Weather and Climate Extremes 41 (2023) 100584

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: www.elsevier.com/locate/wace

Timing and spatial selection bias in rapid extreme event attribution

Ophélia Miralles *, Anthony C. Davison

Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ARTICLE INFO

Dataset link: https://github.com/OpheliaMiralles/timing-bias-extremes, https://github.com/OpheliaMiralles/pykelihood, repository, NOAA, mev

Keywords:
Climate event attribution
Climate extreme event
Likelihood-based inference
Spatial selection
Timing bias

ABSTRACT

Selection bias may arise when data have been chosen in a way that subsequent analysis does not account for. Such bias can arise in climate event attribution studies that are performed rapidly after a devastating "trigger event", whose occurrence corresponds to a stopping rule. Intuition suggests that naïvely including the trigger event in a standard fit in which it is the final observation will bias its importance downwards, and that excluding it will have the opposite effect. In either case the stopping rule leads to bias recently discussed in the statistical literature (Barlow et al., 2020) and whose implications for climate event attribution we investigate. Simulations in a univariate setting show substantially lower relative bias and root mean squared error for estimation of the 200-year return level when the timing bias is accounted for. Simulations in a bivariate setting show that not accounting for the stopping rule can lead to both over- and under-estimation of return levels, but that bias can be reduced by more appropriate analysis. We also discuss biases arising when an extreme event occurs in one of several related time series but this is not accounted for in data analysis, and show that the estimated return period for the "trigger event" based on a dataset that contains this event can be both biased and very uncertain. The ideas are illustrated by analysis of rainfall data from Venezuela and temperature data from India and Canada.

0. Introduction

An important objective of extreme event attribution (EEA) studies is to quantify the change in the probability of an extreme event due to external forcing, such as anthropogenic climate change (Allen, 2003; Stott et al., 2016; Naveau et al., 2020). Most such studies focus on the extent to which increased greenhouse gas (GHG) levels in the atmosphere affect the risk ratio for a specified extreme event (Stott et al., 2004; Fischer and Knutti, 2015, 2016; Jones et al., 2020). The risk ratio is commonly defined as the ratio of the probability p_1 of exceeding an extreme threshold u in the factual world, and the corresponding probability p_0 of doing so in a counterfactual world, often taken to be the pre-industrial era (Naveau et al., 2020).

The National Academies of Sciences, Engineering, and Medicine (2016) report divides EEA studies into two types: observation-based and climate-model-based. The first type uses series of historical and recent data to capture temporal changes in the probabilities and magnitudes of extreme events and thus to infer the effects of anthropogenic climate change. The second type uses a data-generating process to simulate two different worlds that are intended to be identical except for a "treatment" variable, usually GHG levels (Stott et al., 2004; Fischer and Knutti, 2015), or fine particulate matter (Larsen et al., 2020), and thereby assesses how the "treatment" affects phenomena such as temperature, precipitation or wildfires. In this framework causal inference techniques are required to efficiently capture the causality while reducing the signal-to-noise ratio in a complex and noisy climate system (Reich et al., 2021). The literature on causal statistical analysis has greatly evolved in recent decades and now has many applications (Hernàn and Robins, 2020).

Observation-based EEA studies can be further divided into two groups: return-level-based studies intended to assess temporal changes in the data distribution, and meteorology-oriented studies that explore how long-term trends in large-scale circulation patterns affect local extreme events sharing common meteorological characteristics (National Academies of Sciences, Engineering, and Medicine, 2016).

The purpose of this paper is to bridge recent improvements in inference using extreme value theory (EVT) and EEA studies that employ EVT. Thus it focuses on return-level-based EEA studies such as those using the approach developed within the World Weather Attribution (WWA) initiative (see worldweatherattribution.org/about), which performs rapid return-level-based EEA.

Rapid event attribution usually takes place immediately after an extreme event, especially one with high economic or societal impact (Lerch et al., 2017). In Risser and Wehner (2017), changes in the likelihood of extreme rainfall near Houston, Texas, were analyzed in September 2017, a month after Hurricane Harvey struck. Flooding in the United Kingdom (van Oldenborgh et al., 2015) and in

* Corresponding author.
E-mail address: ophelia.miralles@epfl.ch (O. Miralles).

https://doi.org/10.1016/j.wace.2023.100584
Received 8 March 2023; Received in revised form 7 June 2023; Accepted 8 June 2023
Available online 13 June 2023

Louisiana (van der Wiel et al., 2017) triggered immediate forecast evaluation studies for both events. The climate attribution study for the 2017 heatwave in India (van Oldenborgh et al., 2018) was conducted within a year.

Protocols for the attribution of extreme climate events have recently been improved (Philip et al., 2020; van Oldenborgh et al., 2021), but although the motivations for the choices of a relevant area, timescale, trend and distribution are well-documented, the question of whether or not to include the "trigger event" that led to the attribution study is rarely explored. Most recent studies include this event without further comment (van der Wiel et al., 2017; Risser and Wehner, 2017; Philip et al., 2018), but some exclude it because of a putative "positive bias" (van Oldenborgh et al., 2018). Sometimes it is excluded because the analysis takes place so soon after its occurrence that data are unavailable. An example of this is the study of the 2015 flooding in Northern England and Southern Scotland van Oldenborgh et al. (2015) from which the extreme itself was initially excluded, but which was undertaken again after the data became available (Otto et al., 2018).

Guidelines for avoiding pitfalls in climate event attribution studies are provided by van Oldenborgh et al. (2021), who state:

> "There has been discussion on whether to include the event under study in the fit or not. We used not to do this to be conservative, but now realize that the event can be included if the *event definition does not depend on the extreme event itself.*"

Quote 1: Extract from van Oldenborgh et al. (2021) (our emphasis).

There can be confusion in the climate literature between events and realizations (Quote). The "event definition" section of most rapid EEA papers (van der Wiel et al., 2017; van Oldenborgh et al., 2018) relates to the random variable of concern, whereas the "extreme event" refers to the specific realization under study. For instance, in van der Wiel et al. (2017), "event definition" refers to the annual maximum 3-day precipitation average (a random variable in statistical terms) and the extreme event under study is the 3-day precipitation average of 216.1 mm.day^{-1} observed in Livingston, Louisiana, in August 2016 (a realization of the random variable). This raises three issues: potential for linguistic and hence conceptual confusion, in particular between random variables and events; the possibility that the random variables themselves are defined in light of an observed event; and the inclusion or not of the particular realization in the data analysis. In this particular case, the event definition does not depend on the Louisiana level, but the latter is included in the analysis and thus influences the fitted generalized extreme value distribution. In this paper, we differentiate between random variables, their realizations and events using standard notation: realizations of a random variable X are designated by x and we refer to events using the letter $\mathcal{E}$.

We now focus on the third of the issues just mentioned, namely the inclusion or not of the trigger event in analysis. Certain discussions of protocols for extreme event attribution suggest that even if the extreme observation that stimulated the analysis is excluded from the dataset, the corresponding information can be incorporated by constraining the tail of the fitted distribution to be heavy enough to ensure that the return period for that observation is finite:

> "We do use the information that [the extreme event] occurred by demanding that the distribution has a non-zero probability of the observed event occurring [...]. This primarily affects the uncertainty estimates [...], which usually have upper bounds".

Quote 2: Extract from Philip et al. (2020).

We explain below how appropriate statistical methods can account directly for the trigger event, thus removing any need for constraints of this sort.

Recent work has shown that the upward bias observed when an analysis is performed immediately after an extreme event may stem from the timing of the analysis, whereas excluding the trigger event will lead to a downward bias in estimated return levels (Philip et al., 2020; Barlow et al., 2020). Introducing a stopping rule that appropriately reflects the timing of the analysis can account for such biases without requiring a decision to exclude or include the trigger event, though we shall see below that it is best to exclude it if its return period is to be estimated.

Below we sketch notions of inference using stopping rules, use simulation to compare different approaches to data analysis and reanalyze examples from the recent literature on climate event attribution. We also discuss the estimation of return periods for specific events and the effects of spatial selection, which can occur when the trigger event might have taken place in any of several time series in related locations. The extent to which some selection biases influence EEA is well-documented in the recent literature. The selection of the "trigger event" itself produces a bias, since we are mainly interested in extreme events that happened and for which an increased probability due to climate change is expected (Philip et al., 2020; van Oldenborgh et al., 2021). The bias introduced by reducing the spatial area of interest (Stott et al., 2004; Hammerling et al., 2019) or the possible weather conditions (Philip et al., 2020; van Oldenborgh et al., 2021) to those of a specific observed event is also documented in recent EEA studies. However, to our knowledge, there is no comparable study of such biases in the literature on climate event attribution.

Sections 1 and 2 introduce timing and spatial selection biases, and simulation results are then provided to show how such biases affect return level estimation. Three fast event attribution studies are then reanalyzed, accounting for those selection biases. The paper closes with some recommendations for future rapid return-level-based attribution analyses.

1. Accounting for timing bias

1.1. Preliminaries

We precede our discussion of remedies for timing bias by recalling the cumulative distribution functions of the generalized extreme value and generalized Pareto distributions, and of the joint cumulative distribution function associated with a logistic copula in S dimensions, viz

$$\mathrm{GEV}(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right\}, \quad -\infty < x < \infty, \tag{1}$$

$$\mathrm{GPD}^u(x) = 1-\left(1+\xi\frac{x-u}{\sigma_u}\right)_+^{-1/\xi}, \quad x \in [u,\infty), \tag{2}$$

$$C\left(w_1,\ldots,w_S\right) = \exp\left[-\left\{\sum_{s=1}^{S}\left(-\log w_s\right)^{1/\alpha}\right\}^{\alpha}\right], \quad 0 < w_1,\ldots,w_S < 1, \quad 0 < \alpha \le 1, \tag{3}$$

where $a_+ = \max(a,0)$ for real numbers a. Expressions (1) and (2) respectively provide standard models for block (e.g., annual or seasonal) maxima and for the exceedances of a high threshold u. Both models depend on a shape parameter ξ that determines the weight of the distribution tails; the first also depends on location and scale parameters μ and σ, and the second depends on a scale parameter σ_u. The logistic copula (3) is a one-parameter dependence model in which the variables $w_1,\ldots,w_S$ are independent when $\alpha = 1$ and become totally dependent when $\alpha \to 0$. Such a simple dependence model rarely fits real data well, but it is adequate for our purposes. Below we denote the unknown parameters for each of these expressions by θ. The development in the univariate case below is closely based on Barlow et al. (2020). Belzile and Davison (2022) give an alternative derivation of the main results and investigate improved inference based on them.

1.2. Stopping and estimation

When statistical analysis is performed immediately after the occurrence of a trigger event, the joint probability density of the data should be modified. If for simplicity we suppose that the successive observations are independent replicates of a random variable with probability density and distribution functions f and F, and denote the trigger event by $\mathcal{E}$, at which time the available data are $x_1, \dots, x_T$, then the joint density of the data, conditional on the occurrence of $\mathcal{E}$, is

$$f(x_1, \dots, x_T \mid \mathcal{E}) = \frac{\Pr(\mathcal{E} \mid x_1, \dots, x_T) f(x_1, \dots, x_T)}{\Pr(\mathcal{E})} = \frac{I(\mathcal{E} \cap \{x_1, \dots, x_T\}) f(x_1) \cdots f(x_T)}{\Pr(\mathcal{E})},$$

where the first equality follows from Bayes' theorem and the second from the assumption that $x_1, \dots, x_T$ are independent. The indicator function appearing in the last expression ensures that the joint density equals zero unless the configuration of the data $x_1, \dots, x_T$ has led to the trigger event; it can be dropped below, because we assume throughout that this is the case. If $\mathcal{E}$ occurs at time T, when the data series first exceeds some pre-determined high level η, for example, then $\Pr(\mathcal{E}) = F(\eta)^{T-1}\{1 - F(\eta)\}$, leading to

$$f(x_1, \dots, x_T \mid \mathcal{E}) = \prod_{t=1}^{T-1} \frac{f(x_t)}{F(\eta)} \times \frac{f(x_T)}{1 - F(\eta)}, \quad x_1, \dots, x_{T-1} \le \eta < x_T. \tag{4}$$

The first $T - 1$ terms on the right-hand side of (4) correspond to those observations that did not exceed η, and the last term to the value $x_T > \eta$ that caused the trigger event. The observations $x_1, \dots, x_{T-1}$ are right-truncated at η, whereas x_T is left-truncated at η.

The above formulation assumes that the trigger event is generated by the same physical mechanisms as earlier data. In some cases this may be untrue, because of changes in the background climatology or novel conjunctions of circumstances, but in any case one aspect of attribution analysis is to gauge the appropriate degree of surprise at the trigger event, and this involves comparison with the past. Moreover if this event is so unprecedented that relevant data are very limited or even unavailable, statistical analysis is difficult to justify. We therefore maintain this assumption, though rather gingerly.

For simplicity above we have suppressed the parameter vector θ on which an expression such as (4) depends, but in applications the conditional density is used to fit the model, so we henceforth include θ in the notation. Estimation by maximizing the standard log-likelihood function

$$\mathcal{L}^{\mathrm{STD}}(\theta) = \sum_{t=1}^{T} \log f(x_t; \theta) \tag{5}$$

does not account for the fact that T is determined by the data. A naïve correction excludes the final observation from the data, giving the 'exclusion' log-likelihood function

$$\mathcal{L}^{\mathrm{EX}}(\theta) = \sum_{t=1}^{T-1} \log f(x_t; \theta), \tag{6}$$

but although x_T itself does not appear here, it influences the fit because it helps to determine T. Neither (5) nor (6) allows for the fact that T is random, and, as mentioned above, fitting using them can be expected to lead to respective under- and over-estimation of the return period for x_T.

Two difficulties in the statistical formulation of the trigger event and its associated stopping rule is that these are typically only known after this event has occurred and that the event itself may be somewhat vaguely defined, so entirely watertight inferences appear unattainable. However, sensitivity analysis based on plausible stopping rules is certainly feasible, and below we shall see that it can provide useful insights.

One natural formulation of the stopping rule is to define the trigger event so that the preceding data are not regarded as particularly unusual. A simple way to do this is to set $T = \min\{t : x_t > \eta_t\}$, where $\eta_1, \eta_2, \dots$ is a series of thresholds and x_T is the first observation to exceed the corresponding threshold. Thus $x_t < \eta_t$ for $t = 1, \dots, T-1$, and then $x_T > \eta_T$. In many cases, η_t might be constant over time, but this is not essential to the argument. The resulting full conditional log-likelihood function (Barlow et al., 2020) is a generalization of expression (4),

$$\mathcal{L}^{\mathrm{COND}}(\theta) = \sum_{t=1}^{T} \log\left\{\frac{f(x_t;\theta)}{F(\eta_t;\theta)}\right\} + \log\left\{\frac{f(x_T;\theta)}{1 - F(\eta_T;\theta)}\right\}, \tag{7}$$

which incorporates this stopping rule and thus allows for the timing bias. We do not consider the partial conditioning approach suggested by Barlow et al. (2020), but by analogy to $\mathcal{L}^{\mathrm{EX}}(\theta)$ we introduce

$$\mathcal{L}^{\mathrm{CONDEX}}(\theta) = \sum_{t=1}^{T-1} \log\left\{\frac{f(x_t;\theta)}{F(\eta_t;\theta)}\right\}, \tag{8}$$

which excludes the trigger event from $\mathcal{L}^{\mathrm{COND}}$. Eqs. (5), (6), (7) and (8) easily adapt to the non-stationary case by replacing the parameter vector θ by a time-varying parameter vector θ_t.

The use of varying thresholds would be natural in many applications, but allowing them to depend on recent extremes raises computational issues; see the Supplementary Material.

Analyzing correlated time series to predict return levels in a specific area is common in climate studies. For example, van der Wiel et al. (2017) selected 19 out of 324 stations in the state of Louisiana (US), with at least 0.5° of spatial separation among those selected, in order to reduce spatial dependence between time series of annual maximum 3-day precipitation averages. In van Oldenborgh et al. (2018), maximum annual temperature return levels for two correlated time series close to Phalodi (India) are derived from separate event attribution studies. Thus it is useful to extend our discussion above to the multivariate setting. A simple approach uses a copula to model dependence among S-dimensional variables $x_1, \dots, x_T$, where $x_t = (x_{t,1}, \dots, x_{t,S})$ now consists of observations at S spatial locations that we denote collectively by S. Then the log-likelihood functions (5), (6), (7) and (8) for independent $x_1, \dots, x_T$ generalize to

$$\mathcal{L}^{\mathrm{STD}}(\theta) = \sum_{t=1}^{T} \log\left[f(x_t;\theta) c\left\{F(x_t;\theta);\theta\right\}\right], \tag{9}$$

$$\mathcal{L}^{\mathrm{EX}}(\theta) = \sum_{t=1}^{T-1} \log\left[f(x_t;\theta) c\left\{F(x_t;\theta);\theta\right\}\right], \tag{10}$$

$$\mathcal{L}^{\mathrm{COND}}(\theta) = \sum_{t=1}^{T-1} \log\left[\frac{f(x_t;\theta) c\left\{F(x_t;\theta);\theta\right\}}{C\left\{F(\eta_t;\theta);\theta\right\}}\right] + \log\left[\frac{f(x_T;\theta) c\left\{F(x_T;\theta);\theta\right\}}{1 - C\left\{F(\eta_T;\theta);\theta\right\}}\right], \tag{11}$$

$$\mathcal{L}^{\mathrm{CONDEX}}(\theta) = \sum_{t=1}^{T-1} \log\left[\frac{f(x_t;\theta) c\left\{F(x_t;\theta);\theta\right\}}{C\left\{F(\eta_t;\theta);\theta\right\}}\right], \tag{12}$$

where $F(x_t;\theta) = \left\{F_1\left(x_{t,1};\theta\right), \dots, F_S\left(x_{t,S};\theta\right)\right\}$ represents the vector of marginal cumulative distribution functions, $f(x_t) = \prod_{s=1}^{S} f_s(x_{t,s};\theta)$ is the product of their marginal density functions, and C is a copula with uniform margins defined by

$$\mathbb{P}\left(X_1 \le x_1, \dots, X_S \le x_S;\theta\right) = C\left\{F(x;\theta);\theta\right\} \tag{13}$$

with associated density function $c(u;\theta) = \partial^S C(u;\theta)/\partial u_1 \cdots \partial u_S$, with $u = (u_1, \dots, u_S) \in [0,1]^S$.

1.3. Discussion of the stopping rule

In an ideal world the stopping rule would be clearly specified in advance of potential trigger events by listing circumstances exceptional enough to warrant an attribution study. In many practical situations, such a task is impossible, too time-consuming or too restrictive, so

attribution analysis is often performed without the clear prior specification of a trigger event. In many cases contextual information about what events can be treated as extreme in a specific geographical region can be used to "guess" the stopping rule and thus to determine terms appearing in Eqs. (7) and (8). For example, the authors of the attribution study for the August 2016 Louisiana floods give the following quantitative definition of extreme flooding:

> "In places, the 3-day precipitation totals in Louisiana exceeded [...] 3 times the average annual 3-day precipitation maximum"

Quote 3: Extract from van der Wiel et al. (2017).

If this definition was not influenced by the level recorded in August 2016, then a flooding event would be considered as extreme when, somewhere in the region, a 3-day average precipitation annual maximum three times bigger than its historical average was recorded. The data analyzed in van der Wiel et al. (2017) involve $S = 19$ different spatial locations. Assuming that there is no spatial selection, the stopping rule could be defined as the first time at which one or more of these spatial locations records a 3-day average precipitation X_t^s that exceeds three times the historical annual average 3-day maximum. If data for the years 1950–2000 are used to compute the historical average $\bar{X}^s$ at each location s in the set S containing the locations and stopping can only occur in subsequent years, we might take T to be the first time from the year 2001 onwards that such an event occurs at one or more locations in S, i.e.,

$$T = \min\left\{ t \geq 2001 : \bigcup_{s \in S} \left(X_t^s \geq 3\bar{X}^s \right) \right\}. \tag{14}$$

In van Oldenborgh et al. (2018), an extreme heatwave is declared when TXx, the annual maximum daily temperature between May and June, is at least 4 or 5 degrees above its average for 1981–2010. Eq. (15) transcribes this contextual vision of an extreme temperature to a quantitative stopping rule for use in fitting the observed series, i.e.,

$$T = \min\left\{ t \geq 2010 : \mathrm{TXx}_t - \overline{\mathrm{TXx}}_{[1981,2010]} \geq 4 \right\}. \tag{15}$$

When the precise definition of the extreme event is unclear, various plausible stopping rules could be formulated and used as the basis for sensitivity analyses.

2. Accounting for spatial selection

Thus far we have discussed how analysis immediately after a trigger event can influence the estimation of an underlying extremal probability model and thus affect the probability and/or return period associated with that event. Bias can also arise when the trigger event occurs in a single time series that is selected among several related series, and no allowance is made for the selection. We now give a stylized discussion of how this affects estimated return periods for the event in question.

Suppose that S independent time series are monitored and that extreme events occur in the sth series with distribution $\mathrm{GEV}_s(x)$, where the subscript indicates that the parameters that determine the distribution depend on the series. Suppose that analysis takes place when the largest of the corresponding variables $X_1, \ldots, X_S$ exceeds a given return level, and that this selection is ignored. Without loss of generality we further suppose that this largest value occurs in series $s = 1$, and that its value x_1 is associated with a return period of m years based on the distribution $\mathrm{GEV}_1(x)$, i.e.,

$$\mathrm{GEV}_1(x_1) = 1 - 1/m.$$

This calculation ignores the fact that corresponding values $x_2, \ldots, x_S$ in time series $2, \ldots, S$, each such that $\mathrm{GEV}_s(x_s) = 1 - 1/m$, would also have led to the same return period estimate. Taking the selection into account, the true return period m_S is therefore given by

$$1 - 1/m_S = \Pr(X_1 \leq x_1, \ldots, X_S \leq x_S) = \prod_{s=1}^{S} \Pr(X_s \leq x_s) = (1 - 1/m)^S,$$

i.e.,

$$m_S = \left\{ 1 - (1 - 1/m)^S \right\}^{-1}. \tag{16}$$

If $S = 1$, i.e., there is no selection, then $m_S = m$, and if m is large then $m_S \approx m/S$: m-year events will occur S times more frequently in S independent series.

At first sight, these calculations for independent series might appear irrelevant to the analysis of dependent series. For so-called asymptotically dependent series, however, and with the same notation, one can write

$$\Pr(X_1 \leq x_1, \ldots, X_S \leq x_S) = (1 - 1/m)^{\chi} = 1 - 1/m_{\chi}, \tag{17}$$

where the so-called extremal coefficient χ satisfies $1 \leq \chi \leq S$ and can be interpreted as the "number of independent series" contributing to the overall maximum. If $X_1, \ldots, X_S$ are totally dependent, then $\chi = 1$, and if they are fully independent, then $\chi = S$; see expression (31.12) of Davison et al. (2019), for example. Asymptotically dependent models for spatial extremes can be expected to provide reasonable approximations to phenomena such as maxima of temperature time series at S sites in a relatively small spatial region, and such models will then provide upper bounds on m_S. An alternative class of models, often found appropriate for phenomena such as rainfall at spatially separated sites, has the property of asymptotic independence: increasingly rare observations become closer to independence, i.e., $\chi \lesssim S$ for very rare events. The corresponding m_S given by (16) will then provide a lower bound on the true return period.

Fig. 1 shows how m_{χ} is related to m for various values of χ. Each function is roughly linear for $m \geq \chi$, so the approximation $m_{\chi} \approx m/\chi$ seems adequate in most cases.

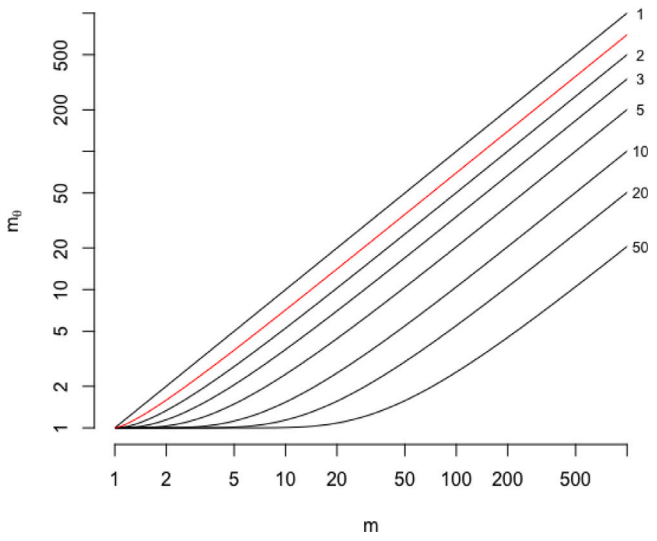

Fig. 1. Dependence of true return period m_{χ} on naïve return period m when the selection of an extreme event in one series among χ "equivalent independent" series is ignored. The figures at the right of the black lines show χ. The red line corresponds to $\chi \approx 1.43$ for the Phalodi analysis in Section 4.2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Simulation studies

3.1. Setup

We now use stochastic simulation of extremal data to compare how fitting based on the various log-likelihood functions described above

affects the estimation of a p-year return level, i.e., the level expected to be crossed by the variable of interest every p years, taking $p = 200$ for illustration. We shall see that not accounting for timing bias can lead to poor estimation in both univariate and bivariate settings. We also consider the association of a return period with a specific observation.

Our Monte Carlo settings were chosen to resemble real uses of extreme event attribution. Many climate variables studied are positive, unbounded and somewhat heavy-tailed, and their annual maxima are commonly fitted with a generalized extreme-value (GEV) distribution. As in Barlow et al. (2020), we defined quantitative stopping rules using a simulated historical sample of $n_h = 10$ maxima, then generated further independent variables from the same GEV distribution, applied different stopping rules, and used the resulting samples of maxima to estimate the three GEV parameters and the p-year return level.

For each run, stopping rules in which the chosen thresholds were return levels η_τ for a GEV (see Eq. (1)) with parameters $\mu = 0$, $\sigma = 1$ and $\xi = -0.2, 0, 0.2$ with different return periods τ were applied, giving stopping times

$$T^\tau = \min\left\{t > n_h : X_t \geq \eta_\tau\right\}. \tag{18}$$

The goal was to evaluate the impact of increasing the unlikeliness of the trigger event on the estimation of the 200-year return level based on the log-likelihood functions (5)–(8). Fits from 1000 simulated datasets were compared to the true 200-year return level in terms of the bias and relative root mean squared error (RRMSE). Confidence interval coverage (CIC) and width (CIW) are derived from confidence bounds for the 200-year return level estimator.

3.2. Timing bias with univariate extremes

We first discuss the effect of timing bias when estimating return levels based on a univariate time series. The results from parameter estimation using the log-likelihood functions (5), (6), (7), and (8) are respectively designated by "Standard", "Excluding Extreme", "Cond. Including Extreme" and "Cond. Excluding Extreme" in the figures and the text.

Barlow et al. (2020) sampled GEV random variables until they first exceeded a threshold η_τ (see Eq. (18)) or until the maximum sample size N was reached, estimated the parameters θ and then estimated the 200-year return level and its 95% confidence bounds. Their simulation studies result in lower relative bias and root mean squared error using the full conditional log-likelihood than using the standard likelihood, whether or not the trigger event is included, with comparable confidence interval coverage and width; see the Supplementary Material. The relative bias decreases when the full conditional log-likelihood includes the extreme event for return periods of $\tau \geq 500$, but this is due to the imposition of a maximum sample size. As τ increases, exceedances of η_τ become less likely, and when no realization exceeds η_τ, sampling stops when the maximum sample size is reached, and it is inappropriate to condition the log-likelihood with regard to a stopping rule that has not been applied.

To avoid the aforementioned problem we performed simulations with the sample size fixed to be $n_C = 200$ and return periods τ exceeding $n_C - n_h$, so that the last event observed is unlikely enough for the stopping rule to make sense. We first generated a "historical sample" of n_h GEV variables, and then, for each return period τ considered (see Eq. (18)), we generated $n_C - n_h - 1$ GEV variables right-truncated at η_τ, followed by a final GEV variable left-truncated at η_τ; these correspond to the conditional densities appearing in (7). We then concatenated the historical sample, the data under the stopping threshold η_τ and the last observation above η_τ to yield a sample of n_C observations, of which the only observation to exceed η_τ was the last, provided η lies above all n_h historical values. Fig. 2 shows the results of this experiment with GEV shape parameter $\xi = 0.2$. The standard fit shows an upward relative bias that increases with the size of the trigger event, and the resulting 200-year return level is less and less reliable (the confidence interval coverage decreases with τ). The differences between results for the other three log-likelihoods are smaller, especially for large η_τ, partly because the conditioning term has little effect on Eqs. (7) and (8) when $F(\eta_\tau) \approx 1$. The coverage of two-sided confidence intervals is most stable for the conditional fits, but this disguises a difference in the one-tailed errors: the intervals tend to be too short in the upper tail and too long in the lower tail. There is little to choose between the results using the conditional fits, though that with all the available information, based on (7), seems slightly preferable for smaller τ.

The corresponding results with $\xi = 0$ and -0.2 reported in the Supplementary Material lead to similar conclusions: conditioning while either including (Eq. (7)) or excluding (Eq. (8)) the "trigger" provides less biased 200-year return level estimates than the standard fit, whether or not the trigger is included in the latter. However, the coverage of the conditioned fit that includes the "trigger" deteriorates when $\xi = -0.2$, and its upper coverage error also significantly increases, as the upper bound for the 95% confidence interval is under-estimated for negative ξ. In this case, excluding the trigger without conditioning leads to underestimation of the upper confidence bound for $\tau < 1000$ and of the lower confidence bound for any τ considered.

3.3. Timing bias with correlated extremes

We now investigate the impact of timing bias in a bivariate setting. The univariate fit for the variable of interest is labeled "Independent", while fits using the log-likelihood functions (9), (10), (11), and (12) are respectively labeled "Including Extreme", "Excluding Extreme", "Cond. Including Extreme" and "Cond. Excluding Extreme".

We suppose that the stopping rule is applied to one variable but the other is merely part of the analysis. This situation can arise when, for instance, a location s_1 is very close to that of the trigger event, but its time series for the variable of interest lacks the data for that event itself. Often a more complete time series is available, and though its location s_2 lies further from that of the trigger event, it can serve as a monitoring reference for data at s_1. If there is strong dependence between time series at the two locations, then observation of an extreme event at s_2 may aid in event attribution for an extreme at s_1.

To explore this setting we generated replicates of two GEV variables X and Y with shape parameters 0.2 and dependence given by the logistic copula (3) with its parameter $\alpha = 0.5$ taken to be known. The maximum sample size was set to N, as in Barlow et al. (2020), and the univariate stopping rule of Eq. (18) was applied to Y: sampling of both series stopped when $Y \geq \eta_\tau$ for some return period τ. Although the stopping rule is applied only to Y, it influences the estimation of quantiles of X because the series are dependent. For every return period considered, the marginal distributions of X and Y were estimated from the time series stopped at T^τ, say, using the log-likelihood functions (9)–(12). The cumulative distribution function for Y is denoted by F_Y. In Eqs. (11) and (12), the bivariate conditional terms $C\left\{F(\eta_t;\theta)\right\}$ for $t \leq T^\tau$ and $1 - C\left\{F(\eta_{T^\tau})\right\}$ are respectively replaced by $F_Y(\eta_\tau;\theta)$ and $1 - F_Y(\eta_\tau;\theta)$. Finally, the 200-year return level for X and its confidence bounds were derived and the summaries used in the univariate case were computed. The standard univariate fit for X from Eq. (5) was used as a benchmark.

When $\alpha = 0.5$, the probability that X is extreme given that Y is extreme can be shown to equal $2 - 2^\alpha \simeq 0.59$. This leads to the following two cases:

$\mathcal{A}$. both X and Y are extreme when the trigger event occurs. In this case, we expect the return levels of X to be overestimated when fitting the time series for X assuming X and Y are independent, like in the standard univariate case. The corresponding simulation results, displayed in Fig. 3, suggest that for every stopping threshold η_τ the relative bias and root mean squared error from the full conditional fit are much lower than for other fits, while confidence intervals have similar coverages and widths. Upper

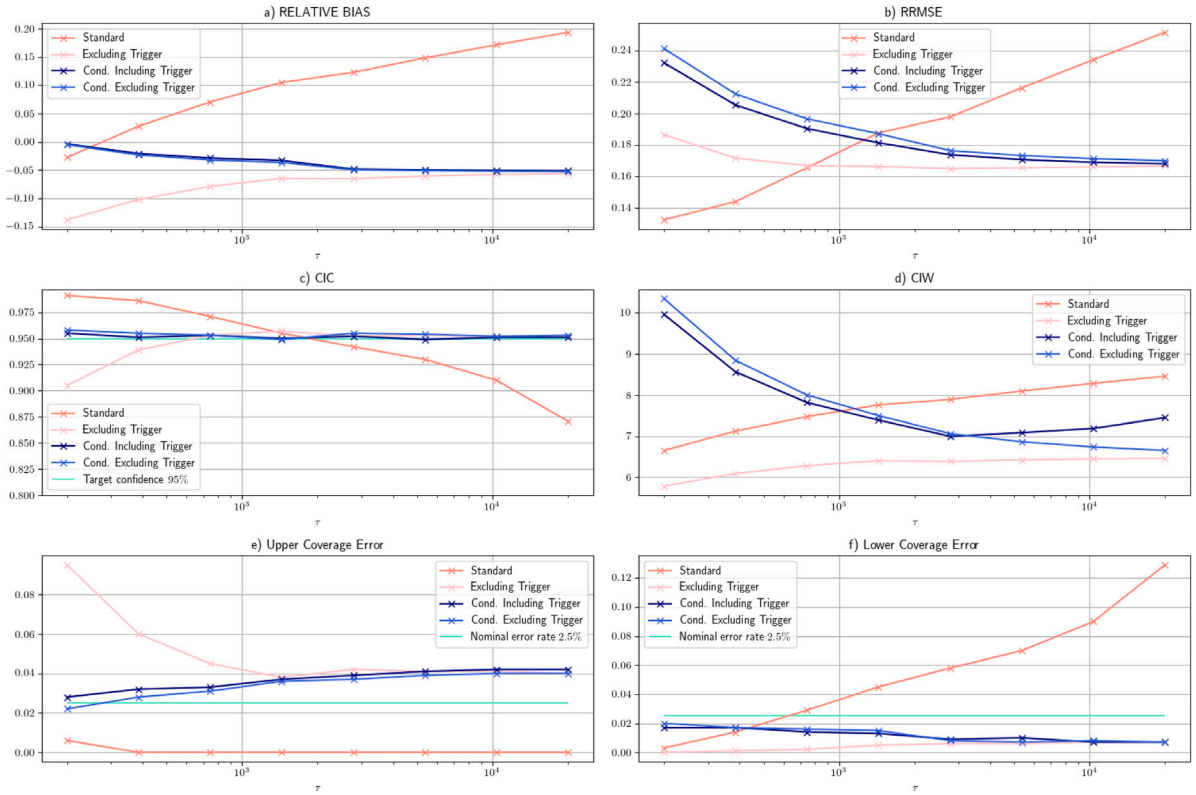

Fig. 2. Summary results for the estimation of a 200-year return level based on simulated GEV random variables with shape parameter 0.2, with stopping thresholds defined from the return periods τ shown on the x-axis as in Eq. (18). The relative bias and mean squared error are shown in Panels a and b, and coverage of 95% confidence intervals and their average widths are shown in Panels c and d. Panels e and f represent the upper and lower coverage errors. The time series are generated so that the first exceedance of the stopping threshold occurs at a specified time.

and lower coverage errors are very similar across methods accounting for the dependence between the series (Fig. 3e and f), though for high stopping thresholds, excluding the trigger with the appropriate conditioning provides the closest upper coverage error to the nominal rate, i.e., the most reliable upper confidence limit for the 200-year return level, while other methods tend to underestimate the upper confidence bound (Fig. 3e);

B. X is not extreme when the trigger event occurs, and we then expect the univariate fit for X to underestimate the return levels for X. Indeed, realizations of X sampled until the trigger event will tend to be low because they are related to those of Y, which lie below η_τ until sampling stops. Fig. 4 shows very reduced relative bias and RRMSE with conditioned bivariate fits, both including and excluding the extreme at s_2, compared with the independent fit for X, which strongly underestimates the 200-year return level at s_1, and with the standard fit including the extreme event, which gives positively biased estimators of the 200-year return level. Excluding the extreme leads to slight downward bias of the estimated return level for X for every stopping threshold τ considered. Of all methods, excluding the trigger with the appropriate conditioning provides the upper coverage error closest to the nominal error rate, especially for high stopping thresholds (Fig. 4e).

The figures show how both cases affect the return level estimates. The improvement due to accounting for the timing bias is much clearer in case $\mathcal{A}$, but case $\mathcal{B}$ better illustrates the situation in which the data are incomplete at the location of interest s_1 but the trigger event is seen only at s_2.

3.4. Bias in return period estimation

Fits of extreme-value models can be highly sensitive to the largest or smallest observations in a sample (Davison and Smith, 1990), so it is natural to wonder whether estimated return periods for particular observations corresponding to rare events might be biased. For concreteness, suppose that the generalized extreme-value distribution (1) has been fitted to a sample whose largest value is $X_{\max}$ and that the return period of $X_{\max}$ is to be estimated from the fit. The fitted distribution is $\widehat{\mathrm{GEV}}(x)$, so the true return period M and its estimate $\widehat{M}$ may be written as

$$M = \left\{1 - \mathrm{GEV}\left(X_{\max}\right)\right\}^{-1}, \quad \widehat{M} = \left\{1 - \widehat{\mathrm{GEV}}\left(X_{\max}\right)\right\}^{-1}.$$

The observation $X_{\max}$ is often described as an "M-year event", but this term applies in relation to $\mathrm{GEV}(x)$. In practice the estimate $\widehat{\mathrm{GEV}}(x)$ is often based on data that include $X_{\max}$, and the latter may strongly influence the estimated distribution. It seems plausible that $\widehat{M} < M$ if $X_{\max}$ is included in the fit, and that $\widehat{M} > M$ if not. The situation here differs from those in previous sections, which concerned the estimation of a return level, i.e., a parameter of the distribution, as the return period M depends on $X_{\max}$ and thus is itself random. To investigate the relation between M and $\widehat{M}$ we performed a further simulation study, which we now describe.

We generated 1000 independent datasets using a simplification of the approach described in Section 3.1, by simulating $n_C - 1$ observations from (1) right-truncated at a fixed threshold η, supplemented by a final observation with return period m. In order to measure the bias as accurately as possible, this final observation is determined by the equation $\mathrm{GEV}(x) = 1 - 1/m$ and thus is fixed. For each such dataset we computed return period estimates $\hat{m}$ using fitted distributions $\widehat{\mathrm{GEV}}(x)$ found using the log-likelihood functions (5)–(8). This process was repeated for different configurations of values of n_C, η and m, with shape parameter $\xi = 0.2$ throughout.

Fig. 5 shows boxplots of the resulting ratios $\hat{m}/m$. The results are extremely variable, with many simulated datasets in which $\hat{m} \gg m$,

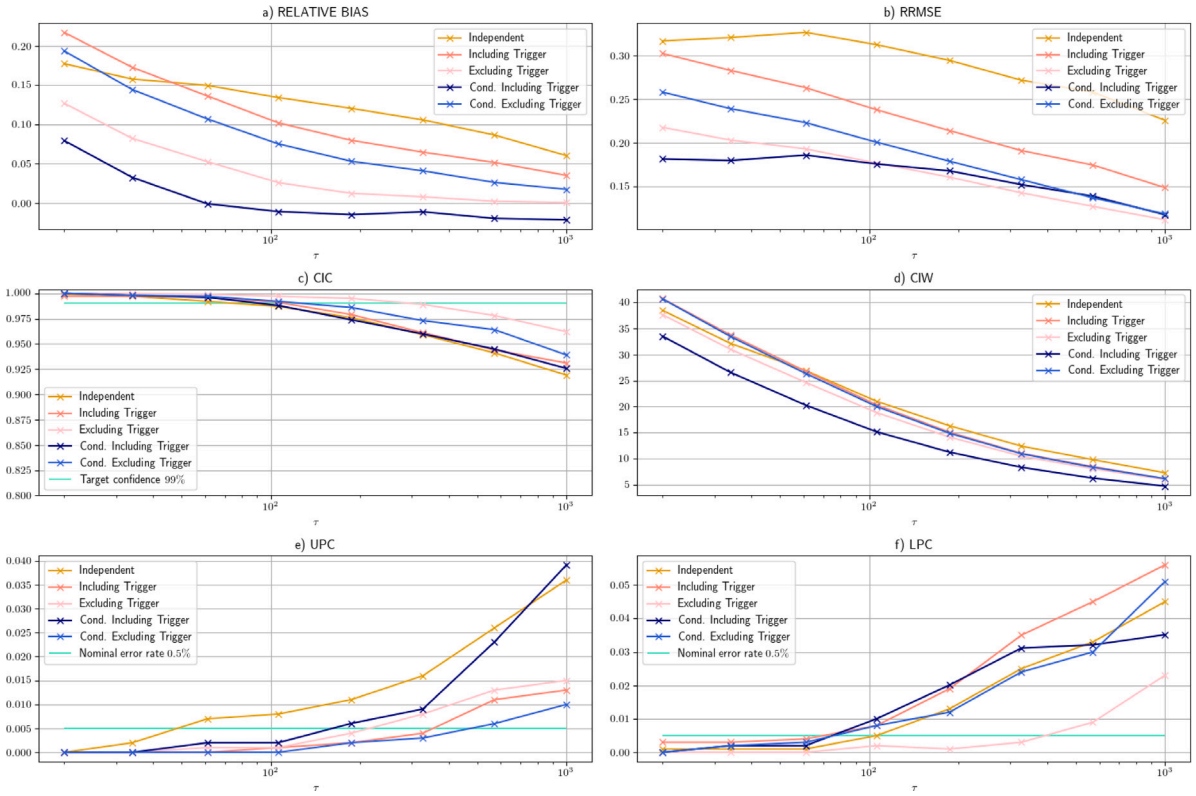

Fig. 3. Summary results for simulated bivariate extremes, case $\mathcal{A}$: X is also extreme when Y is stopped. Simulations of two correlated random variables X and Y following a logistic ($\alpha = 0.5$) copula with GEV ($\mu = 0, \sigma = 1, \xi = 0.2$) margins. The stopping rule is defined for Y as the return level of period τ as in Eq. (18). The return periods τ are shown on the x-axis. The relative bias and relative mean squared error from the theoretical 200-year return level for X are shown in Panels a and b, and 99% confidence interval coverage and width are shown in Panels c and d. Panels e and f represent the upper and lower coverage errors.

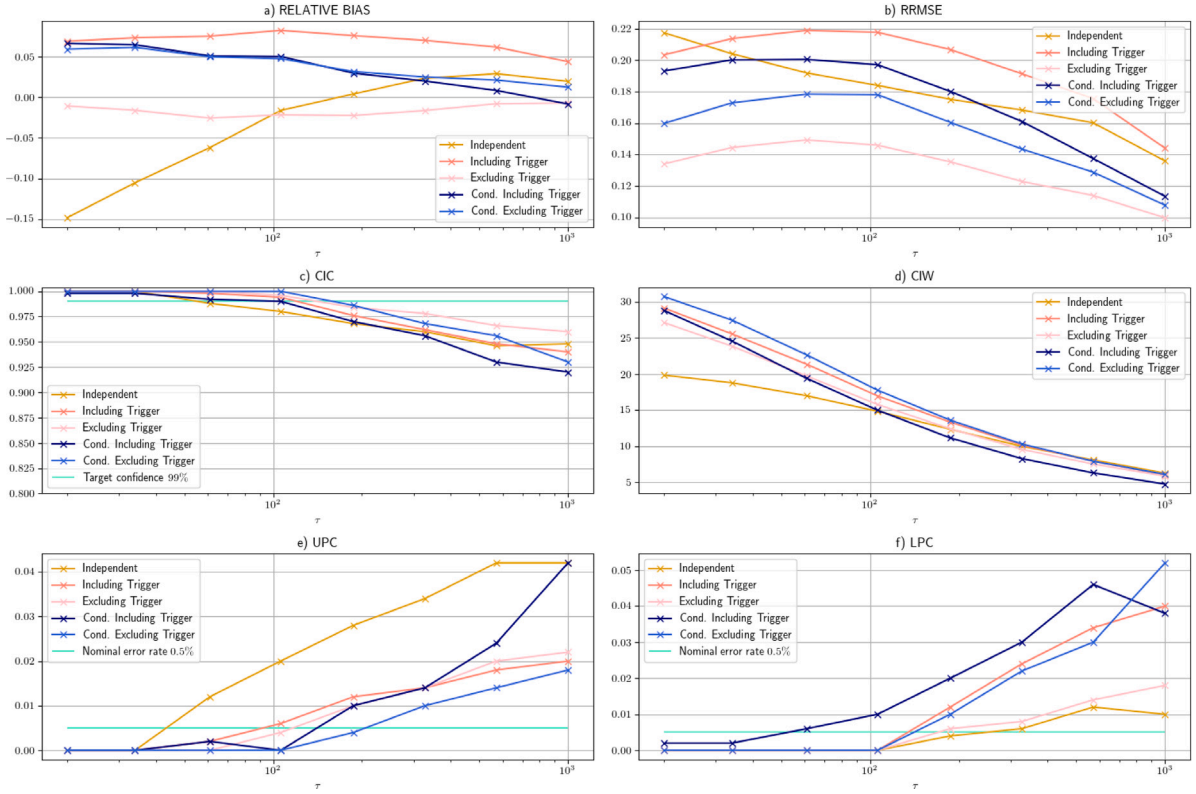

Fig. 4. Summary results for bivariate extremes, case $\mathcal{B}$: X is not extreme when Y is stopped. The simulation setup and stopping rule are the same as in Fig. 3. The relative bias and relative mean squared error from the theoretical 200-year return level for X are shown in Panels a and b, and the 99% confidence interval coverage and width are shown in Panels c and d. Panels e and f represent the upper and lower coverage errors.

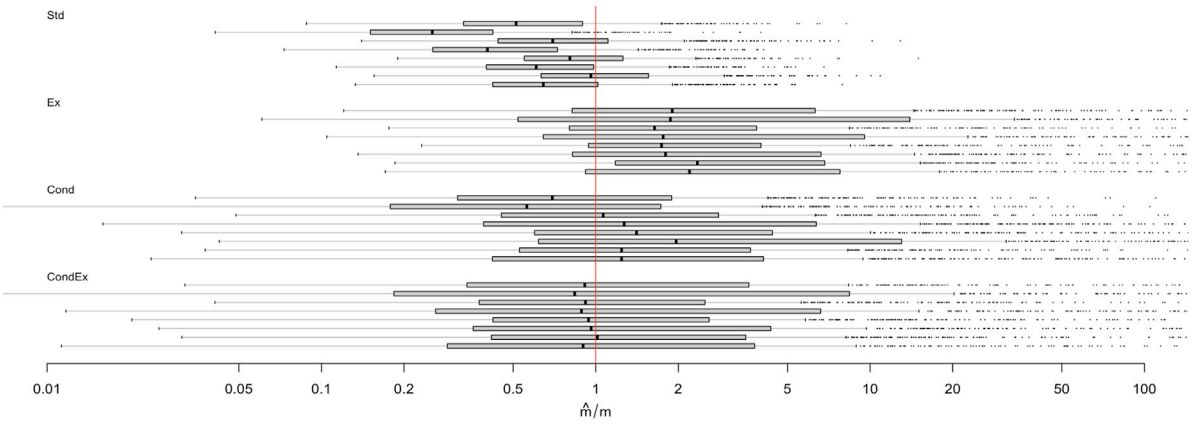

Fig. 5. Ratios of estimated and true return periods $\hat{m}/m$ for various fits to samples of size n_C, with threshold η and true return period m, with parameters estimated from standard and conditional log-likelihoods including and excluding the largest value, labeled Std, Ex, Cond and CondEx. From top in each block of boxes: (η, m, n_C) = (100,200,50), (100,200,80), (150,200,50), (150,200,80), (200,400,50), (200,400,100), (200,1000,50), (200,1000,100).

but some patterns emerge. When GEV is based on the standard log-likelihood and the largest observation is included, we tend to see $\hat{m} < m$, especially when the effect of not allowing for the right-truncation is reinforced by increasing n_C; $\hat{m}$ is systematically too large when the largest observation is excluded. The situation is more variable with the conditional likelihood, which gives the most consistent results when the largest observation is excluded. In itself this is not surprising, as use of log-likelihood (8) is then appropriate and the fitted distribution is independent of the largest observation; thus in this case we expect that $\hat{m}/m \to 1$ as n_C increases, but clearly such convergence is unlikely to be visible for values of n_C seen in applications. Perhaps the most striking feature of the results is that $\hat{m}/m$ is very variable and/or systematically biased in all cases: an event with $m = 1000$, say, might easily have $\hat{m}$ anywhere in the range 250 to 4000. This suggests that extreme caution is required when attributing return periods to particular events; indeed, this should not be attempted without a statement of uncertainty.

4. Real data analyses

4.1. 1999 Flooding in Vargas state, Venezuela

We now consider an extreme flooding event in the Venezuelan state of Vargas in December 1999. According to Méndez et al. (2015), the form of the San Julián basin implies that major rainfall events are extremely rare in this area, but when they do happen the consequences can be very serious. Indeed, these authors observe that this basin has a very wide range of slopes, provoking increased erosion over time, and that its small area (20.68 km^2) implies rapid concentration of surface runoff, so water can very quickly arrive in residential zones. In December 1999, such flooding, combined with a landslide, massive debris transportation and poor infrastructure, caused disastrous damage in the Caraballeda area.

To predict the likelihood of such an event, we estimate return levels for daily maximum hourly precipitation (mm) in Vargas from 1961 to 1999. The San Julián basin is not subject to much seasonality (Méndez et al., 2015), and no long-term trend is perceptible in these data.

We use a generalized Pareto distribution (2) to model daily rainfall amounts over $u = 12$ mm, a choice of threshold justified in the Supplementary Material using the approaches of Northrop and Coleman (2014) and Varty et al. (2021). Let η_p be defined such that

$$\mathbb{P}(X > \eta_p) = \frac{1}{p\lambda}, \tag{19}$$

where λ is the average number of exceedances per year, so $p\lambda$ is the average number in a p-year period. This allows us to interpret the GPD quantile η_p as the p-year return level.

The stopping rule here is ill-defined, so for illustration we took the trigger event to be the first crossing of the historical 100-year return level computed using the first two decades of data. The standard and conditioned fits with and without the extremes from December 1999 are displayed in Fig. 6. When the trigger event is included, the return time for the 1999 event is 464 (95% CI [352, 647]) years for the standard log-likelihood fit but 882 [597, 1447] for the full conditional log-likelihood fit (Fig. 6a). The standard fit changes considerably when the trigger event is excluded, and the return period for the Vargas 1999 event is multiplied by 2.6 to become 1207 [766, 2182] years, but the full conditional results change little except for the upper confidence bound, which increases faster with the return period (Fig. 6b). The uncertainty range for every return period computed is very wide.

An alternative analysis fits the GEV distribution to the annual maxima of daily precipitation values using the same stopping rule. There are fewer annual maxima than exceedances of 12 mm, so each has a larger influence on the fitted model, as we see in Fig. 7, where the standard fit including the extreme event has a heavier upper tail than the other fits. The return level for the 1999 extreme event using a GEV fit is comparable to using a GPD fit when the likelihood is conditioned, but the unconditioned GEV fit predicts a 250-year return level for the Vargas event, around 200 years shorter than the prediction using a GPD; see Coles and Pericchi (2003), Coles et al. (2003) and the Supplementary Material.

Return levels and return periods for a flood as extreme as the trigger estimated with a conditioned likelihood that includes the trigger event are quite different from their unconditioned analog, especially using a block maxima approach. Return level estimates based on the usual likelihood including and excluding the extreme event are very different, whereas inferences based on the conditional likelihoods with or without the trigger event are more stable.

4.2. 2016 Heatwave in Phalodi, India

The importance of accounting for timing bias can be seen by re-considering the attribution analysis for the 2016 heatwave in Phalodi, India, which had disastrous public health consequences. Data sources and methods are detailed in van Oldenborgh et al. (2018), though here we compute likelihood-based confidence intervals rather than use a bootstrap. The Phalodi series is not available in the GHCN-D dataset, but sufficiently complete annual maximum temperature time series are available at two nearby stations, Jodhpur and Bikaner, and we analyze these as a proxy for data at Phalodi. Our findings for a standard fit of the Jodhpur series with a time-related trend, shown in Fig. 8, are similar to those in van Oldenborgh et al. (2018). The location parameter of the fitted GEV distribution slightly decreases over time

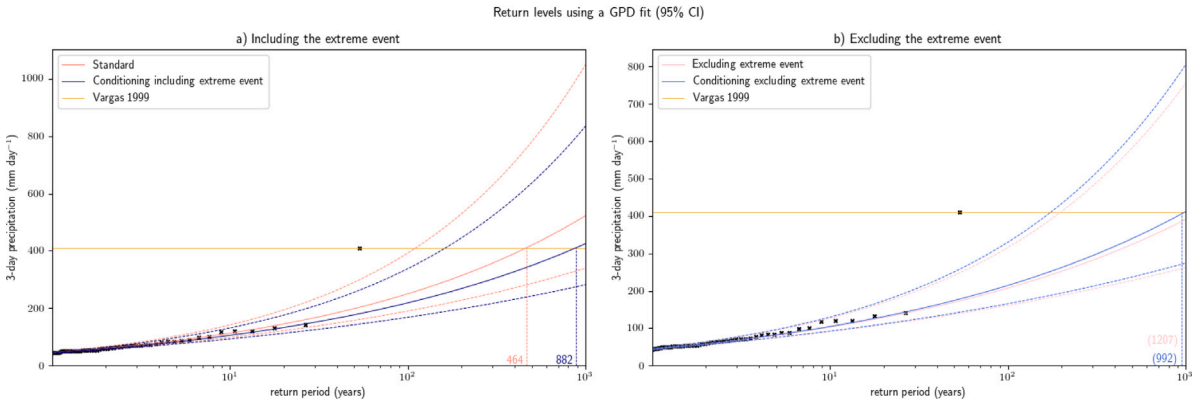

Fig. 6. Vargas data analysis. Estimated p-year return level η_p (see Eq. (19)) and its 95% confidence interval (dashed lines) from a GPD fit. The x-axis shows the return period p (years). Panel (a) shows standard and conditioned fits when the trigger event is included (see Eqs. (5) and (7)), and Panel (b) shows results from standard and conditioned fits when it is excluded (see Eqs. (6) and (8)). Vertical dotted lines show the estimated return periods for the event in Vargas in December 1999.

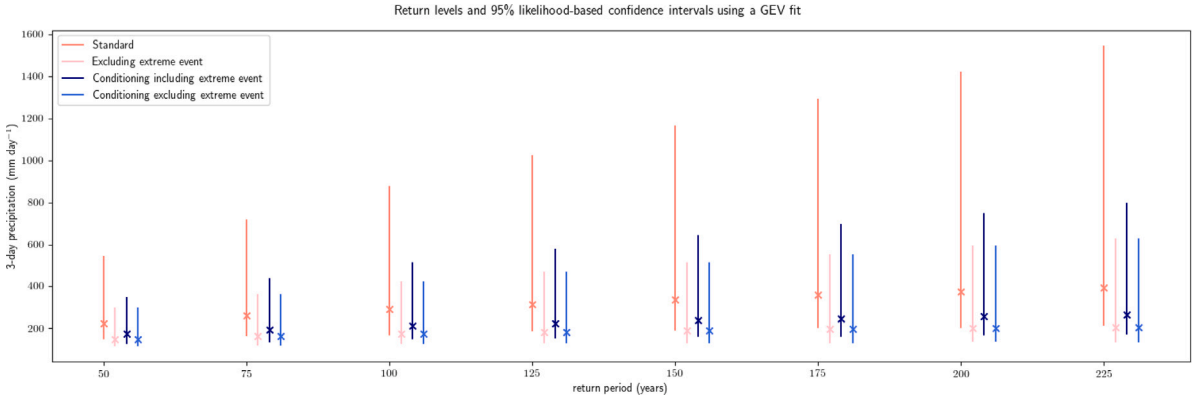

Fig. 7. Vargas data analysis. Estimated p-year event η_p using a GEV fit. The x-axis represents the return period in years p. Marks represent the estimated p-year event, vertical bars denote 95% likelihood-based confidence intervals. Different colors represent results from different fits. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(Fig. 8a) and a risk ratio (see Naveau et al., 2020, for a definition) of 0.511 is found for the occurrence of the trigger event in 2016 relative to 1973 (Fig. 8b). The heterogeneity in the Jodhpur time series could make the slightly negative trend in the location parameter very sensitive to the three observations between 1940 and 1960 (Fig. 8a). However, our aim is to reproduce as closely as possible the work of van Oldenborgh et al. (2018) in order to compare findings when accounting or not for timing and spatial selection bias. When fitting the Jodhpur time series with the fully conditioned log-likelihood (7) and a trend in the location parameter, the estimated risk ratio decreases from 0.511 to 0.4. Return periods for the heatwave as in 1973 and 2016 are given in Fig. 8b for the standard fit and in Fig. 8c for the conditioned fit, although they are very uncertain. The return period for a similar heatwave with the standard fit is 26 (95% CI [14, 150]) years in 1973 and 51 [26, 91] years in 2016. Conditioning slightly increases both return periods and the width of the 95% confidence interval, to reach respectively 32 [15, 302] and 80 [31, 147] years in 1973 and 2016.

Our analysis was performed in two steps, using the fact that the temperature time series at Jodhpur is more complete than that at Bikaner and contains the 2016 extreme event (van Oldenborgh et al., 2018), whereas Bikaner is closer to Phalodi. The stopping rule is defined as in Eq. (15). The first step was an extremal analysis using only the Jodhpur series of annual temperature maxima, TXx. The return levels estimated from univariate fits based on (5)–(8) are shown in Fig. 9. Those obtained using the standard likelihood (5) and including the trigger event are higher than for the other fits, with much higher upper confidence limits. To illustrate how allowing for timing bias can stabilize estimation, we extend the Jodhpur time series with later data and recompute return levels using standard and conditional likelihoods. The latter involves conditioning up to the trigger event year, while standard likelihood contributions are used for data after 2016. The full conditional fits before 2016 use the log-likelihood function (8), since the stopping rule has not yet been applied. Fig. 10 shows that using the standard log likelihood (5) results in a jump in the predicted return levels after the extreme 2016 heatwave, followed by a slow decrease, whereas those from the conditional fits are more stable.

In a second step, we attempt to estimate the return level in Bikaner, where the extreme is not directly observed, by using a logistic copula (3) to model the dependence between the annual maximum temperatures there and at Jodhpur. Figure 18 of the Supplementary Material compares contours of the fitted joint density and cumulative distribution functions with the maxima.

A stopping rule for the Jodhpur series is then defined using the principle described in Section 3. The parameter α for the logistic copula (3), assumed constant, is estimated. Fig. 11 shows how the estimated return levels in Bikaner change over time. Assuming independent data yields lower return level estimates, while using the dependence with the Jodhpur series to incorporate the extreme event into the prediction is stable over time only when the full conditional likelihood is used, as the jump seen in 2016 in the Jodhpur series in Fig. 10 is also visible in the Bikaner return levels estimated with the standard likelihood; see Fig. 11.

The analysis of the Jodhpur TXx series shows that not accounting for timing bias leads to a jump in return level estimates that dissipates slowly for several years after a trigger event, whereas appropriate conditioning avoids this. For bivariate time series, using even a very basic

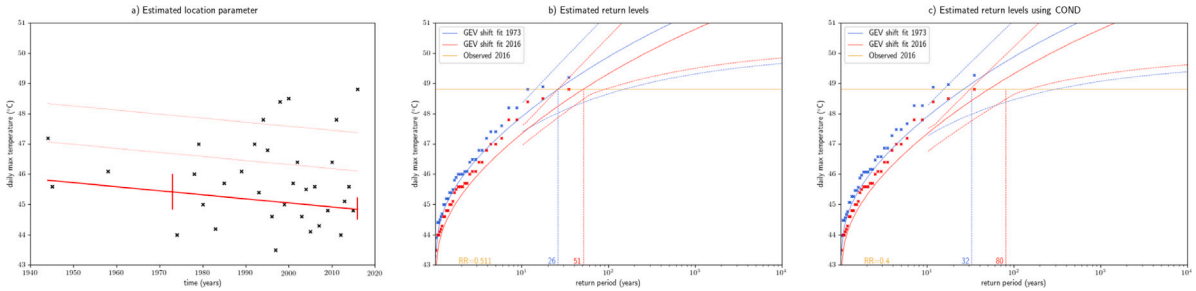

Fig. 8. Phalodi heatwave analysis. Panel (a) shows the estimated location parameter $\hat{\mu}_t$ of the GEV fit (red line) for annual temperature maxima (blue points) at Jodhpur for 1944–2016. The vertical red lines show 95% profile likelihood confidence bounds for μ_t in 1973 and 2016, and the thin red lines denote $\hat{\mu}_t + \hat{\sigma}$ and $\hat{\mu}_t + 2\hat{\sigma}$, where $\hat{\sigma}$ is the estimated scale parameter. Panel (b) displays return level estimates for 1973 (solid blue) and 2016 (solid red) and their 95% confidence intervals (dotted). The observations are shown twice, scaled with the time-related trend (blue and red points). The golden horizontal line represents the extreme temperature observed in Jodhpur in 2016 (48.8 °C), which return periods in 1973 (blue) and 2016 (red) are shown by vertical dotted lines. Panel (c) reproduces plots from Panel (b using the COND log-likelihood (7). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

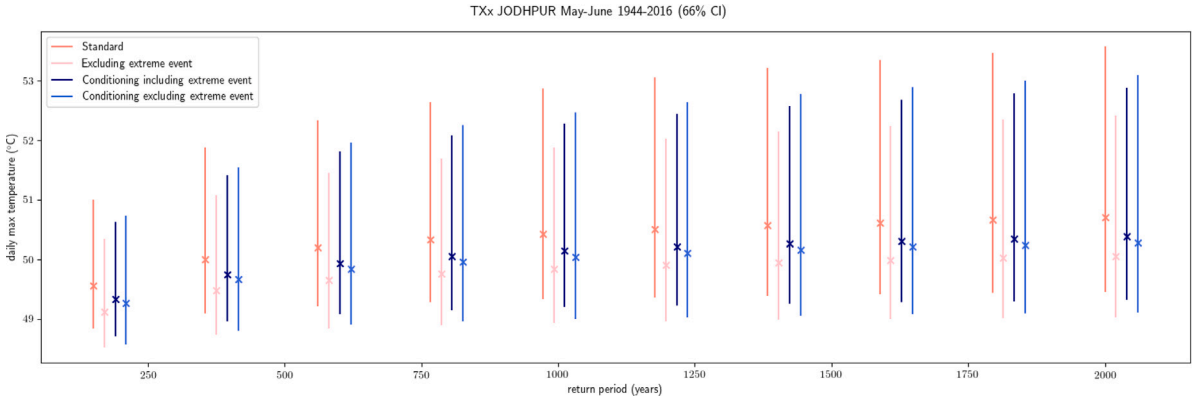

Fig. 9. Phalodi heatwave analysis. Estimated p-year event η_p using GEV fits. The x-axis represents the return period in years. Marks represent the estimated p-year event, for $p = 200, 400, \ldots, 2000$, and vertical bars denote 66% likelihood-based confidence intervals.

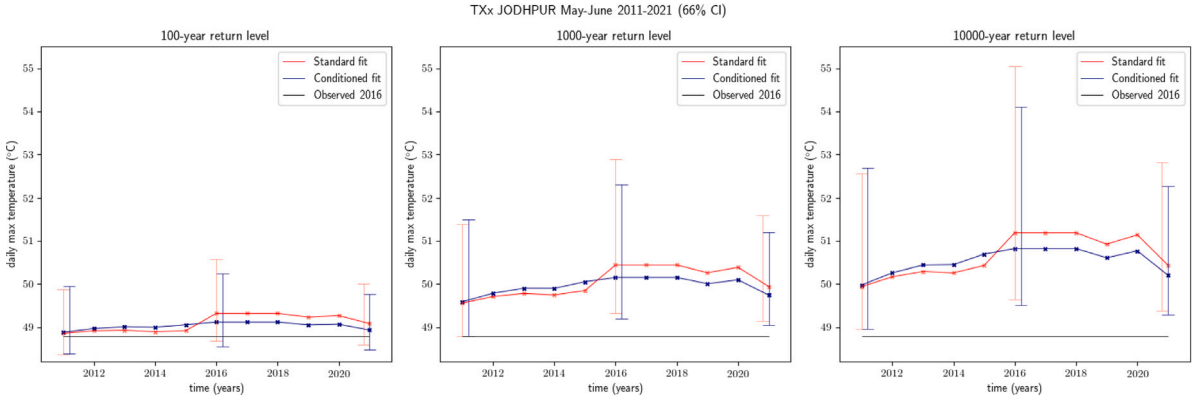

Fig. 10. Phalodi heatwave analysis. Estimated return levels and their 66% confidence intervals (vertical lines finishing with ticks) with the standard and conditioned univariate fits for three return periods for 2011–2021 for TXx in Jodhpur. The horizontal black line indicates the extreme observed in Jodhpur in 2016 (48.8 °C).

correlation model instead of assuming independence has a huge impact on return level estimates, and accounting for timing bias prevents the bias transfer in return level estimation from the stopped series to the nearby series.

We now discuss the impact of spatial selection (Section 2). The logistic copula (3) has $\chi = S^{\alpha}$, and if we assume that we would have performed a similar analysis had an equally extreme event been observed in 2016 at Bikaner rather than at Jodhpur, then $S = 2$ and $\hat{\chi} = S^{\hat{\alpha}} \approx 1.43$. This lies between $\chi = 1$, which would correspond to total dependence between extremes at Bikaner and Jodhpur, and $\chi = 2$, which would correspond to independence. Under this argument the return period of 51 years found in Fig. 8 for the event at Jodhpur, with this location specified before the event occurred, reduces to around 36 years for such an event at one of the two locations, using either the exact formula $m_{\chi} = 1/\{1-(1-1/m)^{\chi}\}$ given by (17) or the approximation $m_{\chi} \approx m/\chi$.

4.3. 2021 Heatwave in the Pacific Northwest

Our third re-analysis concerns the unprecedented "heat dome" event in the United States and Canada in June 2021, which led to wildfires that resulted in the inhabitants of the Canadian town of Lytton becoming climate refugees within a couple of days.

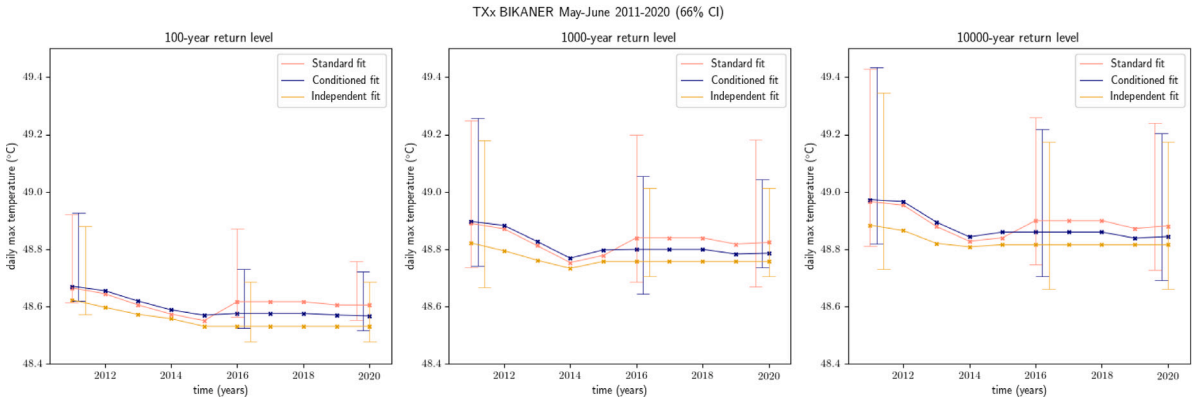

Fig. 11. Phalodi heatwave analysis. Estimated return levels with an independent standard fit (golden line) and with the standard (navy) and conditioned (salmon) fit with a logistic correlation structure for three return periods throughout the 2011–2021 period for TXx in Bikaner. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

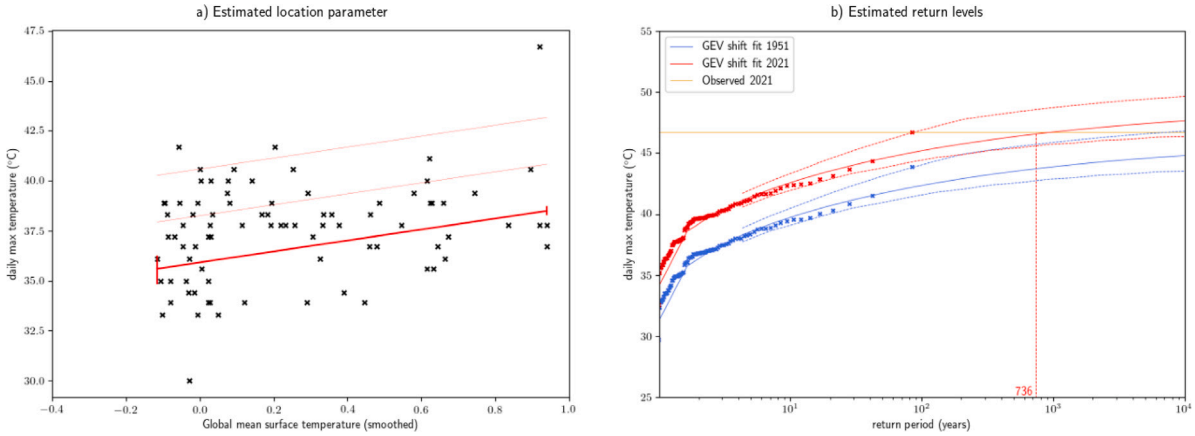

Fig. 12. Portland heatwave analysis. Panel (a) shows the location parameter $\hat{\mu}_t$ of the GEV fit (red line) to annual temperature maxima (blue points) at Portland for 1938–2021, with 95% profile likelihood confidence bounds (vertical red lines). The thin red lines denote $\hat{\mu}_t + \hat{\sigma}$ and $\hat{\mu}_t + \hat{\sigma}$, where $\hat{\sigma}$ is the estimated scale parameter. Panel (b) displays return level estimates for the years 1951 (solid blue) and 2021 (solid red) and their 95% profile likelihood confidence intervals (dotted). The observations are shown twice, scaled with the time-related trend (blue and red points). The golden horizontal line represents the extreme temperature observed in Portland in 2021 (46.7 °C). The return period estimate for this event in 1951 (blue) is shown by a vertical dotted line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We use the Portland International Airport series of observed daily temperatures from the GHCN-D dataset to reproduce part of the attribution study of Philip et al. (2022), wherein data sources and methods are detailed, though we compute likelihood-based confidence intervals rather than use a bootstrap. The chosen stopping rule is the same as in the Phalodi case study; see (15). An increasing trend in the location parameter as a function of the global temperature anomaly (data from NASA-GISS) found in Philip et al. (2022) is shown in Fig. 12a, and Fig. 12b shows return levels using the standard log-likelihood and their 95% confidence intervals. The return period for the 2021 event is displayed in Fig. 12b, though the prediction is very uncertain (95% confidence bounds are available in Table 1).

Table 1 shows how timing bias affects risk ratio estimation. The historical and current probabilities of crossing the previous TXx record of 41.7 °C and its confidence interval are computed by fixing the location parameter of the GEV to the historical/current value for the linear trend in the global temperature anomaly, which is the approach of the WWA (Hammerling et al., 2019). We then parametrize the GEV in terms of the probability of exceeding this level and fit using the different log-likelihoods. Confidence intervals for the risk ratio were obtained using the delta method on its logarithm; see the Supplementary Material. A similar computation applies to the return period for the 2021 Portland event, with the GEV parametrized in terms of its return period and confidence intervals obtained using the profile likelihood.

Including the extreme event without conditioning yields a much shorter return period for the 2021 Portland temperature of 46.7° than when using conditioning, but the latter somewhat increases the risk ratio (Table 1); note that the confidence intervals for the risk ratio based on the standard and conditional fits do not overlap. Excluding the trigger event makes it impossible to estimate its return period, and the estimated risk ratio is less than half that computed by including this event; the same applies for conditional analysis without the trigger event.

5. Discussion

Our results in Sections 3.2 and 3.3 imply that in both univariate and bivariate settings it is generally better to exclude the trigger event if a conditioned fit is not used. In the univariate simulation framework with fixed sample size, the relative bias and relative root mean squared error reduce for $\tau \geq 80$ if the trigger is excluded (Fig. 2). Fitting using a conditional log-likelihood always gives less biased return level estimates, even if the trigger event is not very extreme: simulations for both univariate and bivariate data show much lower bias using the conditioned log-likelihood function for $\tau \leq 200$, and it is increasingly

Table 1
Comparison of estimated risk ratios p_1/p_0 and current return periods for the extreme 2021 temperature $\mathbb{P}\left(\mathrm{TXx}_{2021} > 46.7\ ^\circ\mathrm{C}\right)$ for different log-likelihoods used to fit the Portland TXx series. The factual probability is defined as $p_1 = \mathbb{P}\left(\mathrm{TXx}_{2021} > u\right)$ and the counterfactual (or pre-industrial) probability as $p_0 = \mathbb{P}\left(\mathrm{TXx}_{1951} > u\right)$ for an extreme threshold u, here taken to be the previous record of 41.7 °C. Also given are 95% confidence intervals for the risk ratio and the return period for the 2021 event.

	Risk ratio	Return period for Portland 2021 (years)
Standard	3.31 [3.20, 3.44]	736 [147, 5744]
Excluding	1.41 [0.94, 2.10]	∞ [∞, ∞]
Cond	3.77 [3.68, 3.86]	1830 [183, 16987]
CondEx	1.51 [1.06, 2.14]	∞ [∞, ∞]

important to use an appropriate likelihood when the trigger event becomes more extreme (see the results for $\tau > 500$ in Section 3). Table 1 suggests that although it depends heavily on the trigger event, the estimated risk ratio is much more stable, presumably because it contrasts two probabilities that are typically positively correlated; the same can be expected for functions of the risk ratio, such as the fraction of attributable risk.

The results of Section 3.4 suggest that attributing a return period to a specific observation should if possible be avoided, but if this is essential then the observation itself should be excluded from the fit, which should be performed using a conditional log-likelihood; an uncertainty statement should be included. In any case, the ratio of the estimated and true return periods for a single large observation is extremely uncertain. When the estimated shape parameter $\hat{\xi}$ of the extremal distribution is negative, as often arises for temperature data (see Sections 4.2 and 4.3), the return periods for certain future events may be infinite (see Table 1). This highlights another limitation of the statistical method: when $\hat{\xi} < 0$, excluding the trigger event may make this event effectively impossible. Including the extreme event is then preferable to excluding it, and applying appropriate conditioning will provide roughly unbiased (but very variable) results.

We now summarize the issues that our work raises for the choice of the statistical model for event attribution under an implicit stopping rule.

1. Potential timing bias may be suggested by time series in which the last value is rather unusual.
2. The stopping rule may be difficult to formulate precisely: if obtaining a suitable quantitative definition of an extreme event is impossible, it will be necessary to assemble contextual evidence about what is seen as extreme in the given context and to use that to guess a stopping rule for use in sensitivity analyses.
3. Accounting for timing bias by fitting the data with a conditional log-likelihood is generally desirable, but if for some reason a standard log likelihood must be used, then it is better to exclude the trigger event.
4. A multivariate extremal model allows the analyst to assess the potential effects of spatial selection in the analysis of several related series.
5. Return period estimation for the trigger event can be biased and very uncertain and thus should be avoided, but if it is required then some indication of its uncertainty is essential.

Further work could explore sensitivity analysis on a set of plausible stopping rules with varying thresholds and historical sample sizes. This paper concerns EEA studies that use observations in combination with possibly non-stationary extreme value distributions to estimate return levels, and our simulation studies and examples are specific to the timing bias problem using extreme value models. However, the general framework described in Section 1 could be used for conditioning any type of event with any distribution. Although conceptually straightforward, numerical aspects may become problematic when the computing the probability of the stopping event is complex; see the Supplementary Material. Further work could address selection biases relevant to other EEA methodologies.

6. Conclusion

Existing work on overcoming timing or spatial selection bias in extreme-value statistics has implications for return-level-based extreme event attribution analysis. Indeed, when such a bias exists, not taking it into account in the event attribution can lead to poor, unstable, return level estimates, seriously biased estimates of return periods for extreme observations, and hence to potentially misleading conclusions. Conditioning of the likelihood term uses contextual information more appropriately and hence leads to more reliable findings.

CRediT authorship contribution statement

Ophélia Miralles: Methodology, Software, Validation, Formal analysis, Resources, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Anthony C. Davison:** Methodology, Formal analysis, Software, Supervision, Writing – review & editing, Funding acquisition.

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.

Data availability

The simulations and real case studies in Vargas, Phalodi, and Portland are implemented in Python code available on GitHub (https://github.com/OpheliaMiralles/timing-bias-extremes). The open-source package pykelihood was used for implementation of inference using stopping rule (https://github.com/OpheliaMiralles/pykelihood). A pipeline for downloading and processing the Phalodi data can be found in the same GitHub repository. Daily maximum temperatures were obtained from the NOAA publicly accessible dataset. The Vargas precipitation data may be found in the R package mev.

Acknowledgments

Ophélia Miralles acknowledges funding from the Swiss National Science Foundation (project 200021_178824).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.wace.2023.100584.

References

Allen, M., 2003. Liability for climate change. Nature 421 (6926), 891–892.
Barlow, A.M., Sherlock, C., Tawn, J., 2020. Inference for extreme values under threshold-based stopping rules. J. R. Stat. Soc. Ser. C. Appl. Stat. 69 (4), 765–789.
Belzile, L.R., Davison, A.C., 2022. Improved inference on risk measures for univariate extremes. Ann. Appl. Stat. 16, 1524–1549.
Coles, S.G., Pericchi, L.R., 2003. Anticipating Catastrophes through Extreme Value Modelling, Vol. 52. pp. 405–416.
Coles, S.G., Pericchi, L.R., Sisson, S.A., 2003. A fully probabilistic approach to extreme rainfall modelling. J. Hydrol. 273, 35–50.
Davison, A.C., Huser, R., Thibaud, E., 2019. Spatial extremes. In: Gelfand, A.E., Fuentes, M., Hoeting, J.A., Smith, R.L. (Eds.), Handbook of Environmental and Ecological Statistics. Chapman & Hall/CRC, Boca Raton, pp. 711–744.
Davison, A.C., Smith, R.L., 1990. Models for exceedances over high thresholds (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 52, 393–442.
Fischer, E.M., Knutti, R., 2015. Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. Nature Clim. Change 5 (6), 560–564.
Fischer, E.M., Knutti, R., 2016. Observed heavy precipitation increase confirms theory and early models. Nature Clim. Change 6 (11), 986–991.

Hammerling, D., Katzfuss, M., Smith, R.L., 2019. Climate change detection and attribution. In: Gelfand, A.E., Fuentes, M., Hoeting, J.A., Smith, R.L. (Eds.), Handbook of Environmental and Ecological Statistics. Chapman & Hall/CRC, pp. 789–817.

Hernàn, M.A., Robins, J.M., 2020. Causal Inference: What if. Chapman & Hall/CRC, Boca Raton.

Jones, M.W., Smith, A., Betts, R., Canadell, J.G., Prentice, I.C., Le Quéré, C., 2020. Climate change increases risk of wildfires. ScienceBrief Rev. 116, 117.

Larsen, A., Yang, S., Reich, B.J., Rappold, A.G., 2020. A spatial causal analysis of wildland fire-contributed PM2. 5 using numerical model output. arXiv preprint arXiv:2003.06037.

Lerch, S., Thorarinsdottir, T.L., Ravazzolo, F., Gneiting, T., 2017. Forecaster's dilemma: Extreme events and forecast evaluation. Statist. Sci. 106–127.

Méndez, W., Gil, H.A.P., Ríos, S.C., Montilla, A.M., León, C., 2015. Caracterización hidroclimatológica y morfométrica de la cuenca del río San Julián (estado Vargas, Venezuela): aportes para la evaluación de la amenaza hidrogeomorfológica. Cuadernos de Geografía: Revista Colombiana de Geografía 24 (2), 133–156.

National Academies of Sciences, Engineering, and Medicine, 2016. Attribution of Extreme Weather Events in the Context of Climate Change. National Academies Press.

Naveau, P., Hannart, A., Ribes, A., 2020. Statistical methods for extreme event attribution in climate science. Annu. Rev. Stat. Appl. 7, 89–110.

Northrop, P.J., Coleman, C.L., 2014. Improved threshold diagnostic plots for extreme value analyses. Extremes 17 (2), 289–303.

Otto, F.E., Philip, S., Kew, S., Li, S., King, A., Cullen, H., 2018. Attributing high-impact extreme events across timescales—a case study of four different types of events. Clim. Change 149 (3), 399–412.

Philip, S., Kew, S.F., Jan van Oldenborgh, G., Otto, F., O'Keefe, S., Haustein, K., King, A., Zegeye, A., Eshetu, Z., Hailemariam, K., et al., 2018. Attribution analysis of the Ethiopian drought of 2015. J. Clim. 31 (6), 2465–2486.

Philip, S., Kew, S., van Oldenborgh, G.J., Otto, F., Vautard, R., van der Wiel, K., King, A., Lott, F., Arrighi, J., Singh, R., van Aalst, M., 2020. A protocol for probabilistic extreme event attribution analyses. Adv. Stat. Climatol. Meteorol. Oceanogr. 6 (2), 177–203.

Philip, S.Y., Kew, S.F., van Oldenborgh, G.J., Anslow, F.S., Seneviratne, S.I., Vautard, R., Coumou, D., Ebi, K.L., Arrighi, J., Singh, R., van Aalst, M., Pereira Marghidan, C., Wehner, M., Yang, W., Li, S., Schumacher, D.L., Hauser, M., Bonnet, R., Luu, L.N., Lehner, F., Gillett, N., Tradowsky, J.S., Vecchi, G.A., Rodell, C., Stull, R.B., Howard, R., Otto, F.E.L., 2022. Rapid attribution analysis of the extraordinary heat wave on the Pacific coast of the US and Canada in June 2021. Earth Syst. Dyn. 13 (4), 1689–1713.

Reich, B.J., Yang, S., Guan, Y., Giffin, A.B., Miller, M.J., Rappold, A., 2021. A review of spatial causal inference methods for environmental and epidemiological applications. Internat. Statist. Rev. 89 (3), 605–634, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12452.

Risser, M.D., Wehner, M.F., 2017. Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during Hurricane Harvey. Geophys. Res. Lett. 44 (24), 12–457.

Stott, P.A., Christidis, N., Otto, F.E., Sun, Y., Vanderlinden, J.-P., van Oldenborgh, G.J., Vautard, R., von Storch, H., Walton, P., Yiou, P., et al., 2016. Attribution of extreme weather and climate-related events. Wiley Interdiscip. Rev. Clim. Change 7 (1), 23–41.

Stott, P.A., Stone, D.A., Allen, M.R., 2004. Human contribution to the European heatwave of 2003. Nature 432 (7017), 610–614.

van der Wiel, K., Kapnick, S.B., van Oldenborgh, G.J., Whan, K., Philip, S., Vecchi, G.A., Singh, R.K., Arrighi, J., Cullen, H., 2017. Rapid attribution of the August 2016 flood-inducing extreme precipitation in South Louisiana to climate change. Hydrol. Earth Syst. Sci. 21 (2), 897–921.

van Oldenborgh, G.J., Otto, F.E., Haustein, K., Cullen, H., 2015. Climate change increases the probability of heavy rains like those of storm desmond in the UK–an event attribution study in near-real time. Hydrol. Earth Syst. Sci. Discuss. 12 (12), 13197–13216.

van Oldenborgh, G.J., Philip, S., Kew, S., van Weele, M., Uhe, P., Otto, F., Singh, R., Pai, I., Cullen, H., AchutaRao, K., 2018. Extreme heat in India and anthropogenic climate change. Nat. Hazards Earth Syst. Sci. 18 (1), 365–381.

van Oldenborgh, G.J., van der Wiel, K., Kew, S., Philip, S., Otto, F., Vautard, R., King, A., Lott, F., Arrighi, J., Singh, R., et al., 2021. Pathways and pitfalls in extreme event attribution. Clim. Change 166 (1), 1–27.

Varty, Z., Tawn, J.A., Atkinson, P.M., Bierman, S., 2021. Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. arXiv preprint arXiv:2102.00884.