Richard L. Smith, Department of Statistics and Operations Research,
University of North Carolina at Chapel Hill, Chapel Hill, N.C., U.S.A.

Detection and Attribution of Extreme Weather Events: A Statistical Review

To appear in:
Handbook on Statistics of Extremes
Chapman & Hall / CRC series on Handbook of Modern Statistical Methods
Miguel de Carvalho, Raphael G. Huser, Philippe Naveau and Brian J Reich
(editors)

0.1 Introduction

Detection and Attribution refers to a class of statistical techniques for determining the extent to which observed changes in climate (such as increased temperatures, increased frequency and intensity of hurricane events, and both positive and negative changes in precipitation) may be explained by human influences including, but not limited to, increased greenhouse gas emissions. The methodology was originated by the German physicist Klaus Hasselmann [20, 21, 22], who won the 2021 Nobel Prize in physics, in part because of this work. The subject developed rapidly after 1990, and became a regular feature of the reports of the Intergovernmental Panel on Climate Change (IPCC), most recently in the Sixth Assessment Report [12]. The subject was reviewed by Hammerling *et al.* in a previous Handbook in the Chapman and Hall series [18].

Before moving on to the subject of extreme events, it is worth reviewing the main ideas of this technology:

- The variable of interest, for example, a temperature trend, is estimated simultaneously from observational time series and from climate model simulations. Climate models are large computer programs that simulate the pattern of weather events under different forcing scenarios, such as no forcing (control model runs), natural forcings such as solar fluctuations and volcanic eruptions, and anthropogenic forcings including greenhouse gas emissions, other sources of atmospheric pollution such as particulate matter, and land use changes. Typically these analyses are repeated for data collected over a spatial grid, so we get a spatially correlated set of estimates for both the observations and climate model data.

- The observational estimates are then used to decompose the observational signal into a combination of different model components. The most popular method involves a regression of the observational trends on those derived from climate models, most commonly after a principal components decomposition to reduce the dimension of the problem [36, 23, 27, 2, 1].

- The human signal is said to be *detected* if the regression coefficient corresponding to the anthropogenic effect is statistically significant when all natural forcing factors are included in the analysis — in other words, when the observed change in climate cannot be explained by natural forcings alone. Once the signal is detected, an *attribution* analysis is designed to answer questions of the form "What fraction of the observed warming is due to the human influence?". Such questions are typically answered in terms of the regression coefficients from the fitted statistical model.

- Successive reports of the IPCC have used stronger and stronger language

to describe the strength of the effect, that latest [12] saying, "It is unequivocal that human influence has warmed the atmosphere, ocean and land since pre-industrial times."

Indeed, an argument could be made that, based on present data, the evidence of human influence on climate is so overwhelming that there is no longer any need for detailed statistical analysis to support that claim.

However, the question of *detection and attribution for extreme events* is rather different. In recent years, there have been many notable examples of extreme climate events, a small subset of which we list here:

- Hurricane Harvey, which hit Houston, Texas, in 2017, and resulted in a three-day rainfall that was roughly twice the largest three-day rainfall that had ever been observed previously, with widespread flooding and loss of life as a consequence;

- The Pacific northwest heatwave of June, 2021, which resulted in temperatures that were in some places more than 5°C higher than any that had been observed previously;

- The July 2022 heatwave in southern England and Wales, that resulted in the first ever recorded temperatures over 40° C in London;

- Most recently as of the writing of this chapter, the January 2025 wildfires that devastated large areas around Los Angeles.

Such events are often attributed to "climate change" in the media, but with no specific information about which climate-specific effects are responsible. Hurricanes, wildfires, floods and droughts have always been part of the earth's climate and there is often disagreement even among climate scientists over the extent to which extreme events may have become more frequent, or more extreme, as a result of the human influence. At the other end of the spectrum, one hears statements along the lines "this event was X times more likely in the current climate than it would have been under pre-industrial conditions", but with little detail about how such estimates are reached or even the broad scientific principles on which they are based.

In fact there is a well-developed literature on these topics and methods familiar to extreme value statisticians, including the Generalized Extreme Value and Generalized Pareto distributions, are widely used in making these assessments, but the methods raise many questions about the application of statistical methods to such complex and often ill-defined problems. The objective of this chapter is to review these methods and to make an assessment of how they play into the broader discussion about climate change and its effects on society.

0.2   Overview of GEV/GPD Analyses

It will generally be assumed that readers of this volume are familiar with the basic concepts of extreme value analysis; nevertheless, to fix notation and terminology and make the chapter self-contained, we give a brief introduction.

The Generalized Extreme Value (GEV) distribution is given by the formula

$$\Pr\left\{Y \le y\right\} \quad = \quad \exp\left\{-\left(1 + \xi\frac{y-\mu}{\sigma}\right)_+^{-1/\xi}\right\}$$

where the notation $(\ldots)_+$ denotes positive part (the distribution is restricted to the range $1 + \xi(y-\mu)/\sigma > 0$) and μ, σ, ξ are three parameters known as the location, scale and shape parameters. In environmental applications, this distribution is most often applied to the annual maxima of a variable of interest, though maxima over other time periods may also be considered. The parameters are most often estimated by the method of maximum likelihood (MLE), though other methods are also available; the probability weighted moments (PWM) or L-moments approach is popular in certain sections of the hydrology community and Bayesian methods may also have some advantages as we shall see.

The Generalized Pareto distribution (GPD) is a related distribution for exceedances over a threshold; the exceedance Y is inherently positive and has the distribution

$$\Pr\left\{Y \le y\right\} \quad = \quad 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi}$$

defined on $0 < y < \infty$ if $\xi \ge 0$ or $0 < y < \frac{\sigma}{-\xi}$ when $\xi \le 0$. Here, σ is a scale parameter depending on the threshold and ξ is the same shape parameter as for the GEV. Since this distribution has only two parameters, it is usual to combine it with a separate model for the probability of crossing the threshold [10]

These models are easily extended to include trends; for example, we may assume that the annual maximum in year t has the GEV distribution with parameters μ_t, σ_t, ξ_t; any of these may include covariates, for example, $\mu_t = \beta_0 + \sum_{j=1}^{p} \beta_j x_{tj}$ where $x_{t1}, \ldots, x_{tp}$ are covariates. In climate applications, typical covariates include global mean surface temperature (GMST) or other meteorological variables representing conditions in year t; they may also include circulation indices such as El Niño-Southern Oscillation (ENSO). The covariates may also influence σ_t, though since σ_t is inherently positive, it may be more practical to represent $\log \sigma_t$, rather than σ_t itself, as a linear function of covariates. It is less common to allow ξ_t to be time-dependent; this parameter is in most applications restricted to $|\xi_t| < 1$ and in any case is the hardest of the three parameters to estimate, so even when covariates are included, they are generally not statistically significant.

Similarly with the GPD, the parameters σ and ξ, together with the probability of crossing the threshold, may be made dependent on covariates, though there is an alternative model known as the point process model, which is also based on exceedances over a threshold but uses the $(\mu_t,\ \sigma_t,\ \xi_t)$ parameterization as the GEV, and in some cases makes for a more interpretable fit (for example if μ_t depends on covariates, but not σ_t or ξ_t, this has the interpretation that the effect of covariates is an overall additive shift in the distribution). The interpretation is often convenient for assessing the effect of long-term environmental changes.

The GEV, GPD and point process models have been described in various books that we recommend for background reading, including Coles [8] and Beirlant *et al.* [4].

As an example, Risser and Wehner [34] contructed a dataset representing 7-day total precipitations in an area around Houston, Texas, and used that to examine how extreme was the rainfall event that followed Hurricane Harvey in 2017. They considered statistical models that allowed both μ_t and $\log\sigma_t$ to depend on two covariates: the logarithm of total carbon dioxide in the atmosphere, and the Niño-3.4 index which they took as an indicator for the El Niño–Southern Oscillation effect. The found that "human-induced climate change likely increased the chances of the observed precipitation accumulations during Hurricane Harvey in the most affected areas of Houston by a factor of at least 3.5." The also used the concept of Granger causality [15] to argue that the effect was most likely a causal consequence of climate change, though they did not directly use climate models to substantiate that claim. In the rest of this chapter, we will show how different authors have combined observational data with climate models to validate claims that human-induced climate change is responsible for an increase in the frequency or intensity of extreme events.

0.3   Methods for Extreme Event Attribution

The first paper to attempt a systematic analysis for extremes that mimicked the methods used for conventional detection and attribution analysis was due to Stott, Stone and Allen [37], and was motivated by the devastating European heatwave of 2003 that is widely cited as having killed 70,000 people. In their analysis, they first calculated summer temperature averages (June, July, August) across a wide spatial region: longitudes 10°W to 40°E, latitudes 30° to 50°N. Theu used such a wide region for two reasons: first, to avoid the selection bias that would be an issue if they had focused on a region too closely identified with the more extreme temperatures, and second, to achieve reasonable agreement between observational and climate model data (such agreement is always better when averaged over larger regions of space). They

used five simulations of climate models, four of them including all known anthropogenic and natural forcings and the fifth with just natural forcings (solar variation and volcanic eruptions). As is common in climate studies, both the observational and climate model data were expressed as "anomalies", i.e. differences from the temperatures calculated from a predefined historical period (in this case, 1961–1990).

The authors first performed a conventional detection and attribution analysis on decadal means from the observational and climate model data, essentially estimating regression coefficients for the observational data on the two climate models, and concluding that "the hypothesis that there is no positive anthropogenic influence can be rejected at the 5% level", a result they characterized as "very likely that past anthropogenic forcing is responsible for a significant fraction of the observed European summer warming." They went on to calculate the probability that the 2003 summer mean anomaly was over a threshold of 1.6°C when conditioned on the observed trend, either anthropogenic or natural. Interestingly for an extreme value statistician, they used the Generalized Pareto Distribution (GPD) for this, citing Coles [8] for the technology, though they gave few details about the calculation. One point to note about their paper is that they based their calculations on the probability of exceeding an anomaly of 1.6°C, which was the largest value prior to 2003, rather than the actual 2003 anomaly of 2.3°C, which they justified as "using a threshold that only just exceeds the second warmest summer is relatively conservative." They calculated probabilities p_0 and p_1 for exceeding that threshold, under natural and anthropogenic forcings respectively, estimating $p_1 \approx \frac{1}{250}$ and $p_0 \approx \frac{1}{1000}$. They therefore derived the risk ratio $\frac{p_1}{p_0} \approx 4$ and the "fraction of attributable risk" (FAR), which they defined by the formula

$$FAR \quad = \quad \frac{p_1 - p_0}{p_1}$$

and stated the value 0.75 for that, based on the preceding calculations of p_1 and p_0, though they also provided "normalized estimated likelihood" curves that showed, unsurprisingly, a high level of uncertainty about all of these estimates.

Although the methods for extreme event attribution have advanced considerably since that paper, the basic concepts underlying the paper remain valid. In particular, many recent papers employ methods from extreme value statistics, though mostly based on the Generalized Extreme Value (GEV) distribution rather than the GPD.

During the decade following the paper of Stott, Stone and Allen, a variety of alternative methods was proposed, with considerable disagreements over the results, with some authors claiming that there was no basis for associating extreme events with climate change in every case (for example, [11] claimed that the Russian heatwave of 2010 was primarily caused by a blocking event that was not necessarily connected with climate change, while [24] made similar arguments about the heatwave and drought that hit Texas in

2011). These controversies, and the wide variety of methods being used, led the U.S. National Academies of Science, Engineering and Medicine (NASEM) to commission a review of the whole field [28]. This report was reviewed in more detail in [18], but we reiterate some of the main points here.

In general, the NASEM report did not dwell on statistical details, focusing more on physical explanations of extreme events and the need to use climate models that could represent extreme weather conditions, but they did endorse the use of statistical methods based on either the GEV or GPD distributions, and discussed the difficulties in quantifying the uncertainty in such results using either Bayesian or likelihood-based methods. They also addressed a number of other issues with a statistical flavor:

- **Choice of climate variable.** As noted in [28], "Statements about attribution are sensitive to the way the questions are posed and the context within which they are posed." We already noted that [37] used temperature averages over both a large spatial scale ($50^o \times 20^o$) and a large temporal scale (three-month averages, rather than the approximately one week duration of the most extreme heat), in part to avoid the selection bias issue of choosing an event too closely linked with the actual occurrence. Most subsequent analyses have used more narrowly defined events than that, but the broad issue remains, that it is better in general to select a widely defined event. Another issue is which meteorological variable(s) — for example the 2025 wildfire event in Los Angeles was not so much driven by extreme temperatures as a combination of drought and high winds, which were represented by a number of fire risk indices combining both hazards [3].

- **Framing the question.** The choice is between $FAR = \frac{p_1 - p_0}{p_1}$ and the risk ratio $RR = \frac{p_1}{p_0}$ as the primary measure, where p_1 and p_0 represent the probability of exceeding a high threshold based on all forcings (including anthropogenic) and natural forcings respectively. (The natural forcings result is also sometimes called pre-industrial, because it effectively represent conditions before industrial activity started on a large scale in the mid-nineteenth century.) Although the original paper of Stott, Stone and Allen emphasized the FAR approach, this has a number of disadvantages, especially when it comes to representing uncertainty either through frequentist (e.g. confidence intervals) or Bayesian methods. In some more modern applications of the approach, the pre-industrial probability p_0 is extremely small — indeed, there are some instances where it is claimed to be 0 (a question we shall return to later in this review) — and this makes it hard to represent uncertainty in a meaningful way, but in that context the RR is more easily interpreted than the FAR. The opposite problem, where our estimate of p_0 exceeds that of p_1, occurs less frequently, but it does happen sometimes, and of course the FAR loses its natural meaning if it is negative. Overall, the report recommended using RR rather than FAR.

- **Causality.** A natural question in any study of this nature is whether the observed relationships are causal, but the NASEM report was cautious about the use of a formal causal framework, noting that "a focus on formal analysis of causation may distract attention from important questions about changing probabilities of extreme events and their impacts on risk". The causal framework they adopted was based on Hannart *et al.* [19], which distinguished concepts of "necessary" and "sufficient" causation and proposed probabilistic frameworks for each.

- **Changes in frequency or changes in magnitude?**. The discussion of p_0 and p_1 assumes that the magnitude of the event being considered is the saem in both the present-day scenario and the hypothetical pre-industrial scenario, but it is possible to frame the question the other way round, i.e. for a given pre-industrial probability, what magnitude of present-day event would have the same probability? This can be a more fruitful approach, because different magnitudes of events are directly comparable in terms of their impacts.

- **Conditioning.** A major question addressed by the report was whether probabilities of extreme events should be conditional on broader atmospheric conditions. As an example, Pall *et al.* [29] analyzed the flooding the occurred near Boulder, Colorado, in September 2013, noting "the unusual hydrometeorology of the event" and suggesting that pre-existing conditions may have exacerbated the probability of the extreme flooding that actually occurred. The NASEM report [28] discussed this issue in general terms, noting the elementary probability formula $P(E, N) = P(E|N) \times P(N)$ where E is the event of interest and N is some background event that might have precipitated E. The influence of climate change could be reflected in either $P(N)$, or $P(E|N)$, or both, and the report in effect recommended that researchers consider both types of phenomenon in forming overall conclusions.

- **Selection bias**. This review has already referred to the danger of selection bias, which [28] characterized as "bias from studying only events that occur." This kind of bias may be partially mitigated by expanding the spatial and/or temporal definition of the event, but this is only a partial solution.

0.4 World Weather Attribution

The World Weather Attribution group (https://www.worldweatherattribution.org/) is a collaboration of climate scientists, mostly based in Europe, who have combined to produce rapid attributions of extreme climate events. Their philoso-

phy is that a rapidly produced paper analyzing an event, often within a week or two of the event in question, will have more impact on the general public and the media than a carefully reviewed scientific study which typically takes many months to publish. Nevertheless they are keen to emphasize that their methods have been published in peer reviewed papers, and we summarize some of them here, in particular the two papers [38, 31].

0.4.1 Philip *et al.* (2020).

This paper [38] proposed an "eight-step approach" to the quantitative attribution of extreme events. These are:

1. **Analysis trigger.** The first step is to decide which events to analyze. The authors emphazised that this decision should not be influence by the expected outcome of the analysis, and that the focus should be on high-impact events. Feasibility of the analysis is also a consderation, e.g. whether both observational data and climate model results are available to do the analysis, as well as personnel with the relevant knowledge. They also emphasized the distinction between having a formal procedure to identify suitable events to analyze, and a demand-based procedure based on requests from a national weather service or non-governmental organization. They noted that both types of initiation are used in practice.

2. **Event definition.** The next choice is of a specific event to analyze. Considerations include which meteorological variable to analyze and both the spatial and temporal averaging scales. The objective is to reduce this to a single variable to facilitate calculating the probability of traversing a high or low threshold.

3. **Observed probability and trend.** The next step, and probably the most statistical part of the entire process, is to determine whether there is a trend in the observational data that is distinguishable from natural variability in the process. Considerations include whether to use data from individual weather stations or a gridded data product (e.g. the HadCRUT5 dataset published by the UK Meteorological Office, which calculate historical gridded average over $5^{\circ} \times 5^{\circ}$ grid boxes), and the (spatial) decorrelation length of the variable being studied, e.g. if spatial correlation persists over a large distance it may be necessary to aggregate individual station records to obtain an adequate representation of the event. There is a lengthy discussion of individual statistical methods, including the GEV and GPD methods that are well known in extreme value theory. However they also commented on other methods including Gaussian and gamma distributions. More complicated issues include whether to assume a shift fit or a scale fit or both (in the context of GEV, this essentially means whether the trend is modeled in the

μ or σ parameter, or both) and modes of natural variability, e.g. if an extreme event is influenced by atmospheric or ocean circulation events such as El Niño, it would make sense to include some indicator of those events as a covariate in the analysis. The intention of this entire analysis is to come up with estimates (and appropriate measures of uncertainty, e.g. confidence limits) for the change in probability and intensity of the event of interest.

4. **Model evaluation.** The next step is the selection/verification of an appropriate climate model or models to attribute the trends in observations. The authors emphasized the desirability of starting with a large collection of climate models to maximize the chance of finding at least one model that agrees well with the observations of the variable of interest. This involves statistics tests, e.g. with GEV as the basic statistical model, one could test whether the GEV parameters (μ, σ and ξ), or certain combinations such as σ/μ for precipitation, agree between the observational and model data. They also emphasized the need to verify that the model trends are physically plausible. The authors discussed various procedures of bias correction or calibration to improve the fit between models and observations, though they acknowledged potential challenges to this approach, e.g. the possibility of overfitting if one simply adjusted the model results so that their trends agreed with the observations. The result of this step is a subset of models that adequately represent trends in the data. They suggested that a necessary criterion is for at least two models to pass this test.

5. **Multi-method multi-model attribution.** After determining the observational data and climate models that are used for the analysis, the user proceeds with the attribution itself. The original idea of detection and attribution analysis was to generate two parallel runs of the climate models, one that includes both anthropogenic and natural forcing factors and the other just natural. This method is still widely used, but there is a growing tendency towards using transient model runs, in other words those where the forcing factors change with time, and the comparison is between the present-day climate and some historical period, ideally before the industrial era but if that is not available, some more recent period that predates current warming (e.g. mid twentieth century). They also noted the possibility of using future projections of climate models to estimate probabilities of future extreme events.

6. **Hazard synthesis.** Synthesis here refers to combining the results from different models to produce an overall conclusion for the quantity of interest, which may be a probability ratio between current and historical conditions. Different models typically produce both different means and different measures of uncertainty (e.g. width of

the confidence interval), and they emphasized both ways of combining different measures (e.g. weighted averages) and formal tests such as those based on chi-square statistics. (In other contexts this would be called meta-analysis, but these authors did not use that term). They also discussed different forms of graphical dsplays, and they acknowledged the possibility that if the results are too discrepent between models, one may have to give up on an attribution claim.

7. **Vulnerability and exposure analysis.** The last two steps are less statistical so they are only briefly discussed here. Steps 1–6 are primarily concerned with "hazard", i.e. how the probability of an extreme event has changed because of anthropogenic influences on climate. An overall assessment of risk depends on two other factors: "exposure" essentially means who would be affected by an extreme climate event, and "vulnerability" refers to the magnitude of its impact. Some climate events may be very extreme but less important because the exposure and/or vulnerability are low. An overall assessment of risk depends on a combination of hazard, exposure and vulnerability.

8. **Communication.** The final step they considered was communication of the findings. For this, they emphasized not only publication in the scientific literature, but also providing summaries for policy makers, press releases and social media. Some practitioners of these methods have been very effective in publishing articles in the mainstream media such as *The New York Times* or *The Guardian.*

0.4.2 van Oldenbourgh *et al.* (2021).

Geert Jan van Oldenborgh was a Dutch meteorologist and physicist and one of the founders of the World Weather Attribution project. He was responsible for many of the project's methodological contributions until his untimely death in 2021.

Under "analysis trigger", the authors discussed the challenges of using a societal impact metric to determine which events to analyze. They stated their current criteria of ≥ 100 deaths or $> 1,000,000$ people affected or $\geq 50\%$ of total population affected. However they acknowledged that focusing on high impact events may induce a bias of its own, citing as an example the decreasing frequency of floods induced by snow melt in England; such events did occur in the past, but not recently, therefore the negative impact of climate change in this case will not be represented in attribution studies.

On the question of event definition, the authors acknowledged that "This step has turned out to be one of the most problematic ones in event attribution, both on theoretical and practical grounds". They discussed the tradeoffs involved in choosing a spatial and temporal scale, e.g. averages over large

scales have lower natural variability and are therefore easier to attribute to climate effects, but they may miss the true high-impact event. They also discussed how various framing issues, such as the choice of conditioning variables, may affect the analysis.

Under "observed probability and trend", the authors acknowledged the need to study a time series of sufficent length (they suggested at least 50 years and preferably more than 100) but also highlighted the difficulty of obtaining a sufficiently long record under some circumstances. They also acknowledged sensitivity to the choice of threshold, highlighting the distinction between the Gumbel case ($\xi \approx 0$ in the GEV or GPD model fit) and the cases where ξ is substantially different from 0. This point may be expressed more formally as follows. If we assume a Gumbel distribution with location and scale parameters μ and σ for the annual maxima, then a change in the location parameters from μ_0 to μ_1, holding σ constant, leads to an asymptotic right-hand probability ratio of $e^{-(x-\mu_1)/\sigma}/e^{-(x-\mu_0)/\sigma} = e^{(\mu_1-\mu_0)/\sigma}$ which is independent of the threshold x for which the calculation is being made; but with $\xi \neq 0$ the corresponding ratio is $\left(1 + \xi\frac{x-\mu_1}{\sigma}\right)_+^{-1/\xi} / \left(1 + \xi\frac{x-\mu_0}{\sigma}\right)_+^{-1/\xi}$ which does depend on x and may even be infinite in cases where $\xi < 0$, $\mu_1 - \sigma/\xi > x > \mu_0 - \sigma/\xi$. This is a particular problem with temperature extremes where a GEV fit typically does produce a value $\xi << 0$.

They also commented, "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 realise that the event can be included if the event definition does not depend on the extreme event itself." This has been a somewhat contentious issue; if the observed event is included in the analysis, this eliminates the possibility that the observed event will be beyond the range of the fitted distribution, which as just noted, makes the interpretation problematic. However, from a statistical point of view, it seems more natural to assess the probability of an observation based on a model fit that excludes the observation itself (e.g. this is the whole concept behind cross-validation as it is typically used in statistics). An alternative approach might be to use statistical methods that are less likely to produce zero-probability estimates, e.g. a Bayesian model fit averages over the posterior distribution of the GEV parameters and therefore typically does not lead to a posterior probability of zero for an observed event (though the estimated probability may still be too small to be of practical value, e.g. if very few of the parameter vectors in an MCMC fit result in a non-zero probability for the event in question). They also mentioned the possibility of pooling observations over spatial locations to reduce undertainty in estimating extreme event probabilities, but they did not use formal "spatial statistics" techniques to do this, which have been the focus of much recent work in extreme value theory (see, e.g. [40]). We return to these issues later in this review.

The rest of the paper [31] follows along much the same lines as the earlier paper [38]. They noted the need for "large ensembles or long experiments of multiple climate models and only use the models that represent the ex-

treme under study in agreement with the observed record." They also noted the "need to pay attention that the key global and local forcings are taken into account in the different models to give realistic total trends", mentioning aerosol (particulate matter) emissions and land use changes among local forcing factors that may affect the analysis. They also emphasized the need to communicate in different ways with different groups of potential users of the analysis.

I now summarize the results of two recent WWA analyses.

0.4.3   WWA analysis of the 2021 Pacific northwest heatwave [32].

The authors began by describing the event: temperatures in the Pacific northwest area of the US and the western provinces of Canada that were far above previous records during the last week of June 2021, with an overall highest value in Lytton, B.C., of 49.6°C, which was followed by a devastating wildfire.

The variable they chose was TXx, the annual maximum daily maximum temperature. This was averaged over the region 45–52°N, 119–123°W, which is a large enough region to encompass the three major cities in the region (Portland, Seattle and Vancouver). The source of data was ERA5, a well-known data product produced by the European Center for Medium-range Weather Forecasting. They also looked at individual station records for those three cities. For climate model data, they used data from 18 climate models within the CMIP6 data archive [13], supplemented by simulations from a number of other models. A combination of historical forcings and future emission scenarios was used to run these models from 1850 to present day with future projections up to 2100.

The statistical analysis was based on the GEV distribution with location parameter μ depending on GMST (global mean surface temperature) and σ, ξ held constant. They fitted the same statistical model to climate model data and estimated exceedance probabilities p_1 and p_0 based on GMST for 2021 and for the late nineteenth century respectively, a difference of approximately 1.2°C. They also estimated a future probability of exceedance based on a GMST of 0.8°C above 2021 values (corresponding to 2°C above pre-industrial values, a value that has been determined by IPCC to represent a critical level of warming). Uncertainty was represented by a bootstrap simulation for observational data but (for reasons not fully explained) by a Bayesian-MCMC approach for climate model data.

The initial analysis used gridded observational data up to 2020 and led to a projected upper bound for 2021 of 35.5°C with a confidence interval width (2 standard errors) of 1.3°C; however the observed value was 39.7°C. In other words, even under current climate they estimated a zero probability of achieving the actual observed value. From this they concluded that the GEV analysis is not appropriate to analyze the 2021 heatwave, and they proposed two alternatives: one fitting a GEV without the 2021 value but with a

constraint that the fitted 2021 maximum be greater than the observed value, and the second including the 2021 fully in the analysis. They acknowledged that none of these analyses is fully satisfactory but based on their GEV fits they estimated that the 2021 return value is approximately 1000 years with a confidence interval of $(100, \infty)$. They also estimated a relative risk (RR), compared with preindustrial GMST, of 390 with a confidence interval $(3.2, \infty)$. In addition, the authors performed single-station analyses for Portland, Seattle and Vancouver: all three analyses included 2021 and led to results similar to those for the gridded data.

The remainder of the paper can be summarized more quickly, since it essentially consisted of repeating the same analyses on a wide variety of climate model data. They classified climate models as "good", "reasonable" or "bad" according to how well they agreed with the observational data, as judged by numerous factors but especially their agreement with the GEV model fits. For the most part, only the ones labelled "good" were used subsequently; they computed the RR and a confidence interval for each such model and the proposed a "synthesis" analysis combining different models with weights depending on their uncertainties. Throughout the analysis, however, there is a problem with the RR being infinite (sometimes the point estimate, but almost invariably, the upper bound of the confidence interval is infinite) which they resolved, where needed, truncating the RR at 10,000. However, the need for such an articial fix shows how the issues created by zero-probability estimates, whether these are realistic or not, pervade the whole analysis.

In the end, the authors concluded, "Results for current vs. past climate, i.e. for 1.2°C of global warming vs. pre-industrial conditions (1850–1900), indicate an increase in intensity of about 2.0°C (1.2 to 2.8°C) and a RR of at least 150. Model results for additional future changes if global warming reaches 2°C indicate a further increase in intensity of about 1.3°C (0.8 to 1.7°C) and a RR of at least 3, with a best estimate of 175. This means that an event like the current one, analysed here as having a return period in the current climate of 1000 years, would occur in the future world with 2 °C of global warming roughly every 5 to 10 years according to the best RR estimate, albeit with large uncertainties around it." They discuss various meteorological factors that cold explain the extreme weather conditions, focussing particularly on the large area of very high pressure (the so-called "heat dome" effect) but also discussing the influence of drought conditions and air circulation patterns, but concluding that neither of the last two effects was especially influential. In their final section, they stated "the occurrence of a heat wave of the intensity experienced in the study area would have been virtually impossible without human-caused climate change."

0.4.4 WWA analysis of the January 2025 Los Angeles wildfires [3].

This study analyzed the two large fires, the Palisades and Eaton fires, that hit the city of Los Angeles and surrounding neghborhoods starting on January 7, 2025, that caused over 16,000 lost homes and a financial damage that has yet to be fully quantified. The report was placed online on January 28, thus fulfilling WWA's commitment to make its reports publicly available as quickly as possible after the event. The report is available from WWA's website but has not yet been peer reviewed or published in a scientific journal.

This event was challenging for a climate attribution study because there is no single meteorological variable that was obviously responsible for the event. It was widely reported that 2025 was the hottest January in history, globally, but temperatures were not excessively high in Los Angeles. Instead, two broad factors are blamed for the fire: the exceptionally dry conditions due to drought at the end of 2024, and the Santa Ana winds, that occurred at the same time as the fires and are believed to have greatly exacerbated the damage. However, Santa Ana winds are a common feature of the southern California climate and are not considered to be directly influenced by climate change.

After briefly outlining the background of the fire, the authors defined a number of indices that they used to characterize the severity of the event. The fire weather index (FWI) is computed from temperature, relative humidity and 10-meter windspeed reorded at noon, as well as 24-hour precipitation. The Standardized Precipitation Index (SPI) is based on historical October–December mean precipitations, fitted to a gamma distribution and then transformed to a normal distribution. The drought code (DC) is a subindex of the FWI that is used to determine the date of the end of drought conditions. As covariates, the authors used GMST computed similarly to the earlier report [32] and an index for ENSO. For climate models, they used data from a downscaling experiment that combines several regional climate models (CORDEX) and a high-resolution model (HighResMIP) that uses global sea ice and sea surface temperatures produced by the Hadley Center (part of the U.K. Meteorological Office).

To determine trends in the FWI, they fitted a GEV model with constant σ and ξ whose location parameter μ was modeled as $\mu_0 + \alpha T + \beta I$ where T is a smoothed GMST and I is a detrended ENSO index. They did not consider interactions between T and I but remarked that by removing the climate signal from ENSO they would hope to eliminate any factor of that nature. In calculating the corresponding covariates from the climate models, they used the GMST as given by the model but rescaled the model's ENSO signal to allow for any change in scale between the model and observational data. The FWI for model data was computed from the equivalent meteorological data in the CORDEX experiment.

These analyses showed that the estimated return period for the January 2025 event was 17 years, which they characterized as "unusual but not un-

precedented", but the evidence for a climate effect was equivocal: compared with either pre-industrial GMST or neutral ENSO, the relative risk (RR) was about 1.3 but with a wide confidence interval; they argued that the GMST and ENSO effects were largely independent, so the combined effect of the two is multiplicative; "similarly extreme FWI1X [January max of FWI] values are now 75% more likely to occur than they would have been in a neutral ENSO phase and in a preindustrial climate, and around 12% more intense."

The authors went on to compared these results with those from climate models. They found that only six of the 11 CORDEX models passed the evaluation stage and, of these, two showed no dependence on GMST. Combining all these models, they concluded "the synthesised result is of an increase of 5.7% in peak January FWI intensity (95% interval: -10.4% - +27.4%) and a probability ratio of 1.37 (0.48 - 3.6): a 37% increase in the likelihood of experiencing similarly extreme January FWI in 2025 compared to in a preindustrial climate.... the congruence between the trends in ERA5 and the climate models gives us greater confidence that this is a climate change signal." However they acknowledged that the wide uncertainty bands render this conclusions somewat problematic.

The authors went on to construct similar analyses for the other two indices, SPI and DC, though in these cases they did not use extreme value distributions but instead transformed the raw data to an equivalent normal distribution in order to look at trends through well-known regression techniques. For SPI, they again noted that the very low value in the observational data was "unusual but not unprecedented" (return period about 20 years) and the models gave inconsistent results regarding the climatic contribution to that outcome; they obtained stronger evidence for a climate signal using the more recent HighResMIP models than the older CORDEX models.

For "timing the end of the dry season", they took the day of year on which the 7-day average of DC showed its largest drop (notation, DC-DOY). This variable was found to have a Gaussian distribution so they again used standard regression on GMST or ENSO. Compared with pre-industrial GMST or neutral ENSO, they found a positive shift in DC-DOY (23 days for GMST, 8.6 for ENSO); in both cases, the changes were significant at the 10% level of significance but not at 5%. They continued by making similar calculations based on the climate models; in both cases, they obtained stronger results using the HighResMIP models than the CORDEX models, but the overall results are ambiguous; they concluded "it is likely that the observed trend toward a delayed end of the drought season was promoted by anthropogenic warming, but that due to the small size of the region evaluated in this study, we do not detect a consistent signal in the climate models in our analysis."

In remaining parts of the paper, the authors also analyzed the effect of the Santa Ana winds; they noted that the wind index for January 2025 was very high, but they could not find a significant trend that might be climate induced. They also analyzed "the expected effect of climate change on burned area in the region", concluding that "the potential burned area in December-January

in the Los Angeles area is today substantially higher than it would be in the absence of climate change", but also that models "do not faithfully reproduce observed trends in burned area, and any real-world changes are the combined result of climate change and direct human interventions in the landscape."

Overall, the paper gives an excellent account of the different meteorological influences on this event. The FWI part of the analysis is the most convincing part from an extreme values point of view, but even there, the evidence for a direct climate effect is equivocal.

0.5   Alternative Statistical Approaches

The 2021 Pacific northwest heatwave has stimulated quite a few alternative analyses, some of which we review here.

Bercos-Hickey *et al.* [5] extended the GEV fitting approach of WWA [32] by fitting the GEV model with covariates to data from many individual stations in the Pacific northwest region and also considered Bayesian in addition to maximum likelihood analyses, but in general their conclusions reinforced those of [32]. For many stations, they again found that the observed maximum temperature for 2021 was above the upper bound of the distribution based on data preceding 2021, in this case estimated using Bayesian methods (which in principle should be more robust, since posterior distributions allow for uncertainty in the parameter estimates in ways that the plug-in maximum likelihood approach does not). They also confirmed that including the 2021 value in the GEV analysis frequently leads to data sets that fail simple goodness of fit tests. Therefore, they concluded that the GEV analysis does not lead to a causal attribution for the influence of climate change on these extreme events. In place of the GEV approach, they performed a regional hindcast analysis, essentially running regional climate models to assess how the event of interest would have changed in a world without anthropogenic climate change or with future warming. They found that, compared with pre-industrial conditions, the anthropogenic effect resulted in an increased temperature of around 1°C, while in a hypothetical future world by the end of the 21st century, the increased temperature could be as high as 5°C. This method has been called the "storyline" approach, and is characteristic of methods that use dynamical analysis of the meteorological event in place of the statistical methods emphasized throughout this review. However, even this approach did not reproduce the very extreme temperatures that were actually observed in 2021.

The analysis of Zhang *et al.* [40] took the statistical approach further, in several directions. One was the inclusion of several physical and geographical covariates. Apart from a covariate that combined the influence of the known greenhouse gases, they included: a measure of the El Niño effect called the ELI (El Niño affects large-scale circulation in the atmosphere and ocean and

is often associated with abnormal temperatures); the PDSI (Palmer Drought Severity Index — moisture in the atmosphere generally cools the air, so a lack of moisture could partly explain the very high temperatures); a measure of "urbanicity" (increased land development could create a heat island effect) and geographically based covariates such as long-term mean precipitation, slope and elevation, and distance from the coast, measured at each spatial location.

The raw temperature data used in [40] came from 438 stations with "homogenized" data in the US states of Washington, Oregon and parts of northern California and Nevada, i.e. data that had been post-processed to remove effects due to nonmeteorological factors such as changes in station location or measurement instrument. However for comparisons with the actually observed temperatures for 2021, they used nonhomogenized data on 470 locations. The slight mismatch of spatial locations was not a problem because the spatial model they fitted (to be described) allowed for computing the GEV parameters at any spatial location $\mathbf{s}$, but one disadvatage of proceeding in this way was that they did not obtain estimates for the Canadian part of the region where the most extreme increases in temperature, compared with historical records, were observed.

The incorporation of additional time-dependent covariates (ELI, PDSI and urbanicity) undoubtedly improved the fit of the statistical model, but the main innovation of [40] was a much more sophisticated treatment of the spatial aspects of the analysis. Instead of fitting a single GEV model to a spatially aggregated temperature dataset as in [32], they treated the 438 stations with homogenized records individually, and performed the following analyses:

1. Each station's annual maxima were fitted to a GEV model in which the standard parameters μ, σ and ξ were allowed to depend on both spatial location $\mathbf{s}$ and time t;

2. The temporal effects were modeled through the various covariates already discussed (in the case of PDSI and urbanicity, the variable also varied spatially);

3. Spatial variation was modeled using thin-plate splines: spatial variation in the coefficients was represented by thin-plate splines in the latitude-longitude coordinates and there were additional components from the topographical variabes;

4. However in addition to all of these spatially and temporally dependent effects, the analysis allowed for the spatial coherence of individual events, reflecting the fact that a single extreme heat event may well affect multiple stations simultaneously, an effect not accounted for by simply including spatially dependent covariates. To this end, the authors modified a model from an earlier paper of Zhang and co-authors [39], to lead to a copula model of the form

$$X(\mathbf{s}, t) \quad = \quad R_t \cdot W(\mathbf{s}, t) + \epsilon(\mathbf{s}), \qquad (1)$$

where R_t are independent Pareto variables with index $(1-\delta)/\delta$ (i.e. $\Pr\{R > r\} = r^{-(1-\delta)/\delta}$ for $1 < r < \infty$), $W(\mathbf{s}, t)$ is an isotropic stationary Gaussian process transformed to Pareto with index 1, and $\epsilon(\mathbf{s})$ are independent Gaussian random variables with mean 0 and variance τ^2. The parameter δ controls the degree of dependence between neighboring stations, with $\delta \in (0, 1/2]$ corresponding to asymptotic independence and $\delta \in (1/2, 1)$ to asymptotic dependence.

5. The process $X(\mathbf{s}, t)$ defined by (1) is then transformed marginally to a process $Y(\mathbf{s}, t)$ whose GEV parameters $\mu(\mathbf{s}, t)$, $\sigma(\mathbf{s}, t)$, $\xi(\mathbf{s}, t)$ satisfy the spatial and temporal dependences described in steps 2 and 3.

The analysis of [40] did not use climate models; instead, for their causal attribution statement they relied on the concept of Granger causality [15], following the example of [34] and several subsequent papers. In this analysis, the "factual" scenario corresponded to present-day values of the temporal covariates, while the "counterfactual" scenario considered was for historical (e.g. 1950) greenhouse gases with present-day values of ELI, PDSI and urbanicity. In this way, they isolated the effect of greenhouse gases from the other components of the statistical model. The models were fitted by a hierarchical Bayesian algorithm, and the ultimate objective was defined as calculating the posterior distribution of risk ratios of the form

$$RR(\mathbf{s}) = \frac{p_F(\mathbf{s})}{p_C(\mathbf{s})} \qquad (2)$$

at each site $\mathbf{s}$, where $p_F(\mathbf{s})$ and $p_C(\mathbf{s})$ represent the probability of exceeding the observed 2021 daily maximum temperature at site $\mathbf{s}$, conditional on all data prior to 2021, under the factual and counterfactual scenarios respectively.

In a more recent review, Risser $et\ al.$ [35] have compared the Granger and Pearl [30, 19] approaches to causality. Pearl's approach is more direct because it uses climate models run with different forcing components to compare the factual and counterfactual scenarios, but it adds considerable computational cost. The Granger causality approach is easier to apply because it uses just observational data but it relies on time series models to compare present-day with pre-industrial conditions. It seems likely that both approaches will continue to be used in the future.

We give here only a very brief summary of the results of the analysis from [40]; much more detailed results are in the original paper. Under the full spatial model just described, they found that only 3.4% of stations were "unexplainable" in the sense that the median posterior upper bound of the GEV distribution fitted to 1950–2020 was below the 2021 value. However, that number would have increased about about another 10% under the counterfactual scenario of 1950 greenhouse gases. They stated that the counterfactual risk probability (i.e. $p_C(\mathbf{s})$ in (2) was typically between 0.01 and 0.1, while

the factual risk probability ($p_F(\mathbf{s})$) was typically between 0.05 and 0.3. There were still some stations for which the estimated $p_F(\mathbf{s})$ was 0, but in every such case, $p_C(\mathbf{s})$ was 0 as well, producing an undefined risk ratio of 0/0. A plot of estimated risk ratios for all stations shows very wide variation (including both undefined, i.e. 0/0, and infinite estimated ratios) but risk ratios in the range 3–10 are typical and in many cases statistically significant, in the sense that the lower bound of a 95% Bayesian credible interval is > 1.

In summary, this analysis gives a much more nuanced view of the 2021 heatwave. In some stations the risk ratio is still undefined or infinite, and there are an isolated few where it is < 1, but overall the picture is of risk ratios where both the median posterior estimate and the lower bound of a 95% credible interval are > 1 and in some cases of the order of hundreds or thousands. All of this only reinforces that climate change played a major role in this event. However, we should also note that the analysis was confined to U.S. stations, though Canadian stations were mostly the ones that experienced the most extreme temperatures in excess of past values.

0.6 Examples

In this section, I present some datasets and analyses to illustrate the points made in previous section, and to propose some new interpretations. My intention is not to break new ground research-wise, but rather, to illustrate different possibilities using well established methods of extreme value analysis. The analyses use R [33], in particular the `ismev` package [26].

Because the 2021 Pacific northwest heatwave has attracted so much attention and controversy, the discussion here is focused on that event, though the methods are intended to be applicable to other datasets as well. The main source of data is the Climate Explorer website, https://climexp.knmi.nl/pacificheat_timeseries.cgi. Specifically, I use their reconstruction of Daily Maximum Near-Surface Air Temperature for 1950–2021, derived from ERA5 (a "reanalysis" data product that uses dynamic models to interpolate data between weather stations) and averaged over longitudes 123.125 to 118.875°W, latitudes 44.875 to 52.125°N. The daily data are used to compute annual maxima (also available directly from the website), as well as annual global mean surface temperatures (GMST), smoothed over four years to remove ENSO effects, as described in [32].

0.6.1 GEV analysis

A first analysis uses annual maximum daily maximum temperatures (TXx) and fits the GEV model with location parameter depending linearly on GMST.

Thus the GEV parameters for year t are defined by

$$
\begin{aligned}
\mu_t &= \beta_0 + \beta_1 GMST_t, \\
\sigma_t &= \sigma, \\
\xi_t &= \xi.
\end{aligned} \tag{3}
$$

where $GMST_t$ is GMST for year t.

This model was fitted to data from 1950–2020, so that it could be used retrospectively to predict the maximum temperature for 2021 based on prior data. However, [32] also suggested fitting the model including the 2021 event. I give both models in Table 0.6.1, using the `gev.fit` function in `ismev`.

Parameter	Excluding 2021		Including 2021	
	Estimate	Standard error	Estimate	Standard error
β_0	29.720	0.286	29.242	0.284
β_1	1.758	0.548	2.595	0.641
σ	1.729	0.174	1.694	0.148
ξ	−0.469	0.085	−0.128	0.055
NLLH	132.2	—	145.1	—

TABLE 0.1
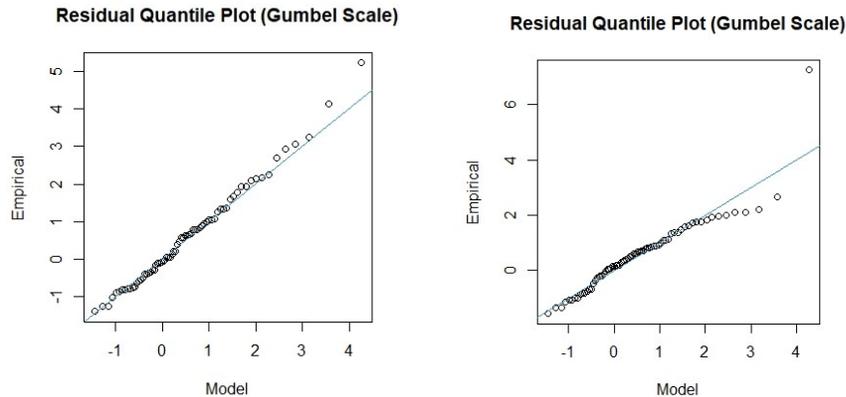Table of GEV parameters for Pacific northwest TXx.

FIGURE 1
QQ plots based on residuals from GEV fit. Left: fitted to 1950–2020 annual maxima. Right: fitted to 1950–2021 annual maxima.

I also used the `gev.diag` function in `ismev` to obtain the QQ plots of

residuals in Figure 1, both with and without 2021. The right hand plot shows clearly the influence of the large outlier, and this is apparent in the model fits also (in particular, the estimates of both β_1 and ξ are substantially different between the two models). This reinforces the conclusion that if the 2021 outlier is included in the analysis, there are significant issues whether the model fits the data. My own conclusion from this is that we really should omit the outlier from the analysis and use the data up to 2020 to characterize the likelihood of the 2021 event.

This conclusion may be reinforced using goodness of fit tests. The fitted GEV model was used to transform each observation to a uniform distribution on [0,1] and standard test statistics computed, specifically, the Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling statistics that are well-established tests of fit [9]. For these datasets, these statistics were computed and p-values calculated using a simulation method that will be described in more detail elsewhere. The simulated p-values for the three tests are respectively 0.77, 0.68 and 0.77 for the dataset without 2021, but 0.21, 0.047, 0.022 for the dataset including 2021. This reinforces that with 2021 included, the GEV model may not fit the data.

The next step is to perform parallel "factual" and "counterfactual" analyses, where by factual we mean using the GMST values as given, and the counterfactual is effectively saying what would happen without human-induced climate change. Although a full analysis of this question requires climate models, a simple proxy for the counterfactual analysis (suggested by [32]) is to assume a GMST that is 1.2°C below present-day temperature, on the grounds that 1.2°C is approximately the overall rise in global mean temperatures since pre-industrial times. So, in effect, we are comparing the projected annual maximum for 2021 using the actual GMST for 2021, with the corresponding projection where the GMST for 2021 is reduced by 1.2°C.

We can estimate the upper bounds of the distribution for 2021 as

$$\omega_F \;=\; \beta_0 + \beta_1 GMST_{2021} - \sigma/\xi, \;\; \omega_C \;=\; \beta_0 + \beta_1(GMST_{2021} - 1.2) - \sigma/\xi$$

where the suffices F and C stand for factual and conterfactual respectively. Substituting the MLEs, we deduce:

$$\hat{\omega}_F \;=\; 35.05 \text{ (standard error 0.61)}, \;\; \hat{\omega}_C \;=\; 32.94 \text{ (standard error 0.40)},$$

where the standard errors are computed by the delta method. The estimate for $\hat{\omega}_F$ differs slightly from that in [32], which may be due to slight revisions in GMST since the data were originally published in 2021, but there is clearly consistency in the overall results.

An alternative approach is a Bayesian analysis. Assuming a prior distribution which is uniform in $(\beta_0, \beta_1, \log \sigma, \xi)$, but restricted to $|\xi| < 1$ because values outside that range are not physically plausible, I ran a Markov chain Monte Carlo algorithm using the adaptive Metropolis procedure of [16], using 20,000,000 iterations but discarding the first half of the sample and storing data every tenth iteration (so the actual MCMC sample is of size 1 million).
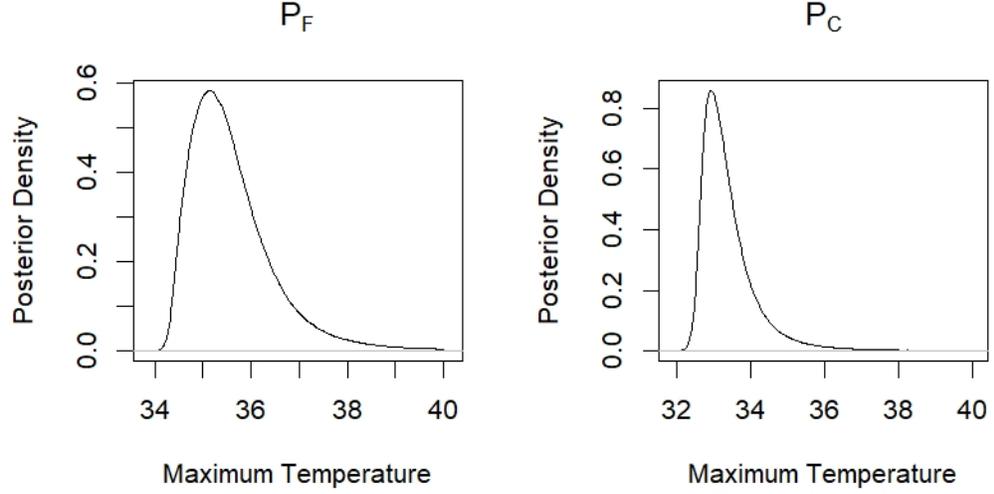
FIGURE 2
Posterior density of the maximum temperature under factual (P_F) and counterfactual (p_C) models.

Figure 2 shows the posterior density of the maximum temperature (ω_F or ω_C) under the factual and counterfactual models. It can be seen that both densities are right-skewed; the posterior probabilities that $\omega_F > 39.6$ or $\omega_C > 39.6$ are 0.0088 and 0.0025 respectively (both numbers subject to slight variation because of the randomness of the MCMC calculation). However, it is not easy to translate this result into a meaningful statement about the relative risk:

1. Let $p_F(x)$ and $p_C(x)$ denote the probability that the maximum temperature in a given year exceeds x under either the factual or counterfactual models ($p_F(x) = 1 - \exp\{-(1 + \xi \frac{x - \beta_0 - \beta_1 GMST_{2021}}{\sigma})_+^{-1/\xi}\}$; $p_C(x)$ is same with $GMST_{2021}$ replaced by $GMST_{2021} - 1.2$). We can estimate $p_F(x)$ and $p_C(x)$ from each row of the MCMC sample, thus obtaining one million pairs from the posterior density.

2. Under this model, the posterior means of $p_F(x)$ and $p_C(x)$, for $x = 39.6$, are 3.1×10^{-6} and 1.8×10^{-7}. These probabilities are much smaller than the corresponding probabilities for ω_F and ω_C, but that reflects the fact that even if the endpoint is > 39.6, the probability that a single annual maximum is > 39.6 is much smaller.

3. Evaluate $RR = p_C(x)/p_F(x)$ for each row of MCMC. Note that I

have reversed the order of numerator and denominator so that I always get the value 0 when $p_C(x) = 0$. However, as a convention, I also take $RR = 0$ when $p_C(x)$ and $p_F(x)$ are both 0. This seems justifiable when $\beta_1 > 0$, because then we always have $p_C(x) < p_F(x)$ when $p_F(x) > 0$.

4. Evaluate $E\{RR \mid \beta_1 > 0\}$. The conditioning on $\beta_1 > 0$ is necessary to exclude a few wild values where $RR >> 1$ in the MCMC sample, but it is justified because the posterior probability that $\beta_1 > 0$ is about 0.9994 and one could question whether $\beta_1 < 0$ is even physically plausible, because it would imply that the Pacific northwest maxima are cooling when the world is warming.

5. I also evaluated the posterior probability that $RR < r$ for various values of $r \in (0, 1)$.

Results of this calculation:

In step 4 I found $E\{RR \mid \beta_1 > 0\}$ is about 7.7×10^{-5}, or about 1 part in 13000. However, apart from the fact that this is very unstable as an estimate, it is almost impossible to put any sensible error bound on this, since 99.8% of the posterior sample values are 0. Alternatively, if we set $r = 0.01$ in step 5, we find the posterior probability that $RR < 0.01$ is about 0.9991. This is still hard to interpret, but in the language used by IPCC [25], an event with greater than 99% probabiity is called "virtually certain". So we could say that it is "virtually certain" that $RR < 0.01$, or in the more conventional framework that RR is defined by p_F/p_C, that the relative risk under anthropogenic compared with natural forcing is greater than 100.

In extreme value theory there has long been a debate about how far outside the range of a sample it is possible to extrapolate. If we did not allow any extrapolation, the theory would have no value, since probability estimates within the range of a sample can be calculated by standard nonparametric estimators. But it stands to reason that there must be a limit how far one can extrapolate. Mathematically, the problem concerns estimating a tail probability or a quantile corresponding to a true tail probability p_n as sample size $n \to \infty$; the question falls within the domain of extreme value theory if $p_n \to 0$, $np_n \to \infty$. There are some elegant theoretical solutions to this problem, for example, [17] proposed a novel variant on bootstrapping for precisely this issue in the case of a Pareto tail (which would not be applicable here). But such theoretical solutions are of little practical value given the asymptotic nature of the procedures and their reliance on second-order assumptions which are hard to verify in practice. Instead, most practical applications of extreme value theory involve assessing the variability of an estimate, essentially assuming that the underlying GEV or GPD model is correct, and then assessing whether the width of the resulting confidence or prediction interval is good enough to allow practical decision making. From this point of view, the preceding analyses are not precise enough to be useful.

As an alternative, I suggest taking a cue from the original paper [37] and

basing the probability estimates on the *second* largest value in the sample, in this case, 33.98°C from 2006. The posterior mean probabilities that the annual maximum is greater than 33.98 are 0.095 and 0.0043 under p_F and p_C; the posterior mean of the estimated risk ratio is 0.0055 (1 part in 180) and the Bayesian 95% credible interval is (0,0.06). We also calculate that the posterior probability that $p_C(33.98)/p_F(33.98) < 0.01$ is about 0.93, greater than 0.9 and therefore "very likely" in IPCC language [25]. Although less dramatic than the calculations based on the observed 2021 event, these estimates are arguably more meaningful, and still confirmatory of the influence of global warming on this event.

0.6.2 Threshold exceedance approach: The point process model

Clarkson *et al.* [7] suggested that a peaks over threshold approach would work better than a block maxima approach for the 2021 heatwave. Their actual analysis involved a novel random effects model applied simultaneously to many stations, but the concept behind their approach is directly applicable using well-established methods. Here, I show how a version of the point process model [8, 4] leads to an improved analysis of the 2021 Pacific northwest dataset of [32].

For this analysis, I use daily data from 1950–2020 and the same annual averages of GMST as previously. All the high temperatures over 29°C occurred between May 15 and September 15 so the analysis is confined to this part of the year. The steps were as follows:

1. Time-varying threshold: a quantile regression was performed with quantile $\tau = 0.9$ and covariates consisting of sinusoidal terms to represent seasonal effects and a linear term for the long-term trend, the fitted values from this model defining the threshold that was used for the subsequent analysis. The 0.9 quantile was selected after trying several others, on the basis that the following analysis leads to excellent agreement between predicted and observed numbers of excesses for all temperatures over 30°C.

2. Declustering of high-level exceedances: there is a clear clustering among the exceedances of this threshold, reflecting the fact that high temperatures tend to persist for several days. Threshold-based analyses should account for this, as argued in [10, 4], for example. I applied the extremal index algorithm of Ferro and Segers [14], implemented through the `extremalIndex` algorithm of the `texmex` R package. The estimated extremal index was about 0.28, and the algorithm was used to identify the peak value in each cluster. Non-peak exceedances were not used in the analysis.

3. The `pp.fit` algorithm from `ismev` was used to fit the peak exceedances over the threshold between days 135–255 each year (ap-

proximately May 15–September 15). The location parameter μ was allowed to depend on annual mean GMST and sinusoidal terms for the annual cycle with periods 12 and 6 months. The scale parameter σ depended just on sinusoidal terms with period 12 months, and the shape parameter ξ was treated as constant. This model was selected after trying several others, comparing their negative log likelihood values and agreement between observed and expected exceedance counts.

4. Based on the fitted model, I computed expected numbers of exceedances over the whole sample for a range of temperatures, and compared them with observed (declustered) counts. The agreement between observed and expected exceedances was excellent ($\chi^2_1 < 1$) for temperatures of 30°C or greater. The analysis was then repeated using the GMST for 2021 to derive projected numbers of exceedances for 2021 under the "factual" model p_F, and similarly, with GMST–1.2 for the "counterfactual" model p_C.

5. I also performed a Bayesian analysis, extracting the `pp.lik` likelihood function from the `pp.fit` algorithm within `ismev`, and running the adaptive MCMC algorithm of [16], similar to the earlier GEV analysis.

Based on the MLE fit, the estimated endpoints, i.e. the largest values for which the expected exceedance count is positive, are 36.4°C and 34.0°C under p_F and p_C respectively. These are larger than the corresponding estimates 35.0 and 32.9 under the GEV model, but still well short of what was observed in 2021. In other words, the threshold analysis improves the estimates a little, but not enough to resolve the basic issue of an observed temperature that was well above the estimated endpoint of the distribution.

The Bayesian analysis follows similar lines to the earlier Bayesian analysis using the GEV model (Table 0.2). For a series of temperatures, we compute the posterior mean of p_F and p_C, and also of the ratio p_C/p_F. Table 0.2 shows the posterior means for all these quantities, as well as a 95% credible interval for p_C/p_F (the lower bound is 0 in every case) and the posterior probability that $p_C/p_F < 0.01$. Based on the latter calculation, we can say that it is "virtually certain" that the relative risk of the 2021 event is > 100 when translated back to the conventional definition of $RR = p_F/p_C$.

0.7 Conclusions

Attribution of extreme weather events is a challenging area for the application of extreme value analysis methods. On the one hand, these applications showcase how the familiar methods based on the GEV and GPD distributions may

Temperature	$E\{p_F\}$	$E\{p_C\}$	$E\{p_C/p_F\}$	$q_{0.025}$	$q_{0.975}$	$\Pr\{p_C/p_F < 0.01\}$
34	0.122	0.0022	0.020	0	0.132	0.65
35	0.031	2.1×10^{-4}	4.8×10^{-3}	0	0.053	0.901
36	4.8×10^{-3}	2.0×10^{-5}	1.4×10^{-3}	0	0.0013	0.971
37	5.0×10^{-4}	2.7×10^{-6}	6.4×10^{-4}	0	0.00083	0.988
38	4.9×10^{-5}	4.8×10^{-7}	3.6×10^{-4}	0	0	0.994
39	5.3×10^{-6}	1.0×10^{-7}	2.1×10^{-5}	0	0	0.996

TABLE 0.2
Expected exceedances of various high temperature in 2021 based on the Bayesian fit to the point process model. Expectations and probabilities are posterior values according to the Bayesian analysis.

be applied to analyze problems of real societal importance. However, there are clear challenges when the methods are applied to very extreme events such as the 2021 heatwave in the Pacific northwest of North America. Temperature data generally lie within the $\xi < 0$ domain of extreme value theory, which means the distributions have finite upper endpoints, and it is problematic when those estimated endpoints lie below the observed extreme values. Alternative statistical analyses, including those based on threshold models or including Bayesian methods of analysis, may resolve some of these issues but typically do not solve the whole problem. More radical alternatives include using different covariates or different methods for handling spatial aspects of the datasets [40, 7]. Finally, this review has not discussed at all the statistical challenges associated with combining data from observations and multiple climate models, but there are recent developments in that field as well [6] and this could also be a fruitful topic for future research.

0.8   Acknowledgement

Bibliography

[1] M.R. Allen, P.A. Stott, and G.S. Jones. Estimating signal amplitudes in optimal fingerprinting, part II: application to a general circulation model. *Climate Dynamics*, 21:493–500, 2003.

[2] Myles R. Allen and Peter A. Stott. Estimating signal amplitudes in optimal fingerprinting, part I: theory. *Climate Dynamics*, 21:477–491, November 2003.

[3] Clair Barnes, Theo Keeping, Gavin Madakumbura, John Abatzoglou, Park Williams, Amir AghaKhouchak, Izidine Pinto, Vikki Thompson, Robert Vautard, Seppe Lampe, Wim Thiery, Rosa Pietroiusti, Friederike Otto, Maja Vahlberg, Roop Singh, Nicole Lambrou, Edward Blakely, Yifang Zhu, Jing Li, Tarik Benmarhnia, Travis Longcore, Miriam Marlier, Emmanuel Raju, Nick Baumgart, and Julie Arrighi. Climate change increased the likelihood of wildfire disaster in highly exposed Los Angeles area. *https://www.worldweatherattribution.org/climate-change-increased-the-likelihood-of-wildfire-disaster-in-highly-exposed-los-angeles-area/*, pages 1–57, January 28, 2025.

[4] Jan Beirlant, Yuri Goegebeur, Jozef Teugels, and Johan Segers. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics, ISBN:9780471976479, Online ISBN:9780470012383, DOI:10.1002/0470012382, 2004.

[5] E Bercos-Hickey, TA O'Brien, MF Wehner, L Zhang, CM Patricola, H Huang, and MD Risser. Anthropogenic contributions to the 2021 pacific northwest heatwave. *Geophys Res Lett*, 49(23):e2022GL099396:doi.org/10.1029/2022GL099396, 2022.

[6] Richard E. Chandler, Clair R. Barnes, and Chris M. Brierley. Characterizing spatial structure in climate model ensembles. *Journal of Climate*, 37:1053–1064, DOI: 10.1175/JCLI–D–23–0089.1, 2024.

[7] Daniel Clarkson, Emma Eastoe, and Amber Leeson. The importance of context in extreme value analysis with application to extreme temperatures in the USA and Greenland. *Applied Statistics*, 72 (4):829–843,https://doi.org/10.1093/jrsssc/qlad020, 2023.

[8] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag, London, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

[9] Ralph B. D'Agostino and Michael A. Stephens. *Goodness-of-Fit Techniques.* Taylor & Francis, https://doi.org/10.1201/9780203753064, eBook ISBN9780203753064, 1986.

[10] A.C. Davison and R.L. Smith. Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society, Series B*, 52(3):393–442, 1990.

[11] R. Dole, M. Hoerling, J. Perlwitz, J. Eischeid, P. Pegion, T. Zhang, X.-W. Quan, T. Xu, and D. Murray. Was there a basis for anticipating the 2010 Russian heat wave? *Geophysical Research Letters*, 38:L06702, doi:10.1029/2010GL046582, 2011.

[12] V. Eyring, N.P. Gillett, K.M. Achuta Rao, R. Barimalala, M. Barreiro Parrillo, N. Bellouin, C. Cassou, P.J. Durack, Y. Kosaka, S. McGregor, S. Min, O. Morgenstern, and Y. Sun. Human Influence on the Climate System. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press*, Chapter 3, doi = 10.1017/9781009157896.005:423–552, 2021.

[13] Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6(CMIP6) experimental design and organization. *Geosci. Model Dev.*, 9:1937–1958, 2016.

[14] C.A.T. Ferro and J. Segers. Inference for clusters of extreme values. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65 (2):545–556, 2003.

[15] C.W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[16] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7 (2):223–242, 2001.

[17] Peter Hall and Ishay Weissman. On the estimation of extreme tail probabilities. *The Annals of Statistics*, 25 (3):1311–1326, 1997.

[18] D. Hammerling, M. Katzfuss, and R.L. Smith. Climate Change Detection and Attribution. *Handbook of Environmental and Ecological Statistics, Chapman and Hall/CRC Handbooks of Modern Statistical Methods*, Chapter 34:789–817, 2019.

[19] A. Hannart, J. Pearl, F.E.L. Otto, and M. Ghil. Causal counterfactual theory for the attribution of weather and climate-related events. *BAMS*, 97:99–110, 2016.

[20] K Hasselmann. On the signal-to-noise problem in atmospheric response studies. In D.B. Shaw, editor, *Meteorology of Tropical Oceans*, pages 251–259. Royal Meteorological Society, 1979.

[21] K Hasselmann. Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, 6(10):1957–1971, 1993.

[22] K. Hasselmann. Multi-pattern fingerprint method for detection and attribution of climate change. *Climate Dynamics*, 13:601–611, 1997.

[23] Gabriele C Hegerl, Hans von Storch, Klaus Hasselmann, Benjamin D Santer, Ulrich Cubasch, and Philip D Jones. Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *Journal of Climate*, 9(10):2281–2306, 1996.

[24] M. Hoerling, A. Kumar, R. Dole, J. Nielsen-Gammon, J. Eischeid, J. Perlwitz, X.-W. Quan, T. Zhang, P. Pegion, and M. Chen. Anatomy of an extreme event. *Preprint*, page http://www.esrl.noaa.gov/psd/csi/pubs/, 2012.

[25] IPCC. Guidance note for lead authors of the IPCC Fifth Assessment Report on consistent treatment of uncertainties. *https://www.ipcc.ch/site/assets/uploads/2017/08/ AR5_Uncertainty_Guidance_Note.pdf*, 2010.

[26] ismev. An introduction to statistical modeling of extreme values. *https://CRAN.R-project.org/package=ismev*, 2018.

[27] Richard A Levine and L. Mark Berliner. Statistical principles for climate change studies. *Journal of Climate*, 12:564–574, 1999.

[28] National Academies of Sciences, Engineering, and Medicine. *Attribution of Extreme Weather Events in the Context of Climate Change*. National Academies Press, Washington, D.C., 2016.

[29] P. Pall, C.M. Patricola, M.F. Wehner, D.A. Stone, C.J. Paciorek, and W.D. Collins. Diagnosing conditional anthropogenic contributions to heavy Colorado rainfall in September 2013. *Weather and Climate Extremes*, 1706.03388, 2018.

[30] Judea Pearl. *Causality: Models, Reasoning and Inference (second edition)*. Cambridge University Press, 2009.

[31] Sjoukje Philip, Sarah Kew, Geert Jan van Oldenborgh, Friederike Otto, Robert Vautard, Karin van der Wiel, Andrew King, Fraser Lott, Julie

Arrighi, Roop Singh, and Maarten van Aalst. A protocol for probabilistic extreme event attribution analyses. *Adv Stat Climatol Meteorol Oceanogr*, 6:177–203, 2020.

[32] Sjoukje Y. Philip, Sarah F. Kew, Geert Jan van Oldenborgh, Wenchang Yang, Gabriel A. Vecchi, Faron S. Anslow, Sihan Li, Sonia I. Seneviratne, Linh N. Luu, Julie Arrighi, Roop Singh, Maarten van Aalst, Mathias Hauser, Dominik L. Schumacher, Carolina Pereira, Marghidan, Kristie L Ebi, Rémy Bonnet, Robert Vautard, Jordis Tradowsky, Dim Coumou, Flavio Lehner, Michael Wehner, Chris Rodell, Roland Stull, Rosie Howard, Nathan Gillett, and Friederike E L Otto. Rapid attribution analysis of the extraordinary heatwave on the Pacific Coast of the US and Canada June 2021. *Earth System Dynamics*, 13 (4):1689–1713, 2022, doi: https://doi.org/10.5194/esd-13-1689-2022.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[34] M. D. Risser and M. F. Wehner. Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during Hurricane Harvey. *Geophysical Research Letters*, 44 (24):12457–12464, 2017.

[35] Mark D. Risser, Mohammed Ombadi, and Michael F. Wehner. Granger causal inference for climate change attribution. *arXiv:2408.16004*, 2024.

[36] B.D. Santer, K.E. Taylor, T.M.L. Wigley, T.C. Johns, P.D. Jones, D.J. Karoly, J.F.B. Mitchell, A.H. Oort, J.E. Penner, V. Ramaswamy, M.D. Schwarzkopf, R.J. Stouffer, and S. Tett. A search for human influences on the threman structure of the atmopshere. *Nature*, 382:39–46, 1996.

[37] P.A. Stott, D.A. Stone, and M.R. Allen. Human contribution to the European heatwave of 2003. *Nature*, 432:610–614, 2004.

[38] Geert Jan van Oldenborgh, Karin van der Wiel, Sarah Kew, Sjoukje Philip, Friederike Otto, Robert Vautard, Andrew King, Fraser Lott, Julie Arrighi, Roop Singh, and Maarten van Aalst. Pathways and pitfalls in extreme event attribution. *Climatic Change*, 166:13:1–27, 2021.

[39] L. Zhang, B.A. Shaby, and J.L. Wadsworth. Hierarchical transformed scale mixtures for flexible modeling of spatial extremes on datasets with many locations. *Journal of the American Statistical Association*, 117:539:1357–1369, 2022.

[40] Likun Zhang, Mark Risser, Michael Wehner, and Travis A. O'Brien. Leveraging extremal dependence to better characterize the 2021 Pacific Northwest heatwave. *Journal of Agricultural, Biological, and Environmental Statistics*, https://doi.org/10.1007/s13253-024-00636-8, 2024.