Dorit Hammerling, Matthias Katzfuss and Richard Smith

Climate Change Detection and Attribution

Contents

List of Figures

List of Tables

Part I

Part I

0.1 Introduction

Climate-change detection and attribution is an important area in the climate sciences, specifically in the study of climate change. Statements such as "It is extremely likely that human influence has been the dominant cause of the observed warming since the mid-20th century" are frequently found in the Assessment Reports of the Intergovernmental Panel on Climate Change (IPCC). These types of statements are largely based on to the synthesis of results from detection and attribution studies [6]. Broadly speaking, the goal of climate-change detection and attribution methods is to differentiate if observed changes in variables quantifying weather (e.g., temperature or rainfall amounts) are consistent with processes internal to the climate system or are evidence for a change in climate due to so-called external forcings [24]. External forcings are often categorized into natural and anthropogenic (human-caused) forcings, where solar and volcanic activity are examples of natural forcings and increased greenhouse gas emissions and land use change are examples of anthropogenic forcings. Figure 1 shows a typical example of a detection and attribution study for long-term temperature change. In this example, natural forcings alone can not explain the observed temperature-change, but a combination of human-caused and natural forcings can.

One key challenge is that (for planet Earth) we can only observe a single realization of climate over space and time. This fact makes it intrinsically difficult to detect changes and to attribute them to specific forcings without further constraining information. This is where climate models play on important role, as they can be used to test the evolution of pathways under different forcings scenarios. The prevailing paradigm is that the climate system is a chaotic system, meaning that minute initial condition changes can lead to varying outcomes, and our observed climate is one specific realization of that system. The variability associated with the chaotic nature of the system, referred to as internal variability, is typically estimated from control runs, which are are climate model runs without any external forcings. [34] provide a more detailed conceptual overview of the problem of climate change detection and attribution from a statistical point of view.

Climate model output needs to be calibrated to agree with observed weather, in that the climate models might be biased or be scaled differently than the observations. For example, this means that it is very difficult to directly compare the global average temperature today to that of, say, 100 years ago, and be able to attribute the increase in temperature to specific forcing scenarios obtained from climate models. What is used (in place of absolute changes) are the patterns of changes in the climate in response to a given forcing, which are referred to as fingerprints [35]. This way, the focus is less on whether an increase in global average temperatures is more consistent with

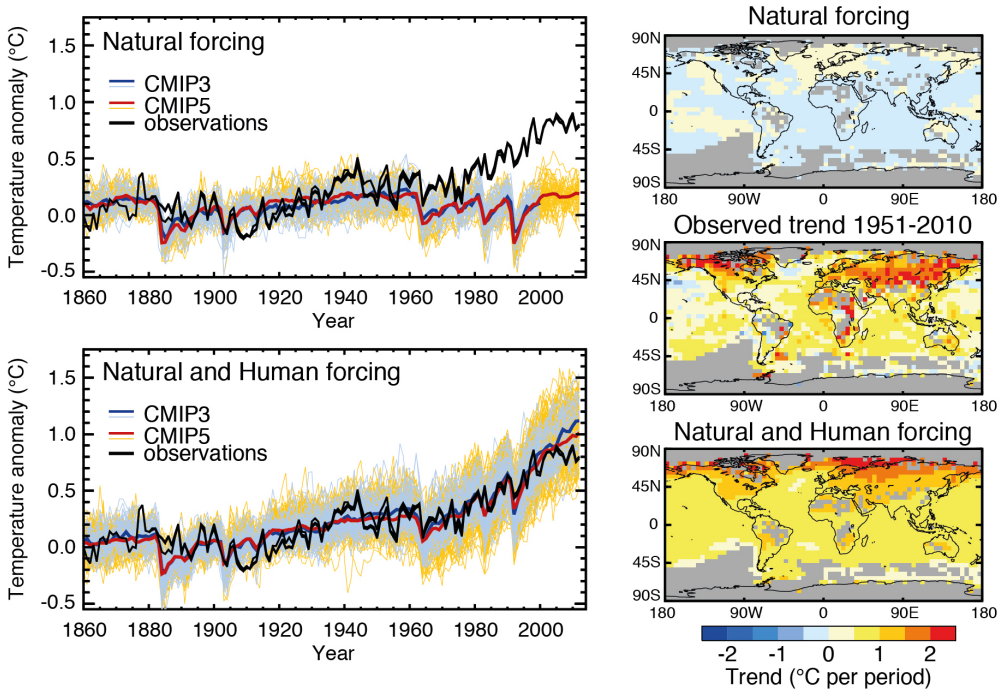

FIGURE 1
Example of a detection and attribution study, reproduced from FAQ 10.1, Figure 1 IPCC 2013: The Physical Science Basis. Time series of global and annual-averaged surface temperature change from 1860 to 2010. The top left panel shows results from two ensemble of climate models driven with just natural forcings, shown as thin blue and yellow lines; ensemble average temperature changes are thick blue and red lines. Three different observed estimates are shown as black lines. The lower left panel shows simulations by the same models, but driven with both natural forcing and human-induced changes in greenhouse gases and aerosols. (Right) Spatial patterns of local surface temperature trends from 1951 to 2010. The upper panel shows the pattern of trends from a large ensemble of Coupled Model Intercomparison Project Phase 5 (CMIP5) simulations driven with just natural forcings. The bottom panel shows trends from a corresponding ensemble of simulations driven with natural + human forcings. The middle panel shows the pattern of observed trends from the Hadley Centre/Climatic Research Unit gridded surface temperature data set 4 (HadCRUT4) during this period.

the observed temperatures, but whether observed changes are greater in a specific region than in another, i.e. how well the patterns of change match.

The most commonly employed framework to address this problem is linear regression, where the observed change is the response variable and a linear combination of the patterns corresponding to the specific external forcing scenarios (obtained from climate models) are the explanatory variables. The inferential goal is the determination of the regression coefficients associated with the different forcings. Their estimated values and uncertainty ranges establish if a change has been detected and to which combination of scenarios it can be attributed.

A different area, discussed in Section 0.4, is extreme event attribution. The focus of extreme event attribution is on assessing specific events such as, for example, an extreme flood. The main goal is to determine if anthropogenic influences have changed the probability of occurrence or magnitude for this particular event. A concept commonly used within this framework is the Fraction of Attributable Risk (FAR), which is defined as $FAR = (p_1 - p_0)/p_1$, where p_1 is the probability of an extreme event with anthropogenic forcings, and p_0 without. [56] and [41] provide recent reviews and we provide a more statistically focussed review here.

Another area, which we will not discuss any further, is the fact that distributions can change in many ways. The simplest, and most commonly considered case, is a change in the mean. But changes in other characteristics of climate, such as the variance, the magnitude or frequency of extreme values, and even changes in the dependence structure over time (e.g., higher likelihood of droughts due to an extended period of no rain) are important and of interest. Here, we will only discuss work focused on changes in the mean.

0.2 Statistical model description

Regression-based climate-change detection and attribution can be viewed as a multivariate spatial or spatio-temporal regression problem, where we express an observed signal as a linear combination of different forcings scenarios. For global studies, the observations and forced responses are typically available as gridded quantities, which are often further aggregated to coarser grids (e.g., to a $2.5^\circ \times 2.5^\circ$ grid, resulting in $144 \times 72 = 10,368$ grid cells or to a $5^\circ \times 5^\circ$ grid, resulting in $72 \times 36 = 2,592$ grid cells). Observations and corresponding forced responses are often averaged in time, e.g. decadal averages, or expressed as estimated slope coefficients from a simple linear regression in each grid cell. Estimating slope coefficients is a straightforward way to smooth out short-term climate variability, which is overlaid on the longer-term trend we are trying to detect. For example, the quantity describing the observations could be slope coefficient estimates based on 30 years of temperature or rainfall observations.

Hence, let $\mathbf{y} = (y_1, \ldots, y_n)'$ be a vector of the true quantity describing the

observations at the n grid cells, and the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_J$ the analogous (true) quantities that would have occurred under the J different forcing scenarios. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$, and define the $n \times n$ covariance matrix characterizing the internal climate variability (without any forcing) as $\mathbf{C}$. We can then write the commonly assumed linear regression model in the form of a conditional distribution,

$$\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \mathbf{C} \sim \mathcal{N}_n\Big(\sum_{j=1}^{J} \beta_j \mathbf{x}_j, \mathbf{C}\Big), \tag{0.1}$$

where $\mathcal{N}_n$ denotes an n-variate normal or Gaussian distribution.

Within this context, climate change detection is viewed as testing whether each of the β_j is equal to zero or not. Assuming that $\mathbf{x}_1$ corresponds to the anthropogenic forcing, the conclusion that $\beta_1 \neq 0$ implies that human-caused climate change (with regard to the specific observed quantity as defined) has been detected. Attribution extends this framework by testing if the β_j are equal to unity, under the assumption that the mean responses for each forcing have been removed and that the responses are additive [e.g., 49]. Under the assumption of normality, the maximum-likelihood estimate for $\boldsymbol{\beta}$ is identical to the generalized-least-squares estimate, which is the solution approach typically pursued within the climate science community.

On first glance, this problem seems trivial. The challenge is, however, that in practice, $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\mathbf{C}$ are all unknown. With the exception of $\boldsymbol{\beta}$, all these unknown quantities are high-dimensional, and our means to learn more about them are rather limited and modeling choices have to be made. Modern understanding further acknowledges that the observations are unknown as well, and can only be observed with measurements errors. Reconstructed and observational data sets are nowadays often provided in the form of ensembles, which allows for the estimation of an observational error covariance matrix to be incorporated in the modeling procedure. The following section describes the development of solution approaches leading up to the most recent formulation using Bayesian hierarchical modeling.

0.3 Methodological development

Climate-change detection and attribution methods have been developed by a variety of groups in the climate science and, to a lesser degree, in the statistics community and notations have varied accordingly. In this section, we apply the notation used in the corresponding original literature.

0.3.1 The beginning: Hasselmann's method and its enhancements

The original name for detection and attribution was *the optimal fingerprint method* [22, 25]. The idea was that human-caused greenhouse gas emissions would not only result in increased temperature overall, but would exhibit distinctive patterns or fingerprints, and with careful analysis, these could be detected in the observational signal. In contrast, other possible causes of warming, such as increases in solar output, would result in quite different patterns. If we were able to detect a pattern that was closer to that associated with greenhouse gases than with changes in solar output, that could be taken as evidence that greenhouse gases, rather than solar variations, were the cause of the changes we saw. The patterns to be expected were taken from climate models in which the different possible forcing factors could be separated into different model runs.

As originally formulated by Hasselmann [21], the setting was as follows:

1. The overall signal (for example, the change over time in a modeled temperature field) is represented by an n-dimensional vector $\bar{\mathbf{\Phi}}$;
2. The estimated change from observation data is written $\mathbf{\Phi}$;
3. We assume $\mathbf{\Phi} - \bar{\mathbf{\Phi}} \sim \mathcal{N}(0, \mathbf{C})$ (multivariate normal with mean 0 and covariance matrix $\mathbf{C}$);
4. $\mathbf{C}$ estimated from data but *treated as known*;
5. The null hypothesis $H_0 : \bar{\mathbf{\Phi}} = 0$ is tested using a χ^2 test.

As Hasselmann showed through a detailed example, this formulation is too simple without further structure on the signal. For example, the amplitude of signal required to be detected at a given level of significance increases with the dimension of the signal itself. Therefore, it is desirable to make use of further information on the anticipated form of the signal.

To bring this idea into the analysis, Hasselmann assumed we could write the signal as a linear combination of individual signals (the fingerprints). Therefore, we write $\bar{\mathbf{\Phi}} = B\bar{\mathbf{\Psi}}$ where B is a $n \times p$ matrix of known basis functions (interpreted as a p-dimensional "signal"). A revised estimate $\tilde{\mathbf{\Phi}}$ is chosen to minimize $|\tilde{\mathbf{\Phi}} - \mathbf{\Phi}|^2$. This in turn is used to construct a revised χ^2 test statistic. A key part of the method is *expansion in principal components*. In the climate literature, principal components are known as Empiricial Orthogonal Functions or EOFs. Hasselmann anticipated that it might in practice be necessary to restrict to a small number of leading EOFs (he suggested between 5 and 20).

The initial paper of Hasselmann was followed by a number of extensions and ramifications in the 1990s, e.g. [22, 23]. North and Stevens [39] presented a particularly simple derivation of the main results using linear algebra and the elementary theory of linear models. Even at this time, however, it was also

implicit that a reduction in dimension (for example, restricting the signal to the leading EOFs) was needed to make the method applicable in practice.

The method started to influence the broader climate community with a series of papers in the mid-1990s applying these ideas to large climate datasets, see in particular [25, 52]. For example, Hegerl *et al.* [25] used a guessed greenhouse gas signal from a climate model, information about natural climate variability derived from control runs, and global near-surface temperature temperature observations. The null hypothesis, that changes in observed temperatures could be explained by natural variability, was rejected with a p-value < 0.05. However, they acknowledged considerable uncertainty about natural variability and did not take into account signals from other forcing factors such as solar variation.

The parallel paper by Santer and co-authors [52] focussed on the vertical structure of temperatures through the atmosphere. One particular issue here is the contrast between warming of the troposphere and cooling of the stratosphere, a pattern that one would expect to be particularly indicative of greenhouse gas warming, whereas other conceivable sources of atmospheric warming, for example if solar radiation were generally increasing, would not lead to such a characteristic vertical pattern of temperature changes. In this paper, they enhanced their conclusions by incorporating other signals besides greenhouse gases (they included stratospheric ozone in their model, as well as sulfate aerosols — small particles in the atmosphere, generally caused by human industrial processes, that have the effect of cooling the atmosphere and thereby partially mitigating the greenhouse gas effect). They also compared results from two climate models to examine the sensitivity of their results to model-dependent uncertainties, and, like [25], used control runs from climate models to assess natural variability, a key step in formulating a statistical significance test. They concluded "it is likely that this trend is partially due to human activities, though many uncertainties remain, particularly relating to estimates of natural variability."

0.3.2 The method comes to maturity: Reformulation as a regression problem; random effects and total least squares

Levine and Berliner [34] showed how Hasselmann's equations could be reformulated as a linear regression problem. The same formulation was proposed independently by Allen and Tett [3] with an observational signal $\mathbf{y}$ regressed on a finite number of covariates $\mathbf{x}_1, ..., \mathbf{x}_J$ representing J signals (for example, greenhouse gases, sulfate aerosols, solar variation, volcanoes) that were supposed to be derived from a climate model. Allen and Stott [4, 2] made the important extension of treating the signals themselves as random quantities, while a paper by Huntingford *et al* [29] showed how to extend the methodology to multiple climate models. The methodology defined in these papers is at the core of many present-day detection and attribution studies. The next

part of our review therefore develops the methodology in these papers in some detail, though we refer to the original papers for full details.

First we outline the paper [3]. The model assumed by them was of the form

$$\mathbf{y} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{0.2}$$

where

- $\mathbf{y}$ is vector of observations ($\ell \times 1$, where ℓ is the number of grid cells — several thousand in a typical climate model);
- $\mathbf{X}$ is matrix of J response patterns ($\ell \times J$ — here J is typically small, for example 4 if the response patterns correspond to greenhouse gases, sulfate aerosols, solar variability and the effects of volcanic eruptions — the first two of these are referred to as *anthropogenic forcing factors* and the latter two as *natural forcing factors*;
- $\mathbf{u}$ is "climate noise", assumed normal with mean 0 and covariance matrix $\mathbf{C}$;
- We assume there exists a normalizing matrix $\mathbf{P}$ such that $\mathbf{PCP}^T = \mathbf{I}$, $\mathbf{C}^{-1} = \mathbf{P}^T\mathbf{P}$.

Then the model (0.2) may be rewritten

$$\mathbf{Py} \quad = \quad \mathbf{PX}\boldsymbol{\beta} + \mathbf{Pu} \tag{0.3}$$

where noise $\mathbf{Pu}$ has covariance matrix $\mathbf{I}$.

The Gauss-Markov Theorem implies that the optimal estimator in (0.3) is

$$\tilde{\boldsymbol{\beta}} \quad = \quad (\mathbf{X}^T\mathbf{P}^T\mathbf{PX})^{-1}\mathbf{X}^T\mathbf{P}^T\mathbf{Py}$$

with covariance matrix

$$V(\tilde{\boldsymbol{\beta}}) \quad = \quad (\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X})^{-1}.$$

A confidence ellipsoid for $\boldsymbol{\beta}$ may be derived from the distributional relationship

$$(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X})^{-1}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_J.$$

The main difficulty in this elegant construction stems from the dimension of the sampling space. Evidently we need an estimate of the covariance matrix $\mathbf{C}$, and the established method [25, 52] is to use control runs of the climate model, but in its general form $\mathbf{C}$ is an $\ell \times \ell$ matrix, and in the best-case scenario we would not have more than about 2,000 years of control-run simulations from which to estimate $\mathbf{C}$. We could have a have a vector of n independent "noise" simulations $\mathbf{y}_N$ and then estimate $\hat{\mathbf{C}} = \frac{1}{n}Y_N Y_N^T$, but typically $n << \ell$ so $\hat{\mathbf{C}}$ is singular.

To resolve this issue, in practice the following steps are followed:

- Restrict to κ EOFs with largest variance (equivalent to replacing $\mathbf{P}$ by $\mathbf{P}^{\kappa}$, consisting of the κ eigenvectors of $\mathbf{C}$ with largest eigenvalues);
- The set of control runs is split into two, one part being used to estimate $\mathbf{C}$ and the other part to estimate $\boldsymbol{\beta}$ and the associated variance estimates and tests of significance;
- These estimates lead to an estimate $\tilde{V}(\tilde{\boldsymbol{\beta}})$ with ν degrees of freedom, where ν corresponds to the *effective sample size* of the control runs. The concept of effective sample size is closely related to the problem of testing for the mean or a trend in an autocorrelated process, which is covered in detail in Chapter 27 of this volume. Allen and Tett referred to the paper by Zwiers and von Storch [60] which has been widely cited in the climate literature.

The end result of these manipulations is a formal test statistic for the significance of $\boldsymbol{\beta}$,

$$(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \tilde{V}(\tilde{\boldsymbol{\beta}})^{-1} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim J F_{J,\nu}$$

which is readily adapted to testing just a subset of the components of $\boldsymbol{\beta}$ or some linear combination of those components.

The final methodological development of Allen and Tett was a procedure for testing the fit of the statistical model. Define

$$\tilde{\mathbf{u}} \;=\; \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}.$$

Then

$$r^2 \;=\; \tilde{\mathbf{u}}^T \mathbf{C}^{-1} \tilde{\mathbf{u}} \;\sim\; \chi^2_{\kappa - J}.$$

With independent control runs

$$\tilde{\mathbf{u}}^T \hat{\mathbf{C}}^{-1} \tilde{\mathbf{u}} \;\sim\; (\kappa - J) F_{\kappa - J, \nu} \text{ approximately.}$$

This can be used as a diagnostic on the model fit and also to guide the choice of κ.

0.3.3 Accounting for noise in model-simulated responses: the total least squares algorithm

The next major methodological development was due to Allen and Stott [4, 2]. They recognized that a flaw in model (0.2) was that it treated the components of the $\mathbf{X}$ matrix as known, whereas in practice, these components (the outputs of a climate model with different forcing factors) are subject to their own random errors due to the internal variability of the climate system. As a first approximation, these random errors should have distributions similar to those of the control runs, so it would be reasonable to assume they have the same means and covariances.

Allen and Stott [4] rewrote (0.2) as follows. First, note that the term $\mathbf{X}\boldsymbol{\beta}$ in (0.2) may also be written $\sum_{j=1}^{J} \mathbf{x}_j \beta_j$ where $\mathbf{x}_j$ is the model-generated signal. Second, assume each observed $\mathbf{x}_j$ is a perturbation of some "true signal" and can therefore be rewritten $\mathbf{x}_j - \mathbf{u}_j$ where $\mathbf{x}_j$ is the true signal and $\mathbf{u}_j$ a random error. This leads to the model

$$\mathbf{y} \quad = \quad \sum_{j=1}^{J} (\mathbf{x}_j - \mathbf{u}_j)\beta_j + \mathbf{u}_0 \tag{0.4}$$

where $\mathbf{u}_0, \mathbf{u}_1, ..., \mathbf{u}_J$ are assumed to be random errors with a common distribution, $\mathcal{N}(0, \mathbf{C})$ in the typical case that we are considering.

To fit the model (0.4), an appropriate algorithm is not ordinary least squares (OLS), but *Total Least Squares* (TLS), which we discuss next.

According to (0.4), the distribution of $\mathbf{y}$ is normal with mean $\sum_j \mathbf{x}_j \beta_j$ and covariance $(1 + \sum_j \beta_j^2)\mathbf{C}$ so the likelihood function is proportional to

$$|\mathbf{C}|^{-1/2}(1 + \sum_j \beta_j^2)^{-\ell/2} \exp\left\{ -\frac{1}{2} \frac{(\mathbf{y} - \sum_j \mathbf{x}_j \beta_j)^T \mathbf{C}^{-1} (\mathbf{y} - \sum_j \mathbf{x}_j \beta_j)}{(1 + \sum_j \beta_j^2)} \right\}.$$

Therefore, one possible estimator of $\boldsymbol{\beta}$, assuming $\mathbf{C}$ known, would choose $\beta_1, ..., \beta_J$ to minimize

$$\frac{(\mathbf{y} - \sum_j \mathbf{x}_j \beta_j)^T \mathbf{C}^{-1} (\mathbf{y} - \sum_j \mathbf{x}_j \beta_j)}{(1 + \sum_j \beta_j^2)} + \ell \ \log(1 + \sum_i \beta_j^2). \tag{0.5}$$

However, there is a practical difficulty with including the second term in (0.5), which arises from the determinant part of the multivariate normal density: it depends critically on the dimension of the signal ℓ, and as we have already seen, in practice the estimation is carried out in only a low-dimensional subset of the true sampling space.

In fact, the model (0.4) and the estimator (0.5) are special cases of the general *errors in variables* (EIV) regression formulation due to Gleser [14], who proposed a *generalized least squares* algorithm minimizing only the quadratic exponent part of the likelihood function, ignoring that part that arises from the determinant. One argument made by Gleser to support this procedure was that it was less dependent on the errors having an exact multivariate normal distribution.

Applied in this context, Gleser's formulation would choose $\beta_1, ..., \beta_J$ to minimize just the first term of (0.5), in other words

$$Q \quad = \quad \frac{(\mathbf{y} - \sum_j \mathbf{x}_j \beta_j)^T \mathbf{C}^{-1} (\mathbf{y} - \sum_j \mathbf{x}_j \beta_j)}{(1 + \sum_j \beta_j^2)}. \tag{0.6}$$

The solution of (0.6) is the TLS estimator of $\boldsymbol{\beta}$.

Allen and Stott [4] discussed several variants of TLS, but noted that "the

differences [among different approaches] are likely to be much less important than the impact of neglecting response-pattern noise altogether." In the simple case of a single regressor $\mathbf{x}$ the formula amounts to minimizing the sum of squares of perpendicular distances from the data points to the best-fit line, instead of the the sum of squares of vertical distances which is the standard OLS procedure. In this form, the method apparently originated in a paper of Adcock [1].

To see that minimizing (0.6) is in fact equivalent to the Allen-Stott solution, we note the following. After applying a pre-whitening operator, Allen and Stott sought constants $v_0, v_1, ..., v_J$ to minimize $(v_0\mathbf{y} - \sum_{j=1}^{J} v_j\mathbf{x}_j)^T\mathbf{C}^{-1}(v_0\mathbf{y} - \sum_{j=1}^{J} v_j\mathbf{x}_j)$ subject to the constraint $\sum_{j=0}^{J} v_j^2 = 1$, and then defined $\beta_j = v_j/v_0$ for $j = 1, ..., J$. However, the two are the same for the following reason. Fix v_0 and write $v_j = \beta_j v_0$ for $j = 1, ..., J$. Then Allen and Stott minimized $v_0^2(\mathbf{y} - \sum_j \beta_j\mathbf{x}_j)^T\mathbf{C}^{-1}(\mathbf{y} - \sum_j \beta_j\mathbf{x}_j)$ subject to the constraint $1 = \sum_{j=0}^{J} v_j^2 = v_0^2(1 + \sum_{j=1}^{J} \beta_j^2)$ which implies $v_0^2 = 1/(1 + \sum_{j=1}^{J} \beta_j^2)$, reducing to (0.6).

0.3.4 Combining multiple climate models

A further extension of this framework was introduced by Huntingford and coauthors [29] to allow for the possibility of multiple climate models. The key assumption here is that, in addition to the internal noise variability between successive runs of any given model, there is also an "inter-model variability" between the output of one model and another. The covariance matrix for the inter-model variability of signal j (denoted $\mathbf{G}_j$ in the discussion to follow) is assumed to be different from the covariance of the internal variability, and is therefore estimated from the model runs. The result is a model that depends on multiple random components but which may also be estimated by *errors in variables* (EIV) methodology, extending the TLS concept.

In more detail, Huntingford *et al.* extended model (0.4) into

$$\mathbf{y} \quad = \quad \sum_{j=1}^{J} (\bar{\mathbf{x}}_j - \mathbf{u}_j - \mathbf{v}_j)\beta_j + \mathbf{u}_0 \tag{0.7}$$

where $\bar{\mathbf{x}}_j$ is the mean over all M climate models and $\mathbf{v}_j$ is an additional noise term that represents the variability among models around $\bar{\mathbf{x}}_j$. In effect, $\mathbf{v}_j$ is treated as an additional random effect with mean $\mathbf{0}$ and a covariance matrix which we write here as $\mathbf{G}_j$ (different for each forcing variable j). Huntingford *et al.* proposed a specific algorithm to estimate $\mathbf{G}_j$ from ensembles of the individual model runs.

The $\mathbf{u}_j$ terms in (0.7) are again assumed to be dominated by the internal variability component of the noise; however, since in this case it is explicit that climate model runs are averaged — we assume a total of M climate models, where the mth climate model has K_m ensemble members, but the model averages $\bar{\mathbf{x}}_j$ are assumed to be the *unweighted* averages of the M model

averages — the natural assumption is to assume $\mathbf{u}_j \sim \mathcal{N}[\mathbf{0}, \kappa\mathbf{C}]$ where $\kappa = M^{-2}\sum_{m=1}^{M} K_m^{-1}$.

This is an instance of the general EIV algorithm where coefficients $\beta_1, ..., \beta_J$ and denoised values $\mathbf{y}^*,\ \mathbf{x}_1^*, ..., \mathbf{x}_J^*$ are chosen to minimize

$$Q^* \quad = \quad (\mathbf{y}-\mathbf{y}^*)^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{y}^*) + \sum_{j=1}^{J}(\mathbf{x}_j-\mathbf{x}_j^*)^T(\mathbf{G}_j+\mathbf{C})^{-1}(\mathbf{x}_j-\mathbf{x}_j^*) \quad (0.8)$$

subject to the constraint

$$\mathbf{y}^* \quad = \quad \sum_{j=1}^{J}\mathbf{x}_j^*\beta_j. \quad (0.9)$$

Huntingford *et al.* [29] cited a paper by Nounou *et al.* [40] for the algorithm used to solve (0.8). Hannart *et al.* [19] pointed out that the method of [40] does not actually solve the correct version of the EIV problem; instead, they proposed an alternative method due to Schaffrin and Wieser [53].

In practice, the whole optimization takes place in the space of PCs of the internal variability and restricted to the first r components, where r is relatively small; thus, $\mathbf{C}$ in (0.8) may be replaced by $\mathbf{I}_r$ and the $\mathbf{G}_j$'s are $r \times r$ sample covariance matrices based on the departures of the individual model runs from the inter-model average.

0.3.5 Recent advances

Recent years have seen a number of new technical developments, especially regarding the estimation of the matrix $\mathbf{C}$, or incorporating the uncertainty of $\mathbf{C}$ into general inference statements about detection and attribution.

It was noted already in Section 0.3.2 that the sample estimate of the matrix $\mathbf{C}$ is typically singular, due to the limited number of control runs available. The traditional approach to this is to restrict attention to κ EOFs, which is equivalent to a low-rank approximation to the covariance matrix. As an alternative, [48] considered a shrinkage estimator of the form

$$\alpha\hat{\mathbf{C}} + \gamma\mathbf{I}, \quad (0.10)$$

where $\hat{\mathbf{C}}$ is the sample covariance matrix and $\mathbf{I}$ denotes the identity matrix. This is known as the *Ledoit-Wolf estimator* and is one of the earliest examples of how to improve the properties of a high-dimensional covariance matrix by regularization [33].

Specifically, [48] used the Ledoit-Wolf estimator to construct an optimal test for detection, and showed it is more powerful than the standard test based on restriction to leading EOFs. They also argued the method was more efficient in the sense that the estimator could be based on small samples of model runs, avoiding the need for long control runs.

[49] extended this approach to attribution. Recall from Section 0.3.2 that the conventional approach to attribution uses two independent estimates of $\mathbf{C}$, one for the initial prewhitening and the second for estimation of $\boldsymbol{\beta}$ and associated covariance estimates and tests. [49] also used two independent estimates, with the Ledoit-Wolf estimator being used for prewhitening but then a regular covariance matrix estimator (without regularization) for the estimation and testing part of the procedure. The latter was primarily motivated by trying to keep the resulting statistical tests and confidence limits relatively simple though ultimately they still recommended Monte Carlo procedures. The methods were worked out both for the Ordinary Least Squares case (Section 0.3.2) and for Total Least Squares (Section 0.3.3).

An entirely different formulation of the detection and attribution problem was given in [50]. Noting that the conventional approach treats the shape of the responses $\mathbf{x}_j$ as known but the magnitudes β_j as unknown, they questioned why that should be a natural assumption, and whether it was more logical to test whether both the shape and magnitude of the climate model response were correct. However, they continued to recognize that both the observations and the climate model responses are subject to error. They also preserved the key assumption of *additivity* that has been a feature of every approach discussed in this chapter. With those points in mind they proposed relating observed $Y, X_1, ..., X_J$ and "true" $Y^*, X_1^*, ..., X_J^*$ by the equations

$$\begin{aligned} Y^* &= \sum_{j=1}^{J} X_j^*, \\ Y &= Y^* + \epsilon_Y,\ \epsilon_Y \sim \mathcal{N}(0, \Sigma_Y), \\ X_j &= X_j^* + \epsilon_{X_j},\ \epsilon_{X_j} \sim \mathcal{N}(0, \Sigma_{X_j}),\ j = 1, ..., J, \end{aligned}$$

where an additional twist is that they assume the covariance matrices Σ_Y and Σ_{X_j} to be known. This assumption appears to have been made largely to permit exact distributional calculations, and they acknowledged that in practice it would be necessary to use a plug-in approach with empirical covariance estimates. Thus, while these developments may well lead to more satisfactory estimation and testing procedures, the full accounting for covariance matrix uncertainty in a context that also allows for climate model variability is still an open problem.

A different approach to these issues was started by Hannart [18], who proposed a hierarchical regression approach, which accounts for uncertainty in the climate covariance matrix by making inference on the covariance matrix and on $\boldsymbol{\beta}$ in a single statistical model. Specifically, the climate covariance matrix $\mathbf{C}$ was assumed to follow an inverse-Wishart prior and subsequently integrated out. In contrast to (0.10), this allowed shrinkage toward target matrices other than the identity, for example covariance matrices that account for spatial dependence. By carrying out the integral with respect to $\mathbf{C}$ analytically, the resulting inference procedure is computationally feasible even without pre-reduction of the dimension of the data.

Katzfuss and co-workers [31] considered an empirical Bayesian hierarchical framework in the context of regression-based detection and attribution. The Bayesian hierarchical formulation ensures that all uncertainties represented by the model are propagated to inference on the regression coefficients of interest. Returning to the traditional expansion of $\mathbf{C}$ in terms of EOFs, their model used a Bayesian model averaging approach to probabilistically infer the optimal number κ of EOFs, instead of choosing a fixed truncation value as in previous approaches. In addition, their model took into account that not only $\mathbf{X}$ but also the observations $\mathbf{y}$ are typically not precisely known. More precisely, they accounted for uncertainty in $\mathbf{y}$ due to a finite number of incomplete and noisy measurements as represented by an ensemble of observations. Their Bayesian hierarchical model was fitted using an efficient Markov chain Monte Carlo (MCMC) procedure that also integrated out analytically all high-dimensional quantities.

Another interesting issue is the treatment of the unknown true mean forcing signals $\mathbf{X}$. The standard practice in most recent approaches [e.g., 18] is to profile (i.e., maximize) out the signals. In contrast, [31] integrated out $\mathbf{X}$ under the assumption of an improper uniform prior. Eliminating unknown nuisance quantities via integration is the standard procedure in Bayesian inference. In the statistics literature on (the simpler) errors-in-variables regression [e.g., 38, 8], it has been noticed that maximization (called a functional approach) can ignore uncertainty and lead to inconsistent estimation of variance parameters. A preliminary simulation study in a simplified detection-and-attribution setting indicated that, for small sample size, integrated likelihood can be conservative with lower power, while the profile likelihood can lead to false positives, but more comprehensive simulations are needed. Also explored should be the effect of an informative spatial or spatio-temporal prior on the unknown signals $\mathbf{X}$, as opposed to the uniform prior used in [18, 31].

0.4 Attribution of extreme events

0.4.1 Introduction

A different kind of question related to detection and attribution concerns the attribution of extreme events. As a concrete example, the extremely active hurricane season in the summer of 2017 led to widespread devastation across the Caribbean, Puerto Rico and the southern United States. A natural question is to what extent such events may be considered to have been "caused by" climate change, where one must be extremely careful about what exactly is meant by "caused by". As an example, as at the time of writing of this chapter, three papers have been published analyzing the influence of anthropogenic climate change on the extreme precipitations produced by Hurricane Harvey

at the end of August 2017 [42, 51, 12]. However, the 2017 hurricane season is only one of numerous instances of extreme weather events in recent years where questions have naturally arisen about the influence of anthropogenic climate change. In response, an extensive literature has grown up.

The subject is usually considered to have begun with the paper of Stott, Stone and Allen [55] which analyzed the European heatwave of 2003. This heatwave produced temperatures more than 10°C above the seasonal norm for several consecutive days across much of Central Europe and, by some estimates, was responsible for as many as 70,000 excess deaths. The paper [55] argued that the probability of such an event was increased by a factor of 4 (with a 90% lower confidence bound of 2) compared with a hypothetical counterfactual world without greenhouse-gas warming. Other papers analyzing the 2003 heatwave such as [5, 54, 30] supported the claim of a strong anthropogenic influence on this event. Later papers such as [26, 43, 20] generally supported the anthropogenic influence on a variety of extreme events though using a wide range of methodologies. However, not every paper in this field conveyed the same message. For example Dole and co-authors [10] argued that the 2010 heatwave that badly affected western Russia was most likely a natural event associated with a blocking pattern in the atmosphere, though they did not address the possibility that the frequency of blocking patterns could itself be increasing as a result of global warming. Hoerling and co-authors [27] made similar arguments in discussing the 2011 Texas drought/heatwave, noting that "the principal factor contributing to the heat wave magnitude was a severe rainfall deficit during antecedent and concurrent seasons related to anomalous sea surface temperatures ... that included a La Niña event" while the human-induced contribution to the probability of a new temperature was much smaller.

The wide variety of methods being used for these assessments, as well as occasional disputes over the results, led the National Academy of Sciences to commission a review of the whole field. Their report [41] appeared in 2016. Our own review follows some of the structure of the National Academy report, though necessarily with much condensation, and focuses specifically on the statistical issues these questions raise.

0.4.2 Framing the question

There are so many different ways of defining the problem that the National Academy report [41] devoted a whole chapter to the "framing question". We follow their approach here, and focus on the specifically statistical issues that they raise.

Many researchers beginning with [55] have used the "fraction of attributable risk" as the primary quantity of interest. Given a specific extreme event, let p_1 be the probability of that event under a scenario that includes all known forcing factors that influence climate, and p_0 the counterfactual probability of the same event under natural forcings only (including random

internal variation). Climate models are needed here because p_0 can only be estimated from a model; typically, parallel runs of a climate model (or several climate models) are used so that p_0 and p_1 can be estimated in a way that makes comparisons possible.

The *Fraction of Attributable Risk* (FAR) is defined to be

$$FAR = 1 - \frac{p_0}{p_1}. \tag{0.11}$$

The reason for the name is that, in the typical case where $p_1 > p_0$, we can partition the total probability of the event (p_1) into two components, p_0 for the natural contribution and $p_1 - p_0$ for the anthropogenic contribution. Thus (0.11) represents the fraction that may be "attributed" to the human influence.

However, FAR is not the only measure used. Another is the risk ratio

$$RR = \frac{p_1}{p_0} \tag{0.12}$$

which, although equivalent to (0.11) in the sense that either formula can be transformed into the other, in some respects has a more natural interpretation — for example, the RR represents the proportion by which insurance claims from extreme weather events would be expected to rise in a world subject to anthropogenic forcings compared with one that is not. Another argument is that RR corresponds to statements of risk that are common in medical research, such as "smoking increases the probability of lung cancer by a factor of X", page 34 of [41].

Our own preference is in favor of RR, and this is reinforced by several arguments made in [41]. For example:

1. FAR can be misleading when it is very close to 1 — for example, FAR=0.99 might not seem much different from FAR=0.999 but the latter represents a ten times greater risk ratio;
2. The "fractional" interpretation of FAR only makes sense when it is between 0 and 1 — indeed, FAR is often truncated at 0, especially in confidence intervals — but there is actually nothing pathological about the possibility that $p_0 > p_1$. Indeed, [41] make the argument that there would probably be many more reported instances with $p_0 > p_1$ were it not for the implicit bias that such events are unlikely to be observed. It is better to put things on an even footing and treat the cases $RR > 1$ and $RR < 1$ as equally interesting and important, at least until the weight of evidence suggests to the contrary;
3. Tests and confidence intervals — an inherent issue with estimating extreme event probabilities is that they are very uncertain, so confidence intervals (or Bayesian credible intervals) tend to be very wide. In the language of hypothesis testing, a formal test of the null hypothesis $p_0 = p_1$ may well result in acceptance of that hypothesis. The Academy report [41] cautions *against* concluding "there is no

> effect" in these cases. To quote the report (p. 35), "Failure to reject the null hypothesis of no effect should not be regarded as evidence in favor of there being no effect." Although one could make a similar assertion with respect to FAR, the issue is more clear-cut when framed in terms of RR.

In recent years, there have been a number of attempts to reformulate detection and attribution in the language of causality research. A particularly influential paper was by Hannart and co-authors [17].

According to the classical theory of Hume [28], quoted by [17], "We may define a cause to be an object followed by another, where, if the first object had not been, the second never had existed." In the language of events, an event X may be said to cause an event Y if Y cannot occur in the absence of X, or in other words, $Y \implies X$. This immediately suggests some probabilistic relationships. [17] review the modern theory of causal inference including the use of graphical relationships to represent causality in system of interacting variables.

Suppose we have (0,1)-valued random variables X and Y where in the present context $X = 1$ is associated with the presence of an anthropogenic effect and $Y = 1$ with the occurrence of an extreme climate event. We may also define Y_x to be the value Y would take if X were fixed at x. In a world of perfect causality where $Y = 1$ if and only if $X = 1$, we would have $Y_0 = 0,\ Y_1 = 1$.

In this context, [17] following [46], defines

$$\begin{aligned} \text{PN} &= \Pr\{Y_0 = 1 \mid Y = 1, X = 1\}, \\ \text{PS} &= \Pr\{Y_1 = 1 \mid Y = 0, X = 0\}, \\ \text{PNS} &= \Pr\{Y_0 = 0,\ Y_1 = 1\}. \end{aligned}$$

Here PN is referred to as "the probability of necessary causation", PS as "the probability of sufficient causation", and PNS as "the probability of necessary and sufficient causation". This subdivision into "necessary" and "sufficient" causation is the main new feature of their approach.

Under two additional assumptions, *monotonicity* and *exogeneity* (of X) they show that the above expressions reduce to

$$\begin{aligned} \text{PN} &= \max\left\{1 - \frac{p_0}{p_1}, 0\right\}, \\ \text{PS} &= \max\left\{1 - \frac{1-p_1}{1-p_0}, 0\right\}, \\ \text{PNS} &= \max\{p_1 - p_0, 0\}, \end{aligned}$$

Here, monotonicity is the property that $Y_1 \geq Y_0$ with probability 1, while exogeneity of X essentially means that X is external to the system, in other words, not changed by any of the other variables being observed.

Thus when $p_1 \geq p_0$, PN corresponds exactly to the FAR. PS and PNS are

new measures which do not seem to have been used previously in the extreme event attribution context.

[17] goes on to consider the implications of these definitions for the European heatwave of 2003. Assuming the same climate variables and probability calculations as [55], for which p_0 was estimated to be $\frac{1}{1000}$ and p_1 to be $\frac{1}{250}$, the probability of necessary causation is 0.75, equal to the FAR discussed earlier. However, PS and PNS are both of the order of 0.003, implying very low evidence for sufficient causation.

However, they also consider other interpretations of the data for which the distinction between PN and PS is less clear-cut. The calculations of [55] were based on temperature anomalies with respect to 1961–1990 averages exceeding the threshold of 1.6°C (the second largest summer-mean anomaly in the dataset). However, had the threshold been set substantially lower, both p_0 and p_1 would be larger and hence so would be both PS and PNS. For thresholds in the lower tail of the distribution, we find PS close to 1; as stated by [17], "anthropogenic CO_2 emissions are virtually certainly a sufficient cause, and virtually certainly not a necessary cause, of the fact that the summer of 2003 was not unusually cold." They also point out that when referred to a much longer time period than one year, even if we (unrealistically) assume stationarity in time, an event of the form "the threshold will be exceeded at least once in the next hundred years" (rather than for one specific year) will lead to much larger p_0 and p_1 and therefore a higher probability for sufficient causation as well as necessary causation.

In summary, the use of causal inference methods in climate research is still a new field but it is growing. Ebert-Uphoff and Deng [11] used Bayesian networks to examine the causal relationships among four large-scale climate circulation indexes and, very recently, Hannart and Naveau [16] have proposed an ambitious reformulation of traditional detection and attribution theory in the language of causal inference. We expect to see much more research of this nature in the next few years.

0.4.3 Other "Framing" Issues

Here, we review more briefly several other framing issues discussed by [41]

1. **Choice of climate variable.** The action of the 2003 European heatwave focussed on a few days in early August on a region of western Europe that included most of France, parts of Germany, Switzerland and northern Italy, but did not extend as far west as the Iberian peninsula or into Scandinavia or Eastern Europe. Yet [55] took as their climate variable of interest the annual summer (June, July, August) temperature means over a large geographical area (30°N to 50°N latitude, 10°W to 40°E longitude). There were a few motivations for defining an event over a substantially larger spatial and temporal scale than the one observed. For example, one

concern was selection bias, discussed further below — focussing on a climatic variable that had very locally extreme behavior in the observational record would attract the criticism that the event had been specifically selected for this reason, whereas by choosing a more generic spatial and temporal scale, the focus was more on the appearance of extreme events in general than this one particular event. However, another reason was that [55] recognized that an extreme event attribution analysis was unlikely to be successful if the considered variables were not well represented in climate models.

Subsequent extreme event analyses have generally focussed on smaller spatial regions, and sometimes smaller time windows as well, than [55], but the general principle remains, that it is better to expand the spatial and temporal coverage to reduce concerns about selection bias and to focus on variables that are well represented in climate models. [41] noted some other considerations: (a) different physical variables, e.g. some analyses of the California drought found no anthropogenic effect if precipitation levels were considered on their own but did find an effect when low precipitation was combined with high temperature: (b) where attribution analysis is based primarily on observations, such as comparing recent records with those of an earlier period, it is important that the observations should be of high quality and consistently measured over the whole time period; (c) robustness of results — "a robust attribution analysis would show that results are qualitatively similar across a range of event definitions, acknowledging that quantitative results are expected to differ somewhat because of difference of definition."

2. **Changes in frequency or changes in magnitude?** The discussion of p_0 and p_1 assumed that the interest is in changes of frequency for an event of fixed magnitude, but one can ask a parallel question with frequency and magnitude interchanged. For example, given a historical estimate of the 100-year return value[1] for a given climatic variable, how would that estimate change if the underlying climate conditions changed? That question, for example, underlies the production of flood risk maps by the Federal Emergency Management Agency (FEMA). [51] give examples of both types of calculations.

3. **Conditioning.** One of the most contentious issues in this field in recent years has been the desirability of *conditioning* as a vital component of the analysis. Trenberth and colleagues [57] argued that when an extreme event depends on the presence of some feature of large-scale atmospheric circulation, it may be problematic to try to attribute the circulation event itself to anthropogenic effects, but, conditional on the appearance of the circulation pattern, other

[1] that value which is exceeded in any given year with probability 0.01

variables that affect the development of an extreme event, such as SST, may be much more clearly attributable to human influence. Examples that they gave included Superstorm Sandy, that caused widespread flooding over New York and New Jersey in 2012, and the Colorado floods of 2013. Therefore, they suggested, the entire analysis should be conducted conditionally on the presence of whatever large-scale circulation feature initiated the event of interest.

With reference to the Colorado floods, a specific example of this kind of analysis was given by Pall and co-workers [44]. The flooding events in question happened during September 2013, and caused over $2 billion damage and nine fatalities. However, the authors noted, "the unusual hydrometeorology of the event...challenges standard frameworks [for attribution]... because they typically struggle to simulate and connect the large-scale meteorology associated with local weather processes." Consequently, these authors developed an approach that was part statistical, part dynamical based on simulations of the local weather conditioned on observed synoptic-scale meteorology. The key meteorological point seems to be that warmer air holds more moisture (the Clausius-Clapeyron relationship) and therefore exacerbates the magnitude of precipitation events within a developing weather system. The authors looked at this from both a "frequency" and "magnitude" point of view, concluding that the magnitude of the extreme event was increased by 30% as a result of anthropogenic climate change, or conversely, the probability of an event of the given magnitude, conditional on the synoptic weather pattern, was increased by a factor of 1.3 compared with what might have been expected in a pre-industrial world.

For the analysis of [44], it appears that an attempt to take into account the probability of the triggering synoptic-scale event would not have been successful because there was no reasonable basis for determining how this event could have been changed by the anthropogenic influence. In other cases, however, there may be a choice: do we base the analysis solely on the conditional probabilities or do we also take into account the probabilities of the conditioning event? As described by [41], the choice lies between considering

$$\frac{\Pr_f\{E \mid N\}}{\Pr_c\{E \mid N\}} \tag{0.13}$$

or

$$\frac{\Pr_f\{E \mid N\}}{\Pr_c\{E \mid N\}} \times \frac{\Pr_f\{N\}}{\Pr_c\{N\}} \quad = \quad \frac{\Pr_f\{E \cap N\}}{\Pr_c\{E \cap N\}}. \tag{0.14}$$

Here E is the event of interest, N is the conditioning event (which may be El Niño, hence the choice of initial) and $\Pr_f$ and $\Pr_c$ denote

probabilities in the observed (factual) and counterfactual worlds. The controversy, such as it is, appears to hinge on the question of whether it is preferable to base the inferences on (0.13) in place of (0.14).

In fact, there is a sound statistical argument for conditioning based on R.A. Fisher's theory of conditional inference and the related concept of *ancillary statistics*. We do not attempt a detailed review here, since there is a extensive and large literature, but we refer to [13] for a relatively recent review.

The key point of this theory is that conditional inference is always indicated when the distribution of the conditioning variable N is independent of the quantity being estimated, which in this case, is the influence of anthropogenic climate change on the probability of the event E. Such a variable N is known as an ancillary statistic. In meteorological terms, if the event N is not affected by climate change, then it is valid to use (0.13) in place of (0.14).

As things stand, this may seem a trivial conclusion, because if N is not influenced by climate change, the second factor on the left side of (0.14) will be 1 and there is no distinction between (0.13) and (0.14). However, there is also an extensive theory of approximate or "local" ancillarity and the overall conclusion is that conditioning is still appropriate in this case [7].

In practice, the more realistic difficulty is that E and N are physical measurements, not random variables satisfying some theoretical distribution, and there may simply not be enough information to determine whether either or both are different under the factual and the counterfactual scenarios. Under such circumstances, arguing conditionally on N seems a logical way to proceed.

4. **Selection bias.** The final "framing" issue we discuss here is that of selection bias — the idea that the selection of an event to study may itself bias the conclusions drawn from it. As noted by [41], selection bias may take various forms, the most pervasive being occurrence bias, "bias from studying only events that occur." As noted already, a partial solution may be to define the climate events of interest on a sufficiently large temporal and spatial scale that extreme local fluctuations do not bias the results. [41] conclude that "selection bias [is] almost inevitable in event attribution applied to individual events" but caution that "Such selection biases interfere with the ability to draw general conclusions about anthropogenic influence on extreme events collectively."

   A possible direction for future methodological development is to adjust statistical event attribution techniques to account explicitly for their tendency to focus on local spatial and/or temporal ex-

tremes, in similar manner to the use of scan statistics in spatial epidemiology, e.g. [32]. In short, we might adjust the extreme event probabilities to allow for a selection bias effect, but then proceed as in earlier analyses with the comparisons of scenarios that do or do not include the anthropogenic component.

0.4.4 Statistical methods

In this section we do not pursue further the various type of framing issues but assume the problem is essentially the basic one that motivated this whole section: given the interest in a specific event E, and the possibility of estimating the event E from parallel runs of either a climate model (under "all forcings" versus "natural forcings" scenarios) or an observational dataset (under "present-day" versus "pre-industrial" conditions), how would we actually estimate the probabilities p_1 and p_0, the respective probabilities under those two scenarios? These probabilities may then be used to estimate the FAR, the RR, or various other measures of interest. We briefly review the main methods for estimation p_1 and p_0.

1. **Methods based on normal distributions.** The simplest methods assume the underlying variables are normally distributed. This assumption was common in the early days of the subject [5, 54, 30] and still surfaces from time to time [47]. In general, we don't recommend normal-theory approaches because, even when the overall distribution is close to normal, as is usually the case with temperatures, deviations from normality in the tails of the distribution can cause serious biases to the estimated probabilities.
2. **Adapting conventional detection and attribution theory.** For example, Min and co-authors [37] fitted the generalized extreme value (GEV) distribution to extremes in spatially averaged precipitation variables and then used a probability integral transform (if G is the CDF fitted to a random variable Y and if Φ denotes standard normal CDF, then $\Phi^{-1}(G(Y))$ is standard normal) to transform the variables to normality. On the resulting normal scale, they then applied conventional detection and attribution theory to determine the statistical significance of the anthropogenic component and to compute estimates and confidence intervals for the various regression coefficients. Similar techniques have been used in other papers such as [61, 59]. Compared with other techniques considered in this section, these methods do not lead to numerical estimates of p_1 and p_0 and hence FAR, RR etc., but in principle they could be: use the fitted GEV distribution to estimate p_1 and then the same transformation applied to the detection and attribution model without the anthropogenic component to estimate p_0, but this would be a decidedly roundabout method of estimating p_0! Overall, this method

seems less well suited to studying the attribution of a single event than to understand the anthropogenic influence on extreme events generally, but that remains an important consideration for climate research.

3. **Methods based on counting exceedances in model simulations.** A nonparametric method for estimating p_1 and p_0 is simply to count the number of exceedances of the threshold of interest in parallel all-forcings and natural-forcings model runs. Tests and confidence intervals may then be based on standard statistical theory for binomial distributions. This method has the considerable advantage of simplicity, when it is applicable; but against that, nonparametric methods cannot be used at all when they involve extrapolating beyond the range of the climate model data. In practice it seems to be used in two situations: (i) when the supposed extreme event is not actually very extreme at all, at least when conditioned on suitable large-scale variables [45], (ii) when the analyst has available a fast-running model for which generating a large ensemble of model runs is not a problem [12]. The "climateprediction.net" experiment (https://www.climateprediction.net/) is a citizen science project to generate very large ensembles of climate model runs through volunteers running climate simulation programs on their laptops, but the emphasis of extreme event attribution analysis in recent years has switched towards trying to get very fast "operational time" results, and the time taken to collect large ensembles through a distributed network would appear to impede that.
4. **Methods based on extreme value theory.** These methods appear, perhaps surprisingly, to be state of the art in this field at the time of writing. For a review of different methods of (univariate) extreme value analysis, see Chapter 8 of this volume. The methods may be divided into two broad categories, (i) block maxima methods, where the Generalized Extreme Value (GEV) distribution is fitted to maxima over blocks of fixed time length (usually the time length of a block is taken to be one year, in which case it is also called the annual maxima method), or (ii) threshold exceedance methods, in the most common form of which the Generalized Pareto Distribution (GPD) is fitted to the exceedances over a fixed threshold. The original paper of [55] used the GPD to characterize the tail distribution of European summer temperatures, but recent applications have more frequently used the GEV applied to annual maxima. A significant issue with these methods is how to characterize uncertainty of the extreme-event probabilities calculated from the fitted distributions — both frequentist (bootstrap) and Bayesian methods have been applied but there seems to be no universal agreement at

this time. We do not review these methods further here because Chapter 8 gives a full account.

0.4.5 Application to precipitation data from Hurricane Harvey

We now return to our earlier discussion of Hurricane Harvey, and specifically the extreme precipitations produced by that event in the region surrounding Houston, by reviewing the three papers that we are aware of that have discussed that event.

1. **The World Weather Attribution Project**. This project is a consortium of researchers in The Netherlands, U.K. and U.S.A. who aim to produce "rapid attributions" of extreme weather events, sometimes within a week of the event in question. Two recent examples are [58] for the August 2016 Louisiana floods and [42] for the flooding associated with Hurricane Harvey. Both studies used the same statistical methods and covered essentially the same geographical region, so they used many of the same meteorological variables as well. The essential idea is to compile several datasets, both observational and model-based, that use different data sources, different spatial resolutions and (with the models) different forcing factors, including historical runs based on all known forcing factors, pre-industrial control runs and "static" experiments with forcings fixed at levels corresponding to various points in time (e.g. 1860, 1940, 1990, 2015) in order to compare equilibrium climate behavior under different forcing scenarios.

   The basic method is to fit generalized extreme value distributions (see Section 0.4.6 below) with adjustment for a covariate which they typically take as global mean temperature for a given year. In the notation of (0.15) below, the model of [58] allows both $\log \eta_t$ and $\log \tau_t$ to be linearly dependent on a global temperature variable T'. The GEV is fitted to each of the observational and model datasets, omitting the extreme event that stimulated the study, and the results compared to evaluate consistency across observations-based and model-based analyses. Typically, p_1 and p_0 are evaluated by setting T' to be its value in the present day and its value at some historical date, e.g. 2017 versus 1900 in [42]. Standard errors and uncertainty bounds are evaluated largely by bootstrapping, with a spatial block bootstrap recommended in spatially aggregated datasets to reduce bias due to spatial dependence.

   For three-day maxima from 85 stations in the Gulf Coast region, the estimated return value associated with the 2017 event is around 9,000 years, and a risk ratio of four compared with the correspond-

ing estimate for 1900. Somewhat lower return level estimates (of the order 200–800 years) are obtained for events based on the spatial maximum over a region than for events at a specific location, but similar risk ratios of the order of 4–6 compared with 1900. In all cases, the range of uncertainties associated with these estimates is very wide. Estimates based on model runs typically show somewhat lower risk ratios but narrower confidence limits, which the authors interpret as evidence of an anthropogenic effect.

2. **Risser and Wehner**. These authors [51] provided an alternative extreme value analysis of precipitations during August 2017. They computed seven-day maximum precipitation values from rain-gauges, aggregated over two regions near Houston: a small and large region of approximately 33,000 km^2 and 105,000 km^2 respectively. The GEV was fitted to annual maximum data from 1950–2016 using a model similar in structure to (0.15) and (0.16) below, but with two covariates: the Niño 3.4 index as a measure of El Niño activity, which the authors identify as a natural variation, and annual global CO_2 measurements. They estimated return values for the 2017 event of the order of 30–100 years (larger for the large region than the small region) and also risk ratios of the order 4–8 (with lower confidence bounds of the order 1.5–4).

   For this paper, a direct "attribution" statement was not possible because climate model runs were not used, but the authors argued that an equivalent if somewhat weaker statement could be made in the language of Granger causality [15]. They argued, in effect, that confidence intervals for risk ratios with lower bounds > 1 are consistent with Granger causality at a "likely" level of uncertainty (66% confidence) though not at a "very likely" level (90% confidence). In this sense, the results provide a valid attribution statement.

3. **Emanuel**. The third author [12] to have examined Hurricane Harvey precipitations took a completely different approach, based on a model for directly simulating hurricanes and their associated storm rainfalls. The model relies on global climate model data to generate the large-scale state of the system and then randomly perturbs that state to create hurricane-like disturbances. Thus, the approach is still dependent on global-scale models for the large-scale variables, but improves significantly on global-scale models in its resolution of specific hurricane events.

   For storm totals at the single point of Houston, Texas, this approach suggests a return value in excess of 2,000 years, though with huge uncertainty as very few of the model runs get close to that level of precipitation. A side comment here is that although this was based on the "counting exceedances" approach (see Section 0.4.4), an extreme value theory approach might also be productive in reducing

the uncertainty of estimation at the very end of the observed range of data. A second calculation based on total rainfall over the state of Texas suggests an annual exceedance probability of around 1% in 1981–2000, increasing to around 18% in 2081–2100 under the representation concentration pathway 8.5 scenario (sometimes called the "business as usual" scenario, since it assumes no significant slowing down in the rate of emissions of greenhouse gases).

This paper was the only one of the three to extrapolate future probabilities of a Harvey-type event, but if the 18% estimate is realistic, for the probability that a Harvey-sized event will occur somewhere in the Gulf region in any given year, that is a disturbing conclusion.

0.4.6 An example

The following example is more limited than the three studies just cited [42, 51, 12] because it only considers precipitation from one station (Houston Hobby Airport) but it nevertheless shows that many of the same effects that they cite as evidence of anthropogenic influence are also present here. In common with [51], seven-day precipitation totals were calculated each year for the hurricane season from July–November. We also calculated annual average sea surface temperatures (SST) for the entire Gulf of Mexico, computed from July of the year preceding the precipitation year through June of the same year; this time window was considered most likely to influence the following hurricane season. Figure 2 illustrates the data. Three points are immediately apparent. First, the 7-day precipitation total associated with Hurricane Harvey is by far the largest in history, more than twice the second-largest value. Second, SSTs have also increased steadily since the 1970s, and the Gulf of Mexico SST mean for 2016-17 was the largest in history for that variable. Third, based on the straight line fit in (c), there is some evidence that 7-day precipitation maxima and SSTs are correlated, though the statistical significance of that is hard to judge from the plot.

A more formal analysis of the latter point may be based on the Generalized Extreme Value (GEV) distribution for annual maxima: see Chapters 8 and 31 of the present volume for a detailed discussion of extreme value theory and the GEV distribution in particular.

If Y_t denotes the maximum precipitation value for year t, we assume Y_t follows the GEV distribution in the form

$$\Pr\{Y_t \le y\} \quad = \quad \exp\left[-\left\{1+\xi\left(\frac{y-\eta_t}{\tau_t}\right)\right\}_+^{-1/\xi}\right], \tag{0.15}$$

where the subscript + denotes positive part, η_t and τ_t are allowed to vary with year and, in accordance with common practice in this field, the shape parameter ξ is treated as a constant. Recall that [51] used two covariates, the Niño 3.4 index and annual global CO_2 means. Here, we are assuming that Gulf

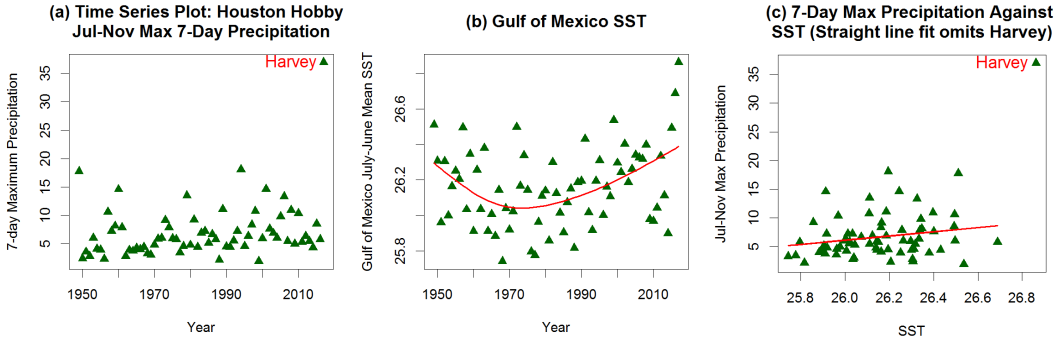

FIGURE 2
Precipitation in Houston and Gulf of Mexico SST. (a) Maximum 7-day precipitation total from Houston Hobby airport, computed from July–November each year. (b) Gulf of Mexico mean July–June sea surface temperature, each year from 1948-49 through 2016-17. The fitted trend curve is based on a spline with 4 DF and is consistent with overall Northern hemisphere temperature trends during this time period. (c) Plot of maximum 7-day precipitation against Gulf mean SST, with a fitted straight line omitting the 2017 outlier. Public data sources: Daily precipitation from the Global Historical Climatological Network (National Centers for Environmental Information, U.S.A.); monthly sea surface temperatures from HadISST (U.K. Meteorological Office.)

of Mexico SST will include any El Niño effect that influences precipitation, but to be consistent with [51], we also included annual global CO_2 means from the RCP database (https://tntcat.iiasa.ac.at/RcpDb).

The following models are considered: each of η_t and log τ_t is a linear function of up to two covariates, where the covariates considered are SST_t (Gulf of Mexico annual mean SST in year t) and $CO2_t$ (global mean CO_2 in year t). For numerical stability, SST_t is expressed as the deviation from 26°C and $CO2_t$ is replaced by $0.01(CO2_t - 350)$. This gives 16 possible models of which the Akaike Information Criterion chooses the following:

$$\begin{aligned} \eta_t &= \theta_1 + \theta_4 SST_t + \theta_5 CO2_t, \\ \log \tau_t &= \theta_2 + \theta_6 SST_t, \\ \xi &= \theta_3. \end{aligned} \tag{0.16}$$

The fitted parameters are given in Table 0.1.

Next, this model is used to calculate exceedance probabilities in various years corresponding to the observed 2017 value due to Hurricane Harvey. First, we smoothed the SST values, using the same smoothing spline as in Figure

Parameter	Estimate	Standard error	t-statistic	p-value
θ_1	4.70	0.29	16.22	0.00
θ_2	0.56	0.13	4.25	0.00
θ_3	0.15	0.09	1.64	0.10
θ_4	3.06	1.49	2.06	0.04
θ_5	1.95	0.82	2.36	0.018
θ_6	1.24	0.50	2.48	0.013

TABLE 0.1
Table of GEV parameters for Houston Hobby precipitation maxima.

2(b). The reason for smoothing is that we are interested in long-term climatic effects, not individual-year fluctuations, and smoothing the SSTs seems a logical way to achieve that. Second, the model defined by (0.15) and (0.16) was refitted using Bayesian methods, assuming a flat prior. The reason for this is to allow the uncertainty of the estimates to be expressed in terms of posterior distributions. The three curves in Figure 3(a) represent the 17th, 50th and 83rd percentiles of the posterior density for the exceedance probability in each year. The reason for the 17th and 83rd percentiles is that the posterior probability between them is 0.66; according to the Intergovernmental Panel on Climate Change uncertainty guidelines [36], it is *likely* that the true value lies between these bounds[2]. For the specific occurrence in 2017, the calculation shows a posterior median exceedance probability of 0.0019 (return value 525 years) with a *likely* range from 0.00022 to 0.00685 (return values 145 to 4472 years).

Climate model data have been downloaded from the CMIP5 model archive and used to calculate annual SST means over the Gulf of Mexico. These are available under three scenarios: (a) historical all-forcings data up to 2005 or 2012; (b) historical natural-forcings data up to 2005 or 2012; (c) future forcings data under the RCP 8.5 scenario, often called the "business as usual" scenario because it does not presume any significant effort to slow down greenhouse gas emissions. All model runs have been converted to anomalies and where natural-forcings data ended before 2017, we simply assumed the last available value (for 2005 or 2012) was also valid up to 2017. We combined the all-forcings and RCP 8.5 data to obtain a continuous record of data from 1949 up to 2080 which was taken as the end-year for this assessment. This exercise was repeated for four climate models; where multiple ensembles were available from the same model, we averaged over ensembles.

The model Gulf of Mexico SSTs do not follow the observational data very closely so, in order to use the regression model fitted previously to observational SSTs, we proceed as follows. The observational SSTs for 1949–2017 are

[2]The stronger terms *very likely* and *virtually certain* are used for events with probability at least 0.9 and 0.99, respectively.

regressed on two covariates: first, the difference between historical-forcings and natural-forcings climate model runs, and, second, the natural-forcings climate model runs on their own. The two components together are then used to define the "all forcings" signal and the second component on its own is used to define the "natural forcings" signal. Both components are represented via smoothing splines to give a smooth signal. This exercise is repeated for each of the four climate models and also with all four models averaged to give the curves in Figure 3(b). A curious feature of these curves, which we are not able to fully explain, is that even the natural-forcings curves seem to show an upwards trend towards the end of the series.

This exercise was repeated to obtain future projections of Gulf of Mexico SST up to 2080; see Figure 3(c). Since there are no natural-forcings projections over this time period, only the RCP 8.5 values are shown.

We now repeat the calculation of the probability of a Harvey-sized event under the circumstances, (a) for 2017 under all forcings, (b) for 2017 under natural forcings, (c) for 2080 under RCP 8.5. The calculation is repeated for all four climate models and for the average over the four models; we used the same posterior density output as before to obtain Bayesian posterior curves. Finally, we took the ratio of (a) to (b) (relative risk for 2017 under the all-forcungs and natural-forcings scenario), and the ratio of (c) to (a) (relative risk for a Harvey-sized event in 2080 compared with 2017). The results are in Table 0.2.

Model	Present			Future		
	Lower	Mid	Upper	Lower	Mid	Upper
CCSM4	1.5	2.0	3.2	9.0	26.2	133
GISS-E2-R	1.8	2.5	4.8	13.5	43.5	244
HadGEM2-ES	1.6	2.1	3.5	23.6	73.3	415
IPSL-CM5A-LR	1.5	2.0	3.3	10.8	33.8	186
Combined	1.7	2.4	4.4	14.3	46.0	254

TABLE 0.2
Relative risks. The columns labelled "Present" refer to relative risks for the 2017 event under an all-forcings scenario versus a natural-forcings scenario, computed under four climate models and with all four models combined. Lower, mid and upper bounds correspond to the 17th, 50th and 83rd percentiles of the posterior distribution. The columns labelled "Future" are relative risks for such an event in 2080 against 2017; same conventions regarding climate models and percentiles.

For the combined-model results, the relative risk of the Harvey precipitation under all-forcings versus natural-forcings scenarios is estimated as 2.4, "likely" between 1.7 and 4.4. For all five sets of model results in Table 0.2, the lower bound exceeds 1, proving that it's "likely" that anthropogenic con-

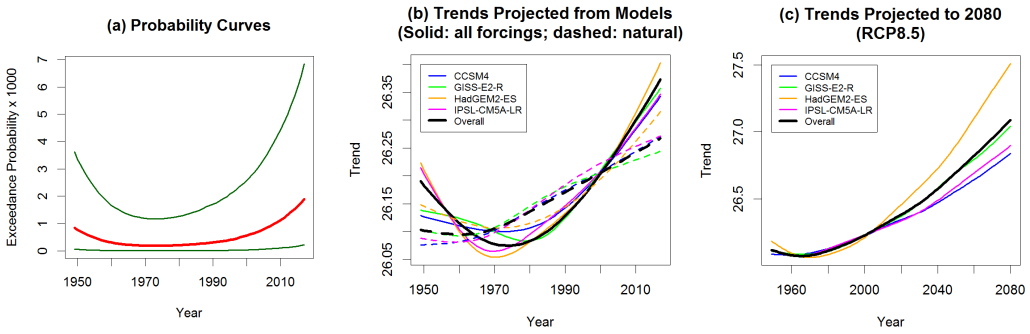

FIGURE 3
Probability Curves and SST Projections. (a) Projected probability (red curve) and 66% confidence bounds (green curves) for the probability of a Harvey-sized event at Houston Hobby airport, 1949-2017. (b) Projected SSTs in the Gulf of Mexico under all-forcings (solid curves) and natural-forcings (dashed curves) for four climate models, and all four models averaged. (c) Projected SSTs through 2080, under the RCP 8.5 scenario for four climate models, and all four models averaged.

ditions affected Harvey. This is consistent with the earlier results reported by [42, 51, 12].

For the relative risks of a Harvey-sized event in 2080 against 2017, the posterior means range from 26 to 73, with "likely" bounds ranging from 9 to 415. Evidently, the uncertainty range for future projections is very wide. Recalling that Emanuel [12] obtained an estimated relative risk of 18 by complete different methods, there seems to be some agreement that a drastic rise in the frequency of this type of event is to be expected.

Further details of these results will be developed elsewhere.

0.4.7 Another approach

Diffenbaugh and co-authors [9] also sought to quantify the increase of extreme event probabilities as a result of global warming, though taking a more global view of the problem in computing probabilities for a number of extreme events related to extreme temperatures, droughts and extreme rain events. They compared results obtained using both observational data and climate models. A particular feature of their approach was the use of some standard statistical goodness of fit procedures (Kolmogorov-Smirnov and Anderson-Darling tests) to assess the agreement between observational and climate-model data after first correcting for the difference in means between pre-industrial cli-

mate model data and detrended observations. They recommend rejecting any climate model which fails the Anderson-Darling test at a p-value of 0.05.

0.5 Summary and open questions

Climate change detection and attribution refers to a set of statistical tools to relate observed changes to external forcings, specifically to anthropogenic influences. While this issue can be viewed in different ways, the most commonly applied framework is linear regression. The problem formulation per se seems straight forward, but the challenges lie in the high dimensionality of the problem and the large number of unknown quantities in the context of limited observations. Current methods differ in their complexity of the problem formulation and what assumptions are being made to reduce the dimensionality of the problem. Most methods implemented so far are of frequentist nature and Bayesian implementations have only recently appeared on the scene.

While many of the approaches discussed address some of the methodological challenges, there is of yet no model framework to address them all comprehensively. For example, most current frameworks assume the different model runs to be independent realizations from a common random quantity. This viewpoint is justifiable in cases where all model runs come from the same climate model or all come from different climate models, but less so if we have multiple, and potentially unequal numbers, of replicates from multiple climate models. In this case a formulation explicitly accounting for inter- and intra-model variability, as considered by [29], is needed. An analogous issue exists with control runs coming from different models. Having a way to use them jointly would drastically increase the amount of information available to estimate internal variability. The assumed covariance structure of observations is also relatively simple in current methods [e.g., 31], if observational uncertainty is considered at all. With the advent of observational products now routinely being provided as ensembles rather than a single data set, which used to render data-driven observational covariance estimation practically impossible, more complex covariance structures can be envisioned. Other directions include joint inference on multiple properties, e.g. different temperature layers in the atmosphere, and incorporating non-linear interactions.

The alternative field of extreme event attribution is still rapidly growing and would seem to offer excellent opportunities for involvement by statisticians. For example, although the standard univariate methods of extreme value theory are becoming standard in this field, none of the references cited in this chapter has made any use of bi/multivariate or spatial extreme value theory, though there is extensive statistical theory in both cases as documented in Chapters 8 and 31 of this volume. Therefore, there are many possibilities for extensions of the methods as they currently exist.

0.6 Acknowledgements

This work was partly supported by NSF grants DMS-1127914 and DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute, and NSF grant DMS-1106862 to the Research Network on Statistics in the Atmosphere and Ocean Sciences (STATMOS). In addition, Katzfuss' research was partially supported by NSF grants DMS-1521676 and DMS-1654083 and Smith's by NSF grant DMS-1242957.

Bibliography

[1] R.J. Adcock. A problem in least squares. *The Analyst*, 5:53–54, 1981.

[2] 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.

[3] M.R. Allen and S.F.B. Tett. Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, 15:419–434, 1999.

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

[5] M. Beniston and H.F. Diaz. The 2003 heatwave as an example of summers in a greenhouse climate? Observations and climate model simulations for Basel, Switzerland. *Global and Planetary Change*, 44:73–81, 2004.

[6] N.L. Bindoff, P.A. Stott, K.M. AchutaRao, M.R. Allen, N. Gillett, D. Gutzler, K. Hansingo, G. Hegerl, Y. Hu, S. Jain, I.I. Mokhov, J. Overland, J. Perlwitz, R. Sebbari, and X. Zhang. *Detection and Attribution of Climate Change: from Global to Regional*, book section 10, pages 867–952. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[7] D.R. Cox. Local ancillarity. *Biometrika*, 67:279–286, 1981.

[8] A. M. Cruddas, N. Reid, and D. R. Cox. A time series illustration of approximate conditional likelihood. *Biometrika*, 76(2):231–237, 1989.

[9] N.S. Diffenbaugh, D. Singh, J.S. Mankin, D.E. Horton, D.L. Swain, D. Touma, A. Charland, Y. Liu, M. Haugen, M. Tsiang, and B. Rajaratnam. Quantifying the influence of global warming on unprecented extreme climate events. *PNAS*, 114(19):4881–4886, 2017.

[10] 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.

[11] Imme Ebert-Uphoff and Yi Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25:5648–5665, 2012.

[12] K. Emanuel. Aseessing the present and future probability of Hurricane Harvey's rainfall. *PNAS*, 114(48):12681–12684, 2017.

[13] M. Ghosh, N. Reid, and D.A.S. Fraser. Ancillary statistics: A review. *Statistica Sinica*, 20:1309–1332, 2010.

[14] L.J. Gleser. Estimation in a multivariate "errors in variables" regression model: Large sample results. *Annals of Statistics*, 9:24–44, 1981.

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

[16] A. Hannart and P. Naveau. Probabilities of causation of climate changes. *Journal of Climate*, doi:10.1175/JCLI-D-17-0304.1, in press, 2018.

[17] 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.

[18] Alexis Hannart. Integrated optimal fingerprinting: Method description and illustration. *Journal of Climate*, 29(6):1977–1998, 2016.

[19] Alexis Hannart, Aurélien Ribes, and Philippe Naveau. Optimal fingerprinting under multiple sources of uncertainty. *Geophysical Research Letters*, 41(4):1261–1268, 2014.

[20] J. Hansen, M. Sato, and R. Ruedy. Perception of climate change. *PNAS PLUS*, page www.pnas.org/cgi/doi/10.1073/pnas.1205276109, 2012.

[21] 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.

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

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

[24] Gabriele Hegerl and Francis Zwiers. Use of models in detection and attribution of climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 2(4):570–591, 2011.

[25] 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.

[26] M. Hoerling, J. Eischeid, X. Quan, and T. Xu. Explaining the record US warmth of 2006. *Geophysical Research Letters*, 34:L17704, doi:10.1029/2007GL030643, 2007.

[27] 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.

[28] David Hume. *An Enquiry Concerning Human Understanding*. Dover, 2004.

[29] Chris Huntingford, Peter A. Stott, Myles R. Allen, and F. Hugo Lambert. Incorporating model uncertainty into attribution of observed temperature change. *Geophysical Research Letters*, 33(5):L05710, 2006.

[30] C.C. Jaeger, J. Krause, A. Haas, R. Klein, and K. Hasselmann. A method for computing the fraction of attributable risk related to climate damages. *Risk Analysis*, 28(4):815–823, 2008.

[31] Matthias Katzfuss, Dorit Hammerling, and Richard L. Smith. A Bayesian hierarchical model for climate change detection and attribution. *Geophysical Research Letters*, 44(11):5720–5728, 2017. 2017GL073688.

[32] M. Kulldorf. A spatial scan statistic. *Communications in Statistics — Theory and Methods*, 26(6):1481–1496, 1997.

[33] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

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

[35] F. C. Lott, P. A. Stott, D. M. Mitchell, N. Christidis, N. P. Gillett, L. Haimberger, J. Perlwitz, and P. W. Thorne. Models versus radiosondes in the free atmosphere: A new detection and attribution analysis of temperature. *Journal of Geophysical Research: Atmospheres*, 118(6):2609–2619, 2013.

[36] M.D. Mastrandrea, C.B. Field, T.F. Stocker, O. Edenhofer, K.L. Ebi, D.J. Frame, H. Held, E. Kriegler, K.J. Mach, P.R. Matschoss, G.-K. Plattner, G.W. Yohe, and F.W. Zwiers. Guidance note for lead authors of the ipcc fifth assessment report on consistent treatment of uncertainties. *Intergovernmental Panel on Climate Change*, www.ipcc.ch, 2010.

[37] S.-K. Min, X. Zhang, F.W. Zwiers, and G.C. Hegerl. Human contribution to more-intense precipitation extremes. *Nature*, 470:378–381, 2011.

[38] J. Neyman and Elizabeth L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948.

[39] G. North and M. Stevens. Detecting climate signals in the surface temperature record. *Journal of Climate*, 11(4), 1998.

[40] Mohamed N. Nounou, Bhavik R. Bakshi, Prem K. Goel, and Xiaotong Shen. Process modeling by Bayesian latent variable regression. *AIChE Journal*, 48(8):1775–1793, August 2002.

[41] National Academy of Sciences. *Attribution of Extreme Weather Events in the Context of Climate Change*. National Academies Press, Washington, D.C., 2016.

[42] G.J. van Oldenborgh, K. van der Wiel, A. Sebastian, R. Singh, J. Arrighi, F. Otto, K. Haustein, S. Li, G. Vecchi, and H. Cullen. Attribution of extreme rainfall from Hurricane Harvey, August 2017. *Environmental Research Letters*, 12:124009, 2017.

[43] P. Pall, T. Aina, D.A. Stone, P.A. Stott, T. Nozawa, A.G.J. Hilberts, Lohmann D., and M.R. Allen. Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. *Nature*, 470:382–386, 2011.

[44] 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.

[45] Pardeep Pall, Tolu Aina, Dáithí A Stone, Peter A. Stott, Toru Nozawa, Arno G J Hilberts, Dag Lohmann, and Myles R. Allen. Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. *Nature*, 470:382–386, March 2011.

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

[47] S. Rahmstorf and D. Coumou. Increase of extreme events in a warming world. *PNAS*, 108(44):17905–17909, 2011.

[48] Aurélien Ribes, Jean-Marc Azaïs, and Serge Planton. Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. *Climate Dynamics*, 33:707–722, May 2009.

[49] Aurélien Ribes, Serge Planton, and Laurent Terray. Application of regularised optimal fingerprinting to attribution. Part I: method, properties and idealised analysis. *Climate Dynamics*, 41(11-12):2817–2836, April 2013.

[50] Aurélien Ribes, Francis W. Zwiers, Jean-Marc Azaïs, and Philippe Naveau. A new statistical approach to climate change detection and attribution. *Climate Dynamics*, 48(1):367–386, Jan 2017.

[51] 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*, 44:1000–1000, 2013.

[52] 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.

[53] B. Schaffrin and A. Wieser. On weighted total least-squares adjustment for linear regression. *J. Geod*, 82:415–421, 2008.

[54] C. Schär, P.L. Vidale, D. Lüthi, C. Frei, C. Häberli, M. Liniger, and C. Appenzeller. The role of increasing temperature variability in European summer heatwaves. *Nature*, 427:332–336, 2004.

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

[56] Peter A. Stott, Nikolaos Christidis, Friederike E. L. Otto, Ying Sun, Jean-Paul Vanderlinden, Geert Jan van Oldenborgh, Robert Vautard, Hans von Storch, Peter Walton, Pascal Yiou, and Francis W. Zwiers. Attribution of extreme weather and climate-related events. *Wiley Interdisciplinary Reviews: Climate Change*, 7(1):23–41, 2016.

[57] K.E. Trenberth, J.T. Fasullo, and T.G. Shepherd. Attribution of climate extreme events. *Nature Climate Change*, 5(8):725–730, 2015.

[58] K. van der Wiel, S.B. Kapnick, G.J. von Oldenborgh, K. Whan, P. Sjoukje, G.A. Vecchi, R.K. Singh, J. Arrighi, and H. Cullen. Rapid attribution of the august 2016 flood-inducing extreme precipitation in south louisana to climate change. *Hydrology and Earth System Sciences*, 21:897–921, 2017.

[59] X. Zhang, J.F. Wang, F.W. Zwiers, and S.-K. Hegerl, G.C.and Min. Attributing intensification of precipitation extremes to human influence. *Geophysical Research Letters*, 40(19):5252–5257, 2013.

[60] F.W. Zwiers and H. von Storch. Taking serial correlation into account in tests of the mean. *Journal of Climate*, 8:336–351, 1995.

[61] F.W. Zwiers, X. Zhang, and Y. Feng. Anthropogenic influence on long return period daily temperature extremes at regional scales. *Journal of Climate*, 24(3):881–892, 2011.