Probabilities of Causation of Climate Changes

Alexis Hannart, Philippe Naveau

Probabilities of causation of climate changes

Alexis Hannart *

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

Philippe Naveau

LSCE, CNRS/CEA, Gif-sur-Yvette, France

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

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

ABSTRACT

Multiple changes in Earth's climate system have been observed over the past decades. Determining how likely each of these changes are to have been caused by human influence, is important for decision making on mitigation and adaptation policy. Here we describe an approach for deriving the probability that anthropogenic forcings have caused a given observed change. The proposed approach is anchored into causal counterfactual theory (Pearl 2000) which has been introduced recently, and was in fact partly used already, in the context of weather and climate-related events attribution. We argue that these concepts are also relevant, and can be straightforwardly extended to, the context of climate change attribution. For this purpose, and in agreement with the principle of *fingerprinting* applied in the conventional D&A framework, a trajectory of change is converted into an event occurrence defined by maximizing the causal evidence associated to the forcing under scrutiny. Other key assumptions used in the conventional D&A framework, in particular those related to numerical models error, can also be adapted conveniently to this approach. Our proposal thus allows to bridge the conventional framework with the standard causal theory, in an attempt to improve the quantification of causal probabilities. An illustration suggests that our approach is prone to yield a significantly higher estimate of the probability that anthropogenic forcings have caused the observed temperature change, thus supporting more assertive causal claims.

1. Introduction

Investigating causal links between climate forcings and the observed climate evolution over the instrumental era represents a significant part of the research effort on climate. Studies addressing these questions in the context of climate change have been providing over the past decades, an ever increasing level of causal evidence that is important for decision-makers in international discussions on mitigation policy. In particular, these studies have produced far-reaching causal claims; for instance the latest IPCC report (Stocker et al. 2013) stated that "*It is extremely likely that human influence has been the dominant cause of the observed warming since the mid-20th century.*" An important part of this causal claim, as well as many related others, regards the associated level of uncertainty. More precisely, the expression "*extremely likely*" in the latter quote has been formally defined by the IPCC (Mastrandrea et al. 2010) to correspond to a probability of 95%. The above quote hence implicitly means that the probability that the observed warming since the mid-20th century was not predominantly caused by human influence but by natural factors, is roughly $1:20$. Based on the current state of knowledge, it means that it is not yet possible to fully rule out that natural factors were the main causes of the observed global warming. This probability of $1:20$, as well as all the probabilities associated to the numerous causal claims that can be found in the past and present climate literature, are critical quantities that are prone to affect the way in which climate change is apprehended by citizens and decision makers, and thereby to affect decisions on the matter. It is thus of interest to examine the method followed to derive them and, potentially, to improve it.

Aforementioned studies buttressing the above claim usually rely on a conventional attribution framework in which "*causal attribution of anthropogenic climate change*" is understood to mean "*demonstration that a detected change is consistent with the estimated responses to anthropogenic and natural forcings combined, but not consistent with alternative, physically plausible explanations that exclude important elements of anthropogenic forcings*" (Hegerl et al. 2010). While this definition has proved to be very useful and relevant, it offers

a description of causality which is arguably overly qualitative for the purpose of deriving a probability. In particular, it comes short of a mathematical definition of the word "*cause*" and incidentally, of the "*probability to have caused*" that we in fact wish to quantify. Hence, beyond these general guidance principles, the actual derivation of these probabilities is left to some extent to the interpretation of the practitioner. In practice, causal attribution has usually been performed by using a class of linear regression models (Hegerl and Zwiers 2011):

$$y = \sum_{f=1}^{p} \beta_f x_f + \varepsilon \tag{1}$$

where the observed climate change y is regarded as a linear combination of p externally forced response patterns x_f with $f = 1, ..., p$ referred to as fingerprints, and where ε represent of internal climate variability and observational error (all variables are vectors of dimension n). The regression coefficient β_f accounts for possible error in climate models in simulating the amplitude of the pattern of response to forcing f. After inference and uncertainty analysis, the value of each coefficient β_f and the magnitude of the confidence intervals determine whether or not the observed response is attributable to the associated forcing. The desired probability of causation, i.e. the probability that the forcing of interest f has caused the observed change y is denoted hereafter $\mathbb{P}(f \to y)$. It has often been equated to the probability that the corresponding linear regression coefficient is positive:

$$\mathbb{P}(f \to y) = \mathbb{P}(\beta_f > 0) \tag{2}$$

A shortcoming of the conventional framework summarized in Equations (1) and (2) above, is that a linear regression coefficient does not have any causal meaning from a formal standpoint. As acknowledged by Pearl (2000), turning an intrinsically deterministic notion such as causality into a probabilistic one, is a difficult general problem which has also long been a matter of debate (Simpson 1951; Suppes 1970; Mellor 1995). Nevertheless, the current approach can be theoretically improved in the context of climate change where the values of the probabilities of causation have such important implications.

Our proposal to tackle this objective is anchored into a coherent theoretical corpus of definitions, concepts and methods of general applicability which has emerged over the past three decades to address the issue of evidencing causal relationships empirically (Pearl 2000). This general framework is increasingly used in diverse fields (e.g. in epidemiology, economics, social science) in which investigating causal links based on observations is a central matter. Recently, it has been introduced in climate science for the specific purpose of attributing weather and climate-related events (Hannart et al. 2015a). The latter article gave a brief overview of causal theory and articulated it with the conventional framework used for the attribution of single weather events, which is also an important topic in climate attribution. In particular, Hannart et al. (2015a) showed that the key quantity referred to as the fraction of attributable risk (FAR) (Allen 2003; Stone and Allen 2005) which buttresses most weather events attribution studies, can be directly interpreted within causal theory.

However, Hannart et al. (2015a) did not address how to extend and adapt this theory in the context of the attribution of climate changes occurring on long timescales. Yet, a significant advantage of the definitions of causal theory — and to start with the very notion of "event" — is precisely that they are relevant no matter the temporal and spatial scale. For instance, from the perspective of a paleoclimatologist studying Earth's climate over the past few hundred millions of years, global warming over the past hundred and fifty years can be considered as a climate event. As a matter of fact, the word "event" is used in paleoclimatology to refer to "rapid" changes in the climate system, but ones that may yet last centuries to millennia. Where to draw the line is thus arbitrary: one person's long term trend is another person's short term event. It should therefore be possible to tackle causal attribution within a unified methodological framework based on shared concepts and definitions of causality. Doing so would allow to bridge the methodological gap that exists between event attribution and trend attribution, thereby covering the full scope of climate attribution studies. Such a unification would present in our view several advantages: enhancing methodological research synergies between D&A topics, improving

the shared interpretability of results, and streamlining the communication of causal claims — in particular when it comes to uncertainty.

Here, we address this issue by adapting some formal definitions of causality and probability of causation to the context of climate change attribution. Technical implementation under standard assumptions in D&A is then detailed. The method is finally illustrated on the warming observed over the 20th century.

2. Causal counterfactual theory

While an overview of causal theory can not be repeated here, it is necessary for clarity and self-containedness to highlight its key ideas and most relevant concepts for the present discussion

Let us first recall the so-called "counterfactual" definition of causality by quoting the 18th century Scottish philosopher David Hume: "*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 other words, an event E (E stands for effect) is caused by an event C (C stands for cause) if and only if E would not occur were it not for C. Note that the word *event* is used here in its general, mathematical sense of *subset* of a sample space Ω. According to this definition, evidencing causality requires a counterfactual approach by which one inquires whether or not the event E would have occurred in an hypothetical world, termed counterfactual, in which the event C would not have occurred. The fundamental approach of causality which is implied by this definition is still entirely relevant in the standard causal theory. It may also arguably be connected to the guidance principles of the conventional climate change attribution framework and to the optimal fingerprinting models, in a qualitative manner. The main virtue of the standard causality theory of Pearl consists in our view in formalizing precisely the above qualitative definition, thus allowing for sound quantitative developments. A prominent feature of this theory consists in first recognizing that causation corresponds to

rather different situations and that three distinct facets of causality should be distinguished: (i) necessary causation, where the occurrence of E requires that of C but may also require other factors; (ii) sufficient causation, where the occurrence of C drives that of E but may not be required for E to occur; (iii) necessary and sufficient causation, where (i) and (ii) both hold. The fundamental distinction between these three facets can be visualized by using the simple illustration shown in Figure 1.

While the counterfactual definition as well as the three facets of causality described above may be understood at first in a fully deterministic sense, perhaps the main strength of Pearl's formalization is to propose an extension of these definitions under a probabilistic setting. The probabilities of causation are thereby defined as follow:

$$\mathrm{PS}(C \to E) = \mathbb{P}(E \mid do(C), \overline{C}, \overline{E})\,, \tag{3a}$$

$$\mathrm{PN}(C \to E) = \mathbb{P}(\overline{E} \mid do(\overline{C}), C, E)\,, \tag{3b}$$

$$\mathrm{PNS}(C \to E) = \mathbb{P}(E \mid do(C), \overline{E} \mid do(\overline{C}))\,. \tag{3c}$$

where $\overline{C}$ and $\overline{E}$ are the complementaries of C and E, and where the notation $do(.)$ means that an *intervention* is applied to the system under causal investigation. For instance PS, the *probability of sufficient causation*, reads from the above: the probability that E occurs when C is interventionally forced to occur, conditional on the fact that neither C nor E were occurring in the first place. Conversely PN, the *probability of necessary causation*, is defined as the probability that E would not occur when C is interventionally forced to not occur, conditional on the fact that both C and E were occurring in the first place. While we omit here the formal definition of the intervention $do(.)$ for brevity, the latter can be understood merely as experimentation: applying these definitions thus requires the ability to experiment. Real experimentation, whether *in situ* or *in vivo*, is often accessible in many fields but it is not in climate research for obvious reasons. In this case, one can thus only rely on numerical *in silico* experimentation: the implications of this constraint are discussed further.

While the probabilities of causation are not easily computable in general, their expression fortunately becomes quite simple under assumptions that are reasonable in the case of external forcings (i.e. exogeneity and monotonicity):

$$\mathrm{PN}(C \to E) = \max\left(1 - \overline{p}/p, 0\right), \tag{4a}$$

$$\mathrm{PS}(C \to E) = \max\left(1 - (1-p)/(1-\overline{p}), 0\right), \tag{4b}$$

$$\mathrm{PNS}(C \to E) = \max\left(p - \overline{p}, 0\right). \tag{4c}$$

where $p = \mathbb{P}(E \mid do(C))$ is the so-called *factual* probability of the event E in the real world where C did occur and $\overline{p} = \mathbb{P}(E \mid do(\overline{C}))$ is its *counterfactual* probability in the hypothetic world as it is would have been had C not occurred. One may easily verify that Equation (4) holds in the three examples of Figure 1 by assuming that the switches are probabilistic and exogenous. In any case, under such circumstances, the causal attribution problem can thus be narrowed down to computing an estimate of the probabilities $\overline{p}$ and p. The latter only requires two experiments: a factual experiment $do(C)$ and a counterfactual one $do(\overline{C})$. Equation (3) then yields PN, PS and PNS from which a causal statement can be formulated.

Each three probability PS, PN and PNS have different implications depending on the context. For instance, two perspectives can be considered: (i) the *ex post* perspective of the plaintiff or the judge who asks "does C bear the responsibility of the event E that did occur?"; and (ii) the *ex ante* perspective of the planner or the policymaker who instead asks "what should be done w.r.t. C to prevent future occurrence of E?". It is PN that is typically more relevant to context (i) involving legal responsibility, whereas PS has more relevance for context (ii) involving policy elaboration. Both these perspectives could be relevant in the context of climate change, and it thus makes sense to trade them off. Note that PS and PN can be articulated with the conventional definition recalled in introduction. Indeed, the "*demonstration that the change is consistent with* (...)" implicitly corresponds to the idea of sufficient causation, whereas "(...) *is not consistent with* (...)" corresponds to that of necessary causation. The conventional definition therefore implicitly requires a high PS and

a high PN to attribute a change to a given cause.

PNS may be precisely viewed as a probability which combines necessity and sufficiency. It does so in a conservative way since we have by construction that $\mathrm{PNS} \leq \min(\mathrm{PN}, \mathrm{PS})$. In particular, this means that a low PNS does not imply the absence of a causal relationship because either a high PN or a high PS may still prevail even when PNS is low. On the other hand, it presents the advantage that any statement derived from PNS asserting the existence of a causal link, holds both in terms of necessity and sufficiency. This property is thus prone to simplify causal communication, in particular towards the general public, since the distinction no longer needs to be explained. Therefore, establishing a high PNS may be considered as a suitable goal to evidence the existence of a causal relationship in a simple and straightforward way. In particular, the limiting case $\mathrm{PNS} = 1$ corresponds to the fully deterministic, systematic and single-caused situation in Figure 1c — i.e. undeniably the most stringent way in which one may understand causality.

3. Probabilities of causation of climate change

We now return to the question of interest: for a given forcing f and an observed evolution of the climate system y, can y be attributed to f? More precisely, what is the probability $\mathbb{P}(f \to y)$ that f has caused y? We propose to tackle this problem by applying the causal counterfactual theory to the context of climate change, and more specifically, by using the three probabilities of causation PN, PS and PNS recalled above. This Section shows that it can be done to a large extent similarly to the approach of Hannart et al. (2015a) for weather event attribution. In particular, as in weather event attribution, the crucial question to be answered as a starting point consists in narrowing down the definitions of the cause event C and of the effect event E associated to the question at stake — where the word "event" is used here in its general mathematical sense of "subset".

a. *Counterfactual setting*

For the cause event C, a straightforward answer is possible: we can follow the exact same approach as in weather attribution by defining C as "presence of forcing f" (i.e. the factual world that occurred) and $\overline{C}$ as "absence of forcing f" (i.e. the counterfactual world that would have occurred in the absence of f). Indeed, forcing f can be switched on and off in numerical simulations of the climate evolution over the industrial period, as in the examples of Fig. 1 and as in standard weather attribution studies. Incidentally, the sample space Ω consists in the set of all possible climate trajectories in the presence and absence of f, including the observed one y. In other words, all forcings other than f are held constant at their observed values as they are not concerned by the causal question.

In practice, the factual runs naturally always correspond to the HIST experiment. The counterfactual runs are obtained from the same setting as HIST but switching off the forcing of interest, and thus correspond to the NAT experiment if f consists of the anthropogenic forcing (i.e. $f = \text{ANT}$), i.e. $\Omega = \{\text{HIST runs}; \text{NAT runs}\}$.

These definitions of C and Ω have an important implication w.r.t. the design of numerical experiments in climate change attribution: the latter are required to be counterfactual (i.e. all forcings except f), in agreement with the design prevailing in weather event attribution, but in contrast with the design prevailing in trend attribution (forcing f only). We elaborate further on this remark in Section 6.

b. *Balancing necessity and sufficiency*

To define the effect event E, we propose to follow the same approach as in weather event attribution, where E is usually defined based on an *ad hoc* climatic index Z exceeding a threshold u:

$$E = \{Z \geq u\} \tag{5}$$

Thus, defining E implies choosing an appropriate variable Z and threshold u that reflect the focus of the question while keeping in mind the implications of the balance between the probabilities of necessary and sufficient causation. We now illustrate this issue and lay out some proposals to address it.

Consider the question "*Have anthropogenic CO_2 emissions caused global warming?*". Following the above, the event "*global warming*" may be loosely defined as a positive trend on global Earth surface temperature, i.e. $E = \{Z \geq 0\}$ where Z is the global surface temperature trend coefficient and the threshold u is zero. In that case, E nearly always occurs in the factual world ($p \simeq 1$) but it is also frequent in the counterfactual one ($\overline{p}$ medium) thus the emphasis is mostly on PS, i.e. on sufficient causation, while PN and PNS will have moderate values. But if global warming is more restrictively defined as a warming trend comparable to or greater than the observed trend, i.e. $E = \{Z \geq z\}$ where $u = z$ is the observed trend, then the event becomes nearly impossible in the counterfactual world ($\overline{p} \simeq 0$) but remains frequent in the factual one ($\overline{p}$ medium) thus the emphasis is on PN, i.e. on necessary causation, while the values of PS and PNS will this time be low. Therefore, the above two extreme definitions both yield a low PNS. But under a more balanced definition of *global warming* as a trend exceeding an intermediate value $u^* \in [0, z]$, then the event nearly always occurs in the factual in the factual world ($p \simeq 1$) and yet remains very rare in the counterfactual one ($\overline{p} \simeq 0$). Hence PNS is then high: both necessary and sufficient causation prevail. We propose to take advantage of this optimal value to define the event "*global warming*" as the global trend index Z exceeding the optimal threshold u^* such that the amount of causal evidence, in a PNS sense, is maximized:

$$u^* = \text{argmax}_{u<z} \, \text{PNS}(C \rightarrow \{Z \geq u\}) \tag{6}$$

where the condition $u < z$ insures that the event has actually occurred. When used on real data (see Section 6), this approach yields a high value of PNS $= 0.95$ for the above question (Figure 2b).

Let us now consider the question "*Have anthropogenic CO_2 emissions caused the Argen-*

tinian heatwave of December 2013?" (Hannart et al. 2015b). Here, the event can be defined as $E = \{Z \geq u\}$ where Z is surface temperature anomaly averaged over an ad-hoc space-time window. Like in the previous case, the causal evidence agains shifts from necessary and not sufficient when u is equal to the observed value of the index $z = 24.5°\mathrm{C}$ (unusual event in both worlds but much more so in the counterfactual one) to sufficient and not necessary when u is small (usual event in both worlds but much more so in the factual one). Like in the previous case, a possible approach here would be to balance both quantities by maximizing PNS in u as in Equation (6). However, this would lead here to a substantially lower threshold which no longer reflects the rare and extreme nature of the event "heatwave" under scrutiny. Furthermore, this would yield a well-balanced, but pretty low level of causal evidence (PNS = 0.35). Thus maximizing PNS is not relevant here. Instead, maximizing PN, even if that is at the expense of PS, is arguably more relevant here since we are dealing with extreme events that are rare in both worlds, thereby inherently limiting the evidence of sufficient causation. This maximization corresponds to $u^* = \mathrm{argmax}_{u<z}\, \mathrm{PN}(C \to \{Z \geq u\})$ which often yields the highest observed threshold $u = z$. Therefore, PN (i.e. the FAR) is an appropriate metric for the attribution of extreme weather events, and a high threshold u matching with the observed value z should be used in order to maximize it. In contrast with weather events, long term changes are prone to be associated with much powerful causal evidence that simultaneously involves necessary and sufficient causation, and may yield high values for PN, PS and PNS. PNS is thus an appropriate summary metric to consider for the attribution of climate changes, in agreement with D&A guidance principles (Hegerl et al. 2010). An optimal intermediate threshold can be chosen by maximizing it.

c. Building an optimal index

In the above example where "*global warming*" is the focus of the question, the variable of interest Z to define the event can be considered as implicitly stated in the question, insofar as the term "*global warming*" implicitly refers to an increasing trend on global temperature.

However, in the context of climate change attribution, we often investigate the cause of "an observed change y" with no precise characterization on the nature of the change thought to be relevant, and where y may be a large dimensional space-time vector. Thus the definition of the index Z in this case is more ambiguous.

We argue that in such a case, the physical characteristics of y which are implicitly considered relevant to the causal question are precisely those which best enhance the existence of a causal relationship in a PNS sense. This indeed corresponds to the idea of "fingerprinting" used thus far in climate change attribution studies (as well as in criminal investigations, hence the name): we seek a fingerprint, i.e. a distinctive characteristic of y which would never appear in the absence of forcing f (i.e. $\overline{p} \simeq 0$) but systematically does in its presence (i.e. $p \simeq 1$). If this characteristic shows up in observations, then the causal evidence is conclusive. A fingerprint may thus be thought of as a characteristic which maximizes the gap between p and $\overline{p}$ and thereby maximizes PNS, by definition.

As an illustration, Marvel and Bonfils (2013) focus on the attribution of changes in precipitation, and subsequently address the question "*Have anthropogenic forcing caused the observed evolution of precipitation at a global level?*". Arguably, this study illustrates our point in the sense that it addresses the question by defining a *fingerprint* index Z which aims precisely at reflecting the features of the change in precipitation that are thought to materialize frequently (if not systematically) in the factual world and yet are expected to be rare (if not impossible) in the counterfactual one, based on physical considerations. In practice, the index Z defined by the authors consists of a non-dimensional scalar summarizing the main spatial and physical features of precipitation evolution w.r.t. dynamics and thermodynamics. The factual and counterfactual PDFs of Z are then derived from the HIST and NAT runs respectively (Fig. 3c). From these PDFs, one can easily obtain an optimal threshold u^* for the precipitation index Z by applying Equation (6). This yields PNS = 0.43, i.e. anthropogenic forcings *have about as likely as not caused the observed evolution of precipitation.*

A qualitative approach driven by physical considerations, such as the one of Marvel and Bonfils (2013), is perfectly possible to define a fingerprint index Z that aims at maximizing PNS. However, a quantitative approach can also help in order to define Z optimally, and to identify the features of y that best discriminate between the factual and counterfactual worlds. Indeed, the qualitative, physical elicitation of Z may be difficult when the joint evolution of the variables at stake is complex or not well-understood a priori. Furthermore, one may also wish to combine lines of evidence by treating several different variables at the same time in y (i.e. precipitation and temperature, Yan et al. (2016)). Let us introduce the notation $Z = \phi(Y)$ where Y is the space-time vectorial random variable of size n which observed realization is y, and ϕ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}$. Extending Equation (6) to the simultaneous determination of the optimal mapping ϕ^* and optimal threshold u^*, we propose the following maximization:

$$(u^*, \phi^*) = \text{argmax}_{u<\phi(y),\, \phi\in\Phi}\ \text{PNS}(C \to \{\phi(Y) \geq u\}) \tag{7}$$

The event $E^* = \{\phi^*(Y) \geq u^*\}$ defined above in Equation (7) may thus be referred to as the *optimal fingerprint* w.r.t. forcing f. The maximization performed in Equation (7) also suggests that our approach shares some similarity with the method of Yan et al. (2016), insofar as the variables of interest are in both cases selected mathematically by maximizing a criterion which is relevant for attribution (i.e. potential detectability in Yan et al. (2016), PNS in the present article), rather than by following qualitative, physics- or impact-oriented, considerations.

4. Implementation under the standard framework

We now turn to the practical aspects of implementing the approach described in Section 3 above, based on the observations y and on climate model experiments.

a. *Generalities*

The maximization of Equation (7) requires the possibility to evaluate the probabilities of occurrence p and $\overline{p}$, in the factual and counterfactual world, of the event $\{\phi(Y) \geq u\}$, for any ϕ and u. For this purpose, it is convenient to derive beforehand the factual and counterfactual PDFs of the random variable Y, denoted $[\,Y \mid f\,]$ and $\left[\,Y \mid \overline{f}\,\right]$ respectively. Assuming their two first moments are finite, we introduce:

$$\begin{aligned} \mathbb{E}\,(Y \mid f) &= \mu\,, \qquad & \mathrm{V}\,(Y \mid f) &= \mathbf{\Sigma} \\ \mathbb{E}\left(Y \mid \overline{f}\right) &= \overline{\mu}\,, \qquad & \mathrm{V}\,(Y \mid f) &= \overline{\mathbf{\Sigma}} \end{aligned} \tag{8}$$

The means μ and $\overline{\mu}$ represent the expected response in the factual and counterfactual worlds; it is intuitive that their difference $\mu - \overline{\mu}$ will be key to the analysis. The covariances $\mathbf{\Sigma}$ and $\overline{\mathbf{\Sigma}}$ represent all the uncertainties at stake, they must be carefully determined based on additional assumptions. To avoid repetition in presenting these assumptions, we will detail them for the factual world only, but they will be applied identically in both worlds.

As recalled above, *in situ* experimentation on the climate system is not accessible, thus we are left with *in silico* experimentation as the only option. While the increasing realism of climate system models renders such an *in silico* approach plausible, it is clear that modeling errors associated to their numerical and physical imperfections should be taken into account into $\mathbf{\Sigma}$. In addition, sampling uncertainty and observational uncertainty, which are commonplace sources of uncertainty in dealing with experimental results in an *in situ* context as well, should also be taken into account. Finally, internal climate variability also needs to be factored. The latter four sources of uncertainty can be represented by decomposing $\mathbf{\Sigma}$ into a sum of four terms:

$$\mathbf{\Sigma} = \mathbf{C} + \mathbf{Q} + \mathbf{R} + \mathbf{S} \tag{9}$$

where the component $\mathbf{C}$ represents climate internal variability; $\mathbf{Q}$ represents model uncertainty; $\mathbf{R}$ represents observational uncertainty; and $\mathbf{S}$ represents sampling uncertainty. Assumptions regarding the latter four sources of uncertainty are also key in the conventional

Gaussian linear regression framework recalled in Section 1. We therefore propose to take advantage of some assumptions, data and procedures that have been previously introduced under the conventional framework, and adapt them to specify μ, $\mathbf{C}$, $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{S}$. The statistical model specification below can thus be viewed as an extension of developments under the conventional framework, in particular those exposed in Hannart (2016). The various parameters and data involved, as well as their conditioning, are summarized in the direct acyclic graph of Figure 3.

b. Model description

The conventional linear regression formulation recalled in Equation (1) implies that the random variable Y is Gaussian with mean $\boldsymbol{x}\beta$ and covariance $\mathbf{C}+\mathbf{R}$:

$$[\, Y \mid \beta, \boldsymbol{x}, \mathbf{C}, \mathbf{R} \,] = \mathcal{N}(\boldsymbol{x}\beta, \mathbf{C}+\mathbf{R}) \tag{10}$$

In the conventional framework, climate models are assumed to correctly represent the response patterns $\boldsymbol{x}$ but to err on their amplitude. Recognizing that the scaling factors β thereby aim at representing the error associated to models, we prefer to treat β as a random variable instead of a fixed parameter to be estimated. The latter factors are also assumed to be Gaussian:

$$[\, \beta \mid \omega \,] = \mathcal{N}(e, \omega^2 \mathbf{I}) \tag{11}$$

where we assume that the expected value of β is $e = (1, ..., 1)'$, and ω is a scalar parameter which will be determined further in this Section. Combining Equations (10) and (11), it comes:

$$[\, Y \mid \mu, \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \omega \,] = \mathcal{N}(\mu, \mathbf{C}+\mathbf{R}+\omega^2 \boldsymbol{x}\boldsymbol{x}') \tag{12}$$

where $\mu = \boldsymbol{x}e = \sum_{i=1}^{p} x_i$. Equation (12) thus shows that the covariance $\mathbf{Q}$ associated to model error can be represented by the component $\omega^2 \boldsymbol{x}\boldsymbol{x}'$, which results from the random scaling of the individual responses $x_1, x_2, ..., x_p$. Furthermore, the expected value of Y, denoted μ, is equal to the sum of the latter individual responses. Under the additivity assumption

prevailing in the conventional framework, μ thus corresponds to the model response under the scenario where the p forcings are present. Hence, μ can be obtained by estimating directly the latter combined response as opposed to estimating the individual responses one by one and summing them up. Such a direct estimation of μ is indeed advantageous from a sampling error standpoint, as will be made clear immediately below.

The PDF of Y in Equation (12) involves three quantities μ, $\boldsymbol{x}$ and $\mathbf{C}$ that needs to be estimated from a finite ensemble of model runs denoted $\mathbf{E}$, which of course introduces sampling uncertainty. Assuming independence among runs, it is straightforward to show that:

$$[\,\mu \mid \mathbf{C}, \mathbf{E}\,] = \mathcal{N}(\widehat{\mu}, \tfrac{1}{r}\mathbf{C})\,, \qquad [\,x_i \mid \mathbf{C}, \mathbf{E}\,] \sim \mathcal{N}(\widehat{x}_i, \tfrac{1}{r_i}\mathbf{C}) \text{ for } i = 1, ..., p \tag{13}$$

where $\widehat{x}_i$ is the ensemble average for the individual response i; $\widehat{\mu}$ is the ensemble average for the combined response; r_i is the number of runs available for the individual response to forcing i; r is the number of combined forcings runs. Combining Equations (12) and (13), and after some algebra, it comes:

$$[\,Y \mid \mathbf{C}, \mathbf{R}, \mathbf{E}, \omega\,] = \mathcal{N}(\widehat{\mu}, \mathbf{C} + \mathbf{R} + \omega^2 \widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}' + \lambda \mathbf{C}) \tag{14}$$

with $\lambda = 1/r + \omega^2 \sum_i 1/r_i$, and where the sampling uncertainty $\mathbf{S}$ on the responses μ and $\boldsymbol{x}$ thus corresponds to the term $\lambda\mathbf{C}$. On the other hand, the internal variability component $\mathbf{C}$ also has to be estimated from the r_0 preindustrial control runs, which introduces additional sampling uncertainty. The sampling uncertainty on $\mathbf{C}$ can be treated by following the approach of Hannart (2016), with an Inverse Wishart PDF:

$$[\,\mathbf{C} \mid \mathbf{E}\,] = \mathcal{IW}(\widehat{\mathbf{C}}, \widehat{\nu}) \tag{15}$$

where the estimated covariance $\widehat{\mathbf{C}}$ consists of a so-called shrinkage estimator:

$$\widehat{\mathbf{C}} = \widehat{a}\widehat{\boldsymbol{\Delta}} + (1 - \widehat{a})\widehat{\boldsymbol{\Omega}} \tag{16}$$

where $\widehat{\boldsymbol{\Omega}}$ is the empirical covariance of the control ensemble; $\boldsymbol{\Delta}$ is a shrinkage target matrix taken here to be equal to $\mathrm{diag}(\widehat{\boldsymbol{\Omega}})$ i.e. $\widehat{\boldsymbol{\Delta}}_{ii} = \widehat{\boldsymbol{\Omega}}_{ii}$ and $\widehat{\boldsymbol{\Delta}}_{ij} = 0$ for $i \neq j$; the shrinkage

intensity $\widehat{a}$ is obtained from the marginal likelihood maximization described in Hannart et al. (2014); and $\widehat{\nu} = 2 + r_0/(1-\widehat{a})$.

Combining Equations (14) and (15), and after some algebra and an approximation, it comes:

$$[\, Y \mid \mathbf{E}, \omega, \sigma \,] = \mathcal{S}t(\widehat{\mu}, \sigma^2\mathbf{I} + \omega^2\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}' + (1+\lambda)\widehat{\mathbf{C}}, \widehat{\nu}) \tag{17}$$

where we adopted the simplified parametric form $\mathbf{R} = \sigma^2\mathbf{I}$ for the covariance of observational error, and where $\mathcal{S}t(\mu, \boldsymbol{\Sigma}, \nu)$ is the multivariate t distribution with mean μ, variance $\boldsymbol{\Sigma}$ and ν degrees of freedom. Equation (17) implies that taking into account the sampling uncertainty on $\mathbf{C}$ does not affect the total variance of Y. Instead, it only affects the shape of the PDF of Y, which has thicker tails than the Gaussian distribution. With these parameterizations, our model for Y is thus a parametric Student t model with two parameters (σ, ω).

After computing the estimators $\widehat{\mu}$, $\widehat{\boldsymbol{x}}$, $\widehat{\mathbf{C}}$ and $\widehat{\nu}$ using the ensemble of model experiments as described above, the parameters (σ, ω) are estimated by fitting the above model to the observation y based on likelihood maximization. The log-likelihood of the model has the following expression:

$$\begin{aligned} \ell(\sigma, \omega; y) = \; & -\tfrac{1}{2}\log|(1+\lambda)\widehat{\mathbf{C}} + \sigma^2\mathbf{I} + \omega^2\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}'| \\ & -\tfrac{1}{2}(\widehat{\nu}+n)\log\left(1 + \tfrac{1}{\widehat{\nu}-2}(y-\widehat{\mu})'\left((1+\lambda)\widehat{\mathbf{C}} + \sigma^2\mathbf{I} + \omega^2\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}'\right)^{-1}(y-\widehat{\mu})\right) \end{aligned} \tag{18}$$

The estimators $\widehat{\sigma}$ and $\widehat{\omega}$ are then obtained numerically using a standard maximization algorithm (e.g. gradient descent). With $\widehat{\mu}$ being obtained from factual runs (i.e. HIST runs) and $\widehat{\boldsymbol{x}}$ containing all the forcings including f, this procedure thus yields the PDF of Y in the factual world:

$$\begin{aligned} & [\, Y \mid f \,] = \mathcal{S}t(\widehat{\mu}, \widehat{\boldsymbol{\Sigma}}, \widehat{\nu}) \\ & \widehat{\boldsymbol{\Sigma}} = (1+\widehat{\lambda})\widehat{\mathbf{C}} + \widehat{\sigma}^2\mathbf{I} + \widehat{\omega}^2\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}' \end{aligned} \tag{19}$$

Next, to obtain $\left[\, Y \mid \overline{f} \,\right]$, one simply needs to change the mean $\widehat{\mu}$ to $\widehat{\overline{\mu}}$ obtained as the ensemble average for the counterfactual experiment "all forcings except f". Some changes also need to be made regarding the covariance. Indeed, since forcing f is absent in the counterfactual world, the model error covariance component $\widehat{\omega}^2\widehat{x}_f\widehat{x}_f'$, corresponding to the random

scaling of the response $\widehat{x}_f$ to forcing f, must be taken out of the covariance. Furthermore, if the number of counterfactual runs $\overline{r}$ differ from the number of factual runs r, the sampling uncertainty $\widehat{\mathbf{C}}/r$ associated to estimating μ also has to be modified. The PDF of Y in the counterfactual world can thus be written:

$$\begin{aligned} \left[\, Y \mid \overline{f} \,\right] &= \mathcal{S}t(\widehat{\overline{\mu}}, \widehat{\overline{\mathbf{\Sigma}}}, \widehat{\nu}) \\ \widehat{\overline{\mathbf{\Sigma}}} &= \widehat{\mathbf{\Sigma}} - \widehat{\omega}^2 \widehat{x}_f \widehat{x}_f' + (\tfrac{1}{\overline{r}} - \tfrac{1}{r})\widehat{\mathbf{C}} \end{aligned} \tag{20}$$

As noted above, when f is anthropogenic forcing, the counterfactual experiment NAT is usually available in CMIP runs, allowing for a straightforward derivation of $\widehat{\overline{\mu}}$. But for other forcings, by the design of CMIP experiments, counterfactual runs are usually not available. A possible workaround then consists in applying the additivity assumption to approximate $\widehat{\overline{\mu}}$ with $\widehat{\mu} - \widehat{x}_f$. However in that case, the sampling uncertainty term $\widehat{\mathbf{C}}/r_f$ corresponding to the estimation of $\widehat{x}_f$ must be added to the covariance $\widehat{\overline{\mathbf{\Sigma}}}$.

c. *Derivation of the probabilities of causation*

With the two PDFs of Y in hand, an approximated solution to the maximization of Equation (7) can be conveniently obtained by linearizing ϕ, yielding a closed mathematical expression for the optimal index $\phi^*(Y)$:

$$\phi^*(Y) = (\widehat{\mu} - \widehat{\overline{\mu}})' \widehat{\mathbf{\Sigma}}^{-1} Y \tag{21}$$

Details of the approximations made and of the mathematical derivation of Equation (21) are given in Appendix. The optimal index $Z^* = \phi^*(Y)$ can thus be interpreted as the projection of Y onto the vector $\widehat{\mathbf{\Sigma}}^{-1}(\widehat{\mu} - \widehat{\overline{\mu}})$ which will be denoted ϕ^* hereinafter, i.e. $\phi^*(Y) \equiv \phi^{*\prime} Y$.

To obtain PNS, we then need to derive the factual and counterfactual CDFs of $Z = \phi^*(Y)$, denoted G and $\overline{G}$ respectively. Since no closed form expression of these CDFs is available, we simulate realizations thereof. Drawing two samples of N random realizations of Y from the Student t distributions $[\, Y \mid f \,]$ and $\left[\, Y \mid \overline{f} \,\right]$ is straightforward, by treating the Student t as a compound Gaussian–Chi-squared distribution. Samples of Z are then immediately

obtained by projecting onto ϕ^* and the desired CDFs can be estimated using the standard kernel estimator, yielding $\widehat{G}(u)$ and $\widehat{\overline{G}}(u)$ for all $u \in \mathbb{R}$. Finally, PNS* follows as:

$$\text{PNS}^* = \widehat{\overline{G}}(u^*) - \widehat{G}(u^*) \tag{22}$$

and:

$$\text{PN}^* = 1 - \frac{1 - \widehat{\overline{G}}(u^*)}{1 - \widehat{G}(u^*)}, \quad \text{PS}^* = 1 - \frac{\widehat{G}(u^*)}{\widehat{\overline{G}}(u^*)} \tag{23}$$

where $u^* = \text{argmax}_{u<z} \{\widehat{\overline{G}}(u) - \widehat{G}(u)\}$.

d. Reducing computational cost

When the dimension of y is large, the above described procedure can become prohibitively costly if applied straightforwardly, due to the necessity to derive the inverse and determinant of $\widehat{\mathbf{\Sigma}}$ at several steps of the procedure. However, the computational cost of these operations can be drastically reduced. Applying the Sherman-Morrison-Woodbury formula (and omitting the notation $\widehat{.}$ for more clarity), we have:

$$\mathbf{\Sigma}^{-1} = \mathbf{A}^{-1} - \omega^2 \mathbf{A}^{-1} \boldsymbol{x} (\mathbf{I} + \omega^2 \boldsymbol{x}' \mathbf{A}^{-1} \boldsymbol{x})^{-1} \boldsymbol{x}' \mathbf{A}^{-1} \tag{24}$$

where $\mathbf{A} = (1+\lambda)\mathbf{C} + \sigma^2 \mathbf{I}$ can be inverted using the same formula:

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} - \tfrac{1}{r_0}(1+\lambda)(1-a)\mathbf{B}^{-1}\boldsymbol{x}_0(\mathbf{I} + \tfrac{1}{r_0}(1+\lambda)(1-a)\boldsymbol{x}_0'\mathbf{B}^{-1}\boldsymbol{x}_0)^{-1}\boldsymbol{x}_0'\mathbf{B}^{-1} \tag{25}$$

where $\mathbf{B} = (1+\lambda)a\mathbf{\Delta} + \sigma^2\mathbf{I}$. Likewise, we apply the Sylvester formula twice to compute the determinant of $\mathbf{\Sigma}$:

$$\begin{aligned} |\mathbf{\Sigma}| &= |\mathbf{A}| \,.\, |\mathbf{I} + \omega^2 \boldsymbol{x}' \mathbf{A}^{-1} \boldsymbol{x}| \\ &= |\mathbf{B}| \,.\, |\mathbf{I} + \tfrac{1}{r_0}(1+\lambda)(1-a)\boldsymbol{x}_0'\mathbf{B}^{-1}\boldsymbol{x}_0| \,.\, |\mathbf{I} + \omega^2 \boldsymbol{x}' \mathbf{A}^{-1} \boldsymbol{x}| \end{aligned} \tag{26}$$

Independently of n, the matrices $\mathbf{I}+\omega^2\boldsymbol{x}'\mathbf{A}^{-1}\boldsymbol{x}$ is of size p, $\mathbf{I}+\frac{1}{r_0}(1+\lambda)(1-a)\boldsymbol{x}_0'\mathbf{B}^{-1}\boldsymbol{x}_0$ is of size r_0, and $\mathbf{B}$ is diagonal. Obtaining their inverse and determinant is therefore computationally cheap. Hence, the inverse and determinant of $\mathbf{\Sigma}$ can be obtained at a low computational cost by applying first Equation (25) to determine $\mathbf{A}^{-1}$ and second Equations (24) and (26).

5. Illustration on temperature change

Our methodological proposal is applied to the observed evolution of Earth's surface temperature during the 20th century, with the focus being restrictively on the attribution to anthropogenic forcings. More precisely, y consists of a spatial-temporal vector of size $n = 54$ which contains the observed surface temperatures averaged over 54 time-space windows. These windows are defined at a coarse resolution: Earth's surface is divided into 6 regions of similar size (3 in each hemisphere) while the period 1910-2000 is divided into 9 decades. The decade 1900-1910 is used as a reference period, and all values are converted to anomalies w.r.t. the first decade. The HadCRUT4 observational dataset (Morice et al. 2012) was used to obtain y. With respect to climate simulations, the runs of the IPSL-CM5A-LR model (Dufresne et al. 2012) for the NAT, ANT, HIST and PIcontrol experiments were used (see Appendix C for details) and converted to the same format as y after adequate space-time averaging.

Following the procedure described in Section 4, we successively derived the estimated factual response $\widehat{\mu}$ using the r HIST runs; the estimated counterfactual response $\widehat{\overline{\mu}}$ using the $\overline{r}$ NAT runs; the estimated individual responses x_1 and x_2 using the r_1 NAT runs and r_2 ANT runs respectively ($p = 2$ and $\boldsymbol{x} = (x_1, x_2)$); the estimated covariance $\widehat{\mathbf{C}}$ from the r_0 PIcontrol runs. Then, we derived $\widehat{\sigma}$ and $\widehat{\omega}$ by likelihood maximization, to obtain $\widehat{\mathbf{\Sigma}}$ and $\widehat{\overline{\mathbf{\Sigma}}}$.

An assessment of the relative importance of the four components of uncertainty was obtained by deriving the trace of each component (i.e. the sum of diagonal terms) normalized to the trace of the complete covariance. The results for the factual and counterfactual covariances are plotted in Figure 3a, showing that climate variability is the dominant contribution, followed by model uncertainty (in the factual world), observational uncertainty and sampling uncertainty. The split between model and observational uncertainty is to some extent arbitrary as we have no objective way to separate them based only on y, i.e. the model could be equivalently formulated as $\mathbf{Q} = \omega^2 \boldsymbol{x}\boldsymbol{x}' + (1-\alpha)\sigma^2\mathbf{I}$ and $\mathbf{R} = \alpha\sigma^2\mathbf{I}$. An objective separation would require an ensemble representing observational uncertainty, allowing for a preliminary

estimation of $\mathbf{R}$.

The optimal vector ϕ^*, designed to capture the space-time patterns that best discriminate the HIST evolution and the NAT one, was then obtained from Equation (21). To identify which features of Y are captured by this optimal mapping, the coefficients $(\phi_1^*, ..., \phi_n^*)$ were averaged spatially and temporally, and were plotted in Figure 3bc. Firstly, it can be noted that the coefficients' global average $\langle \phi^* \rangle = \sum_{i=1}^{n} \phi_i^*$ is large and positive: a major discriminant feature is merely global mean temperature, as expected. Secondly, the coefficients also exhibit substantial variation around their average $\langle \phi^* \rangle$ in both space and time. Spatial variations of ϕ^* unsurprisingly suggest that, beyond global mean temperature, other discriminant features include the warming contrast prevailing between the two hemispheres and/or between low and high latitudes (the low resolution prevent from a clear separation), as well as between ocean and land (Fig. 3b). Temporal variations of ϕ^* suggest that discriminant features includes the linear trend increase as expected, but also higher order temporal variations (Fig. 3c).

The PDFs of the optimal index $Z = \phi^{*'} Y$ were derived, and are plotted in Figure 4, together with PNS as a function of the threshold u. The maximum of PNS determines the desired probability of causation:

$$\mathbb{P}(\text{ANT} \rightarrow y) = 0.9999 \tag{27}$$

In IPCC terminology, this would mean that anthropogenic forcings *have virtually certainly caused the observed evolution of temperature*, according to our approach. More precisely, the probability that the observed evolution of temperature is not caused by anthropogenic forcings is one in then thousands (1:10,000) instead of one in twenty (1:20). Therefore, the level of causal evidence found here is substantially higher than the level assessed in the IPCC report. This discrepancy will be discussed in Section 6.

Before digging into this discussion, it is interesting to assess the relative importance of the "trivial" causal evidence coming from the global increase in temperature, and of the less obvious causal evidence coming from space-time features captured by ϕ^*. For this purpose,

we merely split ϕ^* into the sum of a global average contribution $\sum_{i=1}^{n}\langle\phi^*\rangle Y_i$ and of the remaining variations $\sum_{i=1}^{n}(\phi_i^* - \langle\phi^*\rangle)Y_i$. The PDFs of the resulting indexes are plotted in Figure 4ab. Their bivariate PDF can also be visualized with the scatterplot of Figure 4c. The following two probabilities of causation are obtained:

$$\begin{aligned} \mathbb{P}(\mathrm{ANT} \to \langle y \rangle) &= 0.9781 \\ \mathbb{P}(\mathrm{ANT} \to y - \langle y \rangle) &= 0.9994 \end{aligned} \tag{28}$$

where $\langle y \rangle$ refer to the globally averaged evolution and $y - \langle y \rangle$ refer to other features of evolution. Therefore, while the globally averaged warming provides alone a substantial level of evidence (i.e. $\mathbb{P}(\mathrm{ANT} \to \langle y \rangle) = 0.9781$), these results suggest that the overwhelmingly high overall evidence (i.e. $\mathbb{P}(\mathrm{ANT} \to y) = 0.9999$) is primarily associated to non-obvious space-time features of the observed temperature change.

6. Discussion

a. Comparison with previous statements

The probabilities of causation obtained by using our proposal appear may depart from the levels of uncertainty asserted by the latest IPCC report, and/or by previous work. For instance, when y corresponds to the evolution of precipitation observed over the entire globe during the satellite era (1979-2012), we have shown in Section 3 that, using the dynamic-thermodynamic index built by Marvel and Bonfils (2013), the associated probability of causation $\mathbb{P}(\mathrm{ANT} \to y)$ is found to be 0.43. This probability is thus significantly lower than the one implied by the claim made in this study that "*the changes in precipitation observed in the satellite era are likely to be anthropogenic in nature*" wherein "*likely*" implicitly means $\mathbb{P}(\mathrm{ANT} \to y) \geq 0.66$.

In contrast with the situation prevailing for precipitation, when y corresponds to the observed evolution of Earth's surface temperature during the 20th century, and in spite of using a very coarse spatial resolution, we found a probability of causation $\mathbb{P}(\mathrm{ANT} \to y) =$

0.9999 which considerably exceeds the 0.95 probability implied by the latest IPCC report. Such a gap deserves to be discussed.

Firstly, the probability of causation defined in our approach is of course sensitive to the assumptions that are made on the various sources of uncertainty, all of which are here built into $\mathbf{\Sigma}$. Naturally, increasing the level of uncertainty, for instance through an inflation factor applied to $\mathbf{\Sigma}$, reduces the probability of causation (Figure 5). In the present illustration, uncertainty needs to inflated by a factor 2.4 to obtain $\mathbb{P}(\text{ANT} \to y) = 0.95$ in agreement with the IPCC statement. Therefore, a speculative explanation for the gap is that experts may be adopting a conservative approach by implicitly inflating uncertainty, although not explicitly, perhaps in an attempt to account for additional sources of uncertainty that are not well known. In the present case, such an inflation should amount to 2.4 to explain the gap. This number is arguably too high to provide a satisfactory standalone explanation, yet overall, such a conservativeness may partly contribute to the discrepancy when it comes to temperature. However, no such conservativeness seems to be at play w.r.t. precipitation. This thus highlights the need for a more explicit and consistent use of conservativeness — if any.

Another possible explanation for the discrepancy is more technical. Many previous attribution studies buttressing the IPCC statement regarding temperature, are based on an inference method for the linear regression model of Equation (1) which is not optimal w.r.t. maximizing causal evidence — despite of it being often referred to as "*optimal fingerprinting*". More precisely, the inference on the scaling factors β and the associated uncertainty quantification, are obtained by projecting the observation y as well as the patterns $\boldsymbol{x}$ onto the leading eigenvectors of the covariance $\mathbf{C}$ associated to climate internal variability. Such a projection choice sharply contrasts with the projection used in our approach, which is indeed performed onto the vector $\phi^* = \mathbf{\Sigma}^{-1}(\mu - \overline{\mu})$. Denoting $(\mathrm{v}_1, ..., \mathrm{v}_n)$ the eigenvectors of $\mathbf{\Sigma}$ and

$(\lambda_1, ..., \lambda_n)$ the corresponding eigenvalues, the expression of ϕ^* can be written:

$$\phi^* = \sum_{k=1}^{n} \frac{\langle \mathrm{v}_k \mid \mu - \overline{\mu} \rangle}{\lambda_k} \,.\, \mathrm{v}_k \tag{29}$$

Equation (29) shows that projecting onto ϕ^* does not emphasize the leading eigenvectors of $\mathbf{\Sigma}$, in contrast to the aforementioned standard projection. Instead, it emphasizes the eigenvectors that simultaneously present a low eigenvalue λ_k and a strong alignment with the contrast between the two worlds $\mu - \overline{\mu}$. As a matter of fact, the ratio $\langle \mathrm{v}_k \mid \mu - \overline{\mu} \rangle / \lambda_k$ can be interpreted as the signal-to-noise ratio associated to the eigenvector v_k, where the eigenvalue λ_k quantifies the magnitude of the noise and $\langle \mathrm{v}_k \mid \mu - \overline{\mu} \rangle$ that of the causal signal. Projecting onto ϕ^* thus maximizes the signal-to-noise ratio. In contrast, since $\mathbf{C}$ is a large contribution to $\mathbf{\Sigma}$ (the dominant one in our illustration), a projection onto the leading eigenvectors of $\mathbf{C}$ naturally tends to amplify the noise, and to partly hide the relevant causal signal $\mu - \overline{\mu}$.

In order to assess whether or not these theoretical remarks hold in practice, we revisited our illustration and quantified the impact on $\mathbb{P}(\mathrm{ANT} \to y)$ of using such a projection onto the leading eigenvectors of $\mathbf{C}$. For this purpose, after computing the projection matrix $\mathbf{P}$ on the ten leading eigenvectors of $\mathbf{C}$, our procedure was applied identically, but this time using the projected vector $\phi^+ = \mathbf{P}\phi^*$. Results are shown in Figure 6, again after splitting the contribution of global mean change and patterns of change. Unsurprisingly, the probability of causation associated to the global mean change remains unmodified at 0.978. However, the probability of causation associated to the space-time features of warming drops from 0.9994 to 0.92. Indeed, the features that most discriminate the two worlds, and are summarized in ϕ^*, do not align well with the leading eigenvectors of $\mathbf{C}$. They are thus incompletely reflected by the projected vector ϕ^+ (Figure 7). Furthermore, the uncertainty of the resulting index $Z^+ = \phi^{+\prime} Y$ is high by construction, as the magnitude of climate variability is maximized when projecting onto its leading modes (Figure 6b). This further contributes to reducing $\mathbb{P}(\mathrm{ANT} \to y)$ to 0.992.

b. *Counterfactual experiments*

Our methodological proposal has an immediate implication w.r.t. the design of standardized CMIP experiments dedicated to D&A: a natural option would be to change the present design "forcing f only" into a counterfactual design "all forcings except f". Indeed, $\mathbb{P}(f \to y)$ is driven by the difference $\mu - \overline{\mu}_f$ between the factual response μ (i.e. historical experiment) and the counterfactual response $\overline{\mu}_f$ (i.e. all forcings except f experiment). Under the assumption that forcings do not interact with one another and that the combined response matches with the sum of the individual responses, the difference $\mu - \overline{\mu}_f$ coincides with the individual response x_f (i.e. forcing f only experiment). While this hypothesis is well established for temperature at large scale (Gillett et al. 2004), it appears to break down for other variables (e.g. precipitation, (Shiogama et al. 2013)) or over particular regions (e.g the Southern extratropics, (Morgenstern et al. 2014)) where forcings appear to significantly interplay. Such a lack of additivity would inevitably damage the results of the causal analysis. It is thus important in our view to better understand the domain of validity of the forcing-additivity assumption and to evaluate the drawbacks of the present "one forcing only" design versus its advantages. Such an analysis does require "forcing f only" experiments, but also "all forcings except f" experiments in order to allow for comparison. This effort would hence justify including in the DAMIP set of experiments an "all forcings except f" experiment — which is presently absent even in the lowest priority tier thereof — at least for the most important forcings such as anthropogenic CO_2.

c. *Benchmarking high probabilities*

Section 5 showed that the proposed approach may sometimes yield probabilities of causation that are very close to one. How can we communicate such low levels of uncertainty? This question arises insofar as the term "virtual certainty" applies as soon as PNS exceeds 0.99 under the current IPCC language. Thus, this terminology would be unfit to express in

words a PNS increase from 0.99 to 0.9999, say — even though such an increase corresponds to a large reduction of uncertainty by a factor one hundred. One option to address this issue is to use instead the uncertainty terminology of theoretical physics, in which a probability is translated into an exceedance level under the Gaussian distribution, measured in numbers of σ from the mean (where σ denotes standard deviation), i.e. $F^{-1}(\text{PNS})\sigma$ with F the CDF of the standard Gaussian distribution. Under such terminology, "virtual certainty" thus corresponds to a level of uncertainty of 2.3σ, while $\mathbb{P}(\text{ANT} \rightarrow y) = 0.9999$ found in Section 5 reaches 3.7σ. It is interesting to note that the level of uncertainty officially requested in theoretical physics to corroborate a discovery as such — e.g. the existence of the Higgs Boson — is 5σ. By such high standards, $\mathbb{P}(\text{ANT} \rightarrow y) = 0.9999$ found above can actually still be considered much too low a probability to corroborate that human influence has indeed been the cause of the observed warming. Therefore, further increasing $\mathbb{P}(\text{ANT} \rightarrow y)$ by building more evidence into the analysis, may still be considered to be a relevant goal.

7. Summary and conclusion

We have introduced an approach for deriving the probability that a forcing has caused a given observed change. The proposed approach is anchored into causal counterfactual theory (Pearl 2000) which has been introduced recently in the context of weather and climate-related events attribution. We argued that these concepts are also relevant, and can be straightforwardly extended to the context of climate change attribution. For this purpose, and in agreement with the principle of *fingerprinting* applied in the conventional D&A framework, a trajectory of change is converted into an event occurrence defined by maximizing the causal evidence associated to the forcing under scrutiny. Other key assumptions used in the conventional D&A framework, in particular those related to numerical models error, can also be adapted conveniently to this approach. Our proposal thus allows to bridge the conventional framework with the standard causal theory, in an attempt to improve the

quantification of causal probabilities. Our illustration suggested that our approach is prone to yield a higher estimate of the probability that anthropogenic forcings have caused the observed temperature change, thus supporting more assertive causal claims.

Acknowledgments.

We gratefully acknowledge helpful comments by Aurélien Ribes and inspiring interactions with Judea Pearl and Michael Ghil. This work was supported by the French Agence Nationale de la Recherche grant DADA (AH, PN), and the grants LEFE-INSU-Multirisk, AMERISKA, A2C2, and Extremoscope (PN). The work of PN was completed during his visit at the IMAGE-NCAR group in Boulder, CO, USA.

APPENDIX A

Derivation of the PDF of Y

To obtain Equation (12) from Equation (10) and (11), we integrate out β:

$$[\, Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R} \,] = \int_\beta [\, Y \mid \beta, \boldsymbol{x}, \mathbf{C}, \mathbf{R} \,] \,.\, [\, \beta \mid \omega \,] \, \mathrm{d}\beta \tag{A1}$$

Given the quadratic dependence to β of the two terms under the integral in the right hand side of Equation (A1), it is clear that the PDF of the left hand side is also Gaussian. Thus, instead of computing the above integral, it is more convenient to derive the mean and variance of this PDF by applying the rule of total expectation and total variance:

$$\begin{aligned}
\mathbb{E}(Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}) &= \mathbb{E}\left(\mathbb{E}(Y \mid \beta, \boldsymbol{x}, \mathbf{C}, \mathbf{R}) \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}\right) = \mathbb{E}\left(\boldsymbol{x}\beta \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}\right) = \boldsymbol{x}\mathbb{E}\left(\beta\right) \\
&= \boldsymbol{x}e \\
\mathrm{V}(Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}) &= \mathrm{V}\left(\mathbb{E}(Y \mid \beta, \boldsymbol{x}, \mathbf{C}, \mathbf{R}) \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}\right) + \mathbb{E}\left(\mathrm{V}(Y \mid \beta, \boldsymbol{x}, \mathbf{C}, \mathbf{R}) \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}\right) \\
&= \mathrm{V}\left(\boldsymbol{x}\beta \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}\right) + \mathbb{E}\left(\mathbf{C} + \mathbf{R} \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}\right) \\
&= \boldsymbol{x}\mathrm{V}(\beta)\boldsymbol{x}' + \mathbf{C} + \mathbf{R} = \omega^2 \boldsymbol{x}\boldsymbol{x}' + \mathbf{C} + \mathbf{R} \\
[\, Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R} \,] &= \mathcal{N}(\boldsymbol{x}e, \mathbf{C} + \mathbf{R} + \omega^2 \boldsymbol{x}\boldsymbol{x}')
\end{aligned} \tag{A2}$$

Next, in order to account for the sampling uncertainty on the estimation of μ, we apply Bayes theorem to derive the PDF of μ conditional on the ensemble $\mathbf{E}$. Denote $\mu^{(1)}, ..., \mu^{(r)}$ the r simulated responses in $\mathbf{E}$ which are assumed to be i.i.d. according to a Gaussian with mean μ and covariance $\mathbf{C}$. We have:

$$\begin{aligned}
[\, \mu \mid \mathbf{C}, \mathbf{E} \,] &\propto \Pi_{j=1}^{r} \left[\, \mu^{(j)} \mid \mathbf{C} \,\right] \,.\, [\, \mu \,] \\
&\propto \Pi_{j=1}^{r} \, \mathcal{N}(\mu^{(j)} \mid \mu, \mathbf{C}) \\
&= \mathcal{N}(\mu \mid \widehat{\mu}, \tfrac{1}{r}\mathbf{C})
\end{aligned} \tag{A3}$$

where $\widehat{\mu}$ is the empirical mean of the ensemble, and we use the improper prior $[\,\mu\,] \propto 1$. The exact same approach yields $[\,x_i \mid \mathbf{C}, \mathbf{E}\,] \propto \Pi_{j=1}^{r_i} \mathcal{N}(\boldsymbol{x}_i^{(j)} \mid x_i, \mathbf{C}) = \mathcal{N}(x_i \mid \widehat{x}_i, \frac{1}{r_i}\mathbf{C})$.

To integrate out μ, we proceed by following the same reasoning as above for integrating out β. Since the resulting PDF is clearly Gaussian, it suffices to derive its mean and variance:

$$\begin{aligned}
\mathbb{E}(Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) &= \mathbb{E}\left(\mathbb{E}(Y \mid \mu, \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}\right) = \mathbb{E}\left(\mu \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}\right) \\
&= \widehat{\mu} \\
\mathrm{V}(Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) &= \mathrm{V}\left(\mathbb{E}(Y \mid \mu, \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}\right) + \mathbb{E}\left(\mathrm{V}(Y \mid \mu, \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}\right) \\
&= \mathrm{V}\left(\mu \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}\right) + \mathbb{E}\left(\omega^2 \boldsymbol{x}\boldsymbol{x}' + \mathbf{C} + \mathbf{R} \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}\right) \\
&= \tfrac{1}{r}\mathbf{C} + \omega^2 \boldsymbol{x}\boldsymbol{x}' + \mathbf{C} + \mathbf{R}
\end{aligned} \tag{A4}$$

Likewise, to integrate out $\boldsymbol{x}$, we derive the total mean and total variance:

$$\begin{aligned}
\mathbb{E}(Y \mid \mathbf{C}, \mathbf{R}, \mathbf{E}) &= \mathbb{E}\left(\mathbb{E}(Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) \mid \mathbf{C}, \mathbf{R}, \mathbf{E}\right) = \mathbb{E}\left(\widehat{\mu} \mid \mathbf{C}, \mathbf{R}, \mathbf{E}\right) \\
&= \widehat{\mu} \\
\mathrm{V}(Y \mid \mathbf{C}, \mathbf{R}, \mathbf{E}) &= \mathrm{V}\left(\mathbb{E}(Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) \mid \mathbf{C}, \mathbf{R}, \mathbf{E}\right) + \mathbb{E}\left(\mathrm{V}(Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) \mid \mathbf{C}, \mathbf{R}, \mathbf{E}\right) \\
&= \mathbf{0} + (1 + \tfrac{1}{r})\mathbf{C} + \mathbf{R} + \mathbb{E}\left(\omega^2 \boldsymbol{x}\boldsymbol{x}' \mid \mathbf{C}, \mathbf{E}\right) \\
&= (1 + \tfrac{1}{r})\mathbf{C} + \mathbf{R} + \omega^2 \textstyle\sum_i \mathbb{E}\left(x_i\, x_i' \mid \mathbf{C}, \mathbf{E}\right) \\
&= (1 + \tfrac{1}{r})\mathbf{C} + \mathbf{R} + \omega^2 \textstyle\sum_i \mathrm{V}\left(x_i \mid \mathbf{C}, \mathbf{E}\right) + \omega^2 \textstyle\sum_i \mathbb{E}\left(x_i \mid \mathbf{C}, \mathbf{E}\right) \mathbb{E}\left(x_i \mid \mathbf{C}, \mathbf{E}\right)' \\
&= (1 + \tfrac{1}{r})\mathbf{C} + \mathbf{R} + \omega^2 \textstyle\sum_i \tfrac{1}{r_i}\mathbf{C} + \omega^2 \textstyle\sum_i \widehat{x}_i\, \widehat{x}_j' \\
&= (1 + \tfrac{1}{r} + \omega^2 \textstyle\sum_i \tfrac{1}{r_i})\mathbf{C} + \mathbf{R} + \omega^2 \widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}' \\
&= \mathbf{C} + \mathbf{R} + \omega^2 \widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}' + \lambda \mathbf{C}
\end{aligned} \tag{A5}$$

with $\lambda = 1/r + \omega^2 \sum_i 1/r_i$. Note that $[\,Y \mid \mathbf{C}, \mathbf{R}, \mathbf{E}\,]$ is no longer Gaussian after integrating out $\boldsymbol{x}$, because $\boldsymbol{x}$ appears in the covariance of $[\,Y \mid \boldsymbol{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}\,]$. However, for simplicity, we approximate it to be Gaussian.

The sampling uncertainty on the covariance matrix $\mathbf{C}$ is addressed by using an approach described in Hannart et al. (2014) which main ideas are succinctly recalled here. The reader is referred to the publication for details and explicit calculations. In summary, we apply Bayes theorem in order to derive $[\,\mathbf{C} \mid \mathbf{E}\,]$, as for μ and $\boldsymbol{x}$. However, we use this time an

informative conjugate prior on $\mathbf{C}$, as opposed to an improper prior.

$$[\,\mathbf{C} \mid \boldsymbol{\Delta}, a\,] = \mathcal{IW}(\boldsymbol{\Delta}, a) \tag{A6}$$

where $\boldsymbol{\Delta}$ denotes the a priori mean of $\mathbf{C}$ and a is a scalar parameter that drives the a priori variance. Furthermore, the mean and variance parameters $(\boldsymbol{\Delta}, a)$ of this informative prior are estimated from $\mathbf{E}$ by maximizing the marginal likelihood $\ell(a, \boldsymbol{\Delta})$ associated to this Bayesian model.

$$\begin{aligned} \ell(a, \boldsymbol{\Delta}) &= (\tfrac{a\,r_0}{1-a} + n + 1) \log |\tfrac{a}{1-a}\boldsymbol{\Delta}| - (\tfrac{r_0}{1-a} + n + 1) \log |\widehat{\boldsymbol{\Omega}} + \tfrac{a}{1-a}\boldsymbol{\Delta}| \\ &+ 2 \log \left(\Gamma_n\{\tfrac{1}{2}(\tfrac{r_0}{1-a} + n + 1)\} \,/\, \Gamma_n\{\tfrac{1}{2}(\tfrac{a\,r_0}{1-a} + n + 1)\} \right). \end{aligned} \tag{A7}$$

where Γ_n is the $n-$variate Gamma function and $\widehat{\boldsymbol{\Omega}} = \boldsymbol{x}_0 \boldsymbol{x}_0' / r_0$ is the empirical covariance. The estimators $(\widehat{a}, \widehat{\boldsymbol{\Delta}})$ satisfy to:

$$(\widehat{a}, \widehat{\boldsymbol{\Delta}}) = \text{argmax}_{a \in [0,1],\ \boldsymbol{\Delta} \in \mathcal{F}}\ \ell(a, \boldsymbol{\Delta}), \tag{A8}$$

where $\mathcal{F}$ is a set of definite positive matrices chosen to introduce a regularization constraint on the covariance. Here we choose $\mathcal{F} = \{\text{diag}(\delta_1, ..., \delta_n) \mid \delta_1 > 0, ..., \delta_n > 0\}$ the set of definite positive diagonal matrices, and we derive an approximated solution to Equation (A8) with $\widehat{\boldsymbol{\Delta}} = \text{diag}(\widehat{\boldsymbol{\Omega}})$ and $\widehat{a} = \text{argmax}_{a \in [0,1]}\ \ell(a, \widehat{\boldsymbol{\Delta}})$. Because the prior PDF is fitted on the data, this approach can be referred to as "empirical bayesian". The "fitted" prior $[\,\mathbf{C} \mid \widehat{\boldsymbol{\Delta}}, \widehat{a}\,]$ is then updated using the ensemble $\mathbf{E}$, and the obtained posterior has a closed form expression due to conjugacy:

$$\left[\mathbf{C} \mid \mathbf{E}, \widehat{\boldsymbol{\Delta}}, \widehat{a}\right] \ \propto [\,\mathbf{E} \mid \mathbf{C}\,]\,.\mathcal{IW}(\widehat{\boldsymbol{\Delta}}, \widehat{a}) = \mathcal{IW}(\widehat{\mathbf{C}}, \widehat{a}') \tag{A9}$$

where $\widehat{\mathbf{C}} = \widehat{a}\widehat{\boldsymbol{\Delta}} + (1 - \widehat{a})\widehat{\boldsymbol{\Omega}}$ and $\widehat{a}' = 1/(2 - \widehat{a})$. We can then use the above posterior to integrate out $\mathbf{C}$ in the PDF of Y, in order to obtain $[\,Y \mid \mathbf{E}, \mathbf{R}, \widehat{\boldsymbol{\Delta}}, \widehat{a}\,]$:

$$\left[Y \mid \mathbf{E}, \mathbf{R}, \widehat{\boldsymbol{\Delta}}, \widehat{a}\right] = \int_{\mathbf{C}} [\,Y \mid \mathbf{C}, \mathbf{R}, \mathbf{E}\,]\,.\,\left[\mathbf{C} \mid \mathbf{E}, \widehat{\boldsymbol{\Delta}}, \widehat{a}\right] \mathrm{d}\mathbf{C} \tag{A10}$$

The integral above does not have a closed form expression because the variance $\boldsymbol{\Sigma} = \mathbf{R} + \omega^2 \widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}' + (1 + \lambda)\mathbf{C}$ of $[\,Y \mid \mathbf{C}, \mathbf{R}, \mathbf{E}\,]$ is not proportional to $\mathbf{C}$. To address this issue, we

approximate $[\,\mathbf{\Sigma} \mid \mathbf{E}, \widehat{\mathbf{\Delta}}, \widehat{a}\,]$ by $\mathcal{IW}(\mathbf{R}+\omega^2\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}'+(1+\lambda)\widehat{\mathbf{C}}, \widehat{a}')$. This assumption is conservative in the sense that it extends the sampling uncertainty on $\mathbf{C}$ to $\mathbf{R}+\omega^2\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}'+(1+\lambda)\mathbf{C}$ even though $\mathbf{R}+\omega^2\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}'$ is a constant. It yields a closed form expression of the above integral thanks to conjugacy:

$$\left[\,Y \mid \mathbf{E}, \mathbf{R}, \widehat{\mathbf{\Delta}}, \widehat{a}\,\right] = \mathcal{S}t(\widehat{\mu}, \mathbf{R}+\omega^2\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}'+(1+\lambda)\widehat{\mathbf{C}}, \widehat{\nu}) \tag{A11}$$

APPENDIX B

Optimal index derivation

Let us solve the optimization problem of Equation (7) under the above assumptions. For simplicity, we restrict our search to so called "*half-space*" events which are defined by $E = \{Y \in \Omega_f \mid \phi' Y \geq u\}$ where $\phi' Y$ is a linear index with ϕ a vector of dimension n, and u is a threshold. Let us consider PNS as a function of ϕ and u.

$$\mathrm{PNS}(\phi, u) \quad = \mathbb{P}(\phi' Y \geq u \mid f) - \mathbb{P}(\phi' Y \geq u \mid \overline{f}) \tag{B1}$$

For simplicity, we will use an expression of $\mathrm{PNS}(\phi, u)$ in the treatment of the optimization problem which approximates $[\phi' Y \mid f]$ by a Gaussian PDF, even though it is a Student t PDF from the calculations of Section 4. Note that this approximation is made restrictively here for deriving an optimal index. Once this index is obtained, it is the then the true Student t PDF of Y that will be used to derive the desired value of PNS. Therefore, the implication of this approximation is to yield an index which is suboptimal and thereby underestimates the maximized value PNS^*.

$$\mathrm{PNS}(\phi, u) \quad = F\left(\frac{u - \phi'\overline{\mu}}{\sqrt{\phi' \overline{\boldsymbol{\Sigma}} \phi}}\right) - F\left(\frac{u - \phi'\mu}{\sqrt{\phi' \boldsymbol{\Sigma} \phi}}\right) \tag{B2}$$

where F is the standard Gaussian CDF. The first order condition in u, $\partial \mathrm{PNS}(\phi, u)/\partial u = 0$, thus yields:

$$\exp\left(-\frac{(u - \phi'\overline{\mu})^2}{2\phi' \overline{\boldsymbol{\Sigma}} \phi}\right) = \exp\left(-\frac{(u - \phi'\mu)^2}{2\phi' \boldsymbol{\Sigma} \phi}\right) \tag{B3}$$

Next, we introduce a third approximation $\boldsymbol{\Sigma} \simeq \overline{\boldsymbol{\Sigma}}$ to solve Equation (B3), yielding:

$$\begin{aligned} & u^* = \tfrac{1}{2}\phi'(\mu + \overline{\mu}) \\ & \Rightarrow \mathrm{PNS}(\phi, u^*) = 2F\left(\frac{\phi'(\mu - \overline{\mu})}{2\sqrt{\phi' \overline{\boldsymbol{\Sigma}} \phi}}\right) - 1 \end{aligned} \tag{B4}$$

Then, the first order condition in ϕ, $\partial \text{PNS}(\phi, u^*)/\partial\phi = 0$, yields:

$$\begin{aligned} & (\phi' \boldsymbol{\Sigma} \phi)(\mu - \overline{\mu}) = (\phi'(\mu - \overline{\mu}))\boldsymbol{\Sigma}\phi \\ \Rightarrow \quad & \phi^* = \boldsymbol{\Sigma}^{-1}(\mu - \overline{\mu}) \end{aligned} \tag{B5}$$

which proves Equation (21). Figure 5c illustrates this solution and also shows that the optimization problem of Equation (7) may be viewed as a classification problem. Our proposal to solve Equation (7) is in fact similar to a commonplace classification algorithm used in machine learning and known as Support Vector Machine (SVM) (Cortes and Vapnik 1995).

APPENDIX C

Data used in illustration

As in Hannart (2016), observations were obtained from the HADCRUT4 monthly temperature dataset (Morice et al. 2012), while GCM model simulations were obtained from the IPSL CM5A-LR model (Dufresne et al. 2012), downloaded from the CMIP5 database. An ensemble of runs consisting of two sets of forcings was used, the natural set of forcings (NAT) and the anthropogenic set of forcings (ANT) for which three runs are available in each case over the period of interest and from which an ensemble average was derived. On the other hand, a single preindustrial control run of 1000 years is available and was thus split into ten individual control runs of 100 years. Temperature in both observations and simulations were converted to anomalies by subtracting the time average over the reference period 1960-1991. The data was averaged temporally and spatially using a temporal resolution of ten years. Averaging was performed for both observations and simulations by using restrictively values for which observations were non missing, for a like-to-like comparison between observations and simulations.

REFERENCES

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

Cortes C., V. Vapnik (1995) Support-vector networks, *Machine Learning*, 20, 3, 273.

Dufresne J.-L. et al. (2012). Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. *Clim. Dyn.* 40, 2,123–2,165, doi:10.1007/s00382-012-1636-1.

Gillett N. P., M. F. Wehner, S. F. B. Tett, A. J. Weaver (2004), Testing the linearity of the response to combined greenhouse gas and sulfate aerosol forcing, Geophys. Res. Lett., 31, L14201, doi:10.1029/2004GL020111.

Hannart, A., J. Pearl, F.E.L. Otto, P. Naveau, M. Ghil (2015). Counterfactual causality theory for the attribution of weather and climate-related events. *Bull. Am. Met. Soc.* (in press).

Hannart, A., C. Vera, F.E.L. Otto, B. Cerne (2015). Causal influence of anthropogenic forcings on the Argentinian heat wave of December 2013. *Bull. Am. Met. Soc.*

Hannart, A., P. Naveau (2014). Estimating high dimensional covariance matrices: a new look at the Gaussian conjugate framework. *J. Multiv. Anal.*

Hannart A., A. Carrassi, M. Bocquet, M. Ghil, P. Naveau, M. Pulido, J. Ruiz, P. Tandeo (2015) DADA: Data Assimilation for the Detection and Attribution of Weather and Climate-related Events, Clim. Change., in revision. http://arxiv.org/abs/1503.05236

Hannart A. (2016) Integrated Optimal Fingerprinting: Method description and illustration, *J. Clim.*, 29:6, 1977–1998.

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.

Hegerl G., F. Zwiers (2011) Use of models in detection and attribution of climate change. *Wiley Interdisciplinary Reviews. Clim Change.* doi:10.1002/wcc.121

Karl T. R., A. Arguez, B. Huang, J. H. Lawrimore, J. R. McMahon, M. J. Menne, T. C. Peterson, R. S. Vose, H. M. Zhang (2015) Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, DOI:10.1126/science.aaa5632

Mastrandrea M. D., 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, F. W. Zwiers (2010) Guidance note for Lead Authors of the IPCC Fifth Assessment Report on consistent treatment of uncertainties. *Intergovernmental Panel on Climate Change (IPCC).*

Marvel K., C. Bonfils (2013) Identifying external influences on global precipitation. *Proceed. Nat. Acad. Sci.*, 110(48):19301–19306.

Meehl G. A., A. Hu, J. M. Arblaster, J. Fasullo, K. E. Trenberth (2013) Externally forced and internally generated decadal climate variability associated with the Interdecadal Pacific Oscillation. *J. Clim.* 26, 7298–7310.

Mellor D.H. (1995) *The Facts of Causation*, Routledge, ISBN 0-415-19756-2

Morgenstern O., G. Zeng, S. M. Dean, M. Joshi, N. L. Abraham, A. Osprey (2014), Direct and ozone mediated forcing of the Southern Annular Mode by greenhouse gases, *Geophys. Res. Lett.*, 41, 9050–9057, doi:10.1002/2014GL062140.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset, *J. Geophys. Res.*, 117, D08101, doi:10.1029/2011JD017187.

Pearl J. (2000). *Causality: models, reasoning and inference*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Ribes A., J.-M. Azais, S. Planton (2009). Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate, *Clim. Dyn.*, 33, 707–722.

Sharpe, W.F. (1963), A simplified model for portfolio analysis. *Management Science*, 9, 277–293.

Shiogama H., D. A. Stone, T. Nagashima, T. Nozawa, S. Emori (2013) On the linear additivity of climate forcing-response relationships at global and continental scales, *Int. J. of Climatol.*, 33, 11, 25–42.

Simpson E. H. (1951). The Interpretation of Interaction in Contingency Tables. *J. R. Stat. Soc.*, Series B, 13: 238–241.

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.

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

Suppes P. (1970) *A Probabilistic Theory of Causality*, Amsterdam: North-Holland Publishing.

Yan X., T. DelSole and M. K. Tippett (2016) What Surface Observations are Important for Separating the Influences of Anthropogenic Aerosols From Other Forcings? *J. Clim.*

List of Tables

Term	**Probability**
Virtually certain	≥ 0.99
Extremely likely	≥ 0.95
Very likely	≥ 0.90
Likely	≥ 0.66
About as likely as not	> 0.33 and < 0.66
Unlikely	≤ 0.33
Very unlikely	≤ 0.10
Exceptionally unlikely	≤ 0.01

TABLE 1. Correspondence between language and probabilities in IPCC calibrated terminology (Mastrandrea et al. 2010).

List of Figures

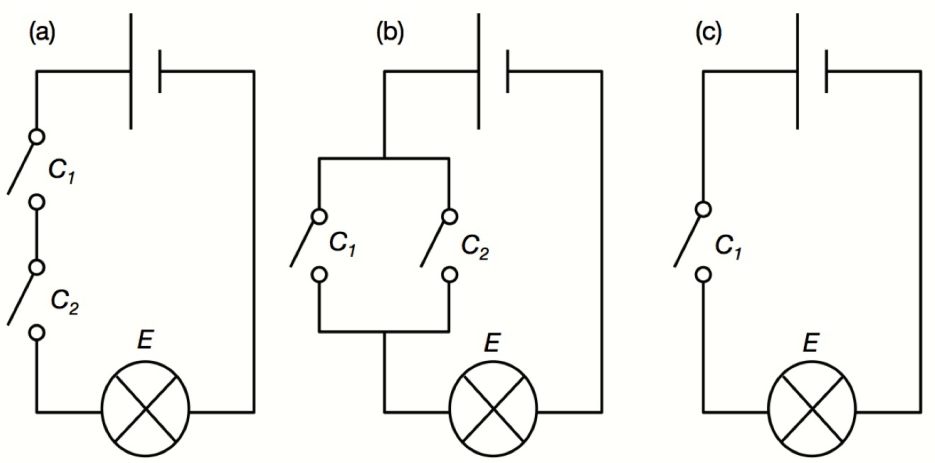

FIG. 1. The three facets of causality. (a) Bulb E can never be lit unless switch C_1 is on, yet activating C_1 does not always result in lighting E as this also requires turning on C_2: turning on C_1 is thus a necessary cause of E lighting, but not a sufficient one. (b) E is lit any time C_1 is turned on, yet if C_1 is turned off E may still be lit by activating C_2: turning on C_1 is thus a sufficient cause of E lighting, but not a necessary one. (c) Turning on C_1 always lights E, and E may not be lighted unless C_1 is on: turning on C_1 is thus a necessary and sufficient cause of E lighting.

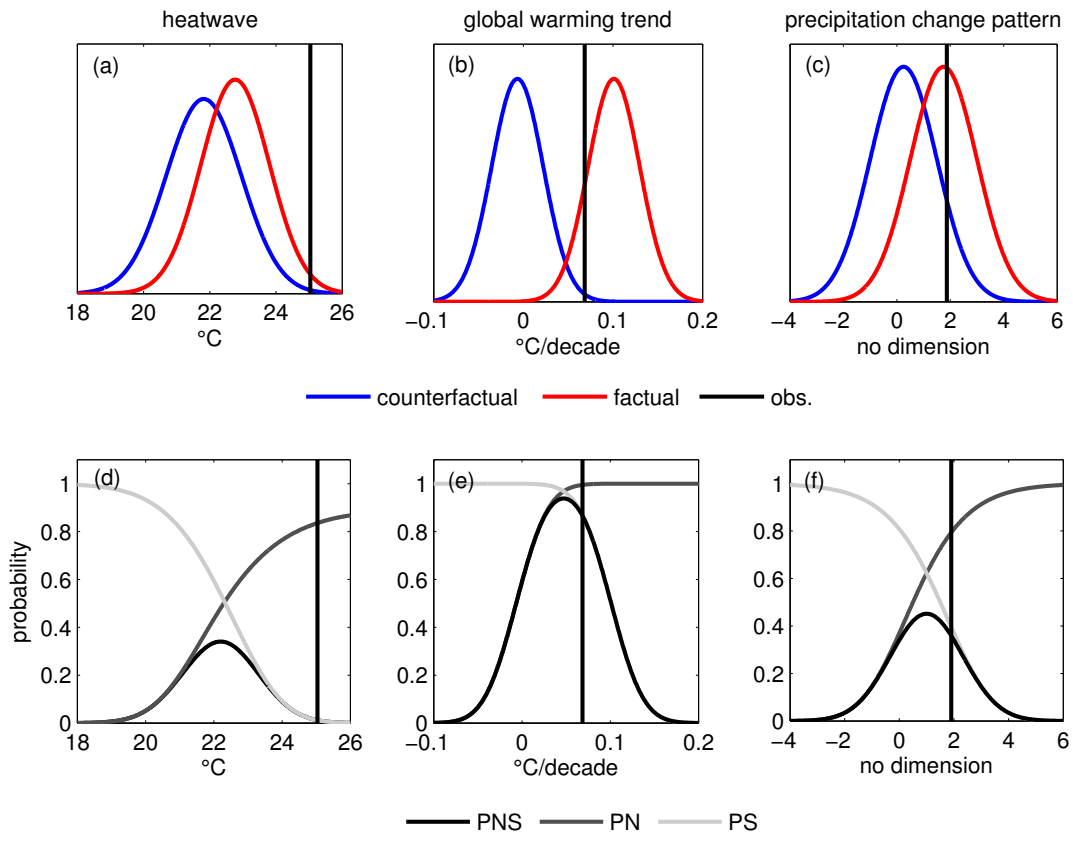

FIG. 2. Probabilities of causation in three different climate attribution situations. Upper panels (a,b,c) : factual PDF (red line) and counterfactual PDF (blue line) of the relevant index Z, observed value z of the index (vertical black line). Lower panels (d,e,f): PN, PS and PNS for the event $\{Z \geq u\}$ as a function of the threshold u. Left column (a,d): attribution of the Argentinian heatwave of December 2013. Middle column (b,e): attribution of the 20th century temperature change. Left column (c,f): attribution of the precipitation change over the satellite era (Marvel and Bonfils 2013).

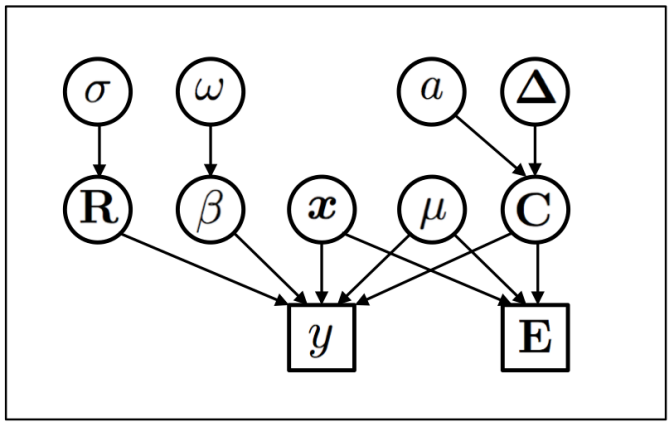

FIG. 3. Structural chart of the statistical model introduced in Section 4: underlying hierarchy of parameters (i.e. unobserved quantities, circles); and data used for inference (i.e. observed quantities, squares).

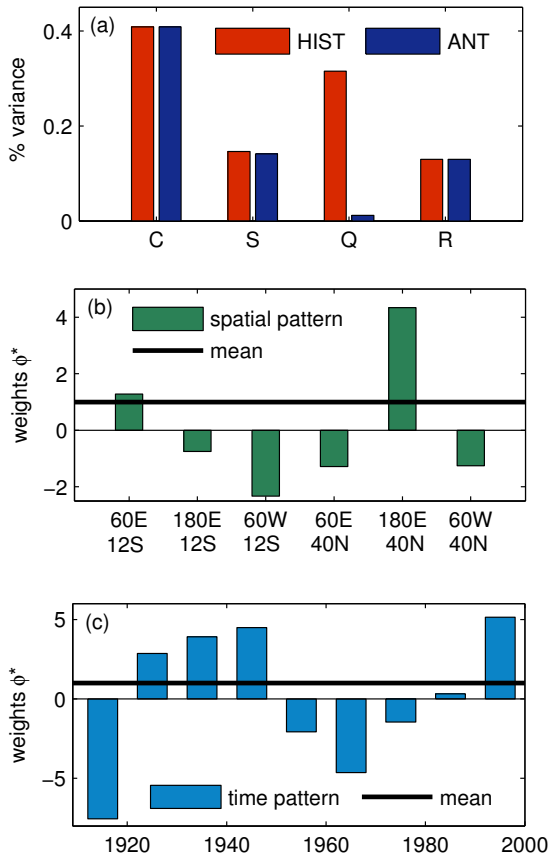

FIG. 4. Illustration on the 20th century temperature change: model fitting. (a) Distribution of the total variance between its four components (%). (b) Coefficients of the optimal mapping ϕ^* averaged spatially. (c) Coefficients of the optimal mapping ϕ^* averaged temporally.

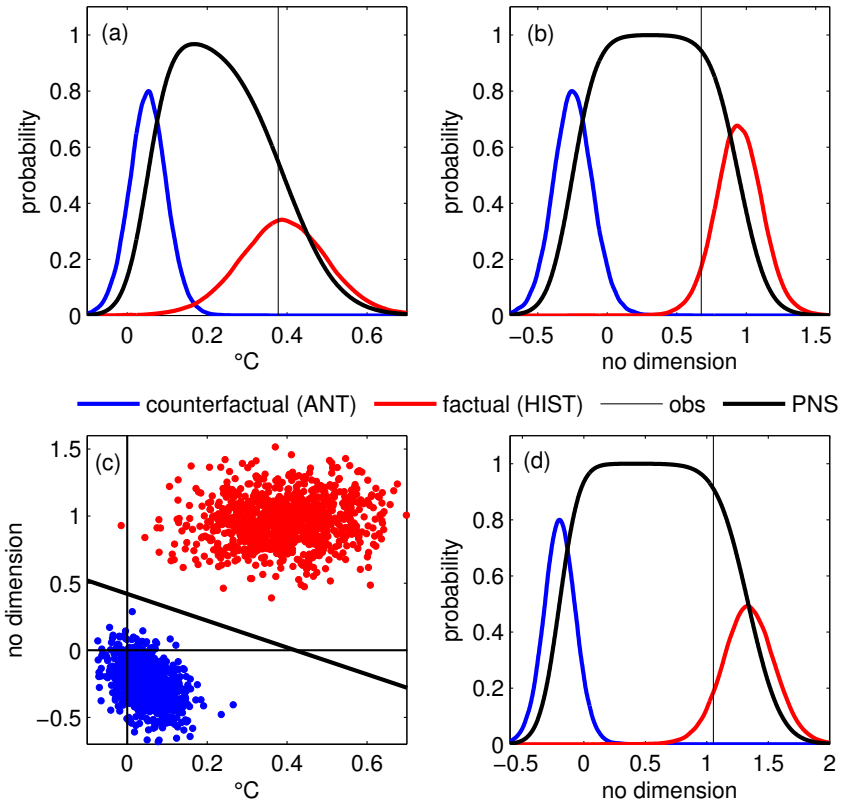

FIG. 5. Illustration on the 20th century temperature change: results. (a) Factual PDF (red line) and counterfactual PDF (blue line) of the optimal index $Z = \phi^*(Y)$, observed value $z = \phi^*(y)$ of the index (thin vertical black line); PNS as a function of the threshold u (thick black line). (b) Same as (a) for the global mean index. (c) Scatterplot of factual (red dots) and counterfactual (blue dots) joint realizations of the global mean index (horizontal axis) and of the space-time pattern index (vertical axis). (d) Same as (a) for the space-time pattern index.

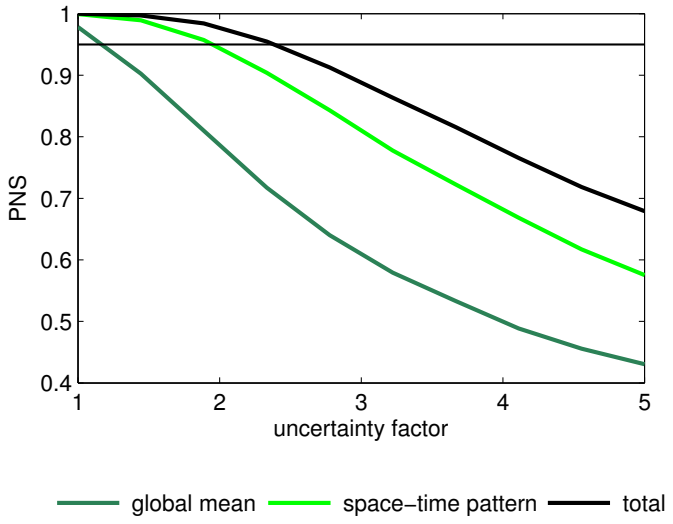

FIG. 6. PNS as a function of the inflation factor applied to all uncertainty sources: global mean alone (light green line), space-time pattern (dark green line), total (thick black line).

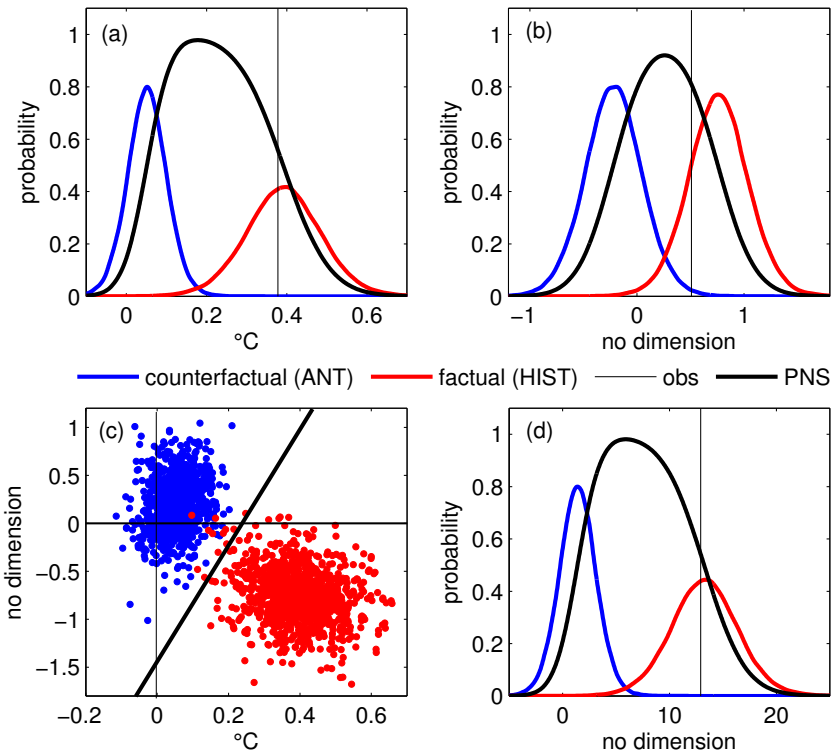

FIG. 7. Same as Figure 4 for the mapping ϕ^+ projected onto the leading eigenvectors of $\mathbf{C}$.

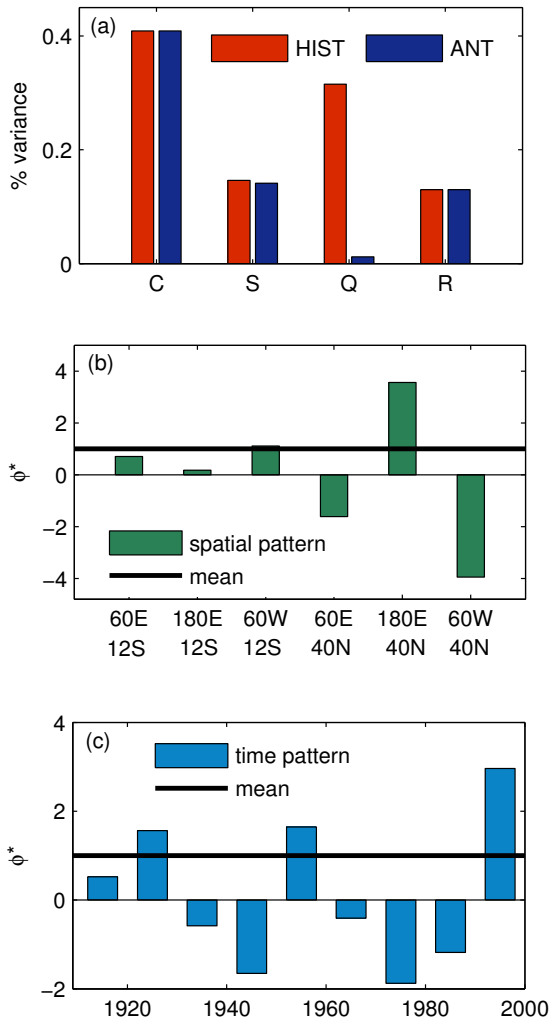

FIG. 8. Same as Figure 3 for the mapping ϕ^+ projected onto the leading eigenvectors of $\mathbf{C}$.