TECHNICAL REPORT
R-451
MARCH 2015

AMERICAN
METEOROLOGICAL
SOCIETY

Bulletin of the American Meteorological Society

EARLY ONLINE RELEASE

This is a preliminary PDF of the author-produced manuscript that has been peer-reviewed and accepted for publication. Since it is being posted so soon after acceptance, it has not yet been copyedited, formatted, or processed by AMS Publications. This preliminary version of the manuscript may be downloaded, distributed, and cited, but please be aware that there will be visual differences and possibly some content differences between this version and the final published version.

The DOI for this manuscript is doi: 10.1175/BAMS-D-14-00034.1

The final published version of this manuscript will replace the preliminary version at the above DOI once it is available.

If you would like to cite this EOR in a separate work, please use the following full citation:

Hannart, A., J. Pearl, F. Otto, P. Naveau, and M. Ghil, 2015: Causal counterfactual theory for the attribution of weather and climate-related events. Bull. Amer. Meteor. Soc. doi:10.1175/BAMS-D-14-00034.1, in press.

TECHNICAL REPORT
R-451
MARCH 2015

Causal counterfactual theory for the attribution of weather and climate-related events

A. Hannart*

IFAECI, CNRS/CONICET/UBA, Buenos Aires, Argentina

J. Pearl

Computer Science Department, UCLA, United States

F. E. L. Otto

Environmental Change Institute, University of Oxford, United Kingdom

P. Naveau

LSCE, CNRS/CEA, France

M. Ghil

École Normale Supérieure, Paris, France

and Dept. of Atmospheric and Oceanic Sciences, UCLA, United States

**Corresponding author address:* Alexis Hannart, IFAECI, Ciudad Universitaria, Pab. II, piso 2, Buenos Aires, Argentina

E-mail: alexis.hannart@cima.fcen.uba.ar

ABSTRACT

The emergence of clear semantics for causal claims and of a sound logic for causal reasoning is relatively recent, with the consolidation over the past decades of a coherent theoretical corpus of definitions, concepts and methods of general applicability (e.g. Pearl [2000]) which is anchored into counterfactuals. The latter corpus has proved to be of high practical interest in numerous applied fields (e.g. epidemiology, economics, social science). In spite of their rather consensual nature and proven efficacy, these definitions and methods are to a large extent not used in Detection and Attribution (D&A). This article gives a brief overview on the main concepts underpinning the causal theory and proposes some methodological extensions for the causal attribution of weather and climate-related events that are rooted into the latter. Implications for the formulation of causal claims and their uncertainty are finally discussed.

CAPSULE SUMMARY

Causal counterfactual theory provides clear semantics and sound logic for causal reasoning. It may help foster research on, and clarify dissemination of, weather and climate-related events attribution.

Background and rationale. A significant and growing part of climate research studies the causal links between climate forcings and observed responses. This part has been consolidated into a separate research topic known as detection and attribution (D&A). The D&A community has increasingly been faced with the challenge of generating causal information about episodes of extreme weather or unusual climate conditions. This challenge arises from the needs for public dissemination, litigation in a legal context, adaptation to climate change or simply improvement of the science associated with these events (Stott et al. 2013). For clarity, we start by introducing a few notations that will be used throughout this article: an event here is associated with a binary variable, say Y, which is equal to 1 when the event occurs and to 0 when it does not, and we use the term "event Y" as an abbreviation for "the event defined by $Y = 1$". In any event attribution study, the precise definition of the event to be studied — i.e., the choice of the variable Y — is crucial. Often, Y is defined *ad hoc* in the aftermath of an observed extreme situation based on exceedance over a threshold u of a relevant climate index Z, where both the index and the threshold are to a large extent arbitrary. In the conventional approach, which was introduced one decade ago by M.R. Allen and colleagues (Allen 2003; Stone and Allen 2005), one evaluates the extent to which a given external climate forcing $f \in \mathscr{F}$ — where $\mathscr{F}$ encompasses for instance solar irradiation, greenhouse gas (GHG) emissions, ozone or aerosol concentrations — has changed the probability of occurrence of the event Y. For this purpose, one compares the probability of occurrence of said event in an ensemble of model simulations representing the observed climatic conditions, which simulates the actual occurrence probability in the real world, with the occurrence probability of the same event in a parallel ensemble of model simulations, which represent an alternative world. The latter world is referred to as *counterfactual,* and it is the one that might have occurred had forcing f been absent. To be precise, we introduce the binary variable X_f to indicate whether or not the forcing f is present. The probability $p_1 = P(Y = 1 \mid X_f = 1)$ of the event occurring in the real world, with f present, is referred to as *factual*, while $p_0 = P(Y = 1 \mid X_f = 0)$ is referred to as counterfactual. Both terms will become clear in the light of what immediately follows. The so-called fraction of attributable risk (FAR) is then defined as:

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

The FAR is interpreted as the fraction of the likelihood of an event which is attributable to the external forcing f. Causal claims follow from the FAR and its uncertainty, associated with model and sampling errors, resulting in statements such as "*It is very likely that over half the risk of European summer temperature anomalies exceeding a threshold of 1.6°C is attributable to human influence.*" (Stott et al. 2004).

This conventional framework and the FAR were initially adapted from best practices in epidemiology (Greenland and Rothman 1998), a field in which causal inference has always been of primary importance. Best practices in epidemiology are themselves to some extent anchored in what can be referred to as the standard theory of causality. Indeed, there exists a theoretical corpus of definitions, concepts and methods to define causality rigorously and to address the issue of evidencing causal relationships empirically, e.g. Pearl (2000). The latter are readily accessible to users and are progressively being implemented in a growing number of fields. As a classic example taken from epidemiology, statements of great importance for public health, such as "*smoking causes lung cancer,*" are often based on these shared definitions and methods to investigate causality. The same is true of many causal studies that can be found in the fields of economics, social science or artificial intelligence, to mention but a few domains of application. One point of entry into the standard theory consists in the following historical definition: "*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.*" (Hume 1748). Or, where X and Y are events: Y is caused by X if and only if (iff), were

X not to occur, then Y would not occur. Despite its dating back to the 18th century, the above counterfactual definition and the general approach to causality that it implies is still relevant. Yet over the past decades, this definition has been further extended and refined within a probabilistic and graph-theoretical framework, allowing for the counterfactual approach to be applied to actual datasets, and to lead to reliable causal inference.

Overall, the current event attribution framework obeys the spirit of counterfactual logic and it is thus loosely connected to the above-mentioned corpus. Yet it would be beneficial to tighten this connection by adding several important concepts, definitions and mathematical results of causal counterfactual theory which, to the best of our knowledge, are lacking in the current event attribution framework. Among other lacking items, perhaps the most important one regards the absence of definition for the word "*cause*". Several recurrent controversial arguments in the realm of event attribution may possibly be related to this lacking definition of causality: for instance, an argument often made (Trenberth (2012)) is that any single event has multiple causes, so one can never assert that CO_2 emissions, nor any other factors, have actually caused the event. Following this logic, single events are thus inherently never causally attributable at all. It is arguably difficult to clearly address this objection — nor possibly many others — without a precise definition of causality in hand.

The purpose of this paper is to propose a set of definitions and methodological extensions to the current event attribution framework that are rooted in recent developments of causal counterfactual theory. We start with a brief overview of the counterfactual theory, emphasizing the most relevant concepts, and then proceed to illustrate the proposed extensions by revisiting the historical case study of the European heat wave of 2003. Implications for causal claims are finally discussed.

A brief overview of the theory of causality. We all deal with cause and effect in our everyday life. Yet, the notion of causality has long been shrouded in controversy, and the field of climate science is no exception in this respect. One may argue that the main reason for this state of affairs is the lack of clear semantics for causal claims: scientists and philosophers have indeed struggled to define precisely when one event truly *causes* another, and conversely when it does not. For instance, while we all understand that barometers do not cause rain, even such a simple fact cannot be easily translated into a precise formalization or a mathematical equation. Beside this semantic difficulty, a fundamental question is to determine what evidence is required to justify the causal claim: "*the falling barometer did not cause the rainy episode*" and how such evidence may be extracted from observations.

Consider a naive observer O who knows nothing about either meteorology or barometers. By recording the movements of the barometer's needle together with the changes in weather during a few weeks, O may be tempted to infer from the repeated observation of rainy episodes being preceded by a barometer fall and of sunny ones being preceded by a rise, that the needle's movement actually did cause the weather to change — even without a clue with respect to (w.r.t.) the physical mechanism that may account for this causal relationship. However, O's causal hypothesis will be quickly ruined if she/he has the flash of inspiration to start experimenting with the barometer: forcing its needle up and down will soon convince O that acting on the barometer does not induce a weather change. This simple example illustrates two aspects of causality: first, that causal investigation relies crucially on observations; and second, that two different types of observations may be used by the causal investigator: experimental and natural (i.e. non experimental). While both of these aspects may seem obvious, the difficulty starts with the implementation: given a piece of data, experimental or not, what causal conclusions can be drawn from it? And what is the level of confidence associated with such causal conclusions? Over the past decades, a rigorous theory of causality has emerged and been consolidated, with the purpose of addressing these questions. Its main ideas and concepts are exposed next.

The mathematical basis of causal theory. The counterfactual definition of causality given by David Hume and spelled out above — i.e. Y is caused by X iff Y would not have occurred were it not for X — can be used to introduce this brief overview. For instance, let R be a rainy episode and B a downward move of the barometer's needle; then observing R while impeding B — i.e. by holding the barometer's needle — provides counterfactual evidence that falling barometers do not cause rain. Applying this approach to data requires a few mathematical concepts from the theory of probability and from graph theory. The former entails the notion of *dependence* between random variables which is, of course, different from that of *causal dependence* but proves instrumental in the formalization of causality. In the rainy episode example above, it is clear that the variables B and R are dependent, which of course does not imply anything about their causal relationship. If we now introduce the variable W to denote whether or not a road near O is wet, then the rain R and the wet road W are clearly dependent and this is also the case of the barometer B and the wet road W. Once we know, however, that it has rained, we can deduce that the road is certainly wet no matter the evolution of the barometer, so that W is independent of B conditionally on R. This important property is called *conditional independence*:

$$P(W \mid B,R) = P(W \mid R)\,; \tag{2}$$

this equation basically expresses that R *screens off* B from W. If we further complement our illustration by introducing L, which denotes whether or not a low-pressure meteorological system is present above O, one can see by following a similar reasoning that $P(R \mid B,L) = P(R \mid L)$ and $P(W \mid R,L) = P(W \mid R)$, i.e. that L screens off B from R and that R screens off L from W.

Oriented graphs are a very useful tool to visualise these considerations and can be considered as the second building block of causal theory (Pearl 2000). Skipping the rigorous definitions, a graph can be described as a mapping of the conditional dependence relationships prevailing within a given joint probability distribution $P(Z_1, Z_2, ..., Z_n)$ under study (Pearl 2000; Ihler et al. 2007). Each variable Z_k is thus represented by a node, which is connected to one or more nodes by arrows, and each arrow points from a *parent* to a *child*. It is thus intuitive that graphs complement the purely probabilistic notion of dependence, which is symmetric and non-causal, by introducing an asymmetry in the connections between variables, which is suited to encode causal relationships. The graph associated with $(Z_1, Z_2, ..., Z_n)$ may be understood as a visual representation of the following factorization:

$$P(Z_1, Z_2, ..., Z_n) = \Pi_{k=1}^{n} P(Z_k \mid \mathscr{P}_k)\,, \tag{3}$$

where $\mathscr{P}_k$ denotes the parents of variable Z_k. The graph representing causality in our illustrative wet-road example is shown in Fig. 1a and visually encodes the following factorization:

$$P(B, R, W, L) = P(L)\, P(B \mid L)\, P(R \mid L)\, P(W \mid R)\,. \tag{4}$$

Causal relationships among a set of variables can thus conveniently be represented by their joint probability distribution, provided conditional dependence relationships are fully specified; such specification is conveniently encoded by using an oriented graph in which each arrow represents a causal relationship. The existence of causal relationships has various implications on the joint dependence structure: e.g. independent causes become dependent conditional on their common effect and dependent effects become independent conditional on their common cause. From the moment we have access to enough observations to infer the dependence structure, we are able to detect these signatures and thereby to evidence causal relationships. Algorithms such as those described in Spirtes et al. (2000) and Shimizu et al. (2006) basically follow this strategy, and could perfectly be applied to the natural observations of R, B and L collected by O.

An important limitation of using natural data though, is that several graphs can be compatible with the same joint distribution and hence with the same observations: identifiability is an issue.

For instance, simultaneous changes in X and Y are compatible with both the causal relationships $X \to Y$ and $Y \to X$ whenever only these two variables are observed (e.g. when observing R, B but not L). The experimental approach is thus required for disambiguation of the causal relationship between X and Y. Several outcomes Y are thereby experimentally collected, for each tested value of X. The value of X is thus chosen by the experimenter, and treating it as a random variable is no longer relevant in this experimental context. However, a probabilistic treatment of the response Y is still relevant, because other factors potentially affecting Y may not be controlled in the experimental set-up. The notion of *intervention* was hence introduced to describe the situation where X is set by the experimenter at a chosen value x; it is denoted $\mathrm{do}(X = x)$. The notion of *interventional* probability then corresponds to the distribution of Y obtained in an experiment under the intervention $\mathrm{do}(X = x)$. It is denoted $P(Y \mid \mathrm{do}(X = x))$ or alternatively $P(Y_x)$, where Y_x denotes the new random variable obtained for Y subject to the intervention $\mathrm{do}(X = x)$. The set $\{P(Y_x = y) \mid x, y = 0, 1\}$ obtained by collecting all the interventional probabilities of Y for every possible value of X is termed the *causal effect* of X on Y. It is important to note that, in general:

$$P(Y \mid \mathrm{do}(X = x)) \neq P(Y \mid X = x), \tag{5}$$

which is why the notation $\mathrm{do}(X = x)$ is required. Indeed, $P(R = 1 \mid B = 1)$ reads in our example "*the probability of rain knowing that the barometer is decreasing*" in a non-experimental context in which the barometer evolution is left unconstrained, whereas $P(R = 1 \mid \mathrm{do}(B = 1))$ reads "*the probability of rain forcing the barometer to decrease*" in an experimental context in which the barometer is manipulated. The two probabilities are obviously distinct and it is their difference that allows for disambiguation, as it reveals the absence of a causal link between B and R.

Nonetheless, confusion is still possible because $P(Y \mid \mathrm{do}(X = x))$ and $P(Y \mid X = x)$ may also sometimes be equal. This is the case when X satisfies a property called *exogeneity* w.r.t. Y. Without going into details, a sufficient condition for X to be exogenous w.r.t. any variable is to be a top node of a causal graph. In the present context, radiative forcings under causal scrutiny are actually modeled in a physical setting, such as a general circulation model (GCM), as prescribed conditions that are external to the climate system; they are thus exogenous by construction. Provided D&A keeps on focusing on causal relationships between variables that are exogenous, the otherwise critical distinction between conditional and interventional probability is therefore not of utmost importance here because both quantities are actually the same.

Necessity, sufficiency and probabilities of causation. In order to assess how likely it is that one event *was the cause* of another, the *probability* PN *of necessary causality* is defined, in agreement with the counterfactual principle, as the probability that the event Y would *not* have occurred in the *absence* of the event X given that both events Y and X did *in fact* occur. The probability PN thus quantifies how likely it is that X has caused Y in a *necessary causation* sense; here "X is a *necessary cause* of Y" means that X is required for Y to occur but that other factors might be required as well. In other words, it means that Y would not occur were it not for X. Sufficient causation, on the other hand, as in "X is a *sufficient cause* of Y," means that X always triggers Y but that Y may also occur for other reasons without requiring X. The probability PS of sufficient causation is defined to be the probability that Y would have occurred in the presence of X, given that Y and X did not occur. Note that PN and PS are thus simultaneously interventional and conditional probabilities. To complete the probabilistic setting, PNS is the probability of necessary and sufficient causation. It is defined as the probability that Y would have occurred in the presence of X, and that Y would not have occurred in the absence of X. These three definitions are formally expressed as follows (Pearl (2000) p. 286):

$$\begin{aligned}
\mathrm{PN} &=_{\mathrm{def}} P(Y_0 = 0 \mid Y = 1, X = 1)\,, \\
\mathrm{PS} &=_{\mathrm{def}} P(Y_1 = 1 \mid Y = 0, X = 0)\,, \qquad (6)\\
\mathrm{PNS} &=_{\mathrm{def}} P(Y_0 = 0, Y_1 = 1)\,.
\end{aligned}$$

The three probabilities PN, PS and PNS are of utmost importance because they provide a complete characterization of the causal relationship between X and Y, as well as of the associated uncertainties. Their estimation can thus be viewed as the ultimate purpose of a causal attribution study. Before addressing the issue of deriving them in practice, it is enlightening to discuss which of the three probabilities are most relevant for causal attribution, in which context, and how they should be interpreted.

On the one hand, PN closely matches the reasoning used in lawsuits, where legal responsibility is understood counterfactually, i.e. in the sense of necessary causation. In such a context, PN equals the probability that the damage Y suffered by the plaintiff would not have occurred were it not for the defendant's action X, and the latter is declared guilty whenever it can be proven that PN is high enough: the threshold is explicitly set to $1/2$ in a civil case ("preponderance of the evidence") and to an unspecified value that is supposedly very close to one in a criminal case ("beyond reasonable doubt"). Assume for instance that an individual A fires a gun (X) in a seemingly desert but public place. Unluckily, an individual B who happens to be standing one kilometer away is hit and injured (Y). Legally speaking, A is an obvious culprit for the injury of B and will likely be convicted in case of a trial, because PN is very close to unity here: B would be safe and sound had it not been for A shooting. Nevertheless, the probability of the bullet hitting someone from such a long distance is very low, the lightest wind gust could possibly have deviated its trajectory and saved B. The probability of sufficient causation PS is thus close to zero here but this is not important in a legal context, in which it is only PN that matters, while PS does not.

In contrast, consider the case of a policymaker who aims at reducing the number of casualties from accidental shootings (Y) through a policy (X). An abrupt policy prohibiting gun sales altogether will clearly be sufficient but arguably not necessary, since a smoother policy based on tightly regulated sales may achieve a similar result. In parallel, improving the dissemination of safety information to gun owners is arguably necessary but will likely not be sufficient. In any case, it is a high PS that guarantees that the desired objective Y will be met by the policy X, not a high PN: PS therefore tends to be more important than PN in the context of elaborating and assessing policies.

Even though all three probabilities relate to counterfactual worlds, it is worthwhile underlining that these quantities are not nebulous metaphysical notions: the definitions are precise and unambiguously implementable, as long as a fully specified probabilistic model of the world is postulated. This being said, it is still a difficult task to derive them under general assumptions, and one that remains an active and challenging research topic in causal theory at present. Important results were obtained, however, by introducing some additional assumptions. For instance, under the assumption of monotonicity, the following exact expressions hold:

$$\begin{aligned}
\mathrm{PN} &= 1 - \frac{p_0}{p_1} + \frac{p_0 - P(Y_0 = 1)}{P(X = 1, Y = 1)}\,, \\
\mathrm{PS} &= 1 - \frac{1 - p_1}{1 - p_0} - \frac{p_1 - P(Y_1 = 1)}{P(X = 0, Y = 0)}\,, \\
\mathrm{PNS} &= P(Y_1 = 1) - P(Y_0 = 1)\,;
\end{aligned} \tag{7}$$

where variable Y is said to be monotonic w.r.t. variable X iff for any realization ω in the probability space Ω, $Y_x(\omega)$ is a monotonic function of x. Furthermore, when assuming exogeneity of X w.r.t. Y in addition to monotonicity, the expressions given in Eq. (7) simplify because interventional and conditional probabilities are then equal, i.e. $p_x = P(Y_x = 1)$ for $x \in \{0, 1\}$, and thus

$$\mathrm{PN} = 1 - \frac{p_0}{p_1}, \quad \mathrm{PS} = 1 - \frac{1 - p_1}{1 - p_0}, \quad \text{and } \mathrm{PNS} = p_1 - p_0\,. \tag{8}$$

Note that under such conditions, PN matches with the FAR — we elaborate on this coincidence further in this article. Another important result of causal theory which is linked to to Equation (8) is that under exogeneity and releasing the assumption of monotonicity, the probabilities of causation are then no longer identifiable, but the three quantities $1 - p_0/p_1$, $1 - (1 - p_1)/(1 - p_0)$ and $p_1 - p_0$ provide lower bounds respectively for PN, PS and PNS. Figure 2 shows a plot of the expressions given in Eq. (8): it can be seen that PN is more sensitive to p_0 than to p_1, and conversely that PS is more sensitive to p_1 than to p_0: necessary causation is enhanced further by an event being rare in the counterfactual world, whereas sufficient causation is enhanced further by its being frequent in the real one. This being said, PN and PS are clearly not independent and coincide under two situations: (i) when $p_0 + p_1 = 1$ (e.g. in a deterministic context where $p_1 = 1$ and $p_0 = 0$, then both PN and PS $= 1$); and (ii) when $p_0 = p_1$ (e.g. where the counterfactual and real worlds' responses are identical, then both PN and PS $= 0$).

Causal attribution of climate-related events. Choosing to focus on PN or PS is a matter of point of view. To illustrate this issue, we can consider two typical perspectives: the *ex post* perspective of the plaintiff — or the judge, or the insurance contract holder — and the *ex ante* perspective of the planner — or the policymaker, or the campaigner. In the first case, the question "who is to blame for the event that occured?" — with the potentially many implications of its answer — is central. The problem of climatic event attribution can thus be compared to a lawsuit, and actually does already appear in courts (Adam 2011): we may primarily seek to determine responsibilities for the event and its aftermaths, where responsibility is understood in a legal sense i.e. in a necessary causation sense. Event attribution thus requires the adversarial debate typical of a lawsuit in order to cautiously balance incriminating versus exonerating evidence, i.e. to evaluate the main cause under scrutiny, e.g. anthropogenic forcings, as well as each and every possible alternative explanations, e.g. natural forcings or internal variability of the climate system, which may have led to the same outcome. If the resulting PN is high enough, then human responsibility is established and a ruling may in theory follow, as it does in litigation cases. In any case, as in the imprudent shooter example, PS does not matter here, only PN does.

By contrast, the planner is looking forward and may ask instead the general type of question "what should be done today w.r.t. events that may occur in the future?" For instance, in the context of mitigation, two causal questions are at stake: on the one hand, what is the, expectedly beneficial, effect of limiting CO_2 emissions? and, on the other hand, what is the, expectedly costly, effect of not limiting them? The first question seeks a causal guarantee that removing the forcing will make the event less frequent and the concern is thus predicated on necessary causality. Conversely, the second question seeks a causal guarantee that maintaining the forcing will maintain the event frequency and the concern is thus predicated on sufficient causality. Therefore, PS is the appropriate focus for the planner when assessing the future costs that inaction will imply, but PN is at stake when assessing the future benefits of enforcing strong mitigation actions. Policy elaboration requires both sides of this assessment; thus both PN and PS are of interest here. To summarize, depending on context, PN, PS or both may be relevant and can help answer different causal questions.

Methodological proposal. Our methodological proposal for the attribution of weather and climate-related events is rather straightforward and it is derived from the previous considerations. It consists of deriving the probabilities of necessary and of sufficient causality, PN^f and PS^f associated with the causal relationship between each forcing $f \in \mathscr{F}$ and an event Y of interest. As outlined in the introduction, the choice of Y is based on a climate variable Z and a threshold u; this choice depends on the causal focus of the study and is otherwise rather arbitrary. Once Y has been duly defined, the causal chain to be investigated is actually quite simple, notwithstanding the complexity of the climate system. It can be represented by the single, standard graph of

Fig. 1b, independently of the specificities of the event Y under scrutiny. A set of binary variables $\{X_f : f \in \mathscr{F}\}$ that represent the external forcings occupy the top nodes in this graph and are thus exogenous. The event variable Y has parents $\mathscr{P} = \{X_f : f \in \mathscr{F}\}$ and it is also influenced by internal climate variability v which is treated here as random terms (Ghil et al. 2008).

Next, we can apply Eq. (8) because all the forcings are exogenous and one may also assume that the event Y is monotonous w.r.t. the forcing. Indeed, assuming that the latter does not hold would imply that despite the event being more frequent in the factual world than in the counterfactual one (i.e. $p_1 > p_0$), there exists some realizations $\omega \in \Omega$ such that $Y_0(\omega) = 1$ and $Y_1(\omega) = 0$. That is, one can find some conditions under which the event does occur when the forcing is turned off but no longer occurs only by turning it on — other conditions being held unchanged. Such conditions are arguably not realistic physically for a broad class of events and for the forcings usually considered in D&A. We thus derive $\mathrm{PN} = 1 - p_0/p_1$ and $\mathrm{PS} = 1 - (1-p_1)/(1-p_0)$ for each forcing f and omit hereinafter for simplicity the index f. Hence, the challenge is now to estimate the causal effects $\{p_0, p_1\}$. In many fields, experimental and/or natural observations of a response Y — say, in epidemiology, a disease — and of a factor X — say, a bad habit or a treatment — are available for a sample of individuals, allowing for a direct estimation of p_1 and p_0. Most unfortunately, in the climate sciences, no such sample of “Earth-like climate systems” is accessible to natural observation, and even less so to experimental testing. The paleoclimatic record may in theory palliate this difficulty by considering several remote episodes of Earth’s climatic history as a sample (National Research Council 1995). An important limitation of this approach, however, is the limited size and high uncertainty of the indirect paleoclimatic estimates of both the response Y and the forcings X_f over the distant past. Furthermore, such non-experimental analysis is inherently restricted to forcings that can be traced to paleoclimatic perturbations that did occur and for which exogeneity is guaranteed. With such strong limitations on the natural observation side and with *in* *situ* experimentation inaccessible, we are left with the only remaining alternative: so-called *in silico* experimentation. This option is rendered plausible by the increasing realism of climate system models that were developed partly for this purpose. Estimates of the causal effects $\{p_0, p_1\}$ can be obtained from an ensemble of numerical experiments consisting of r_1 and r_0 runs under factual and counterfactual conditions, respectively, w.r.t. one or more forcings f. An obvious estimation strategy is to use the empirical frequencies $\widehat{p}_x = \sum_{k=1}^{r_x} Y_x^{(k)}/r_x$ for $x \in \{0,1\}$, where $Y_x^{(k)}$ is the event occurrence in the k-th run of the factual or counterfactual experiment. This option presents a major shortcoming since $\widehat{p}_x$, as well as PN and PS, are affected by high sampling uncertainty. In practice, due to restrictions on computer resources, r_x is typically in the range of 10 to 100, while asymptotic convergence requires r_x to be large compared to the return period $T_x \simeq 1/p_x$ of the event; the latter is clearly out of reach for the rare events usually at stake. Another serious difficulty is that climate models, including the most detailed GCMs, are simplified representations of reality that are affected by both numerical and physical modeling errors. Thus the real causal effects may differ from the model causal effects. While both these difficulties are serious, they can be addressed by introducing additional assumptions on the distribution of the climate variable Z, and by treating model error as an additional random term influencing the response variable Y. Discussing such approaches is beyond the scope of this paper. The probabilities PN and PS are then derived from the estimates $\widehat{p}_1$ and $\widehat{p}_0$ so obtained.

Causal claims are eventually formulated from these probabilities, translated into words based on standardized uncertainty wording, such as the one used in IPCC (2013). Summarizing, the general methodological approach proposed herewith consists of the following:

- Define a response variable of interest Y based on a climate index Z and threshold u;
- Infer the causal effects associated with Y, based on *in silico* experimentation;

• Derive PN and PS for each forcing and formulate associated causal claims, by using for instance the IPCC (2013) uncertainty terminology.

2003 European heatwave. We illustrate our approach by revisiting one of the first counterfactual event attribution studies (Stott et al. 2004), which focused on the European heat wave of the summer of 2003. Applying our notation and the above three steps to this study:

• Z is the mean summer temperature anomaly over Europe, and u is set at 1.6°C;

• The factual and counterfactual probability density functions (pdfs) of Z are obtained from the corresponding two ensembles by fitting a generalized Pareto distribution to each one, cf. Fig. 3a. The inference procedure yields two ranges of values for the return periods: $350 \leq T_0 \leq 2500$ and $100 \leq T_1 \leq 1000$. For the sake of clarity, we choose to concentrate here on two values which are arbitrarily chosen within these ranges: $T_0 = 1250$ years and $T_1 = 125$ years, implying $p_0 = 0.0008$ and $p_1 = 0.008$;

• These values of p_0 and p_1 yield PN $= 0.9$ and PS $= 0.0072$, by applying Eq. (8).

It follows that CO_2 emissions *are very likely to be a necessary cause, but are virtually certainly not a sufficient cause, of the summer of 2003 heat wave*. This statement highlights a distinctive feature of unusual events: several necessary causes may often be supported by the data, but rarely a sufficient one. To further illustrate this point, we plot PN, PS and PNS as a function of the threshold u in Fig. 3b. It is clear from this figure that the causal evidence shifts from necessary and not sufficient when u is large (unusual event) to sufficient and not necessary when u is small (usual event). This shift occurs because in the latter case, it is the nonoccurrence of event Y that becomes an unusual event. But this rare "non-event" tends to be less unusual in the counterfactual world than in the factual one, which implies necessity for the "non-event" and thus sufficiency for the event, by the definitions of PN and PS, respectively, in Eq. (6).

In any case, a low threshold conversely yields PN $\simeq 0$ and PS $\simeq 1$: it follows that anthropogenic CO_2 emissions *are virtually certainly a sufficient cause, and are virtually certainly not a necessary cause, of the fact that the summer of 2003 was not unusually cold*. Therefore, this symmetrically illustrates that the occurrence of a usual event — or equivalently, the non-occurrence of a rare event — is thus often prone to have a sufficient cause but rarely necessary ones.

The above analysis defines the occurrence of the event "2003 European heatwave" w.r.t. to the particular year when it occurred. Such a definition of the event inherently considers that the particular year of occurrence 2003 is a relevant feature thereof, and consequently builds this feature into the causal analysis. This approach is particularly relevant in the context, say, of an insurance contract, which may often apply only to a single specified year. But a broader perspective focusing on longer timescales is arguably more relevant in other contexts, such as elaborating adaptation and mitigation policy, which has no reason to grant any particular importance to the year 2003. In such a context, one would release the year 2003 as an event feature and focus instead on the fact that a severe European heatwave did occur. The meaningful temporal feature retained here would be "occurrence during the industrial period" instead of "occurrence during year 2003". It is straightforward to translate this approach into our proposed framework by going through the same three steps again. In what follows, we denote for clarity by an asterisk the new variables Y^*, Z^* and u^*:

- Z^* is defined to be the number of occurrences of European heatwaves over a time period of length τ ending in 2003, where in any given year a heatwave occurrence is defined as above by $Z \geq u$, and the threshold u^* is set to 1. The event Y^* thus occurs if at least one heatwave took place in Europe during the time interval $2004 - \tau \leq t \leq 2003$.
- Deriving the new causal effects $\{p_0^*, p_1^*\}$ is straightforward, subject to assuming stationarity w.r.t.

time (see discussion immediately below), based on the previous causal effects $\{p_0, p_1\}$:

$$p_x^* = P(Z_x^* \geq 1) = 1 - (1 - p_x)^\tau \; . \tag{9}$$

For $\tau = 1$, this equation reduces to $p_x^* = p_x$, since $Y^* = Y$ in this case. For τ large compared to the return period $T_x \simeq 1/p_x$ of event Y, it implies $p_x^* \simeq 1$; this is also unsurprising because in either the factual or the counterfactual world, the occurrence of a heatwave, no matter how rare in any given year, is certain over a sufficiently long period.

• Plotting in Fig. 3c PN* and PS* as a function of τ, based on Eq. (9), we see that the causal evidence shifts from necessary and not sufficient in the limiting case $\tau = 1$ (since $Y^* = Y$), to sufficient and not necessary when τ gets asymptotically large. For $\tau = 200$ years — i.e. the industrial period, which matches approximately the instrumental record length — we find from Eq. (9) that $p_0 = 0.14$ and $p_1 = 0.80$, and next that PN$^* \simeq$ PS$^* \simeq 0.8$.

It follows that anthropogenic CO_2 emissions *are likely to be both a necessary cause and a sufficient one for a 2003-like heatwave to have occurred at least once over the industrial period.* Summarizing, sufficient causality does not apply to the event occurrence on the particular year when it did occur, but it does for such an event to have occurred at least once over the entire period. Evidence of necessary causality, on the other hand, is strong in both cases. This illustrative example thus shows that whether one considers something as fortuitous as its particular year of occurrence to be a relevant feature of the event under scrutiny, or not, has crucial implications for the associated level of causal evidence. Replacing the feature "year of occurrence" by the feature "occurrence during the industrial period" may be more relevant to the analysis in many situations, and yield more powerful causal evidence.

This being said, the stationarity hypothesis underlying Eq. (9) is unrealistic because mean temperature did change over the period considered, and so did extremes. This convenient assumption was made here for the sake of illustrating in a simple and qualitative way the effect on PN and PS of defining the event occurrence on a longer period of length τ. While a realistic non stationary treatment of this case study is beyond our scope, it is important to underline that including assumptions of non-stationarity into a causal inference study presents no particular difficulties in general. For instance, in the present case study, this may be done merely by using the more general expression:

$$p_x^* = 1 - \prod_{t=1}^{\tau} (1 - p_{x,t}) \,. \tag{10}$$

in place of Eq. (9) in order to determine the causal effects $\{p_0^*, p_1^*\}$. In Eq. (10), $p_{x,t}$ denotes the probability of occurrence of a heatwave in year t and is thereby allowed to change over time. In practice, $(p_{x,t})_{t=1}^{\tau}$ may be estimated based on an ad-hoc statistical model accounting for non-stationarity. For instance, a commonplace choice for the latter is to specify the PDF of the index Z in year t conditionally on a covariate which changes in time (e.g. mean temperature) and/or an explicit parametric dependence to time t (e.g. a linear trend). Note that Eq. (10) would clearly be required for the estimation of p_1^* because the factual world has undeniably changed. Yet Eq. (9) may still be considered acceptable for the estimation of p_0^* since the counterfactual world would arguably have suffered limited changes. Accordingly, one may expect that when moving to a non stationary treatment, (i) p_0^* would only be marginally affected, (ii) p_1^* would potentially be substantially affected. More precisely, one would expect p_1^* to have a lower value because $p_{x,t}$ is expected to be lower than its value in year 2003, for any year t preceding it. Therefore, based on above considerations and on Fig. 2, accounting for non-stationarity would expectedly translate here into a slight decrease in PN, a potentially pronounced decrease in PS, and a lower level of causal evidence overall — as compared to the values given above for illustration.

In any case, each of the different perspectives taken above addresses a causal question about the 2003 heatwave that is different, and may be of interest for distinct purposes. But while the questions only differ slightly, the answers vary greatly. The answer to such an open question as "*have CO_2 emissions caused the 2003 European heatwave?*" is thus dramatically affected by (*i*) how one defines the event "2003 European heatwave"; and (*ii*) whether causality is understood in a necessary or sufficient sense. Precise causal answers about climate events thus require precise causal questions.

Concluding remarks. We have provided an introduction to causal theory, as used in causal studies across several disciplines, and proposed a simple methodology for its application to D&A studies. We hope that this methodological framework — along with the more precise vocabulary it relies on — will help clarify discussions between D&A experts, as well as communication to wider audiences.

We have shown, with simple examples, that it is important to distinguish between necessary and sufficient causality. Such a distinction is, at present, lacking in the conventional event attribution framework. Any time a causal statement is being made about a weather or climate-related event, part of the audience understands it in a necessary-causation sense, while another part understands it in a sufficient-causation sense — which can give rise to many potential misunderstandings. Introducing the clear distinction may thus clarify discussions. Specifically, it may for instance help address the claim recalled in introduction, according to which single events are never attributable since they are multi-caused. In light of what precedes, this claim intrinsically postulates that a cause qualifies as such only if it is both necessary and sufficient. The latter is arguably far too restrictive an approach of causation.

Our revisiting the well-known case study of the European heatwave of 2003 should clarify an apparent paradox in the interpretation of such studies. Even in the few such cases where evidence supporting necessary causation is strong, assertive causal statements appear to have been shied away from, possibly by the perception that sufficiency was lacking. A statement such as "*CO_2*

emissions have not caused the particular event Y: they have only caused the probability of occurrence of Y-like events to increase" may actually often be too conservative and even wrong: as in the above example, it may indeed be the case that *CO_2 emissions did cause event Y* — although in a restrictively necessary causation sense. Further, by defining the event to mean not just occurrence in a particular year but during the entire industrial era, it may be possible to establish that *event Y was in fact caused by increased CO_2 emissions* — this time w.r.t. both necessity and sufficiency.

Our proposed methodology, like the conventional one, relies on *in silico* experimentation to derive both the factual and the counterfactual probabilities p_1 and p_0, respectively, use the two to obtain the quantity $1 - p_0/p_1$, and then translate it into a causal statement. Our extended framework, however, has important distinctive features. First, we have shown that $1 - p_0/p_1$ is associated only with the first facet of causality, that of necessity, and we have introduced its second facet, that of sufficiency, which is associated to the symmetric quantity $1-(1-p_1)/(1-p_0)$. Both have been shown to be relevant depending on the context. Second, the interpretation given to $1 - p_0/p_1$ differs under both frameworks, which has deep implications for the formulation of causal statements and the treatment of uncertainty. The quantity $1 - p_0/p_1$ was coined as the fraction of attributable risk upon being introduced in event attribution — and similarly in other applied fields, terms like *excess risk ratio*, *attributable fraction* or *attributable proportion* are also used to name the same quantity. The FAR, as well as these similar terms, is used to communicate the idea — particularly relevant in epidemiology from which it originates — that the exposition to a given risk factor X translates into an increase of, say, the frequency of a given disease Y. In this terminology, the quantity $1 - p_0/p_1$ is a frequency increase index: it corresponds to a statistical monitoring approach, which is more descriptive than structural, in the sense that it does not embed any precisely defined causal meaning. For this reason, Pearl (2000) has argued that the term *attributable risk* is a misnomer: because such a precise causal meaning is lacking, the associated statement can only address the increase in frequency. Accordingly, uncertainty analysis conducted on the FAR by deriving its probability distribution cannot be easily translated into uncertainty on the causal link at stake — instead, the focus on the frequency increase and its uncertainty yields statements like "*There is a 90% confidence level that CO_2 emissions have increased the frequency of occurrence of Y-like events by a factor at least two*".

In causal theory, the probability of necessary causation PN formally embeds the notion of causal attribution in its definition, given by Equation (6). While PN is not easily computable in general, it coincides with $1 - p_0/p_1$ under exogeneity and monotonicity. These two rather restrictive conditions are fortunately met in the context of D&A, thus the quantity $1 - p_0/p_1$ usually referred as FAR now has a precise causal meaning, instead of being merely an index of frequency increase. This shift in interpretation affects the associated causal claim, which can now address more directly the actual causal link. Moreover, this shift has an immediate implication in terms of assessing the uncertainty of the claim: the latter is indeed already quantified because PN is a probability, which inherently measures uncertainty. Therefore, based on the same supporting data, the new interpretation translates into "*CO_2 emissions are likely to have caused event Y in a necessary causation sense*," a claim that is more direct, assertive and clear from a causal attribution standpoint than the previous one.

Finally, at a more practical level, attribution studies applying causal theory require the availability of counterfactual model simulations. This carries an immediate implication w.r.t. the design of standardized Coupled Modeling Intercomparison Project (CMIP) experiments that specifically address D&A purposes. The present analysis suggests moving towards a fully counterfactual design in the future — i.e., all forcings except f being 'on'— instead of the mostly factual one prevailing at present — i.e., forcing f only being on. Generalizing this design would be a significant step forward in attribution studies of weather and climate-related events.

Acknowledgments. Part of this work was supported by the French Agence Nationale de la Recherche grants DADA (AH, PN and MG), MCSim (PN), and MOPERA (PN), U.S. National Science Foundation grants DMS-1049253 and OCE-1243175 (MG).

References

Adam D. (2011) Climate change in court. *Nature Climate Change,* 1:127–130.

Allen M. R. (2003) Liability for climate change. *Nature*, 421:891–892.

Ghil M., M. D. Chekroun, and E. Simonnet (2008) Climate dynamics and fluid mechanics: Natural variability and related uncertainties, *Physica D*, 237:2111–2126.

Greenland S., and K. J. Rothman (1998) Measures of effect and measures of association, Chapter 4 in Rothman, K. J., and Greenland, S. (eds.), *Modern Epidemiology*, 2nd edn., Lippincott-Raven, Philadelphia, USA.

Hegerl, G.C., O. Hoegh-Guldberg, G. Casassa, M.P. Hoerling, R.S. Kovats, C. Parmesan, D.W. Pierce, P.A. Stott (2010): Good Practice Guidance Paper on Detection and Attribution Related to Anthropogenic Climate Change. In: *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Detection and Attribution of Anthropogenic Climate Change* [Stocker, T.F., C.B. Field, D. Qin, V. Barros, G.-K. Plattner, M. Tignor, P.M. Midgley, and K.L. Ebi (eds.)]. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland.

Hume D. (1748) *An Enquiry Concerning Human Understanding.* Reprinted Open Court Press (1958), LaSalle, IL, USA.

Ihler A. T., S. Kirshner, M. Ghil, A. W. Robertson and P. Smyth (2007) Graphical models for statistical inference and data assimilation, *Physica D*, 230:72–87, 2007.

IPCC (2013) Summary for Policymakers. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung,

A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

National Research Council (1995) *Natural Climate Variability on Decade-to-Century Time Scales,* D. G. Martinson, K. Bryan, M. Ghil et al. (Eds.), National Academy Press, Washington, D.C., USA.

Otto F. E. L., Massey N., van Oldenborgh G. J., Jones R. G., Allen M. R. (2012) Reconciling two approaches to attribution of the 2010 Russian heatwave. *Geophys. Res. Lett.* 39:L04702.

Pearl J. (2000) *Causality: Models, Reasoning and Inference*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Shimizu S., P. Hoyer, A. Hyvarinen, A. Kerminen (2006) A linear, non-gaussian acyclic model for causal discovery. *J. Machine Learning Res.*, 7:2003–2030.

Spirtes P., C. Glymour and R. Scheines. *Causation, Prediction, and Search*, 2nd ed., MIT Press, Cambridge, MA.

Stone D. A., and M. R. Allen (2005) The end-to-end attribution problem: from emissions to impacts. *Clim. Change*, 71:303–318.

Stott P. A., et al. (2013) Attribution of weather and climate-related events, in *Climate Science for Serving Society: Research, Modelling and Prediction Priorities*, G. R. Asrar and J. W. Hurrell (Eds.), Springer, in press.

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

Trenberth K. E. (2012) Framing the way to relate climate extremes to climate change. *Clim. Change*, 115(2):283–290.

LIST OF FIGURES

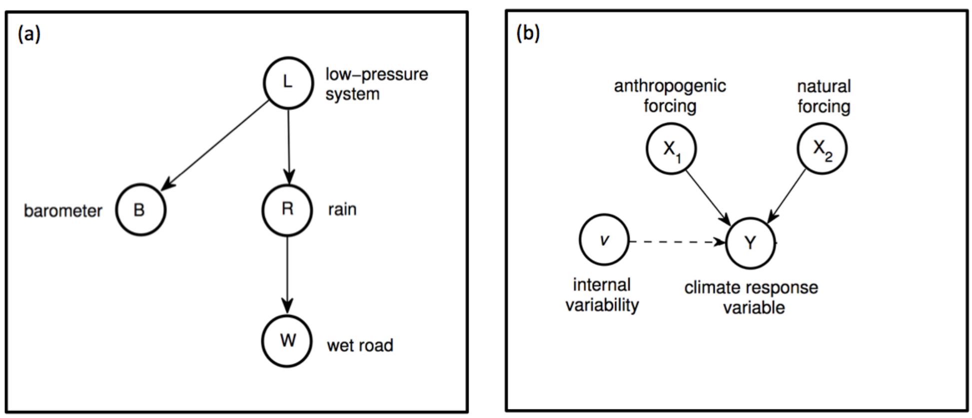

FIG. 1. Graphs representing dependencies: (a) among the four variables (R, B, W, L) used in our illustrative example, (b) among forcings (X_1, X_2) and climate response Y. Dotted arrows represent dependency upon the unobserved variable v.

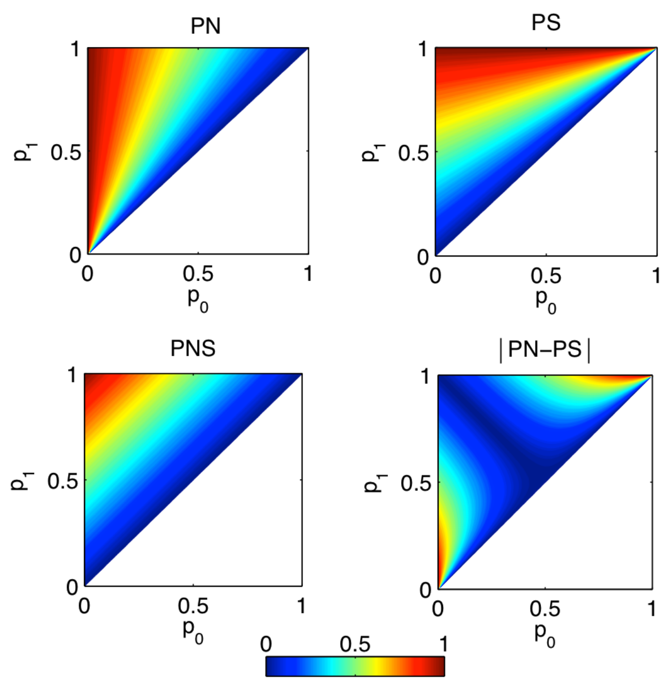

FIG. 2. Contour plots of PN (upper left), PS (upper right), PNS (lower left), and |PN − PS| (lower right) as functions of the counterfactual probability p_0 (horizontal axis) and of the factual probability p_1 (vertical axis).

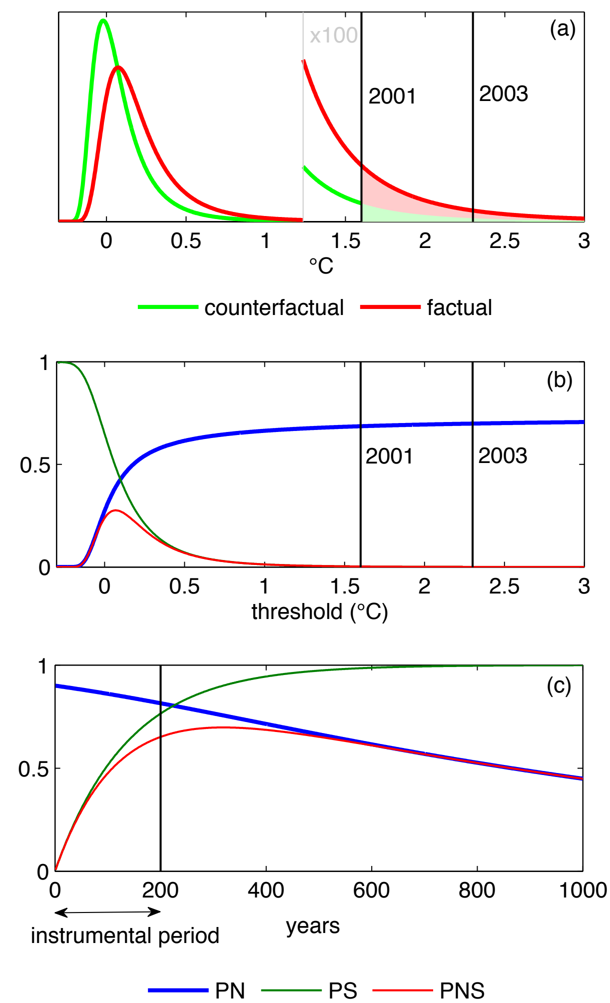

FIG. 3. Causal inference for the 2003 European heat wave. (a) Counterfactual and factual probability density functions (pdfs) of the temperature anomaly index, using a generalized Pareto distribution fit after Stott et al. (2004); (b) probabilities PN, PS and PNS as a function of the threshold u; (c) PN, PS and PNS as a function of the length of the observation period τ.