, the asymptotic nature of the GPDH theorem means that the exceedance distribution is better approximated by the GP distribution as the threshold becomes more extreme (i.e. as the threshold level a_u is lowered to 0). Diagnostic tools exist for the latter issue (Scarrott & MacDonald, 2012), and these are often used to inform threshold selection in this time-homogeneous case: for instance, in the construction of the GARCH-EVT model when fitting GP tails to the estimated innovations $\hat{e}_t$ (Echaust & Just, 2020). However, the introduction of Hawkes-type arrival dynamics adds a second layer of importance to threshold selection, because this also determines the sample of arrival times $t_k^{\leftrightharpoons}$. This affects not only the expected distribution of the conditional intensity $\lambda^{\leftrightarrow}$, but also the distribution in time of the threshold-exceeding events that provide information to the Hawkes process. While these additional effects invalidate the threshold-selection procedures used in the stationary case, they may at the same time introduce distinct phases along a_u in which the 2T-POT Hawkes model identifies signals from the generating mechanisms of the underlying system. This could be evidenced through a phase transition along a_u in the fitted parameters or in the forecasting accuracy, either of which could provide a physically meaningful definition of an extreme event in the combined context of arrivals and magnitudes. To explore this, we calibrate the threshold-based models across a wide range of threshold levels: $a_u = 0.0125 k_u$, where $k_u \in \mathbb{Z} \cup [1, 20]$. We believe this is a novel contribution in the application of POT Hawkes models to financial returns, since, to our knowledge, previous literature has only considered a single (arbitrary) threshold level per study, typically in the range $0.025 \leq a_u \leq 0.1$.

Multiple specifications of the 2T-POT Hawkes model were calibrated across this range of threshold levels using the estimation procedure described in Appendix B. Likelihood ratio tests were used to select which variant of the model would be used in the later forecasting validation: this selected for the common intensity model with the $a_u^{\leftrightarrow} = 2a_u d_t^{-1}$ constraint applied and for a Student-t distributed bulk. Tables containing the results of these likelihood ratio tests are included in Appendix B along with a graphical presentation of the estimated parameters (with standard errors) of the selected $H_2^S(a_u)$ model as a function of a_u. The fit of the exceedance model to both the arrivals process and the distribution of excess magnitudes is tested through residual analysis (Tomlinson et al., 2021). For the arrivals process, the Kolmogorov–Smirnov (KS) test is used to test the null hypothesis that the exceedance arrivals are unit Poisson in residual time,

$$\tilde{t}^{\leftrightharpoons} = \int_0^t \lambda^{\leftrightharpoons}(t')dt'. \tag{27}$$

For the excess distributions, the null hypothesis that the residual excess magnitudes

$$\tilde{m}_k^{\leftrightharpoons} = \frac{1}{\xi^{\leftrightharpoons}} \ln\left[1 + \xi^{\leftrightharpoons} \frac{m_k^{\leftrightharpoons}}{\sigma_{t_k}^{\leftrightharpoons}}\right], \tag{28}$$

are unit exponential distributed is subject to the KS test. These tests are shown graphically in Fig. 2 for the H_2 ($a_u = 0.025$) exceedance model fitted to the DAX daily log returns. Equivalent plots for the other five indices are included in the supplementary material.

A further aim of fitting the Hawkes exceedance models across a wide range of threshold levels a_u is to verify whether the asymmetric Hawkes arrival dynamics reported in Tomlinson et al. (2021) persist across this range. They fitted the 2T-POT Hawkes model to the daily log returns of the SPX over the in-sample period, 1959-10-02 to 2008-09-01, at $a_u = 0.025$ and found significant asymmetries in both the branching vector ($\hat{\gamma}_{\leftrightarrow\leftarrow}/\hat{\gamma}_{\leftrightarrow\rightarrow} = 2.2 \pm 0.5$) and the decay vector ($\hat{\beta}_{\leftarrow}/\hat{\beta}_{\rightarrow} = 4.6 \pm 1.2$) that together described a temporal leverage effect. It is anticipated that the parameters will converge to symmetry as $a_u \to 0.5$, since, in this limiting case, every value in X_t will qualify as an exceedance event. Thus, the conditional intensity would be fixed and entirely attributed to the exogenous background, making excitation from either tail equally redundant. Fig. 3 shows that, with some exceptions, the estimated asymmetries in the arrival parameters are remarkably stable along a_u and between the different indices. A slight convergence towards symmetry in both parameters is generally observed as a_u increases, but there is no evidence of a sharp transition within the investigated range. The left-over-right component ratios are quantitatively similar to those reported in Tomlinson et al. (2021) in the vast majority of cases and are never
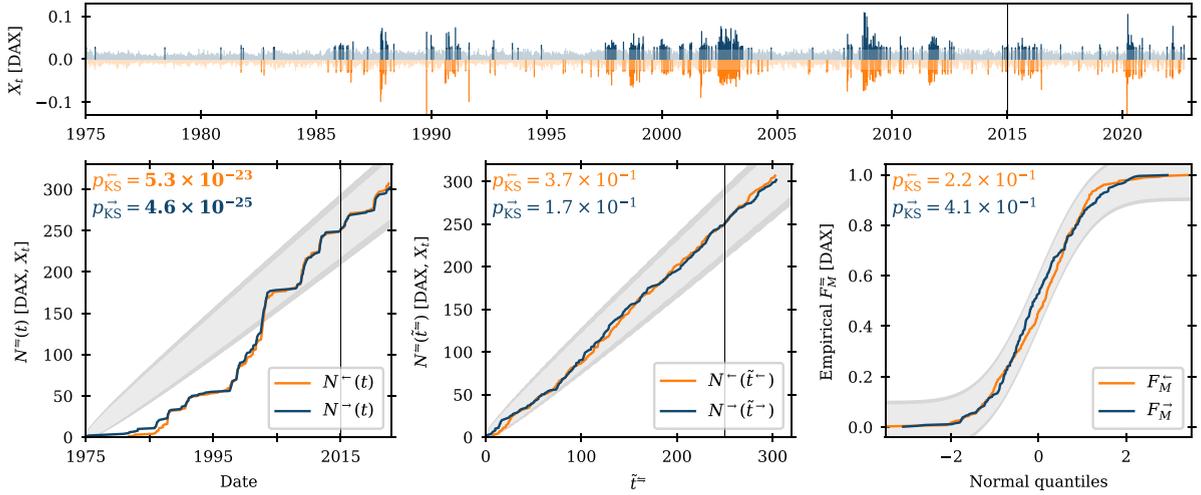
Fig. 2. $H_2(a_u = 0.025)$ Hawkes exceedance model fitted to the DAX daily log returns (top panel). The bottom panels show the KS test performed on the in-sample arrival process in time (bottom-left panel, $\mathcal{H}_0$: exceedance arrivals are Poisson in time t), the in-sample arrivals process in residual time (bottom-middle panel, $\mathcal{H}_0$: exceedance arrivals are unit Poisson in residual time $\tilde{t}^{\leftrightarrows}$), and the residual excess magnitudes (bottom-right panel, $\mathcal{H}_0$: residual excess magnitudes $\tilde{m}_k^{\leftrightarrows}$ are unit exponential random variables). The grey shaded areas show the 95% (lighter) and 99% (darker) KS confidence intervals; the KS p-values for the left- (light orange) and right-tail (dark blue) processes are also shown, with rejections at the 95% confidence level highlighted in bold. The vertical black lines mark the end of the in-sample period. Equivalents of this figure for the other indices are included in the supplementary material.
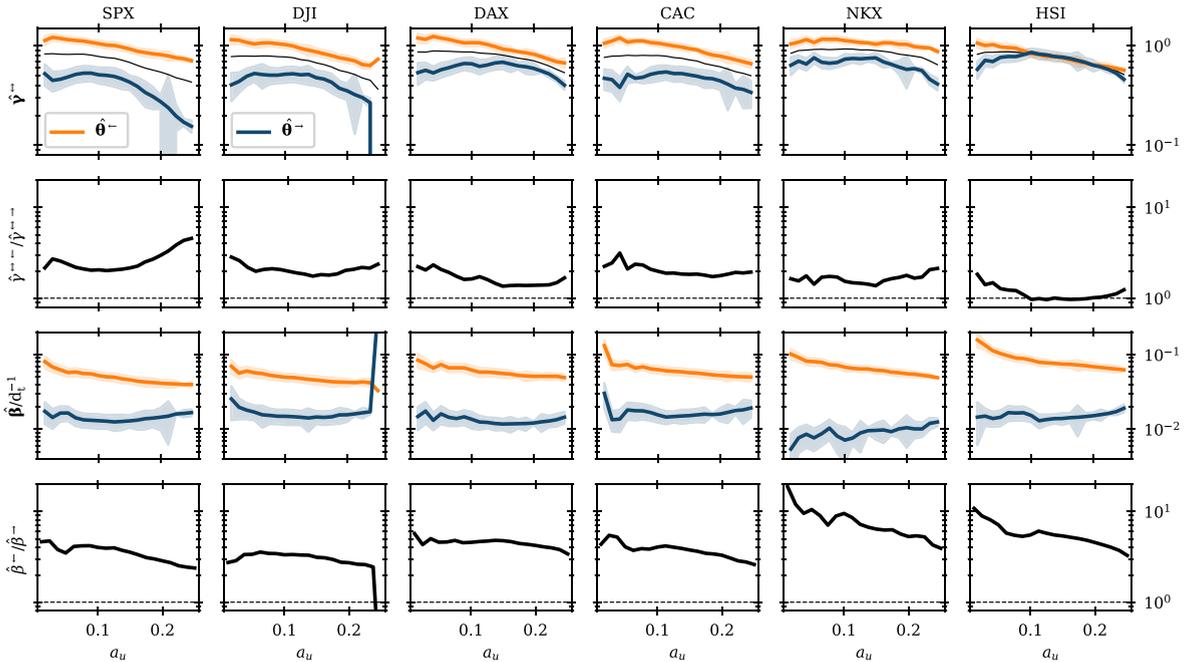
Fig. 3. Asymmetries in the Hawkes arrival parameters for $H_2(a_u)$ estimated over the in-sample period, 1975-01-01 to 2015-01-01. Estimated left- (light orange) and right-tail (dark blue) components of the branching vector $\gamma^{\leftrightarrow}$ (first row; the thin black line is the combined branching ratio) and decay constant β (third row) are shown above their respective left-over-right component ratios (second and fourth rows). Vertical axes are displayed on a log scale.

observed to invert. A notable outlier is the HSI branching vector $\hat{\gamma}^{\leftrightarrows,\text{HSI}}$, which converges to symmetry at $a_u = 0.1$.

This possibly reflects that the HSI had a notably higher average volatility in the in-sample period compared to the

other indices (as shown in Table 1), which corresponds to the fact that the HSI was a less developed market in this period and, therefore, should possibly be regarded as a separate class. It is also noted that the NKX and HSI decay constant asymmetries $\hat{\beta}^{\leftarrow}/\hat{\beta}^{\rightarrow}$ are much larger at low a_u compared to the other indices.

Overall, these results add further empirical support to hypothesis of Tomlinson et al. (2021) that there is a temporal aspect to the leverage effect observed in financial daily log returns, such that the impact of losses on future extremes is not only greater but also more immediate. This important structural feature contributes to the forecasting of exceedance arrivals, which will be seen in the relative forecasting performance of $H_2(a_u)$ compared to the fully symmetric $H_1(a_u)$ model in Section 4.

4. Backtesting of out-of-sample quantile forecasts

In this section, we use the models calibrated on the in-sample data (1975-01-01 to 2015-01-01) in Section 3 to produce next-step-ahead forecasts of the left- and right-tail conditional quantile $\hat{Q}_{a_q,t}^{\leftrightarrows}$ and conditional violation expectation $\hat{E}_{a_q,t}^{\leftrightarrows}$ in the out-of-sample period (2015-01-01 to 2022-09-10). The accuracy of these forecasts is then validated and compared through backtesting methods. The forecasts produced by the $H_2^S(a_u)$ Hawkes model are compared against those produced by the $G_1^S(a_u)$ GARCH-EVT model in order to determine which conditional EVT approach – Hawkes-type arrival dynamics, or GARCH-type volatility dynamics – provides the best description of extreme log returns. We include $H_1^S(a_u)$ in this comparison to demonstrate explicitly the benefits of incorporating asymmetry in the 2T-POT Hawkes model. We also take into account three standard GARCH models of increasing complexity: $G_0^{\mathcal{N}}(0)$, $G_0^S(0)$, and $G_1^S(0)$. Forecasts are evaluated across the range of coverage levels: $a_q = 0.0025 k_q$, where $k_q \in \mathbb{Z} \cup [1, 60]$. This is a much wider and finer range of coverage levels than has been investigated with these backtesting methods in previous financial risk literature (Bień-Barkowska, 2020; Echaust & Just, 2020; Jalal & Rockinger, 2008; Taylor, 2020). This is intended to more fully compare where the relative advantages of each model lie and to explore how this depends on the threshold level a_u. Since each of the backtesting methods are defined identically for the left and right tails, the $\leftrightarrows$ superscript is dropped from $I_{a_q,t}^{\leftrightarrows}$, $Q_{a_q,t}^{\leftrightarrows}$, and $E_{a_q,t}^{\leftrightarrows}$ in the remainder of Section 4 for clarity, unless explicitly required.

4.1. Conditional quantile backtesting methods

The first set of backtesting methods evaluates the accuracy of the conditional quantile forecasts $\hat{Q}_{a_q,t}$ by considering the series of observed violations $\hat{I}_{a_q,t}$. If the conditional quantile forecasts are perfectly accurate (i.e. $\hat{Q}_{a_q,t} = Q_{a_q,t} \forall t$), then the arrival process of violations is a Poisson point process:

$$\Pr\left\{\hat{I}_{a_q,t} = 1 \big| t\right\} = \Pr\{I_{a_q,t} = 1 \big| t\} \equiv a_q \forall t. \quad (29)$$

All of the backtesting methods in Section 4.1 derive their respective null hypotheses from Eq. (29).

4.1.1. Unconditional coverage (UC) test

The unconditional convergence (UC) test (Kupiec, 1995) assesses the null hypothesis that the observed proportion of violations $\hat{\pi}_1$ is equal to the assumed coverage level a_q ($\mathcal{H}_0 : \hat{\pi}_1 = a_q$). The UC test is formulated as a likelihood ratio test which compares two Bernoulli likelihood functions. Asymptotically, as the total number of observations in the sample, T, goes to infinity, the test statistic LR_{UC} is distributed as χ^2 with one degree of freedom:

$$\text{LR}_{\text{UC}} = -2 \ln\left[\frac{a_q^{\hat{T}_1}\left(1 - a_q\right)^{1-\hat{T}_1}}{\hat{\pi}_1^{\hat{T}_1}\left(1 - \hat{\pi}_1\right)^{1-\hat{T}_1}}\right] \sim \chi_1^2, \quad (30)$$

where $\hat{T}_1 = \sum_t \hat{I}_{a_q,t}$ is the observed number of violations in the sample of length T and $\hat{\pi}_1 = \hat{T}_1/T$. $\mathcal{H}_0$ is rejected if the conditional quantile forecasts $\hat{Q}_{a_q,t}$ have a significant bias, such that they consistently under- or overestimate the true conditional quantiles $Q_{a_q,t}$ in the sample.

4.1.2. Conditional coverage (CC) test

The conditional coverage (CC) test (Christoffersen, 1998) seeks to verify both the correct coverage and the independence of violations over consecutive observations. The process of violations is described by a first-order Markov model, with the CC test based on the 2×2 estimated transition matrix, with elements $\hat{\pi}_{ij}$ that give the estimated conditional probability of there being a violation ($i = 1$) or no violation ($i = 0$) at t given that there was ($j = 1$) or was not ($j = 0$) a violation at $t - 1$:

$$\hat{\pi}_{ij} = \hat{T}_{ij} / \left(\hat{T}_{i0} + \hat{T}_{i1}\right), \quad (31)$$

where $\hat{T}_{ij}$ is the number of observations where $\hat{I}_{a_q,t} = i | \hat{I}_{a_q,t-1} = j$. The null hypothesis of the CC test is that the conditional probabilities of a violation at t with and without a violation at $t - 1$ are both equal to the assumed coverage level, i.e. $\mathcal{H}_0 : \hat{\pi}_{10} = \hat{\pi}_{11} = a_q$. As the total number of observations T goes to infinity, the test statistic LR_{CC} is asymptotically distributed as a χ^2 with two degrees of freedom:

$$\text{LR}_{\text{CC}} = -2 \ln\left[\frac{a_q^{\hat{T}_1}\left(1 - a_q\right)^{1-\hat{T}_1}}{\left(1 - \hat{\pi}_{10}\right)^{\hat{T}_{00}} \hat{\pi}_{10}^{\hat{T}_{10}} \left(1 - \hat{\pi}_{11}\right)^{\hat{T}_{01}} \hat{\pi}_{11}^{\hat{T}_{11}}}\right] \sim \chi_2^2. \quad (32)$$

$\mathcal{H}_0$ is rejected if either the conditional quantile forecasts at time t that follow a violation at $t - 1$ ($\hat{Q}_{a_q,t} | \hat{I}_{a_q,t-1} = 1$) or that follow no violation at $t - 1$ ($\hat{Q}_{a_q,t} | \hat{I}_{a_q,t-1} = 0$) have a significant overall bias in the sample.

4.1.3. Dynamic quantile (DQ_J) test

The dynamic quantile (DQ$_J$) test (Engle & Manganelli, 2004) is designed to detect higher-order autocorrelation in the series of violations as well as dependence on other explanatory variables. The test is based on the hit function:

$$\hat{\text{Hit}}_{a_q,t} = \hat{I}_{a_q,t} - a_q. \quad (33)$$

If follows from Eqs. (29) and (33) that, under the correctly specified model, the series $\hat{\text{Hit}}_{a_q,t}$ should be i.i.d. and have

zero mean. Accordingly, $\hat{\text{Hit}}_{a_q,t}$ should be independent of lagged values of itself and also independent of the contemporaneous conditional quantile $\hat{Q}_{a_q,t}$, such that the conditional expectation of $\hat{\text{Hit}}_{a_q,t}$ should be 0, regardless of any such information available at $t-1$. The DQ_J test used here is derived as the Wald statistic from an auxiliary regression,

$$\hat{\text{Hit}}_{a_q,t} = \phi_0 + \sum_{j=1}^{J} \phi_j \hat{\text{Hit}}_{t-j} + \phi_{1+J}\hat{Q}_{a_q,t} + \epsilon_t, \qquad (34)$$

where the DQ_J null hypothesis is $\mathcal{H}_0 : \phi_k = 0 \; \forall k \in \mathbb{Z}_0 \cup [0, 1+J]$. In other words, the null states that the observed violation coverage probability is equal to the assumed coverage level ($\phi_0 = 0$) and that there is no dependence of $\hat{\text{Hit}}_{a_q,t}$ on the $1+J$ explanatory variables. The DQ test statistic is asymptotically χ^2 distributed with $2+J$ degrees of freedom:

$$\text{DQ}_J = \frac{\hat{\mathbf{Hit}}_{a_q}' \mathbf{A} \left(\mathbf{A}'\mathbf{A}\right)^{-1} \mathbf{A}' \hat{\mathbf{Hit}}_{a_q}}{a_q \left(1 - a_q\right)} \sim \chi^2_{2+J}, \qquad (35)$$

where $\hat{\mathbf{Hit}}_{a_q}$ is a $1 \times T$ vector containing the series $\hat{\text{Hit}}_{a_q,t}$ observed in the sample, and $\mathbf{A}$ is a $T \times (2+J)$ matrix comprising the observed sample series of each explanatory variable.

4.2. Conditional violation expectation backtesting methods

4.2.1. Zero mean discrepancy (ZMD) test

The conditional violation expectation forecasts $\hat{E}_{a_q,t}$ are evaluated by testing the null hypothesis that the standardised discrepancies

$$\hat{D}_{a_q,t} = \frac{X_t - \hat{E}_{a_q,t}}{\hat{Q}_{a_q,t} - \hat{Q}_{0.5,t}} \qquad (36)$$

at violations (i.e. when $\hat{I}_{a_q,t} = 1$) have zero mean (McNeil & Frey, 2000), i.e. $\mathcal{H}_0 : \sum_t \hat{D}_{a_q,t}\hat{I}_{a_q,t}/\hat{T}_1 = 0$. Assumptions about the distributions of the standardised discrepancies are avoided by employing the dependent circular block bootstrap used in Jalal and Rockinger (2008) and Politis and White (2004). The null hypothesis is subject to a two-tailed test against the alternative hypothesis, $\mathcal{H}_1 : \sum_t \hat{D}_{a_q,t}\hat{I}_{a_q,t}/\hat{T}_1 = 0$. Thus, the null is rejected when there is a significant bias in the conditional violation expectation forecasts $\hat{E}_{a_q,t}$ over the observed violations ($\hat{I}_{a_q,t} = 1$) in the sample.

4.3. Backtesting results and discussion

We present the backtesting results synoptically in Tables 2 and 3 and as full visualisations in Figs. 4–6. Tables 2 and 3 summarise the results for each backtesting method by giving the proportion of null hypothesis rejections at the 95% confidence level ($p < 0.05$, which appears as dark blue in Figs. 4–6) aggregated across all six indices within six bands of the coverage level, $a_q^0 < a_q \le a_q^1$. For the EVT models ($a_u > 0$), these tables include the results at three representative values of the threshold level, $a_u \in \{0.05, 0.1, 0.2\}$. In Table 2, the results at these

three threshold levels are aggregated together to give a single overall proportion of null rejections for each of the EVT models. These are given with the equivalent results for the standard ($a_u = 0$) GARCH models. In Table 3, the results for the EVT models at the three representative threshold levels are given independently. These tables are designed so that the relative performance of the models at each test and coverage level can be more easily compared. We are primarily focused on the performance within the lower (i.e. more extreme) coverage levels, $a_q \le 0.05$. In addition to these tables, we provide full visualisations of the original p-values as a field within the (a_q, a_u)-space. Due to the large size of these figures, we only include three in the main paper: Figs. 4 and 5 show the full results of the UC test for each of six indices, while Fig. 6 shows the results of the ZMD test for the DJI, CAC, and HSI. The equivalent visualisations for the other tests and data series are included in the supplementary material.

We first observe that the asymmetric $H_2^{\mathcal{S}}(a_u)$ model consistently produces more accurate forecasts – as measured by lower proportions of null rejections – than the symmetric $H_1^{\mathcal{S}}(a_u)$. This is true both for the aggregated results in Table 2 and when the two models are compared at the same threshold level a_u in Table 3. In the left tail, this advantage is mostly restricted to the lower coverage bands – specifically, to $a_q \le 0.05$ in the UC and DQ_4 tests, to $a_q \le 0.075$ in the CC test, and to the full range, $a_q \le 0.15$, in the ZMD test. In the right tail, $H_2^{\mathcal{S}}(a_u)$ tends to hold an advantage over $H_1^{\mathcal{S}}(a_u)$ at all values of a_q in the UC, CC, and ZMD tests, but this advantage is larger in lower coverage bands and there is no consistent trend for the DQ_4 test. It can be taken from this that there is a demonstrable gain in predictive power when asymmetries are incorporated in the 2T-POT Hawkes model, and this gain is strongest for the most extreme future events. This adds further empirical evidence for the role of these asymmetries – including those that describe the temporal leverage effect – in the generating process for extreme daily log returns in large-cap equity indices.

Now, considering the GARCH-type models in Table 2, it is observed for the left tail that increasing model complexity does tend to yield more accurate forecasts, especially within the lowest coverage level band, $0 < a_q \le 0.025$. This shows that each of the additional features – the leptokurtic innovation distribution [$G_0^{\mathcal{S}}(0)$], the leverage effect in the variance dynamics [$G_1^{\mathcal{S}}(0)$], and the asymmetric GP tails [$G_1^{\mathcal{S}}(a_u)$] – capture significant aspects of the generating process for left-tail extremes. However, for the right-tail extremes, the simplest $G_0^{\mathcal{N}}(0)$ model outperforms the two more complex standard ($a_u = 0$) GARCH models in all of the tests based on the conditional quantile (UC, CC, and DQ_4) within this lowest coverage band. This is because the Student-t distribution used for the innovations ϵ_t in $G_0^{\mathcal{S}}(0)$ and $G_1^{\mathcal{S}}(0)$ overestimates the heaviness of the right tail and so provides a worse approximation than the normal distribution. This problem is to be expected when a symmetric innovation distribution is fitted to the full distribution of observations in one step, since the fit to the right tail is compromised by the significantly heavier left tail. In contrast, the asymmetric GP tails of the GARCH-EVT model are fitted to each

Table 2

Summarised backtesting results in the out-of-sample period, 2015-01-01 to 2022-09-10. The proportions of null hypothesis rejections at the 95% confidence level ($p < 0.05$) across all six indices within coverage level bands $a_q^0 < a_q \leq a_q^1$ are shown for each model, with the results for the EVT models aggregated over three representative values of the threshold level, $a_u \in \{0.05, 0.1, 0.2\}$. Lower rejection proportions correspond to better performance on the test. Intra-tail performance rankings along each row are given in parentheses, with the first- and second-best performing models highlighted in dark and light grey, respectively.

Test	a_q^0	a_q^1	p, left-tail ←						p, right-tail →					
			$G_0^N(0)$	$G_0^S(0)$	$G_1^S(0)$	$G_1^S(a_u)$	$H_2^S(a_u)$	$H_1^S(a_u)$	$G_0^N(0)$	$G_0^S(0)$	$G_1^S(0)$	$G_1^S(a_u)$	$H_2^S(a_u)$	$H_1^S(a_u)$
UC	.000	.025	(6) 1.0	(5) .87	(4) .82	(2) .48	(1) .35	(3) .68	(2) .43	(5) .68	(6) .72	(3) .47	(1) .36	(4) .62
	.025	.050	(4) .52	(6) .77	(3) .42	(2) .40	(1) .38	(5) .64	(6) .77	(5) .72	(4) .72	(3) .66	(1) .47	(2) .56
	.050	.075	(1) .12	(5) .47	(3) .40	(2) .37	(6) .49	(4) .46	(6) 1.0	(5) .83	(4) .82	(3) .76	(1) .24	(2) .27
	.075	.100	(1) .02	(4) .38	(2) .08	(3) .22	(5) .43	(6) .44	(6) 1.0	(3) .60	(5) .75	(4) .68	(1) .15	(2) .23
	.100	.125	(4) .28	(2) .12	(1) .00	(3) .19	(6) .34	(5) .34	(6) 1.0	(4) .50	(5) .53	(3) .50	(1) .22	(2) .36
	.125	.150	(6) .47	(3) .13	(1) .03	(2) .09	(5) .44	(4) .44	(6) 1.0	(4) .50	(5) .57	(3) .50	(1) .27	(2) .37
CC	.000	.025	(6) 1.0	(5) .87	(4) .73	(2) .44	(1) .37	(3) .69	(2) .37	(5) .55	(6) .62	(3) .38	(1) .31	(4) .47
	.025	.050	(4) .65	(6) .83	(3) .47	(1) .42	(2) .43	(5) .76	(6) .68	(4) .57	(5) .60	(3) .53	(1) .37	(2) .40
	.050	.075	(1) .47	(6) .90	(2) .58	(3) .69	(4) .72	(5) .82	(6) .90	(3) .57	(5) .72	(4) .58	(2) .20	(1) .14
	.075	.100	(2) .57	(4) .80	(1) .55	(3) .63	(5) .81	(6) .84	(6) 1.0	(3) .65	(5) .82	(4) .77	(1) .11	(2) .14
	.100	.125	(2) .55	(4) .75	(1) .53	(3) .59	(6) .83	(5) .79	(6) 1.0	(4) .50	(5) .55	(3) .46	(1) .13	(2) .18
	.125	.150	(4) .53	(3) .53	(2) .50	(1) .50	(6) .77	(5) .75	(6) .83	(4) .50	(5) .53	(3) .48	(1) .19	(2) .28
DQ$_4$	.000	.025	(6) .93	(4) .83	(3) .73	(2) .58	(1) .57	(5) .84	(1) .15	(2) .17	(6) .32	(4) .28	(5) .32	(3) .28
	.025	.050	(4) .83	(6) .93	(3) .78	(2) .72	(1) .69	(5) .91	(6) .62	(4) .48	(5) .58	(1) .43	(2) .45	(3) .46
	.050	.075	(2) .83	(3) .88	(1) .83	(5) .89	(6) .92	(4) .89	(6) .70	(1) .25	(3) .53	(2) .39	(5) .64	(4) .62
	.075	.100	(2) .83	(4) .90	(1) .83	(3) .87	(5) .92	(6) .92	(6) .95	(1) .42	(3) .63	(2) .54	(5) .78	(4) .68
	.100	.125	(2) .85	(6) .95	(1) .83	(3) .88	(4) .94	(5) .94	(6) 1.0	(3) .50	(2) .50	(1) .50	(4) .78	(5) .79
	.125	.150	(4) .85	(3) .83	(2) .67	(1) .67	(5) .86	(6) .91	(5) .83	(3) .50	(2) .50	(1) .49	(4) .79	(6) .91
ZMD	.000	.025	(6) 1.0	(4) .37	(5) .45	(1) .12	(2) .13	(3) .29	(6) .33	(5) .28	(2) .20	(1) .13	(3) .24	(4) .28
	.025	.050	(6) 1.0	(4) .77	(5) .78	(3) .46	(1) .27	(2) .41	(1) .20	(5) .35	(3) .28	(2) .24	(4) .29	(6) .44
	.050	.075	(6) 1.0	(5) .83	(4) .83	(2) .48	(1) .38	(3) .58	(1) .13	(5) .53	(4) .42	(2) .33	(3) .39	(6) .74
	.075	.100	(6) 1.0	(5) .83	(4) .83	(2) .58	(1) .49	(3) .69	(1) .27	(4) .50	(5) .62	(2) .36	(3) .37	(6) .67
	.100	.125	(6) 1.0	(5) .83	(4) .83	(2) .64	(1) .52	(3) .73	(3) .35	(4) .50	(6) .65	(2) .34	(1) .33	(5) .65
	.125	.150	(6) 1.0	(5) .83	(4) .83	(2) .72	(1) .59	(3) .74	(1) .33	(4) .50	(5) .67	(2) .34	(3) .39	(6) .76

tail independently, and so the negative impact on the forecasting accuracy of $G_1^S(a_u)$ in both tails is mitigated, especially at the best performing threshold level in Table 3. This independence demonstrates a key advantage of conditional EVT methods to forecasters of extreme events.

We now compare the asymmetric 2T-POT Hawkes model, $H_2^S(a_u)$, with the GARCH-EVT model, $G_1^S(a_u)$, focusing first on the results for the UC and CC tests. We reiterate that these are simple convergence tests that verify whether the fraction of conditional quantile violations in the out-of-sample period is equal to the coverage level a_q (UC test) and also if the series $\hat{I}_{a_q,t}$ is serial independent over a single time step (CC test). These tests therefore provide the most basic measures of overall accuracy and serial independence for the next-step-ahead forecasts of the conditional quantile $\hat{Q}_{a_q,t}$. For the threshold-aggregated UC and CC test results in Table 2, $H_2^S(a_u)$ is by far the best performing model within the lowest left-tail coverage band ($a_q \leq 0.025$) and in all right-tail coverage bands ($a_q \leq 0.15$). Looking at the equivalent threshold-disaggregated results in Table 3, we see that, for the left tail, $G_1^S(0.2)$ and $H_2^S(0.05)$ stand out as the first- and second-best performing models within the lower half of coverage bands, $a_q \leq 0.075$. Over the same range of coverage bands in the right tail, $H_2^S(a_u)$ consistently outperforms $G_1^S(a_u)$ irrespective of threshold selection, though $H_2^S(0.2)$ is by far the best performing.

To better understand the summarised UC and CC test results in Tables 2 and 3, we examine the visualisation of

the full results of the UC test in Figs. 4 and 5. The equivalent visualisations for the CC test (which can be found in the supplementary material) display very similar features, meaning that the following observations are robust with respect to lag-1 violation independence. A consistent tuning relationship is observed in the left tail for the 2T-POT Hawkes models across the six indices, wherein there is an approximately linear relationship between coverage level a_q and the optimal threshold level a_u for forecasting accuracy at that coverage level, as observed by the highest average p-values and, accordingly, by the lowest density of $p < 0.05$ null hypothesis rejections within the (a_q, a_u)-space. For the DAX, NKX, and HSI, this trend is flatter, such that very accurate forecasts are produced across almost the full range of coverage levels a_q at a single (low) value of the threshold level a_u. This is reflected in Table 3, where $H_2^S(0.05)$ is the best performing Hawkes model up to $a_q \leq 0.125$. This predictable and stable dependency of forecasting accuracy on threshold selection is ideal for practical applications, since it means that the correctly tuned 2T-POT Hawkes model should produce reliably accurate forecasts of the left-tail conditional quantile $\hat{Q}_{a_q,t}^{\leftarrow}$ across a broad range of coverage levels a_q. In contrast, the pattern of forecasting accuracy for $G_1^S(a_u)$ in the left tail is much more unpredictable. Phases of poor forecasting accuracy – as identified by the rejection of the null hypothesis – show erratic dependence on threshold selection and instead tend to concentrate within ranges of coverage levels a_q that are different for each of the indices. One consistent feature (observed in all but the HSI
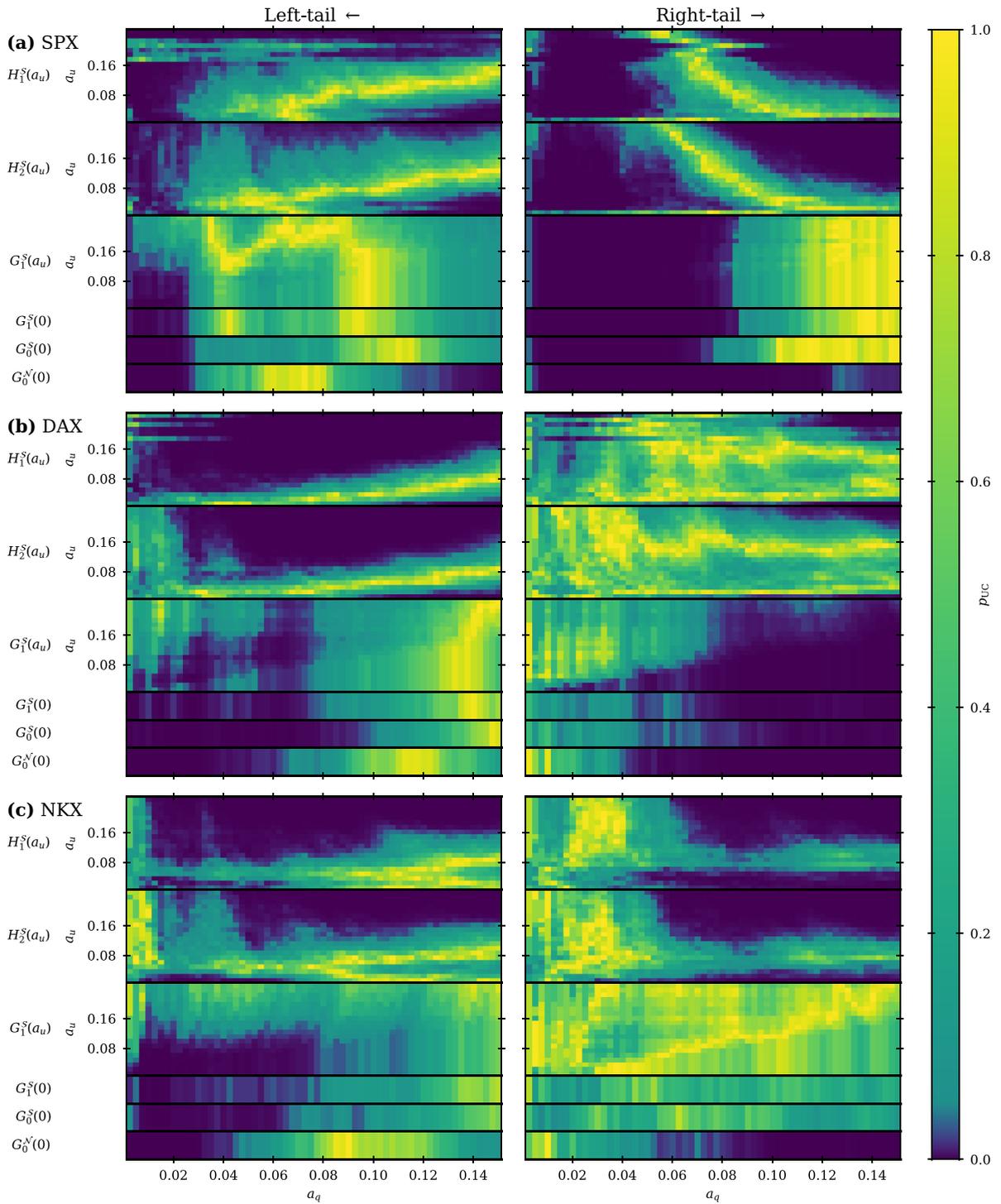
Fig. 4. Unconditional convergence test p-values p_{UC} at coverage level a_q for three of the six indices (see Fig. 5 for the other three) in the out-of-sample period, 2015-01-01 to 2022-09-10. The left- and right-column panels show the results for the left and right tails, respectively, with each panel row giving the results for a particular index: (a) SPX, (b) DAX, and (c) NKX. Within each panel, the p-values under six different models (leftmost labels) are shown: the top three (thick) strips show the p-values under the EVT models, $H_1^S(a_u)$, $H_2^S(a_u)$, and $G_1^S(a_u)$, as a function of (a_q, a_u); the bottom three (thin) strips show the p-values under the standard GARCH models, $G_1^S(0)$, $G_0^S(0)$, and $G_0^{\mathcal{N}}(0)$, as a function of a_q.
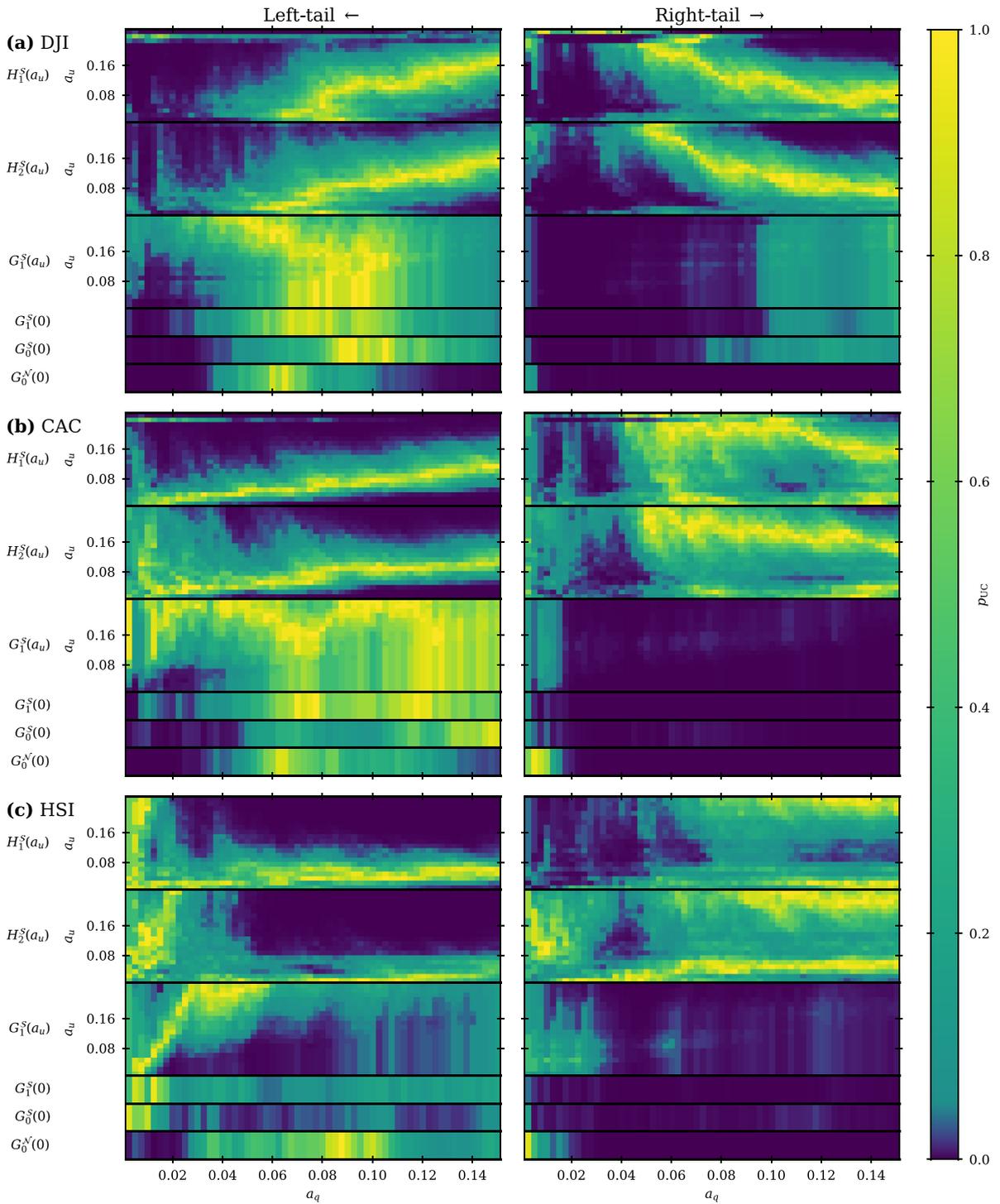
Fig. 5. Unconditional convergence test p-values p_{UC} at coverage level a_q for three of the six indices (see Fig. 4 for the other three) in the out-of-sample period, 2015-01-01 to 2022-09-10. The left- and right-column panels show the results for the left and right tails, respectively, with each panel row giving the results for a particular index: (a) DJI, (b) CAC, and (c) HSI. Within each panel, the p-values under six different models (leftmost labels) are shown: the top three (thick) strips show the p-values under the EVT models, $H_1^S(a_u)$, $H_2^S(a_u)$, and $G_1^S(a_u)$, as a function of (a_q, a_u); the bottom three (thin) strips show the p-values under the standard GARCH models, $G_1^S(0)$, $G_0^S(0)$, and $G_0^N(0)$, as a function of a_q.
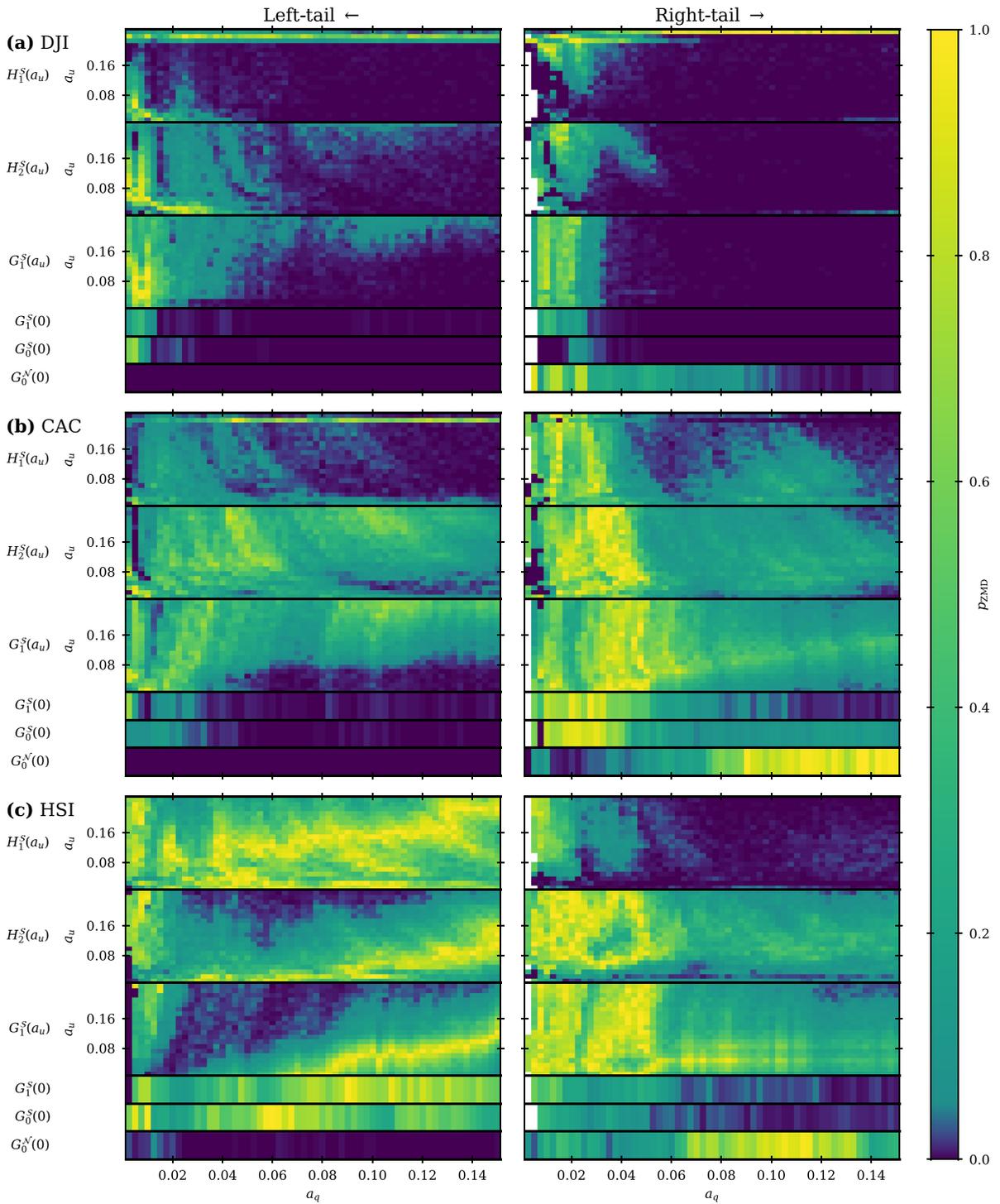
Fig. 6. Zero mean discrepancy test *p*-values p_{ZMD} at coverage level a_q for three of the six indices (see the supplementary material for the other three) in the out-of-sample period, 2015-01-01 to 2022-09-10. The left- and right-column panels show the results for the left and right tails, respectively, with each panel row giving the results for a particular index: (a) DJI, (b) CAC, and (c) HSI. Within each panel, the *p*-values under six different models (leftmost labels) are shown: the top three (thick) strips show the *p*-values under the EVT models, $H_1^S(a_u)$, $H_2^S(a_u)$, and $G_1^S(a_u)$, as a function of (a_q, a_u); the bottom three (thin) strips show the *p*-values under the standard GARCH models, $G_1^S(0)$, $G_0^S(0)$, and $G_0^N(0)$, as a function of a_q.

Table 3

Summarised backtesting results for the EVT models in the out-of-sample period, 2015-01-01 to 2022-09-10. For each of the EVT models at three values of the threshold level, $a_u \in \{0.05, 0.1, 0.2\}$, the proportions of null hypothesis rejections at the 95% confidence level ($p < 0.05$) across all six indices within coverage level bands $a_q^0 < a_q \leq a_q^1$ are shown. Lower rejection proportions correspond to better performance on the test. Intra-tail performance rankings along each row are given in parentheses, with the first- and second-best performing models highlighted in dark and light grey, respectively.

Test	a_q^0	a_q^1	p, left-tail ←								
			$G_1^S(a_u)$			$H_2^S(a_u)$			$H_1^S(a_u)$		
			$G_1^S(0.05)$	$G_1^S(0.1)$	$G_1^S(0.2)$	$H_2^S(0.05)$	$H_2^S(0.1)$	$H_2^S(0.2)$	$H_1^S(0.05)$	$H_1^S(0.1)$	$H_1^S(0.2)$
UC	.000	.025	(9) .83	(6) .57	(1) .05	(2) .28	(3) .35	(4) .42	(5) .50	(8) .77	(7) .77
	.025	.050	(6) .70	(5) .47	(1) .03	(2) .10	(4) .30	(8) .73	(3) .28	(7) .72	(9) .93
	.050	.075	(6) .48	(7) .50	(3) .13	(2) .08	(5) .42	(9) 1.0	(1) .08	(4) .42	(8) .87
	.075	.100	(4) .30	(5) .32	(3) .05	(2) .05	(6) .33	(8) .92	(1) .00	(7) .37	(9) 1.0
	.100	.125	(7) .25	(6) .25	(3) .08	(2) .05	(5) .22	(8) .77	(1) .02	(4) .15	(9) .87
	.125	.150	(3) .10	(2) .10	(1) .07	(7) .47	(5) .17	(8) .70	(6) .38	(4) .12	(9) .83
CC	.000	.025	(7) .70	(5) .45	(1) .17	(2) .33	(3) .38	(4) .40	(6) .57	(8) .73	(9) .77
	.025	.050	(7) .77	(5) .48	(1) .00	(2) .22	(3) .40	(6) .68	(4) .48	(8) .82	(9) 1.0
	.050	.075	(4) .82	(7) .87	(2) .40	(1) .30	(5) .85	(9) 1.0	(3) .60	(6) .87	(8) 1.0
	.075	.100	(3) .63	(5) .70	(1) .55	(2) .58	(6) .83	(9) 1.0	(4) .67	(7) .87	(8) 1.0
	.100	.125	(3) .60	(2) .60	(1) .58	(5) .65	(7) .85	(9) 1.0	(4) .62	(6) .77	(8) 1.0
	.125	.150	(3) .50	(2) .50	(1) .50	(5) .62	(7) .70	(9) 1.0	(6) .63	(4) .62	(8) 1.0
DQ4	.000	.025	(7) .73	(2) .55	(1) .45	(4) .57	(3) .57	(5) .58	(6) .72	(8) .87	(9) .93
	.025	.050	(7) .88	(5) .77	(1) .52	(3) .68	(2) .65	(4) .75	(6) .82	(8) .90	(9) 1.0
	.050	.075	(6) 1.0	(5) .90	(2) .80	(1) .78	(7) 1.0	(9) 1.0	(4) .83	(3) .83	(8) 1.0
	.075	.100	(4) .90	(5) .92	(2) .80	(1) .75	(9) 1.0	(8) 1.0	(3) .83	(6) .93	(7) 1.0
	.100	.125	(5) .88	(4) .88	(3) .88	(1) .82	(9) 1.0	(8) 1.0	(2) .83	(7) 1.0	(6) 1.0
	.125	.150	(3) .67	(2) .67	(1) .67	(5) .78	(4) .78	(9) 1.0	(7) .88	(6) .85	(8) 1.0
ZMD	.000	.025	(4) .13	(2) .10	(3) .12	(1) .03	(6) .20	(5) .15	(7) .27	(9) .32	(8) .28
	.025	.050	(9) .60	(7) .40	(3) .37	(1) .18	(2) .25	(5) .38	(8) .45	(6) .40	(4) .38
	.050	.075	(9) .70	(4) .38	(3) .35	(2) .35	(1) .32	(6) .47	(8) .70	(7) .58	(5) .45
	.075	.100	(9) .83	(6) .67	(1) .25	(5) .67	(2) .33	(4) .47	(8) .83	(7) .78	(3) .47
	.100	.125	(9) .83	(5) .72	(2) .38	(6) .78	(3) .42	(1) .35	(8) .83	(7) .83	(4) .53
	.125	.150	(9) .83	(6) .82	(2) .50	(5) .72	(4) .60	(1) .45	(8) .83	(7) .83	(3) .55

Test	a_q^0	a_q^1	p, right-tail →								
			$G_1^S(a_u)$			$H_2^S(a_u)$			$H_1^S(a_u)$		
			$G_1^S(0.05)$	$G_1^S(0.1)$	$G_1^S(0.2)$	$H_2^S(0.05)$	$H_2^S(0.1)$	$H_2^S(0.2)$	$H_1^S(0.05)$	$H_1^S(0.1)$	$H_1^S(0.2)$
UC	.000	.025	(5) .43	(4) .43	(6) .53	(3) .40	(2) .40	(1) .28	(7) .58	(9) .65	(8) .62
	.025	.050	(9) .68	(6) .63	(8) .67	(3) .55	(5) .62	(1) .25	(7) .67	(4) .60	(2) .40
	.050	.075	(9) .83	(8) .73	(7) .70	(6) .37	(3) .22	(1) .15	(5) .35	(4) .23	(2) .22
	.075	.100	(9) .68	(8) .68	(7) .68	(1) .07	(2) .08	(5) .30	(4) .18	(3) .13	(6) .38
	.100	.125	(9) .50	(8) .50	(7) .50	(1) .05	(2) .13	(5) .48	(3) .27	(4) .32	(6) .50
	.125	.150	(7) .50	(6) .50	(5) .50	(1) .07	(3) .17	(8) .57	(2) .17	(4) .33	(9) .62
CC	.000	.025	(5) .37	(4) .37	(6) .40	(3) .37	(2) .32	(1) .23	(8) .47	(9) .50	(7) .43
	.025	.050	(8) .53	(7) .50	(9) .55	(6) .48	(5) .48	(1) .15	(4) .48	(3) .43	(2) .28
	.050	.075	(9) .67	(8) .57	(7) .52	(6) .33	(4) .13	(3) .13	(5) .23	(1) .08	(2) .10
	.075	.100	(9) .77	(8) .77	(7) .77	(3) .05	(2) .03	(6) .23	(4) .17	(1) .02	(5) .23
	.100	.125	(9) .52	(8) .52	(5) .35	(3) .02	(1) .00	(7) .37	(4) .17	(2) .02	(6) .37
	.125	.150	(9) .50	(8) .50	(5) .43	(1) .00	(2) .12	(6) .45	(4) .20	(3) .15	(7) .48
DQ4	.000	.025	(7) .28	(6) .28	(5) .28	(2) .25	(4) .28	(9) .42	(1) .23	(3) .27	(8) .33
	.025	.050	(8) .47	(2) .40	(4) .43	(1) .40	(7) .47	(9) .48	(6) .47	(5) .47	(3) .43
	.050	.075	(3) .42	(1) .35	(2) .42	(4) .43	(7) .60	(9) .88	(6) .57	(5) .57	(8) .72
	.075	.100	(5) .58	(4) .52	(3) .52	(2) .47	(7) .87	(9) 1.0	(1) .42	(6) .63	(8) 1.0
	.100	.125	(4) .50	(3) .50	(2) .50	(1) .50	(7) .83	(9) 1.0	(5) .55	(6) .82	(8) 1.0
	.125	.150	(3) .50	(2) .50	(1) .47	(4) .65	(6) .72	(9) 1.0	(5) .72	(7) .83	(8) 1.0
ZMD	.000	.025	(3) .12	(2) .10	(5) .17	(8) .38	(6) .22	(4) .13	(9) .40	(7) .35	(1) .10
	.025	.050	(3) .23	(4) .27	(2) .22	(7) .37	(5) .33	(1) .17	(9) .50	(8) .45	(6) .37
	.050	.075	(3) .33	(2) .33	(1) .33	(6) .42	(4) .37	(5) .40	(7) .57	(8) .70	(9) .95
	.075	.100	(3) .33	(6) .42	(2) .33	(1) .33	(4) .37	(5) .42	(7) .50	(8) .55	(9) 1.0
	.100	.125	(5) .33	(6) .35	(4) .33	(3) .33	(2) .33	(1) .33	(8) .50	(7) .50	(9) .95
	.125	.150	(4) .35	(2) .33	(3) .35	(5) .40	(1) .33	(6) .43	(8) .63	(7) .63	(9) 1.0

data) is that $G_1^S(a_u)$ produces a phase of poor forecasting in the left tail when both a_q and a_u are at their lowest. Consequently, the forecasting accuracy of the GARCH-EVT model with respect to the left-tail conditional quantile $\hat{Q}_{a_q,t}^{\leftarrow}$ (i.e. value at risk) at the most extreme coverage

levels ($a_q \leq 0.025$) is much more sensitive to threshold selection than that of the 2T-POT Hawkes model. Hence, it is the latter that has the lowest proportion of null hypothesis rejections in the corresponding rows of Table 2. From this, we infer that the 2T-POT Hawkes model is

more reliably accurate than the GARCH-EVT model in this most crucial case.

In the right tail, the tuning relationship for the 2T-POT Hawkes models is found to be inverted, such that the most accurate forecasts for high coverage levels are achieved when the threshold level a_u is low, and vice versa. Considering the asymmetries in the arrival parameters reported in Section 3.2, this likely reflects that less extreme ($a_q > 0.05$) right-tail observations are frequently triggered by the preceding arrivals of more extreme ($a_u < 0.05$) left-tail observations. Despite this unexpected pattern, the right-tail forecasting accuracy of $H_2^S(a_u)$ retains some of the practical advantages compared to $G_1^S(a_u)$, as observed in the left tail, namely greater consistency between the different indices and approximate tuning relationships that allow for maximally effective forecasting over a wider range of coverage levels a_q. Indeed, it is noted that the aggregated right-tail performance of the GARCH-based models in the UC and CC tests in Tables 2 and 3 would be significantly worse across all coverage bands but for the outlier that is the NKX data.

Next, we consider the results for the DQ$_4$ test. As defined in Section 4.1.3, this test verifies whether the intensity of conditional quantile violations (and, therefore, the accuracy of conditional quantile forecasts $\hat{Q}_{a_q,t}$) is dependent on a set of explanatory series, including lagged values of $\hat{Q}_{a_q,t}$ and $\hat{I}_{a_q,t}$. It is expected that the DQ$_4$ test might be less favourable to the 2T-POT Hawkes model, due to its inherently less smooth response to extreme log returns. Violations at time $t-1$ ($\hat{I}_{a_q,t-1} = 1$) are also more likely to be exceedance events in $N^{\leftrightarrow}$ than are non-violations ($\hat{I}_{a_q,t-1} = 0$). This would tend to cause larger increases in the forecast conditional quantile $\hat{Q}_{a_q,t}$ at time t, due to the jump-like response of the Hawkes model to exceedances, which might then cause a negative bias in the coefficient ϕ_1 in Eq. (34). In the threshold-aggregated DQ$_4$ test results (Table 2), we find in both tails that $H_2^S(a_u)$ and $G_1^S(a_u)$ share nearly equal proportions of null hypothesis rejections in the two lowest coverage bands ($a_q \leq 0.05$), with those for the Hawkes model being slightly lower in the left tail and slightly higher in the right tail. For the higher coverage bands ($0.05 < a_q$), $H_2^S(a_u)$ begins to perform notably worse on this test relative to $G_1^S(a_u)$. In the threshold-disaggregated DQ$_4$ test results (Table 3), we again observe that the performance of $G_1^S(a_u)$ in the extreme left tail is much more sensitive to threshold selection compared with $H_2^S(a_u)$. The main implication of these results is that, at the 5% coverage level or lower, the left- and right-tail conditional quantile forecasts $\hat{Q}_{a_q,t}^{=}$ produced by the 2T-POT Hawkes model show no greater conditional bias than the forecasts produced by the GARCH-EVT model. This is an important point for practical forecasting applications, where such conditional biases might be correlated with risk, and so might inflict a disproportionate penalty. When taken with the UC and CC test results, this shows that the 2T-POT Hawkes model presents a superior alternative to GARCH-EVT for these forecasts.

Finally, we examine the results for the ZMD test, which are summarised in Tables 2 and 3 and are visualised in Fig. 6. The ZMD test is the only test that measures the accuracy of the conditional violation expectation forecasts $\hat{E}_{a_q,t}$. Consequently, it places more emphasis on the accuracy of the tail distribution – especially at the extremities – compared to the other tests. This is evident when comparing $H_2^S(a_u)$ to $H_1^S(a_u)$: the proportions of null hypothesis rejections under the former are universally lower in both tails because of the significant asymmetries found in the shape parameter $\hat{\xi}$ of the GP tail distributions (see Fig. B.7 in Appendix B). In comparison to the UC test visualised in Figs. 4 and 5, the ZMD test results visualised in Fig. 6 show much weaker sensitivity to the threshold level a_u in all cases. It is also apparent that the test is affected by the low sample size at the smallest values of a_u. Indeed, there is sometimes no defined value for the test statistic at the lowest values of a_q because there are not enough violations at this coverage level within the out-of-sample period to perform the circular block bootstrap. Notably, this is only observed in the right tail – indicating an asymmetric bias in all models that results in a systemic overestimation of the more extreme ($a_q \leq 0.05$) right-tail conditional quantiles $\hat{Q}_{a_q,t}^{\rightarrow}$ (and perhaps also a systemic underestimation of $\hat{Q}_{a_q,t}^{\leftarrow}$ within this range as well).

For the threshold-aggregated left-tail results in Table 2, ZMD is the test in which the performance of the 2T-POT Hawkes model is strongest relative to the GARCH-EVT model. The two models perform almost equally well in the lowest band ($a_q \leq 0.025$), and $H_2^S(a_u)$ is the best performing model by far in the other coverage bands. Conversely, in the right tail, $G_1^S(a_u)$ is found to have slightly lower proportions of null rejections than $H_2^S(a_u)$ in five out of six coverage bands. This is the opposite of what was observed in the UC and CC tests, where the latter held a clear advantage. One explanation is that the tests are technically performed over different samples: the UC and CC tests (which examine the conditional quantile forecasts $\hat{Q}_{a_q,t}$) consider all observations in X_t, whereas the ZMD test only considers observations that are also violations ($\hat{I}_{a_q,t} = 1$). In particular, we note that the right-tail results in the lowest coverage band ($a_q \leq 0.025$) are affected by the failure of the circular block bootstrap observed in Fig. 6. Nevertheless, these results show that the 2T-POT Hawkes model presents a very competitive forecasting tool for financial risk analysts to consider.

5. Summary and concluding remarks

We developed the 2T-POT Hawkes model of Tomlinson et al. (2021) as a conditional EVT model adapted for the forecasting of conditional quantile-based risk measures in both the left and right tails of the same real univariate discrete time series X_t. Our reparameterisation of the exceedance model in terms of the expected average intensity a_λ more than halved the optimisation time, and, when constrained by the threshold level a_u, achieved a dimension reduction of 1 at a negligible cost to the goodness of fit, as verified by in- and out-of-sample likelihood ratio tests. The resulting improvements to the speed and reliability of the optimisation procedure enabled us to greatly expand our application of the 2T-POT Hawkes

model to historic market data: (i) the model was fitted to the daily log returns of six international large-cap equity indices over the in-sample period, 1975-01-01 to 2015-01-01; and (ii) we independently fitted the model to each series using a wide range of exceedance thresholds, from the 1.25% to 25.00% mirrored in-sample quantiles. The significant asymmetries in the estimated branching vector $\hat{\gamma}^{\leftrightarrow}$ and the decay constant $\hat{\beta}$ reported in Tomlinson et al. (2021) were reproduced here and were quantitatively similar across the six indices. Moreover, these asymmetries were found to be stable over the majority of the tested threshold levels a_u. This adds further empirical support for a temporal leverage effect in which the impact of losses is not only greater but also more immediate.

By introducing a subordinate bulk distribution that is conditional on the Hawkes exceedance process, we extended the support of the 2T-POT Hawkes model to the full distribution of X_t; this guaranteed that forecasts of conditional quantile-based risk measures were always defined at all coverage levels, $a_q \in [0, 1]$. The fully supported 2T-POT Hawkes model was used to produce next-step-ahead forecasts of the left- and right-tail conditional quantiles $\hat{Q}_{a_q,t}^{\leftrightarrow}$ (value at risk) and the conditional violation expectations $\hat{E}_{a_q,t}^{\leftrightarrow}$ (expected shortfall). The accuracy and serial independence of these forecasts in the out-of-sample period, 2015-01-01 to 2022-09-10, were assessed through backtesting methods. These results were compared with those for the symmetric 2T-POT Hawkes model and for a set of GARCH models that included GARCH-EVT. This greatly expanded upon similar analysis in previous literature (Bień-Barkowska, 2020; Echaust & Just, 2020; Jalal & Rockinger, 2008; Taylor, 2020), both by extending the analysis to the right tail and by evaluating forecasts over a much wider and finer range of coverage levels, from 0.25% to 15.00%. Our asymmetric 2T-POT Hawkes model [$H_2^S(a_u)$] was found to produce the most reliably accurate conditional quantile forecasts $\hat{Q}_{a_q,t}^{\leftrightarrow}$ in both tails at the 5% coverage level or lower. Within this same coverage range, our model was also found to produce the most accurate forecasts of the left-tail conditional violation expectation $\hat{E}_{a_q,t}^{\leftarrow}$, though the right-tail forecasts $\hat{E}_{a_q,t}^{\rightarrow}$ were slightly less accurate than those produced by the GARCH-EVT model [$G_1^S(a_u)$]. The comparison with the symmetric 2T-POT Hawkes model [$H_1^S(a_u)$] confirms that the incorporation of left–right asymmetries provides a demonstrable increase in predictive power, which adds further empirical support to the significance of the temporal leverage effect. The comparison with the GARCH-EVT model indicates that asymmetric Hawkes-type arrival dynamics provide a better approximation of the true generating process for extreme log returns within large-cap equity indices than GARCH-type variance dynamics. This successfully extends the foundational principle of extreme value theory – that 'extreme events should speak for themselves' – to the conditional case.

In future work, the comparison between Hawkes- and GARCH-based EVT methods could be extended to include time-inhomogeneous processes. These could be based, for example, on the MSGARCH process developed in Ardia et al. (2018), which describes a single GARCH process that stochastically switches between multiple regimes according to a Markov transition matrix. If developed, a regime-switching variant of the 2T-POT Hawkes model could identify whether the asymmetries between the tails reported here are constant or change under different market states. In order to extend the comparison of conditional EVT methods done in this paper to the like-for-like time-inhomogeneous case, a novel merger of the MSGARCH model of Ardia et al. (2018) and GARCH-EVT model of McNeil and Frey (2000) could be developed. Other possible developments of this work include the extension of the analysis to multi-step-ahead aggregate forecasts of the same quantile-based risk measures. This is a natural application for the 2T-POT Hawkes model, since accurate single-step forecasts for one tail would have a compounding impact on the multi-step aggregate forecasts of both tails. To further improve forecasting accuracy, the 2T-POT Hawkes model could also incorporate explanatory variables as sources of non-constant exogenous intensity. Candidates relevant to the equity indices used here include volatility indices such as the CBOE volatility index (VIX). Finally, possible generating mechanisms for the universal asymmetries observed across the six indices may be explored by fitting the 2T-POT Hawkes model to the output of relevant heterogeneous agent-based models (hABMs).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix A. 2T-POT Hawkes model reparameterisation

In the univariate Hawkes process, $\lambda = \mu + \gamma\chi$, each mother event spawns γ daughter events on average. If the original (0-th) generation of events is triggered by the exogenous background intensity μ and every subsequent generation by the endogenous excitement produced by the preceding generation, then the n-th generation events will arrive with an average intensity $\gamma^n\mu$. Thus, by summation, the expected average intensity for the whole process is

$$a_\lambda \equiv \mathbb{E}[\lambda] = \sum_{n=0}^{\infty} \gamma^n\mu = \frac{\mu}{1 - \gamma}. \tag{A.1}$$

Table B.4

Likelihood ratio test *p*-values comparing the goodness of fit to threshold-exceeding log returns of the common-intensity model $H_2(a_u)$ versus the bivariate model $H_{2,\text{bi}}(a_u)$. $\mathcal{H}_0 : \ell_u[H_2(a_u)] = \ell_u[H_{2,\text{bi}}(a_u)]$. $\mathcal{H}_1 : \ell_u[H_2(a_u)] < \ell_u[H_{2,\text{bi}}(a_u)]$. Rejections of $\mathcal{H}_0$ at the 95% and 99% confidence levels are highlighted in light and dark grey, respectively.

a_u	p_{LR} | In-sample (1975-01-01 – 2015-01-01)						p_{LR} | Out-of-sample (2015-01-01 – 2022-09-10)					
	SPX	DJI	DAX	CAC	NKX	HSI	SPX	DJI	DAX	CAC	NKX	HSI
0.0125	2.5E−01	2.6E−01	4.5E−01	4.3E−01	7.3E−01	7.2E−02	5.5E−01	1.5E−01	1.0E+00	7.5E−01	9.1E−01	4.4E−01
0.0250	2.6E−01	9.8E−02	2.2E−01	1.0E−01	2.8E−01	4.1E−01	1.9E−01	1.3E−01	2.9E−01	3.7E−01	1.6E−01	7.8E−01
0.0375	2.0E−01	1.5E−02	8.9E−02	3.9E−02	7.2E−02	2.6E−02	3.8E−01	3.9E−02	1.0E−01	3.8E−01	1.0E+00	2.2E−01
0.0500	6.5E−02	1.8E−02	1.3E−01	2.1E−01	8.3E−02	2.5E−02	3.5E−01	4.5E−02	2.1E−01	2.5E−01	1.0E+00	1.4E−01
0.0625	2.5E−02	2.4E−02	2.9E−01	2.7E−01	1.8E−01	1.6E−01	4.8E−02	4.2E−02	4.2E−01	2.1E−01	9.7E−01	2.1E−01
0.0750	1.8E−02	1.1E−01	4.2E−01	1.3E−01	1.1E−02	7.7E−02	3.9E−02	1.8E−01	9.0E−01	1.4E−01	1.0E+00	1.6E−01
0.0875	8.1E−02	8.3E−02	5.9E−01	2.2E−01	5.1E−02	8.0E−02	1.2E−01	1.7E−01	7.8E−01	1.6E−01	1.0E+00	1.5E−01
0.1000	6.0E−02	1.3E−01	7.2E−01	1.7E−01	1.3E−01	2.7E−01	1.4E−01	2.7E−01	1.0E+00	9.9E−02	7.7E−01	1.5E−01
0.1125	1.9E−01	1.1E−01	8.0E−01	1.4E−01	2.8E−01	6.2E−01	2.2E−01	2.4E−01	1.0E+00	1.1E−01	9.5E−01	4.4E−01
0.1250	1.7E−01	7.5E−02	9.3E−01	1.4E−01	6.9E−01	6.8E−01	1.8E−01	2.4E−01	1.0E+00	1.6E−01	1.0E+00	6.1E−01
0.1375	8.1E−02	3.7E−02	6.9E−01	1.7E−01	6.5E−01	8.2E−01	1.6E−01	1.6E−01	5.5E−01	2.3E−01	1.0E+00	7.0E−01
0.1500	2.4E−01	6.1E−02	7.1E−01	1.8E−01	4.4E−01	8.5E−01	3.2E−01	3.0E−01	6.4E−01	1.6E−01	1.0E+00	1.0E+00
0.1625	2.6E−01	6.7E−02	6.6E−01	2.5E−01	1.8E−01	8.8E−01	5.1E−01	4.1E−01	5.6E−01	2.0E−01	1.0E+00	8.2E−01
0.1750	2.2E−01	1.0E−01	8.3E−01	3.2E−01	4.7E−01	8.4E−01	9.0E−01	3.3E−01	9.0E−01	2.3E−01	1.0E+00	1.0E+00
0.1875	1.5E−01	2.5E−01	8.1E−01	3.4E−01	3.7E−01	7.0E−01	8.5E−01	5.7E−01	8.2E−01	2.3E−01	1.0E+00	1.0E+00
0.2000	1.1E−01	1.5E−01	7.7E−01	3.1E−01	3.4E−01	7.7E−01	1.0E+00	5.2E−01	7.8E−01	2.2E−01	1.0E+00	1.0E+00
0.2125	8.2E−02	1.2E−01	7.1E−01	3.0E−01	6.3E−01	7.6E−01	1.0E+00	8.4E−01	7.8E−01	2.8E−01	1.0E+00	1.0E+00
0.2250	1.0E−01	1.3E−01	4.9E−01	2.7E−01	6.0E−01	6.8E−01	1.0E+00	6.1E−01	5.9E−01	3.2E−01	1.0E+00	1.0E+00
0.2375	1.0E+00	2.4E−01	4.6E−01	2.2E−01	9.9E−01	5.9E−01	1.0E+00	5.7E−01	5.4E−01	1.9E−01	9.9E−01	1.0E+00
0.2500	8.4E−02	1.2E−06	6.3E−01	2.4E−01	9.9E−01	5.1E−01	1.0E+00	3.4E−01	7.3E−01	2.6E−01	1.0E+00	1.0E+00

In Eq. (A.1), the infinite sum $\sum_{n=0}^{\infty} \gamma^n = (1-\gamma)^{-1}$ is the average number of events in a full lineage. Equivalently, it may be said that $1 - \gamma$ is the proportion of all events that are in the 0-th generation. Hence,

$$\mu = (1 - \gamma)\, a_\lambda. \tag{A.2}$$

In the multivariate case, $\boldsymbol{\lambda} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\chi}$, the full lineage of the branching matrix is

$$\sum_{n=0}^{\infty} \boldsymbol{\Gamma}^n = \sum_{n=0}^{\infty} \left(\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}\right)^n = \sum_{n=0}^{\infty} \mathbf{Q}\mathbf{D}^n\mathbf{Q}^{-1}$$
$$= \mathbf{Q}\left(\mathbf{I} - \mathbf{D}\right)^{-1}\mathbf{Q}^{-1}$$
$$= \left(\mathbf{I} - \boldsymbol{\Gamma}\right)^{-1}, \tag{A.3}$$

where $\mathbf{Q}$ and $\mathbf{D}$ are the matrices of eigenvectors and eigenvalues of $\boldsymbol{\Gamma}$, respectively, and $\mathbf{I}$ is the identity matrix. By analogy with Eq. (A.1),

$$\mathbf{a}_\lambda \equiv \mathbb{E}[\boldsymbol{\lambda}] = \sum_{n=0}^{\infty} \boldsymbol{\Gamma}^n \boldsymbol{\mu} = (\mathbf{I} - \boldsymbol{\Gamma})^{-1}\,\boldsymbol{\mu}, \tag{A.4}$$

and, therefore,

$$\boldsymbol{\mu} = (\mathbf{I} - \boldsymbol{\Gamma})\,\mathbf{a}_\lambda. \tag{A.5}$$

The common intensity model is implemented as a constrained case of the bivariate process,

$$\begin{pmatrix} \lambda^\leftarrow \\ \lambda^\rightarrow \end{pmatrix} = \begin{pmatrix} \mu^\leftarrow \\ \mu^\rightarrow \end{pmatrix} + \begin{pmatrix} \gamma^{\leftarrow\leftarrow} & \gamma^{\leftarrow\rightarrow} \\ \gamma^{\rightarrow\leftarrow} & \gamma^{\rightarrow\rightarrow} \end{pmatrix} \begin{pmatrix} \chi^\leftarrow \\ \chi^\rightarrow \end{pmatrix}, \tag{A.6}$$

in which $a_\lambda^\leftarrow = a_\lambda^\rightarrow = a_\lambda^\leftrightarrow/2$, $\mu^\leftarrow = \mu^\rightarrow = \mu^\leftrightarrow/2$, and $\gamma^{\leftarrow\leftarrow} = \gamma^{\rightarrow\rightarrow} = \gamma^{\leftrightarrow\leftrightarrow}/2$. Substituting these into Eq. (A.5) yields

$$\mu^\leftrightarrow = [2 - (\gamma^{\leftrightarrow\leftarrow} + \gamma^{\leftrightarrow\rightarrow})]\, a_\lambda^\leftrightarrow. \tag{A.7}$$

Appendix B. Model estimation and selection

The log likelihood of the 2T-POT Hawkes exceedance model under the parameters $\boldsymbol{\theta}_u$ over the data $X_{0:T} = \{X_t | t \in \mathbb{Z} \cup [0, T]\}$ can be expressed as the sum over the tails:

$$\ell_u^\leftrightarrow(\boldsymbol{\theta}_u | X_{0:T}) = \sum_{i \in \{\leftarrow, \rightarrow\}} \ell_u^i(\boldsymbol{\theta}_u | X_{0:T}). \tag{B.1}$$

The log-likelihood components for each tail can be further separated:

$$\ell_u^\leftarrow(\boldsymbol{\theta}_u | X_{0:T}) = \ell_\lambda^\leftarrow(\boldsymbol{\theta}_u | X_{0:T}) + \ell_M^\leftarrow(\boldsymbol{\theta}_u | X_{0:T}), \tag{B.2}$$

where

$$\ell_\lambda^\leftarrow(\boldsymbol{\theta}_u | X_{0:T}) = -\int_0^{T-1} \lambda^\leftarrow\left(t' | \boldsymbol{\theta}_u; \mathcal{I}_{t'}\right) dt'$$
$$+ \sum_{k: t_k^\leftarrow < T} \ln\left[\lambda^\leftarrow\left(t_k^\leftarrow | \boldsymbol{\theta}_u; \mathcal{I}_{t_k^\leftarrow}\right)\right], \tag{B.3}$$

$$\ell_M^\leftarrow(\boldsymbol{\theta}_u | X_{0:T}) = \sum_{k: t_k^\leftarrow < T} \ln\left[f_{M,t_k^\leftarrow}^{\mathcal{P},\leftarrow}\left(M_k^\leftarrow | \boldsymbol{\theta}_u; \mathcal{I}_{t_k^\leftarrow}\right)\right], \tag{B.4}$$

are the log-likelihood components of the arrivals process and of the conditional GP tail distributions, respectively. The estimated exceedance model parameters $\hat{\boldsymbol{\theta}}_u$ are then estimated by ML estimation using the SLSQP method in SciPy (Nocedal & Wright, 2006). Any unconstrained parameters of the parametric bulk distribution (denoted by the vector $\boldsymbol{\theta}_B^{\mathcal{D}}$) are obtained in a second step, through maximisation of the log likelihood:

$$\ell_B^{\mathcal{D}}\left(\boldsymbol{\theta}_B^{\mathcal{D}} | X_{0:T}\right) = \sum_{t \notin \{t_k^\leftrightarrow\} < T} \ln\left[f_{B,t}^{\mathcal{D}}\left(X_t | \boldsymbol{\theta}_B^{\mathcal{D}}\right)\right]. \tag{B.5}$$

Fig. B.7 shows the estimated parameters of the $H_2^{\mathcal{S}}(a_u)$ model as a function of a_u. The standard errors of the
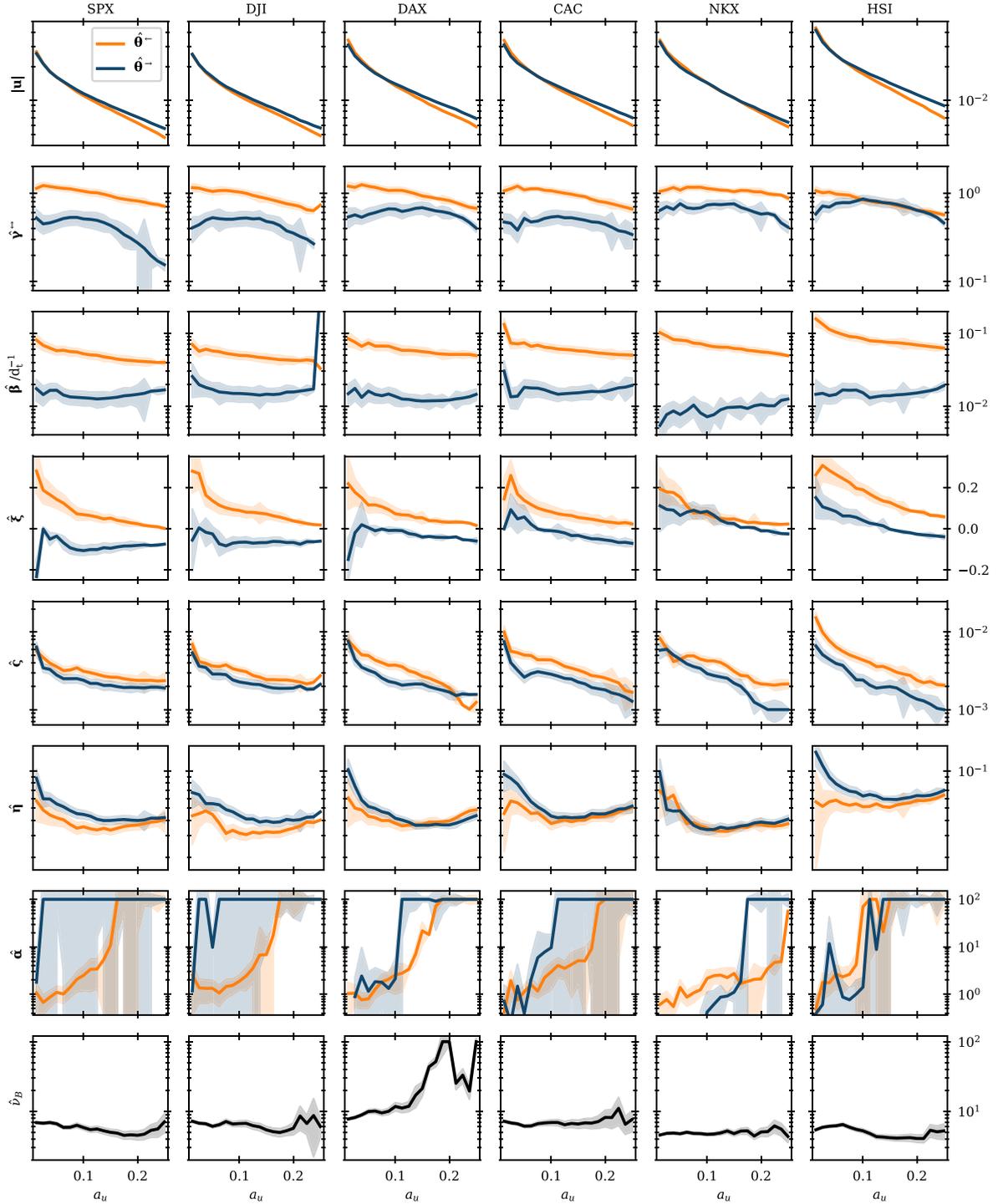
Fig. B.7. $H_2^{\mathcal{S}}(a_u)$ parameter estimates (lines) and standard errors (shaded areas) calibrated over the in-sample period, 1975-01-01 to 2015-01-01. The left- and right-tail components of the vector parameters are shown in light orange and dark blue, respectively. Vertical axes (except for those on the row describing the estimated GP shape parameter vector $\hat{\boldsymbol{\xi}}$) are displayed on a log scale.

estimated parameters are obtained by finite difference approximation of the Hessian matrix.

Alternative parameterisations of the 2T-POT Hawkes model are compared through likelihood ratio tests in order to select the appropriate parameterisation for the analysis in Section 4. At almost all values of a_u for all indices, Table B.4 shows no significant difference between the goodness of fit of the common intensity model compared with the bivariate model. This greatly expands upon the same finding in Tomlinson et al. (2021), which only

Table B.5

Likelihood ratio test *p*-values comparing the goodness of fit to threshold-exceeding returns of $H_2(a_u)$ with and without the constraint $a_\lambda^\leftrightarrow = 2a_u d_t^{-1}$.

$\mathcal{H}_0 : \ell_u\left[H_2\left(a_u \middle| a_\lambda^\leftrightarrow = 2a_u d_t^{-1}\right)\right] = \ell_u\left[H_2\left(a_u \middle| a_\lambda^\leftrightarrow \neq 2a_u d_t^{-1}\right)\right]$. $\mathcal{H}_1 : \ell_u\left[H_2\left(a_u \middle| a_\lambda^\leftrightarrow = 2a_u d_t^{-1}\right)\right] < \ell_u\left[H_2\left(a_u \middle| a_\lambda^\leftrightarrow \neq 2a_u d_t^{-1}\right)\right]$. Rejections of $\mathcal{H}_0$ at the 95% and 99% confidence levels are highlighted in light and dark grey, respectively.

a_u	p_{LR} | In-sample (1975-01-01 – 2015-01-01)						p_{LR} | Out-of-sample (2015-01-01 – 2022-09-10)					
	SPX	DJI	DAX	CAC	NKX	HSI	SPX	DJI	DAX	CAC	NKX	HSI
0.0125	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	7.9E−01	1.0E+00	8.9E−01	8.2E−01	7.2E−01
0.0250	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	8.3E−01	8.0E−01	1.0E+00	8.6E−01	8.3E−01	7.5E−01
0.0375	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	8.5E−01	7.8E−01	9.2E−01	8.6E−01	9.0E−01	7.3E−01
0.0500	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	8.2E−01	7.7E−01	1.0E+00	7.2E−01	9.3E−01	7.8E−01
0.0625	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	7.4E−01	8.9E−01	6.2E−01	1.0E+00	7.4E−01
0.0750	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	8.6E−01	6.3E−01	1.0E+00	7.1E−01
0.0875	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	5.9E−01	9.6E−01	7.1E−01
0.1000	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	9.1E−01	6.1E−01	9.4E−01	7.8E−01
0.1125	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	6.6E−01
0.1250	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	8.7E−01	1.0E+00	5.3E−01	1.0E+00	5.0E−01
0.1375	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	7.3E−01	5.3E−01	1.0E+00	1.0E+00	1.0E+00
0.1500	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	4.4E−01	1.0E+00	4.7E−01
0.1625	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	4.4E−01	1.0E+00	1.0E+00	1.0E+00	4.4E−01
0.1750	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	4.3E−01	1.0E+00	3.1E−01	1.0E+00	3.8E−01
0.1875	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	3.2E−01	1.0E+00	2.8E−01	1.0E+00	3.2E−01
0.2000	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	2.7E−01	1.0E+00	2.7E−01	1.0E+00	2.5E−01
0.2125	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	3.4E−01	8.8E−01	1.9E−01	1.0E+00	2.3E−01
0.2250	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	8.3E−01	1.9E−01	1.0E+00	1.9E−01
0.2375	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	6.3E−01	1.5E−01	1.0E+00	1.5E−01
0.2500	1.0E+00	6.4E−06	1.0E+00	1.0E+00	1.0E+00	1.0E+00	5.3E−01	5.6E−02	1.0E+00	1.3E−01	1.0E+00	1.1E−01

Table B.6

Likelihood ratio test *p*-values comparing the goodness of fit to bulk returns of $H_2^{\mathcal{N}}(a_u)$ against $H_2^{\mathcal{S}}(a_u)$. $\mathcal{H}_0 : \ell_B[H_2^{\mathcal{N}}(a_u)] = \ell_B[H_2^{\mathcal{S}}(a_u)]$. $\mathcal{H}_1 : \ell_B[H_2^{\mathcal{N}}(a_u)] < \ell_B[H_2^{\mathcal{S}}(a_u)]$. Rejections of $\mathcal{H}_0$ at the 95% and 99% confidence levels are highlighted in light and dark grey, respectively.

a_u	p_{LR}^t | In-sample (1975-01-01 – 2015-01-01)						p_{LR}^t | Out-of-sample (2015-01-01 – 2022-09-10)					
	SPX	DJI	DAX	CAC	NKX	HSI	SPX	DJI	DAX	CAC	NKX	HSI
0.0125	5.2E−174	2.7E−143	1.8E−153	1.7E−142	0.0E+00	4.3E−260	2.7E−50	8.2E−51	3.0E−33	1.4E−56	8.8E−37	1.5E−66
0.0250	4.9E−102	4.1E−90	1.0E−79	1.3E−84	5.0E−196	3.1E−132	2.1E−42	1.6E−39	6.7E−23	5.4E−36	2.9E−22	9.2E−35
0.0375	1.8E−61	6.6E−60	6.4E−48	3.2E−60	7.8E−142	6.8E−82	1.5E−31	1.1E−32	2.3E−16	7.5E−28	8.2E−18	1.7E−22
0.0500	1.0E−46	5.4E−49	7.6E−28	1.9E−51	1.1E−110	2.4E−56	2.0E−26	6.4E−28	2.1E−11	2.7E−24	1.5E−15	1.5E−18
0.0625	1.7E−43	1.6E−32	1.1E−19	7.9E−38	1.3E−82	1.8E−38	5.3E−24	1.2E−22	2.5E−09	6.1E−20	2.0E−13	1.5E−13
0.0750	1.8E−32	4.1E−20	8.2E−17	1.7E−25	1.4E−69	9.7E−35	2.3E−20	9.3E−18	1.2E−08	3.8E−15	9.5E−13	1.6E−13
0.0875	2.6E−21	5.0E−18	6.4E−12	8.5E−19	3.8E−57	1.9E−30	1.2E−15	2.8E−16	1.9E−06	6.6E−12	8.8E−10	2.2E−11
0.1000	7.9E−19	5.7E−14	2.4E−07	1.0E−14	2.6E−48	3.5E−26	4.1E−14	1.2E−13	1.8E−06	6.3E−11	9.9E−08	6.2E−10
0.1125	4.8E−17	4.9E−11	1.7E−06	1.1E−11	2.9E−37	1.1E−23	3.5E−13	1.1E−11	2.4E−06	1.7E−09	4.3E−07	4.1E−08
0.1250	3.7E−13	4.1E−10	6.5E−05	9.5E−11	6.3E−27	3.2E−23	2.0E−10	2.1E−10	3.7E−05	4.2E−08	5.9E−05	1.0E−07
0.1375	7.0E−12	5.0E−09	1.5E−02	3.2E−08	1.0E−19	1.2E−19	8.6E−10	2.4E−09	1.3E−03	5.7E−07	4.3E−04	2.2E−06
0.1500	1.1E−10	7.1E−09	8.9E−02	5.6E−07	1.3E−18	1.3E−16	3.7E−09	2.2E−08	8.6E−03	1.5E−05	6.3E−04	3.2E−05
0.1625	3.3E−09	2.7E−08	5.5E−01	1.6E−05	1.5E−12	7.5E−14	5.7E−09	1.5E−07	1.1E−01	2.4E−04	1.3E−03	7.3E−05
0.1750	8.1E−09	2.7E−07	8.2E−01	1.2E−04	5.1E−10	3.9E−11	5.7E−09	7.3E−07	1.5E−01	1.4E−03	1.3E−03	3.9E−04
0.1875	3.6E−07	1.5E−05	1.0E+00	1.0E−03	3.6E−11	3.9E−09	5.4E−08	3.6E−05	3.9E−01	5.1E−03	1.6E−02	7.0E−04
0.2000	2.6E−06	3.4E−04	1.0E+00	1.0E−02	3.3E−07	5.1E−08	7.1E−08	4.2E−04	4.5E−01	1.8E−02	6.5E−02	3.3E−03
0.2125	4.3E−05	3.8E−02	4.3E−01	2.1E−02	5.9E−06	7.1E−07	2.2E−06	4.4E−03	1.5E−01	9.9E−02	4.8E−01	3.0E−03
0.2250	1.3E−03	1.4E−02	5.9E−01	1.3E−01	9.7E−04	9.4E−04	1.2E−05	2.9E−03	2.5E−01	1.5E−01	1.0E+00	3.6E−02
0.2375	5.3E−03	9.3E−02	4.1E−01	1.7E−02	1.0E−03	1.6E−03	9.0E−05	1.6E−02	1.7E−01	1.7E−01	1.0E+00	3.2E−02
0.2500	5.1E−02	2.5E−02	1.0E+00	6.9E−02	1.3E−04	5.8E−03	2.9E−03	1.2E−02	5.8E−01	4.6E−01	1.0E+00	3.6E−02

tested the SPX at $a_u = 0.025$. This suggests that a common intensity arrivals process can be universally assumed for the extreme daily log returns of large-cap stock indices. Table B.5 shows that the $a_\lambda^\leftrightarrow = 2a_u d_t^{-1}$ constraint produces a negligible cost to the goodness of fit in all but one case out of 120 in the in-sample data. Table B.6 shows that the Student-*t* distributed bulk achieves a significantly better fit than the normal distribution in most cases. Given these results, we use the common intensity exceedance model with the $a_\lambda^\leftrightarrow = 2a_u d_t^{-1}$ constraint applied and with a Student-*t* distributed bulk in Sections 3 and 4.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2023.03.003.

References

Adamopoulos, L. (1976). Cluster models for earthquakes: Regional comparisons. *Journal of the International Association for Mathematical Geology, 8*(4), 463–475. http://dx.doi.org/10.1007/BF01028982.

Aït-Sahalia, Y., Cacho-Diaz, J., & Laeven, R. J. A. (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics, 117*(3), 585–606. http://dx.doi.org/10.1016/

j.jfineco.2015.03.002, URL https://www.sciencedirect.com/science/article/pii/S0304405X15000264.

Ardia, D., Bluteau, K., Boudt, K., & Catania, L. (2018). Forecasting risk with Markov-switching GARCH models: A large-scale performance study. *International Journal of Forecasting*, 34(4), 733–747. http://dx.doi.org/10.1016/j.ijforecast.2018.05.004, URL https://www.sciencedirect.com/science/article/pii/S0169207018300840.

Bacry, E., Dayri, K., & Muzy, J. F. (2012). Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *The European Physical Journal B*, 85(5), 157. http://dx.doi.org/10.1140/epjb/e2012-21005-8.

Bacry, E., Mastromatteo, I., & Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(1), Article 1550005. http://dx.doi.org/10.1142/S2382626615500057, URL https://www.worldscientific.com/doi/abs/10.1142/S2382626615500057.

Balkema, A. A., & de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5), 792–804. http://dx.doi.org/10.1214/aop/1176996548, URL https://projecteuclid.org/journals/annals-of-probability/volume-2/issue-5/Residual-Life-Time-at-Great-Age/10.1214/aop/1176996548.full.

Basel Committee (2016). Minimum capital requirements for Market Risk Contents. URL https://www.bis.org/bcbs/publ/d352.htm.

Bień-Barkowska, K. (2020). Looking at extremes without going to extremes: A new self-exciting probability model for extreme losses in financial markets. *Entropy*, 22(7), 789. http://dx.doi.org/10.3390/e22070789, URL https://www.mdpi.com/1099-4300/22/7/789.

Chavez-Demoulin, V., Davison, A. C., & McNeil, A. J. (2005). Estimating value-at-risk: a point process approach. *Quantitative Finance*, 5(2), 227–234. http://dx.doi.org/10.1080/14697680500039613.

Chen, J. J., Zheng, B., & Tan, L. (2013). Agent-based model with asymmetric trading and herding for complex financial systems. *PLoS ONE*, 8(11), Article e79531. http://dx.doi.org/10.1371/journal.pone.0079531, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079531.

Chiang, W.-H., Liu, X., & Mohler, G. (2022). Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *International Journal of Forecasting*, 38(2), 505–520. http://dx.doi.org/10.1016/j.ijforecast.2021.07.001, URL https://www.sciencedirect.com/science/article/pii/S0169207021001126.

Chicheportiche, R., & Chakraborti, A. (2014). Copulas and time series with long-ranged dependencies. *Physical Review E*, 89(4), Article 042117. http://dx.doi.org/10.1103/PhysRevE.89.042117, URL https://link.aps.org/doi/10.1103/PhysRevE.89.042117.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841. http://dx.doi.org/10.2307/2527341, URL http://www.jstor.org/stable/2527341.

Coles, S. (2001). *Springer series in statistics*: *vol. 208*, *An introduction to statistical modeling of extreme values*. London: Springer, http://dx.doi.org/10.1007/978-1-4471-3675-0, URL http://link.springer.com/10.1007/978-1-4471-3675-0.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2), 223–236. http://dx.doi.org/10.1080/713665670, URL https://www.tandfonline.com/doi/abs/10.1080/713665670.

Davies, L., & Krämer, W. (2016). Stylized facts and simulating long range financial data. arXiv:1612.05229, URL http://arxiv.org/abs/1612.05229.

de Haan, L., & Ferreira, A. (2006). *Springer series in operations research and financial engineering*, *Extreme value theory*. New York: Springer, http://dx.doi.org/10.1007/0-387-34471-3, URL http://link.springer.com/10.1007/0-387-34471-3.

Echaust, K., & Just, M. (2020). Value at risk estimation using the GARCH-EVT approach with optimal tail selection. *Mathematics*, 8(1), 114. http://dx.doi.org/10.3390/math8010114, URL https://www.mdpi.com/2227-7390/8/1/114.

Embrechts, P., Liniger, T., & Lin, L. (2011). Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A), 367–378. http://dx.doi.org/10.1017/S0021900200099344, URL https://www.cambridge.org/core/product/identifier/S0021900200099344/type/journal_article.

Engle, R. F., & Manganelli, S. (2004). CAViaR. *Journal of Business & Economic Statistics*, 22(4), 367–381. http://dx.doi.org/10.1198/073500104000000370.

Filimonov, V., & Sornette, D. (2012). Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5), Article 056108. http://dx.doi.org/10.1103/PhysRevE.85.056108, URL https://link.aps.org/doi/10.1103/PhysRevE.85.056108.

Fujita, K., Medvedev, A., Koyama, S., Lambiotte, R., & Shinomoto, S. (2018). Identifying exogenous and endogenous activity in social media. *Physical Review E*, 98(5), Article 052304. http://dx.doi.org/10.1103/PhysRevE.98.052304, URL https://link.aps.org/doi/10.1103/PhysRevE.98.052304.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5), 1779–1801. http://dx.doi.org/10.1111/j.1540-6261.1993.tb05128.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1993.tb05128.x.

Gresnigt, F., Kole, E., & Franses, P. H. (2015). Interpreting financial market crashes as earthquakes: A new Early Warning System for medium term crashes. *Journal of Banking & Finance*, 56, 123–139. http://dx.doi.org/10.1016/j.jbankfin.2015.03.003, URL http://www.sciencedirect.com/science/article/pii/S0378426615000643.

Grothe, O., Korniichuk, V., & Manner, H. (2014). Modeling multivariate extreme events using self-exciting point processes. *Journal of Econometrics*, 182(2), 269–289. http://dx.doi.org/10.1016/j.jeconom.2014.03.011, URL http://www.sciencedirect.com/science/article/pii/S030440761400089X.

Hardiman, S. J., Bercot, N., & Bouchaud, J.-P. (2013). Critical reflexivity in financial markets: a Hawkes process analysis. *The European Physical Journal B*, 86(10), 442–450. http://dx.doi.org/10.1140/epjb/e2013-40107-3, URL https://link.springer.com/article/10.1140/epjb/e2013-40107-3.

Hardiman, S. J., & Bouchaud, J.-P. (2014). Branching-ratio approximation for the self-exciting Hawkes process. *Physical Review E*, 90(6), Article 062807. http://dx.doi.org/10.1103/PhysRevE.90.062807, URL https://link.aps.org/doi/10.1103/PhysRevE.90.062807.

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 33(3), 438–443. http://dx.doi.org/10.1111/j.2517-6161.1971.tb01530.x, URL http://doi.wiley.com/10.1111/j.2517-6161.1971.tb01530.x.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1), 83–90. http://dx.doi.org/10.2307/2334319, URL https://academic.oup.com/biomet/article-abstract/58/1/83/224809.

Hawkes, A. G. (2018). Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2), 193–198. http://dx.doi.org/10.1080/14697688.2017.1403131.

Hawkes, A. G. (2020). Hawkes jump-diffusions and finance: a brief history and review. *The European Journal of Finance*, 1–15. http://dx.doi.org/10.1080/1351847X.2020.1755712, URL https://www.tandfonline.com/doi/full/10.1080/1351847X.2020.1755712.

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. http://dx.doi.org/10.1016/j.jempfin.2008.02.004, URL https://www.sciencedirect.com/science/article/pii/S0927539808000200.

Johnson, N., Hitchman, A., Phan, D., & Smith, L. (2018). Self-exciting point process models for political conflict forecasting. *European Journal of Applied Mathematics*, 29(4), 685–707. http://dx.doi.org/10.1017/S095679251700033X, URL https://www.cambridge.org/core/product/identifier/S095679251700033X/type/journal_article.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2), 73–84. http://dx.doi.org/10.3905/jod.1995.407942, URL https://jod.pm-research.com/content/3/2/73.

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. http://dx.doi.org/

10.1016/S0927-5398(00)00012-8, URL https://www.sciencedirect.com/science/article/pii/S0927539800000128.

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting*, 30(3), 491–497. http://dx.doi.org/10.1016/j.ijforecast.2014.01.004, URL https://www.sciencedirect.com/science/article/pii/S0169207014000284.

Mohler, G., Carter, J., & Raje, R. (2018). Improving social harm indices with a modulated Hawkes process. *International Journal of Forecasting*, 34(3), 431–439. http://dx.doi.org/10.1016/j.ijforecast.2018.01.006, URL https://www.sciencedirect.com/science/article/pii/S0169207018300311.

Nocedal, J., & Wright, S. J. (2006). Sequential Quadratic Programming. In *Numerical optimization* (pp. 529–562). New York: Springer, http://dx.doi.org/10.1007/978-0-387-40065-5_18.

Pernice, V., Staude, B., Cardanobile, S., & Rotter, S. (2012). Recurrent interactions in spiking networks with arbitrary topology. *Physical Review E*, 85(3), Article 031916. http://dx.doi.org/10.1103/PhysRevE.85.031916, URL https://link.aps.org/doi/10.1103/PhysRevE.85.031916.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1), 119–131. http://dx.doi.org/10.1214/aos/1176343003, URL https://projecteuclid.org/journals/annals-of-statistics/volume-3/issue-1/Statistical-Inference-Using-Extreme-Order-Statistics/10.1214/aos/1176343003.full.

Politis, D. N., & White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, 23(1), 53–70. http://dx.doi.org/10.1081/ETC-120028836.

Rambaldi, M., Pennesi, P., & Lillo, F. (2015). Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach. *Physical Review E*, 91(1), Article 012819. http://dx.doi.org/10.1103/PhysRevE.91.012819, URL https://link.aps.org/doi/10.1103/PhysRevE.91.012819.

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3), 299–318. http://dx.doi.org/10.1214/17-STS629, URL https://projecteuclid.org/journals/statistical-science/volume-33/issue-3/A-Review-of-Self-Exciting-Spatio-Temporal-Point-Processes-and/10.1214/17-STS629.full.

Ruppert, D., & Matteson, D. S. (2015). *Springer texts in statistics, Statistics and data analysis for financial engineering* (2nd ed.). New York: Springer, http://dx.doi.org/10.1007/978-1-4939-2614-5, URL https://link.springer.com/book/10.1007/978-1-4939-2614-5.

Scarrott, C., & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *Revstat - Statistical Journal*, 10(1), 33–60, URL https://www.ine.pt/revstat/pdf/rs120102.pdf.

Shcherbakov, R., Zhuang, J., Zöller, G., & Ogata, Y. (2019). Forecasting the magnitude of the largest expected earthquake. *Nature Communications*, 10(1), 4051. http://dx.doi.org/10.1038/s41467-019-11958-4.

Short, M. B., Mohler, G. O., Brantingham, P. J., & Tita, G. E. (2014). Gang rivalry dynamics via coupled point process networks. *Discrete & Continuous Dynamical Systems - B*, 19(5), 1459–1477. http://dx.doi.org/10.3934/dcdsb.2014.19.1459, URL http://aimsciences.org/article/doi/10.3934/dcdsb.2014.19.1459.

Sornette, D. (2006). *Critical phenomena in natural sciences* (2nd ed.). Berlin: Springer, http://dx.doi.org/10.1007/3-540-33182-4, URL https://link.springer.com/book/10.1007/3-540-33182-4.

Stooq (2021). Stooq. URL https://stooq.com/.

Tannenbaum, N. R., & Burak, Y. (2017). Theory of nonstationary Hawkes processes. *Physical Review E*, 96(6), Article 062314. http://dx.doi.org/10.1103/PhysRevE.96.062314, URL https://link.aps.org/doi/10.1103/PhysRevE.96.062314.

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, 36(2), 428–441. http://dx.doi.org/10.1016/j.ijforecast.2019.05.014, URL https://www.sciencedirect.com/science/article/pii/S0169207019301918.

Tomlinson, M. F., Greenwood, D., & Mucha-Kruczyński, M. (2021). Asymmetric excitation of left- and right-tail extreme events probed using a Hawkes model: Application to financial returns. *Physical Review E*, 104(2), Article 024112. http://dx.doi.org/10.1103/PhysRevE.104.024112, URL https://link.aps.org/doi/10.1103/PhysRevE.104.024112.

Tsay, R. S. (2010). *Wiley series in probability and statistics, Analysis of financial time series* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc., http://dx.doi.org/10.1002/9780470644560, URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470644560.

Wheatley, S., Wehrli, A., & Sornette, D. (2019). The endo–exo problem in high frequency financial price fluctuations and rejecting criticality. *Quantitative Finance*, 19(7), 1165–1178. http://dx.doi.org/10.1080/14697688.2018.1550266.