



HAL
open science

Probabilities of Causation of Climate Changes

Alexis Hannart, Philippe Naveau

► **To cite this version:**

Alexis Hannart, Philippe Naveau. Probabilities of Causation of Climate Changes. *Journal of Climate*, 2018, 10.1175/JCLI-D-17-0304.1 . hal-02410902

HAL Id: hal-02410902

<https://hal.science/hal-02410902v1>

Submitted on 3 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Probabilities of causation of climate changes**

2 ALEXIS HANNART *

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

3 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

4
5 Multiple changes in Earth's climate system have been observed over the past decades. De-
6 termining how likely each of these changes are to have been caused by human influence, is
7 important for decision making on mitigation and adaptation policy. Here we describe an ap-
8 proach for deriving the probability that anthropogenic forcings have caused a given observed
9 change. The proposed approach is anchored into causal counterfactual theory (Pearl 2000)
10 which has been introduced recently, and was in fact partly used already, in the context
11 of weather and climate-related events attribution. We argue that these concepts are also
12 relevant, and can be straightforwardly extended to, the context of climate change attribu-
13 tion. For this purpose, and in agreement with the principle of *fingerprinting* applied in the
14 conventional D&A framework, a trajectory of change is converted into an event occurrence
15 defined by maximizing the causal evidence associated to the forcing under scrutiny. Other
16 key assumptions used in the conventional D&A framework, in particular those related to
17 numerical models error, can also be adapted conveniently to this approach. Our proposal
18 thus allows to bridge the conventional framework with the standard causal theory, in an
19 attempt to improve the quantification of causal probabilities. An illustration suggests that
20 our approach is prone to yield a significantly higher estimate of the probability that an-
21 thropogenic forcings have caused the observed temperature change, thus supporting more
22 assertive causal claims.

23 1. Introduction

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

44 Aforementioned studies buttressing the above claim usually rely on a conventional attri-
45 bution framework in which “*causal attribution of anthropogenic climate change*” is under-
46 stood to mean “*demonstration that a detected change is consistent with the estimated re-*
47 *sponses to anthropogenic and natural forcings combined, but not consistent with alternative,*
48 *physically plausible explanations that exclude important elements of anthropogenic forcings*”
49 (Hegerl et al. 2010). While this definition has proved to be very useful and relevant, it offers

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

$$y = \sum_{f=1}^p \beta_f x_f + \varepsilon \quad (1)$$

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

$$\mathbb{P}(f \rightarrow y) = \mathbb{P}(\beta_f > 0) \quad (2)$$

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

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

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

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

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

106 2. Causal counterfactual theory

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

110 Let us first recall the so-called “counterfactual” definition of causality by quoting the
111 18th century Scottish philosopher David Hume: “*We may define a cause to be an object
112 followed by another, where, if the first object had not been, the second never had existed.*” In
113 other words, an event E (E stands for effect) is caused by an event C (C stands for cause)
114 if and only if E would not occur were it not for C . Note that the word *event* is used here in
115 its general, mathematical sense of *subset* of a sample space Ω . According to this definition,
116 evidencing causality requires a counterfactual approach by which one inquires whether or
117 not the event E would have occurred in an hypothetical world, termed counterfactual, in
118 which the event C would not have occurred. The fundamental approach of causality which
119 is implied by this definition is still entirely relevant in the standard causal theory. It may
120 also arguably be connected to the guidance principles of the conventional climate change
121 attribution framework and to the optimal fingerprinting models, in a qualitative manner.
122 The main virtue of the standard causality theory of Pearl consists in our view in formalizing
123 precisely the above qualitative definition, thus allowing for sound quantitative developments.
124 A prominent feature of this theory consists in first recognizing that causation corresponds to

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

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

$$\text{PS}(C \rightarrow E) = \mathbb{P}(E \mid do(C), \overline{C}, \overline{E}), \quad (3a)$$

$$\text{PN}(C \rightarrow E) = \mathbb{P}(\overline{E} \mid do(\overline{C}), C, E), \quad (3b)$$

$$\text{PNS}(C \rightarrow E) = \mathbb{P}(E \mid do(C), \overline{E} \mid do(\overline{C})). \quad (3c)$$

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

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

$$\text{PN}(C \rightarrow E) = \max(1 - \bar{p}/p, 0), \quad (4a)$$

$$\text{PS}(C \rightarrow E) = \max(1 - (1 - p)/(1 - \bar{p}), 0), \quad (4b)$$

$$\text{PNS}(C \rightarrow E) = \max(p - \bar{p}, 0). \quad (4c)$$

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

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

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

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

184 **3. Probabilities of causation of climate change**

185 We now return to the question of interest: for a given forcing f and an observed evolution
186 of the climate system y , can y be attributed to f ? More precisely, what is the probability
187 $\mathbb{P}(f \rightarrow y)$ that f has caused y ? We propose to tackle this problem by applying the causal
188 counterfactual theory to the context of climate change, and more specifically, by using the
189 three probabilities of causation PN, PS and PNS recalled above. This Section shows that it
190 can be done to a large extent similarly to the approach of Hannart et al. (2015a) for weather
191 event attribution. In particular, as in weather event attribution, the crucial question to be
192 answered as a starting point consists in narrowing down the definitions of the cause event
193 C and of the effect event E associated to the question at stake — where the word “event”
194 is used here in its general mathematical sense of “subset”.

195 *a. Counterfactual setting*

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

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

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

214 *b. Balancing necessity and sufficiency*

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

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

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

222 Consider the question “*Have anthropogenic CO₂ emissions caused global warming?*”.
 223 Following the above, the event “*global warming*” may be loosely defined as a positive trend
 224 on global Earth surface temperature, i.e. $E = \{Z \geq 0\}$ where Z is the global surface
 225 temperature trend coefficient and the threshold u is zero. In that case, E nearly always
 226 occurs in the factual world ($p \simeq 1$) but it is also frequent in the counterfactual one (\bar{p}
 227 medium) thus the emphasis is mostly on PS, i.e. on sufficient causation, while PN and PNS
 228 will have moderate values. But if global warming is more restrictively defined as a warming
 229 trend comparable to or greater than the observed trend, i.e. $E = \{Z \geq z\}$ where $u = z$ is
 230 the observed trend, then the event becomes nearly impossible in the counterfactual world
 231 ($\bar{p} \simeq 0$) but remains frequent in the factual one (\bar{p} medium) thus the emphasis is on PN, i.e.
 232 on necessary causation, while the values of PS and PNS will this time be low. Therefore, the
 233 above two extreme definitions both yield a low PNS. But under a more balanced definition
 234 of *global warming* as a trend exceeding an intermediate value $u^* \in [0, z]$, then the event
 235 nearly always occurs in the factual in the factual world ($p \simeq 1$) and yet remains very rare
 236 in the counterfactual one ($\bar{p} \simeq 0$). Hence PNS is then high: both necessary and sufficient
 237 causation prevail. We propose to take advantage of this optimal value to define the event
 238 “*global warming*” as the global trend index Z exceeding the optimal threshold u^* such that
 239 the amount of causal evidence, in a PNS sense, is maximized:

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

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

243 Let us now consider the question “*Have anthropogenic CO₂ emissions caused the Argen-*

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

266 *c. Building an optimal index*

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

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

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

283 As an illustration, Marvel and Bonfils (2013) focus on the attribution of changes in pre-
284 cipitation, and subsequently address the question “*Have anthropogenic forcing caused the*
285 *observed evolution of precipitation at a global level?*”. Arguably, this study illustrates our
286 point in the sense that it addresses the question by defining a *fingerprint* index Z which
287 aims precisely at reflecting the features of the change in precipitation that are thought to
288 materialize frequently (if not systematically) in the factual world and yet are expected to
289 be rare (if not impossible) in the counterfactual one, based on physical considerations. In
290 practice, the index Z defined by the authors consists of a non-dimensional scalar summa-
291 rizing the main spatial and physical features of precipitation evolution w.r.t. dynamics and
292 thermodynamics. The factual and counterfactual PDFs of Z are then derived from the
293 HIST and NAT runs respectively (Fig. 3c). From these PDFs, one can easily obtain an
294 optimal threshold u^* for the precipitation index Z by applying Equation (6). This yields
295 $\text{PNS} = 0.43$, i.e. anthropogenic forcings *have about as likely as not caused the observed*
296 *evolution of precipitation.*

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

$$(u^*, \phi^*) = \operatorname{argmax}_{u < \phi(y), \phi \in \Phi} \operatorname{PNS}(C \rightarrow \{\phi(Y) \geq u\}) \quad (7)$$

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

316 4. Implementation under the standard framework

317 We now turn to the practical aspects of implementing the approach described in Section
 318 3 above, based on the observations y and on climate model experiments.

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

$$\begin{aligned} \mathbb{E}(Y | f) &= \mu, & \mathbb{V}(Y | f) &= \Sigma \\ \mathbb{E}(Y | \bar{f}) &= \bar{\mu}, & \mathbb{V}(Y | \bar{f}) &= \bar{\Sigma} \end{aligned} \tag{8}$$

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

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

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

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

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

349 *b. Model description*

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

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

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

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

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

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

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

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

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

$$[\mu | \mathbf{C}, \mathbf{E}] = \mathcal{N}(\hat{\mu}, \frac{1}{r}\mathbf{C}), \quad [x_i | \mathbf{C}, \mathbf{E}] \sim \mathcal{N}(\hat{x}_i, \frac{1}{r_i}\mathbf{C}) \text{ for } i = 1, \dots, p \quad (13)$$

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

$$[Y | \mathbf{C}, \mathbf{R}, \mathbf{E}, \omega] = \mathcal{N}(\hat{\mu}, \mathbf{C} + \mathbf{R} + \omega^2 \hat{\mathbf{x}}\hat{\mathbf{x}}' + \lambda\mathbf{C}) \quad (14)$$

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

$$[\mathbf{C} | \mathbf{E}] = \mathcal{IW}(\hat{\mathbf{C}}, \hat{\nu}) \quad (15)$$

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

$$\hat{\mathbf{C}} = \hat{a}\hat{\mathbf{\Delta}} + (1 - \hat{a})\hat{\mathbf{\Omega}} \quad (16)$$

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

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

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

$$[Y | \mathbf{E}, \omega, \sigma] = \mathcal{St}(\hat{\mu}, \sigma^2 \mathbf{I} + \omega^2 \hat{\mathbf{x}} \hat{\mathbf{x}}' + (1 + \lambda) \hat{\mathbf{C}}, \hat{\nu}) \quad (17)$$

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

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

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

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

$$\begin{aligned} [Y | f] = & \mathcal{St}(\hat{\mu}, \hat{\Sigma}, \hat{\nu}) \\ \hat{\Sigma} = & (1 + \hat{\lambda}) \hat{\mathbf{C}} + \hat{\sigma}^2 \mathbf{I} + \hat{\omega}^2 \hat{\mathbf{x}} \hat{\mathbf{x}}' \end{aligned} \quad (19)$$

403 Next, to obtain $[Y | \bar{f}]$, one simply needs to change the mean $\hat{\mu}$ to $\bar{\hat{\mu}}$ obtained as the en-
 404 semble average for the counterfactual experiment “all forcings except f ”. Some changes also
 405 need to be made regarding the covariance. Indeed, since forcing f is absent in the counter-
 406 factual world, the model error covariance component $\hat{\omega}^2 \hat{\mathbf{x}}_f \hat{\mathbf{x}}_f'$, corresponding to the random

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

$$\begin{aligned} [Y | \bar{f}] &= \mathcal{St}(\widehat{\mu}, \widehat{\Sigma}, \widehat{\nu}) \\ \widehat{\Sigma} &= \widehat{\Sigma} - \widehat{\omega}^2 \widehat{x}_f \widehat{x}_f' + (\frac{1}{\bar{r}} - \frac{1}{r}) \widehat{\mathbf{C}} \end{aligned} \quad (20)$$

411 As noted above, when f is anthropogenic forcing, the counterfactual experiment NAT is
 412 usually available in CMIP runs, allowing for a straightforward derivation of $\widehat{\mu}$. But for other
 413 forcings, by the design of CMIP experiments, counterfactual runs are usually not available.
 414 A possible workaround then consists in applying the additivity assumption to approximate
 415 $\widehat{\mu}$ with $\widehat{\mu} - \widehat{x}_f$. However in that case, the sampling uncertainty term $\widehat{\mathbf{C}}/r_f$ corresponding to
 416 the estimation of \widehat{x}_f must be added to the covariance $\widehat{\Sigma}$.

417 *c. Derivation of the probabilities of causation*

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

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

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

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

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

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

431 and:

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

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

433 *d. Reducing computational cost*

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

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

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

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

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

$$\begin{aligned} |\Sigma| &= |\mathbf{A}| \cdot |\mathbf{I} + \omega^2 \mathbf{x}' \mathbf{A}^{-1} \mathbf{x}| \\ &= |\mathbf{B}| \cdot |\mathbf{I} + \frac{1}{r_0}(1 + \lambda)(1 - a)\mathbf{x}'_0\mathbf{B}^{-1}\mathbf{x}_0| \cdot |\mathbf{I} + \omega^2 \mathbf{x}' \mathbf{A}^{-1} \mathbf{x}| \end{aligned} \quad (26)$$

442 Independently of n , the matrices $\mathbf{I} + \omega^2 \mathbf{x}' \mathbf{A}^{-1} \mathbf{x}$ is of size p , $\mathbf{I} + \frac{1}{r_0}(1 + \lambda)(1 - a)\mathbf{x}'_0\mathbf{B}^{-1}\mathbf{x}_0$ is of size
 443 r_0 , and \mathbf{B} is diagonal. Obtaining their inverse and determinant is therefore computationally
 444 cheap. Hence, the inverse and determinant of Σ can be obtained at a low computational
 445 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 $\hat{\mu}$ using the r HIST runs; the estimated counterfactual response $\hat{\bar{\mu}}$ using the \bar{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 $\mathbf{x} = (x_1, x_2)$); the estimated covariance $\hat{\mathbf{C}}$ from the r_0 PIcontrol runs. Then, we derived $\hat{\sigma}$ and $\hat{\omega}$ by likelihood maximization, to obtain $\hat{\Sigma}$ and $\hat{\bar{\Sigma}}$.

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 $\mathbf{Q} = \omega^2 \mathbf{x}\mathbf{x}' + (1 - \alpha)\sigma^2 \mathbf{I}$ and $\mathbf{R} = \alpha\sigma^2 \mathbf{I}$. An objective separation would require an ensemble representing observational uncertainty, allowing for a preliminary

473 estimation of \mathbf{R} .

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

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

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

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

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

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

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

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

508 6. Discussion

509 a. Comparison with previous statements

510 The probabilities of causation obtained by using our proposal appear may depart from
511 the levels of uncertainty asserted by the latest IPCC report, and/or by previous work. For
512 instance, when y corresponds to the evolution of precipitation observed over the entire globe
513 during the satellite era (1979-2012), we have shown in Section 3 that, using the dynamic-
514 thermodynamic index built by Marvel and Bonfils (2013), the associated probability of cau-
515 sation $\mathbb{P}(\text{ANT} \rightarrow y)$ is found to be 0.43. This probability is thus significantly lower than the
516 one implied by the claim made in this study that “*the changes in precipitation observed in*
517 *the satellite era are likely to be anthropogenic in nature*” wherein “*likely*” implicitly means
518 $\mathbb{P}(\text{ANT} \rightarrow y) \geq 0.66$.

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

522 0.9999 which considerably exceeds the 0.95 probability implied by the latest IPCC report.
523 Such a gap deserves to be discussed.

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

538 Another possible explanation for the discrepancy is more technical. Many previous at-
539 tribution studies buttressing the IPCC statement regarding temperature, are based on an
540 inference method for the linear regression model of Equation (1) which is not optimal w.r.t.
541 maximizing causal evidence — despite of it being often referred to as “*optimal fingerprint-*
542 *ing*”. More precisely, the inference on the scaling factors β and the associated uncertainty
543 quantification, are obtained by projecting the observation y as well as the patterns \boldsymbol{x} onto
544 the leading eigenvectors of the covariance \mathbf{C} associated to climate internal variability. Such a
545 projection choice sharply contrasts with the projection used in our approach, which is indeed
546 performed onto the vector $\phi^* = \Sigma^{-1}(\mu - \bar{\mu})$. Denoting (v_1, \dots, v_n) the eigenvectors of Σ and

547 $(\lambda_1, \dots, \lambda_n)$ the corresponding eigenvalues, the expression of ϕ^* can be written:

$$\phi^* = \sum_{k=1}^n \frac{\langle v_k | \mu - \bar{\mu} \rangle}{\lambda_k} \cdot v_k \quad (29)$$

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

558 In order to assess whether or not these theoretical remarks hold in practice, we revisited
 559 our illustration and quantified the impact on $\mathbb{P}(\text{ANT} \rightarrow y)$ of using such a projection onto
 560 the leading eigenvectors of \mathbf{C} . For this purpose, after computing the projection matrix \mathbf{P} on
 561 the ten leading eigenvectors of \mathbf{C} , our procedure was applied identically, but this time using
 562 the projected vector $\phi^+ = \mathbf{P}\phi^*$. Results are shown in Figure 6, again after splitting the
 563 contribution of global mean change and patterns of change. Unsurprisingly, the probability
 564 of causation associated to the global mean change remains unmodified at 0.978. However, the
 565 probability of causation associated to the space-time features of warming drops from 0.9994
 566 to 0.92. Indeed, the features that most discriminate the two worlds, and are summarized in
 567 ϕ^* , do not align well with the leading eigenvectors of \mathbf{C} . They are thus incompletely reflected
 568 by the projected vector ϕ^+ (Figure 7). Furthermore, the uncertainty of the resulting index
 569 $Z^+ = \phi^{+'}Y$ is high by construction, as the magnitude of climate variability is maximized
 570 when projecting onto its leading modes (Figure 6b). This further contributes to reducing
 571 $\mathbb{P}(\text{ANT} \rightarrow y)$ to 0.992.

572 *b. Counterfactual experiments*

573 Our methodological proposal has an immediate implication w.r.t. the design of stan-
574 dardized CMIP experiments dedicated to D&A: a natural option would be to change the
575 present design “forcing f only” into a counterfactual design “all forcings except f ”. Indeed,
576 $\mathbb{P}(f \rightarrow y)$ is driven by the difference $\mu - \bar{\mu}_f$ between the factual response μ (i.e. historical
577 experiment) and the counterfactual response $\bar{\mu}_f$ (i.e. all forcings except f experiment). Un-
578 der the assumption that forcings do not interact with one another and that the combined
579 response matches with the sum of the individual responses, the difference $\mu - \bar{\mu}_f$ coincides
580 with the individual response x_f (i.e. forcing f only experiment). While this hypothesis is
581 well established for temperature at large scale (Gillett et al. 2004), it appears to break down
582 for other variables (e.g. precipitation, (Shiogama et al. 2013)) or over particular regions
583 (e.g the Southern extratropics, (Morgenstern et al. 2014)) where forcings appear to signifi-
584 cantly interplay. Such a lack of additivity would inevitably damage the results of the causal
585 analysis. It is thus important in our view to better understand the domain of validity of
586 the forcing-additivity assumption and to evaluate the drawbacks of the present “one forcing
587 only” design versus its advantages. Such an analysis does require “forcing f only” experi-
588 ments, but also “all forcings except f ” experiments in order to allow for comparison. This
589 effort would hence justify including in the DAMIP set of experiments an “all forcings except
590 f ” experiment — which is presently absent even in the lowest priority tier thereof — at least
591 for the most important forcings such as anthropogenic CO₂.

592 *c. Benchmarking high probabilities*

593 Section 5 showed that the proposed approach may sometimes yield probabilities of cau-
594 sation that are very close to one. How can we communicate such low levels of uncertainty?
595 This question arises insofar as the term “virtual certainty” applies as soon as PNS exceeds
596 0.99 under the current IPCC language. Thus, this terminology would be unfit to express in

597 words a PNS increase from 0.99 to 0.9999, say — even though such an increase corresponds
598 to a large reduction of uncertainty by a factor one hundred. One option to address this issue
599 is to use instead the uncertainty terminology of theoretical physics, in which a probability is
600 translated into an exceedance level under the Gaussian distribution, measured in numbers
601 of σ from the mean (where σ denotes standard deviation), i.e. $F^{-1}(\text{PNS})\sigma$ with F the CDF
602 of the standard Gaussian distribution. Under such terminology, “virtual certainty” thus
603 corresponds to a level of uncertainty of 2.3σ , while $\mathbb{P}(\text{ANT} \rightarrow y) = 0.9999$ found in Section
604 5 reaches 3.7σ . It is interesting to note that the level of uncertainty officially requested in
605 theoretical physics to corroborate a discovery as such — e.g. the existence of the Higgs Boson
606 — is 5σ . By such high standards, $\mathbb{P}(\text{ANT} \rightarrow y) = 0.9999$ found above can actually still be
607 considered much too low a probability to corroborate that human influence has indeed been
608 the cause of the observed warming. Therefore, further increasing $\mathbb{P}(\text{ANT} \rightarrow y)$ by building
609 more evidence into the analysis, may still be considered to be a relevant goal.

610 7. Summary and conclusion

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

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

625 *Acknowledgments.*

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

APPENDIX A

631

632

633

Derivation of the PDF of Y

634

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

$$[Y | \mathbf{x}, \mathbf{C}, \mathbf{R}] = \int_{\beta} [Y | \beta, \mathbf{x}, \mathbf{C}, \mathbf{R}] \cdot [\beta | \omega] d\beta \quad (\text{A1})$$

635

Given the quadratic dependence to β of the two terms under the integral in the right hand

636

side of Equation (A1), it is clear that the PDF of the left hand side is also Gaussian. Thus,

637

instead of computing the above integral, it is more convenient to derive the mean and variance

638

of this PDF by applying the rule of total expectation and total variance:

$$\begin{aligned} \mathbb{E}(Y | \mathbf{x}, \mathbf{C}, \mathbf{R}) &= \mathbb{E}(\mathbb{E}(Y | \beta, \mathbf{x}, \mathbf{C}, \mathbf{R}) | \mathbf{x}, \mathbf{C}, \mathbf{R}) = \mathbb{E}(\mathbf{x}\beta | \mathbf{x}, \mathbf{C}, \mathbf{R}) = \mathbf{x}\mathbb{E}(\beta) \\ &= \mathbf{x}e \end{aligned}$$

$$\begin{aligned} V(Y | \mathbf{x}, \mathbf{C}, \mathbf{R}) &= V(\mathbb{E}(Y | \beta, \mathbf{x}, \mathbf{C}, \mathbf{R}) | \mathbf{x}, \mathbf{C}, \mathbf{R}) + \mathbb{E}(V(Y | \beta, \mathbf{x}, \mathbf{C}, \mathbf{R}) | \mathbf{x}, \mathbf{C}, \mathbf{R}) \\ &= V(\mathbf{x}\beta | \mathbf{x}, \mathbf{C}, \mathbf{R}) + \mathbb{E}(\mathbf{C} + \mathbf{R} | \mathbf{x}, \mathbf{C}, \mathbf{R}) \\ &= \mathbf{x}V(\beta)\mathbf{x}' + \mathbf{C} + \mathbf{R} = \omega^2\mathbf{x}\mathbf{x}' + \mathbf{C} + \mathbf{R} \end{aligned}$$

$$[Y | \mathbf{x}, \mathbf{C}, \mathbf{R}] = \mathcal{N}(\mathbf{x}e, \mathbf{C} + \mathbf{R} + \omega^2\mathbf{x}\mathbf{x}') \quad (\text{A2})$$

639

Next, in order to account for the sampling uncertainty on the estimation of μ , we apply

640

Bayes theorem to derive the PDF of μ conditional on the ensemble \mathbf{E} . Denote $\mu^{(1)}, \dots, \mu^{(r)}$

641

the r simulated responses in \mathbf{E} which are assumed to be i.i.d. according to a Gaussian with

642

mean μ and covariance \mathbf{C} . We have:

$$\begin{aligned} [\mu | \mathbf{C}, \mathbf{E}] &\propto \prod_{j=1}^r [\mu^{(j)} | \mathbf{C}] \cdot [\mu] \\ &\propto \prod_{j=1}^r \mathcal{N}(\mu^{(j)} | \mu, \mathbf{C}) \\ &= \mathcal{N}(\mu | \hat{\mu}, \frac{1}{r}\mathbf{C}) \end{aligned} \quad (\text{A3})$$

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

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

$$\begin{aligned}
 \mathbb{E}(Y | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) &= \mathbb{E}(\mathbb{E}(Y | \mu, \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) = \mathbb{E}(\mu | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) \\
 &= \hat{\mu} \\
 \mathbb{V}(Y | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) &= \mathbb{V}(\mathbb{E}(Y | \mu, \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) + \mathbb{E}(\mathbb{V}(Y | \mu, \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) \\
 &= \mathbb{V}(\mu | \mathbf{x}, \mathbf{C}, \mathbf{R}) + \mathbb{E}(\omega^2 \mathbf{x} \mathbf{x}' + \mathbf{C} + \mathbf{R} | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) \\
 &= \frac{1}{r} \mathbf{C} + \omega^2 \mathbf{x} \mathbf{x}' + \mathbf{C} + \mathbf{R}
 \end{aligned} \tag{A4}$$

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

$$\begin{aligned}
 \mathbb{E}(Y | \mathbf{C}, \mathbf{R}, \mathbf{E}) &= \mathbb{E}(\mathbb{E}(Y | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) | \mathbf{C}, \mathbf{R}, \mathbf{E}) = \mathbb{E}(\hat{\mu} | \mathbf{C}, \mathbf{R}, \mathbf{E}) \\
 &= \hat{\mu} \\
 \mathbb{V}(Y | \mathbf{C}, \mathbf{R}, \mathbf{E}) &= \mathbb{V}(\mathbb{E}(Y | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) | \mathbf{C}, \mathbf{R}, \mathbf{E}) + \mathbb{E}(\mathbb{V}(Y | \mathbf{x}, \mathbf{C}, \mathbf{R}, \mathbf{E}) | \mathbf{C}, \mathbf{R}, \mathbf{E}) \\
 &= \mathbf{0} + (1 + \frac{1}{r}) \mathbf{C} + \mathbf{R} + \mathbb{E}(\omega^2 \mathbf{x} \mathbf{x}' | \mathbf{C}, \mathbf{E}) \\
 &= (1 + \frac{1}{r}) \mathbf{C} + \mathbf{R} + \omega^2 \sum_i \mathbb{E}(x_i x_i' | \mathbf{C}, \mathbf{E}) \\
 &= (1 + \frac{1}{r}) \mathbf{C} + \mathbf{R} + \omega^2 \sum_i \mathbb{V}(x_i | \mathbf{C}, \mathbf{E}) + \omega^2 \sum_i \mathbb{E}(x_i | \mathbf{C}, \mathbf{E}) \mathbb{E}(x_i | \mathbf{C}, \mathbf{E})' \\
 &= (1 + \frac{1}{r}) \mathbf{C} + \mathbf{R} + \omega^2 \sum_i \frac{1}{r_i} \mathbf{C} + \omega^2 \sum_i \hat{x}_i \hat{x}_j' \\
 &= (1 + \frac{1}{r} + \omega^2 \sum_i \frac{1}{r_i}) \mathbf{C} + \mathbf{R} + \omega^2 \hat{\mathbf{x}} \hat{\mathbf{x}}' \\
 &= \mathbf{C} + \mathbf{R} + \omega^2 \hat{\mathbf{x}} \hat{\mathbf{x}}' + \lambda \mathbf{C}
 \end{aligned} \tag{A5}$$

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

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

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

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

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

$$\begin{aligned} \ell(a, \mathbf{\Delta}) &= \left(\frac{ar_0}{1-a} + n + 1\right) \log \left| \frac{a}{1-a} \mathbf{\Delta} \right| - \left(\frac{r_0}{1-a} + n + 1\right) \log \left| \widehat{\mathbf{\Omega}} + \frac{a}{1-a} \mathbf{\Delta} \right| \\ &+ 2 \log \left(\Gamma_n \left\{ \frac{1}{2} \left(\frac{r_0}{1-a} + n + 1 \right) \right\} / \Gamma_n \left\{ \frac{1}{2} \left(\frac{ar_0}{1-a} + n + 1 \right) \right\} \right). \end{aligned} \quad (\text{A7})$$

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

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

664 where \mathcal{F} is a set of definite positive matrices chosen to introduce a regularization constraint
 665 on the covariance. Here we choose $\mathcal{F} = \{\operatorname{diag}(\delta_1, \dots, \delta_n) \mid \delta_1 > 0, \dots, \delta_n > 0\}$ the set of definite
 666 positive diagonal matrices, and we derive an approximated solution to Equation (A8) with
 667 $\widehat{\mathbf{\Delta}} = \operatorname{diag}(\widehat{\mathbf{\Omega}})$ and $\widehat{a} = \operatorname{argmax}_{a \in [0,1]} \ell(a, \widehat{\mathbf{\Delta}})$. Because the prior PDF is fitted on the data,
 668 this approach can be referred to as “empirical bayesian”. The “fitted” prior $[\mathbf{C} \mid \widehat{\mathbf{\Delta}}, \widehat{a}]$ is
 669 then updated using the ensemble \mathbf{E} , and the obtained posterior has a closed form expression
 670 due to conjugacy:

$$[\mathbf{C} \mid \mathbf{E}, \widehat{\mathbf{\Delta}}, \widehat{a}] \propto [\mathbf{E} \mid \mathbf{C}] \cdot \mathcal{IW}(\widehat{\mathbf{\Delta}}, \widehat{a}) = \mathcal{IW}(\widehat{\mathbf{C}}, \widehat{a}') \quad (\text{A9})$$

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

$$[Y \mid \mathbf{E}, \mathbf{R}, \widehat{\mathbf{\Delta}}, \widehat{a}] = \int_{\mathbf{C}} [Y \mid \mathbf{C}, \mathbf{R}, \mathbf{E}] \cdot [\mathbf{C} \mid \mathbf{E}, \widehat{\mathbf{\Delta}}, \widehat{a}] \, d\mathbf{C} \quad (\text{A10})$$

673 The integral above does not have a closed form expression because the variance $\mathbf{\Sigma} = \mathbf{R} +$
 674 $\omega^2 \widehat{\mathbf{x}} \widehat{\mathbf{x}}' + (1 + \lambda) \mathbf{C}$ of $[Y \mid \mathbf{C}, \mathbf{R}, \mathbf{E}]$ is not proportional to \mathbf{C} . To address this issue, we

675 approximate $[\boldsymbol{\Sigma} \mid \mathbf{E}, \widehat{\boldsymbol{\Delta}}, \widehat{a}]$ by $\mathcal{IW}(\mathbf{R} + \omega^2 \widehat{\boldsymbol{x}} \widehat{\boldsymbol{x}}' + (1 + \lambda) \widehat{\mathbf{C}}, \widehat{a}')$. This assumption is conservative
 676 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
 677 though $\mathbf{R} + \omega^2 \widehat{\boldsymbol{x}} \widehat{\boldsymbol{x}}'$ is a constant. It yields a closed form expression of the above integral
 678 thanks to conjugacy:

$$\left[Y \mid \mathbf{E}, \mathbf{R}, \widehat{\boldsymbol{\Delta}}, \widehat{a} \right] = St(\widehat{\mu}, \mathbf{R} + \omega^2 \widehat{\boldsymbol{x}} \widehat{\boldsymbol{x}}' + (1 + \lambda) \widehat{\mathbf{C}}, \widehat{\nu}) \tag{A11}$$

679

680

681

Optimal index derivation

682 Let us solve the optimization problem of Equation (7) under the above assumptions.
 683 For simplicity, we restrict our search to so called “*half-space*” events which are defined by
 684 $E = \{Y \in \Omega_f \mid \phi'Y \geq u\}$ where $\phi'Y$ is a linear index with ϕ a vector of dimension n , and u
 685 is a threshold. Let us consider PNS as a function of ϕ and u .

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

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

$$\text{PNS}(\phi, u) = F\left(\frac{u - \phi'\bar{\mu}}{\sqrt{\phi'\bar{\Sigma}\phi}}\right) - F\left(\frac{u - \phi'\mu}{\sqrt{\phi'\Sigma\phi}}\right) \quad (\text{B2})$$

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

$$\exp\left(-\frac{(u - \phi'\bar{\mu})^2}{2\phi'\bar{\Sigma}\phi}\right) = \exp\left(-\frac{(u - \phi'\mu)^2}{2\phi'\Sigma\phi}\right) \quad (\text{B3})$$

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

$$\begin{aligned} u^* &= \frac{1}{2}\phi'(\mu + \bar{\mu}) \\ \Rightarrow \text{PNS}(\phi, u^*) &= 2F\left(\frac{\phi'(\mu - \bar{\mu})}{2\sqrt{\phi'\bar{\Sigma}\phi}}\right) - 1 \end{aligned} \quad (\text{B4})$$

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

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

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

APPENDIX C

701

702

Data used in illustration

703

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

REFERENCES

- 719 Allen M. R. (2003). Liability for climate change. *Nature*, 421:891–892.
- 720 Cortes C., V. Vapnik (1995) Support-vector networks, *Machine Learning*, 20, 3, 273.
- 721 Dufresne J.-L. et al. (2012). Climate change projections using the IPSL-CM5 Earth System
722 Model: from CMIP3 to CMIP5. *Clim. Dyn.* 40, 2,123–2,165, doi:10.1007/s00382-012-1636-
723 1.
- 724 Gillett N. P., M. F. Wehner, S. F. B. Tett, A. J. Weaver (2004), Testing the linearity of the
725 response to combined greenhouse gas and sulfate aerosol forcing, *Geophys. Res. Lett.*, 31,
726 L14201, doi:10.1029/2004GL020111.
- 727 Hannart, A., J. Pearl, F.E.L. Otto, P. Naveau, M. Ghil (2015). Counterfactual causality
728 theory for the attribution of weather and climate-related events. *Bull. Am. Met. Soc.* (in
729 press).
- 730 Hannart, A., C. Vera, F.E.L. Otto, B. Cerne (2015). Causal influence of anthropogenic
731 forcings on the Argentinian heat wave of December 2013. *Bull. Am. Met. Soc.*
- 732 Hannart, A., P. Naveau (2014). Estimating high dimensional covariance matrices: a new
733 look at the Gaussian conjugate framework. *J. Multiv. Anal.*
- 734 Hannart A., A. Carrassi, M. Bocquet, M. Ghil, P. Naveau, M. Pulido, J. Ruiz, P. Tandeo
735 (2015) DADA: Data Assimilation for the Detection and Attribution of Weather and
736 Climate-related Events, *Clim. Change.*, in revision. <http://arxiv.org/abs/1503.05236>
- 737 Hannart A. (2016) Integrated Optimal Fingerprinting: Method description and illustration,
738 *J. Clim.*, 29:6, 1977–1998.

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

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

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

751 Mastrandrea M. D., C. B. Field, T. F. Stocker, O. Edenhofer, K. L. Ebi, D. J. Frame, H.
752 Held, E. Kriegler, K. J. Mach, P. R. Matschoss, G. K. Plattner, G. W. Yohe, F. W. Zwiers
753 (2010) Guidance note for Lead Authors of the IPCC Fifth Assessment Report on consistent
754 treatment of uncertainties. *Intergovernmental Panel on Climate Change (IPCC)*.

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

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

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

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

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

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

769 Ribes A., J.-M. Azais, S. Planton (2009). Adaptation of the optimal fingerprint method for
770 climate change detection using a well-conditioned covariance matrix estimate, *Clim. Dyn.*,
771 33, 707–722.

772 Sharpe, W.F. (1963), A simplified model for portfolio analysis. *Management Science*, 9,
773 277–293.

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

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

779 IPCC, 2013: Summary for Policymakers. In: *Climate Change 2013: The Physical Science*
780 *Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergov-*
781 *ernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor,
782 S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge
783 University Press, Cambridge, United Kingdom and New York, NY, USA.

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

786 Suppes P. (1970) *A Probabilistic Theory of Causality*, Amsterdam: North-Holland Publish-
787 ing.

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

790 **List of Tables**

791 1 Correspondence between language and probabilities in IPCC calibrated ter-
792 minology (Mastrandrea et al. 2010). 40

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

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

793 List of Figures

- 794 1 The three facets of causality. (a) Bulb E can never be lit unless switch C_1 is
795 on, yet activating C_1 does not always result in lighting E as this also requires
796 turning on C_2 : turning on C_1 is thus a necessary cause of E lighting, but not
797 a sufficient one. (b) E is lit any time C_1 is turned on, yet if C_1 is turned off
798 E may still be lit by activating C_2 : turning on C_1 is thus a sufficient cause of
799 E lighting, but not a necessary one. (c) Turning on C_1 always lights E , and
800 E may not be lighted unless C_1 is on: turning on C_1 is thus a necessary and
801 sufficient cause of E lighting. 43
- 802 2 Probabilities of causation in three different climate attribution situations. Up-
803 per panels (a,b,c) : factual PDF (red line) and counterfactual PDF (blue line)
804 of the relevant index Z , observed value z of the index (vertical black line).
805 Lower panels (d,e,f): PN, PS and PNS for the event $\{Z \geq u\}$ as a function of
806 the threshold u . Left column (a,d): attribution of the Argentinian heatwave
807 of December 2013. Middle column (b,e): attribution of the 20th century tem-
808 perature change. Left column (c,f): attribution of the precipitation change
809 over the satellite era (Marvel and Bonfils 2013). 44
- 810 3 Structural chart of the statistical model introduced in Section 4: underlying
811 hierarchy of parameters (i.e. unobserved quantities, circles); and data used
812 for inference (i.e. observed quantities, squares). 45
- 813 4 Illustration on the 20th century temperature change: model fitting. (a) Distri-
814 bution of the total variance between its four components (%). (b) Coefficients
815 of the optimal mapping ϕ^* averaged spatially. (c) Coefficients of the optimal
816 mapping ϕ^* averaged temporally. 46

817	5	Illustration on the 20th century temperature change: results. (a) Factual PDF	
818		(red line) and counterfactual PDF (blue line) of the optimal index $Z = \phi^*(Y)$,	
819		observed value $z = \phi^*(y)$ of the index (thin vertical black line); PNS as a	
820		function of the threshold u (thick black line). (b) Same as (a) for the global	
821		mean index. (c) Scatterplot of factual (red dots) and counterfactual (blue	
822		dots) joint realizations of the global mean index (horizontal axis) and of the	
823		space-time pattern index (vertical axis). (d) Same as (a) for the space-time	
824		pattern index.	47
825	6	PNS as a function of the inflation factor applied to all uncertainty sources:	
826		global mean alone (light green line), space-time pattern (dark green line),	
827		total (thick black line).	48
828	7	Same as Figure 4 for the mapping ϕ^+ projected onto the leading eigenvectors	
829		of \mathbf{C} .	49
830	8	Same as Figure 3 for the mapping ϕ^+ projected onto the leading eigenvectors	
831		of \mathbf{C} .	50

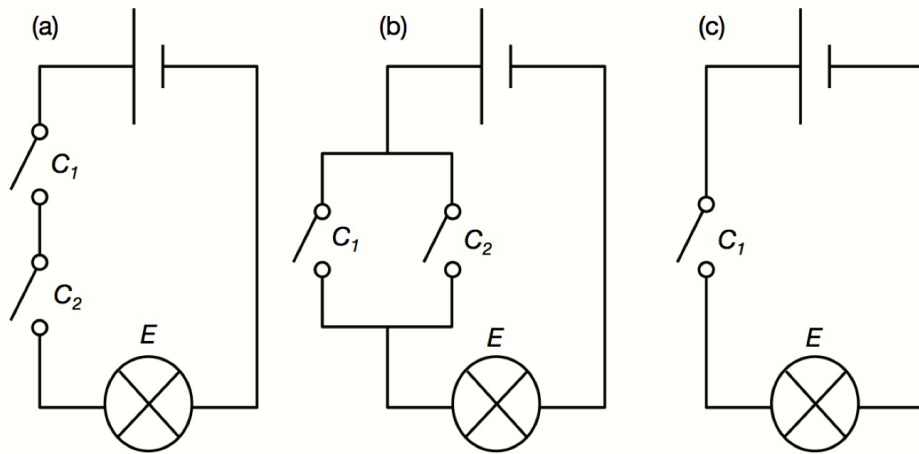


FIG. 1. The three facets of causality. (a) Bulb E can never be lit unless switch C_1 is on, yet activating C_1 does not always result in lighting E as this also requires turning on C_2 : turning on C_1 is thus a necessary cause of E lighting, but not a sufficient one. (b) E is lit any time C_1 is turned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

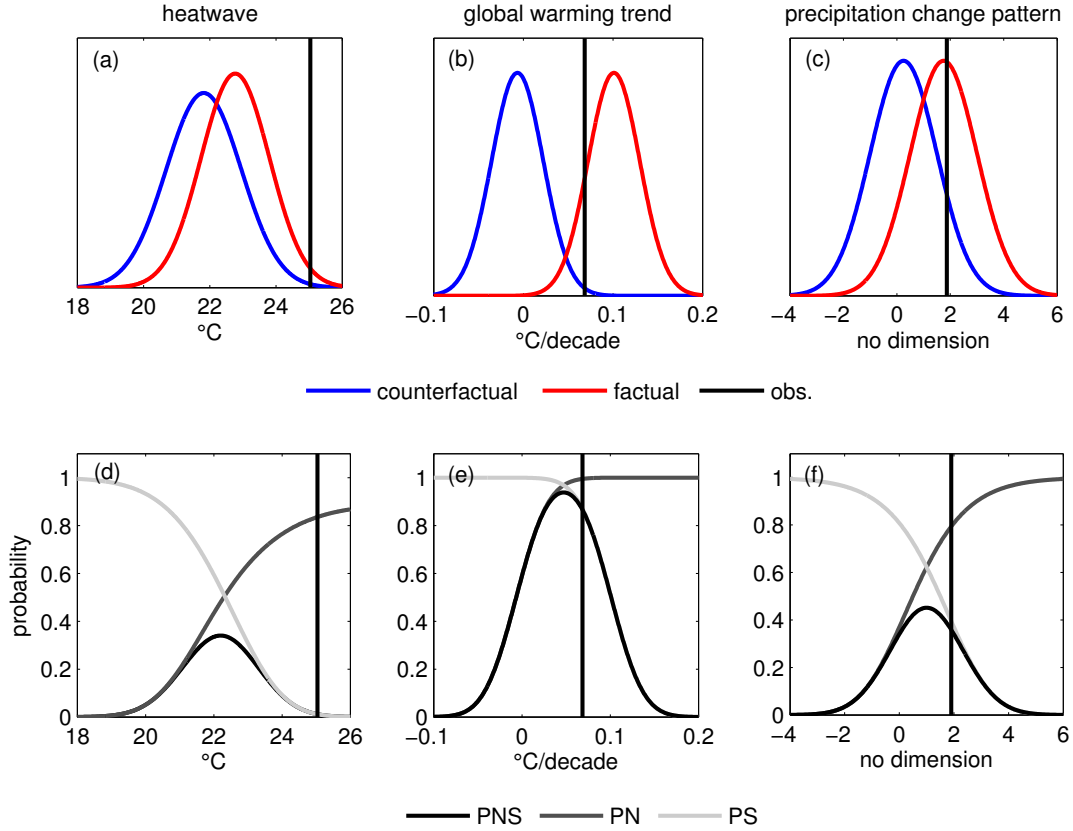


FIG. 2. Probabilities of causation in three different climate attribution situations. Upper panels (a,b,c) : factual PDF (red line) and counterfactual PDF (blue line) of the relevant index Z , observed value z of the index (vertical black line). Lower panels (d,e,f): PN, PS and PNS for the event $\{Z \geq u\}$ 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).

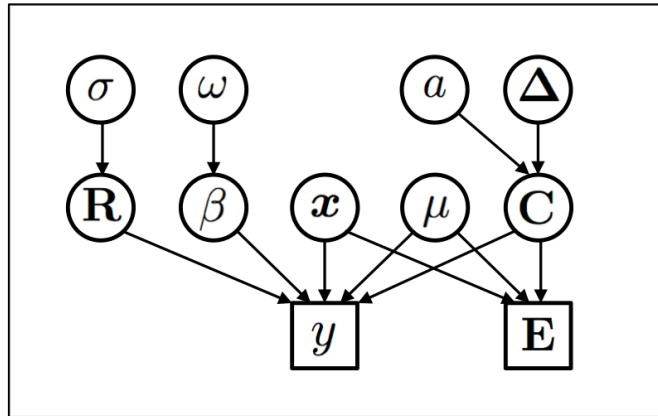


FIG. 3. Structural chart of the statistical model introduced in Section 4: underlying hierarchy of parameters (i.e. unobserved quantities, circles); and data used for inference (i.e. observed quantities, squares).

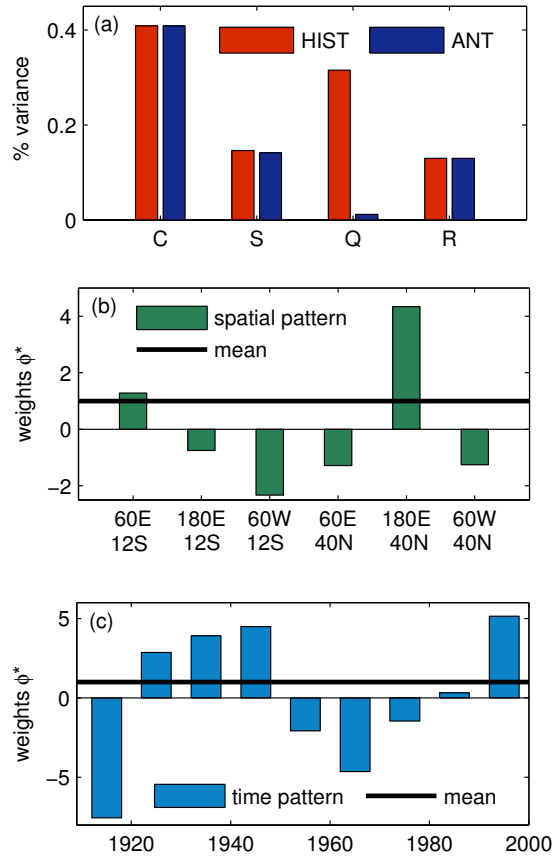


FIG. 4. Illustration on the 20th century temperature change: model fitting. (a) Distribution of the total variance between its four components (%). (b) Coefficients of the optimal mapping ϕ^* averaged spatially. (c) Coefficients of the optimal mapping ϕ^* averaged temporally.

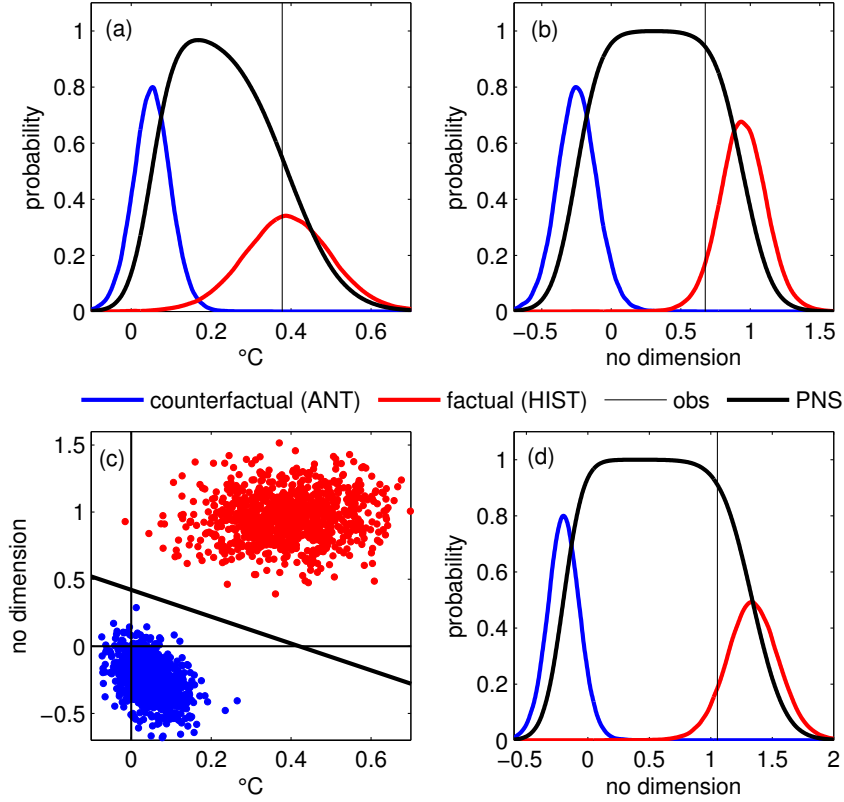


FIG. 5. Illustration on the 20th century temperature change: results. (a) Factual PDF (red line) and counterfactual PDF (blue line) of the optimal index $Z = \phi^*(Y)$, observed value $z = \phi^*(y)$ of the index (thin vertical black line); PNS as a function of the threshold u (thick black line). (b) 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.

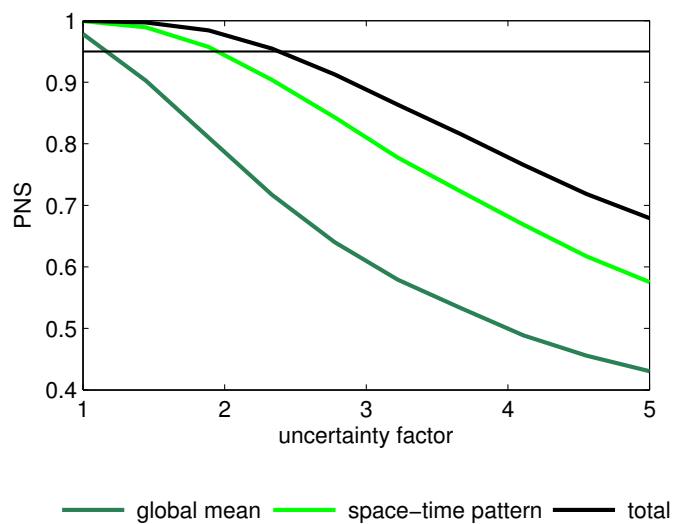


FIG. 6. PNS as a function of the inflation factor applied to all uncertainty sources: global mean alone (light green line), space-time pattern (dark green line), total (thick black line).

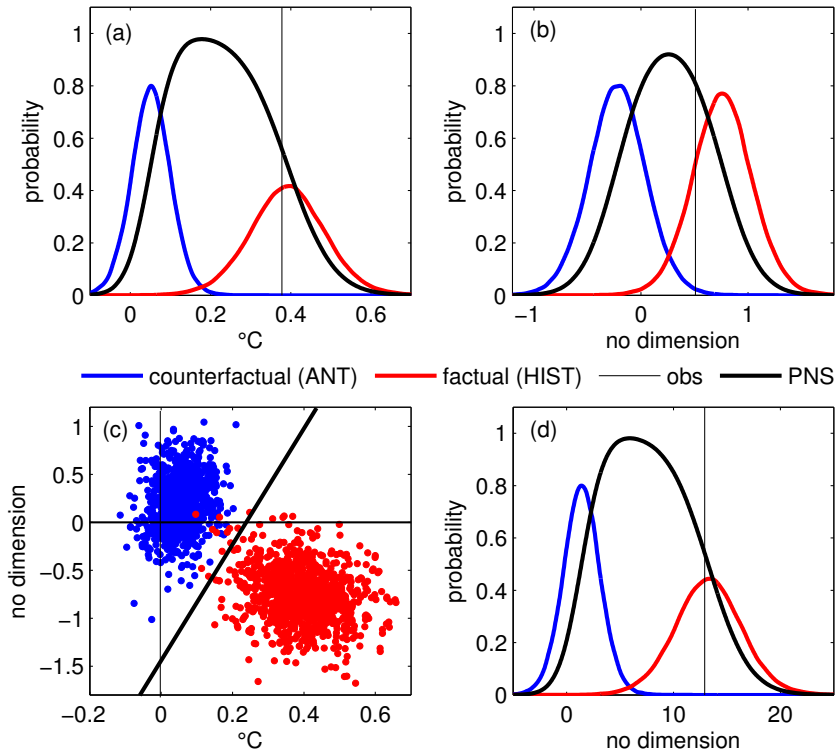


FIG. 7. Same as Figure 4 for the mapping ϕ^+ projected onto the leading eigenvectors of \mathbf{C} .

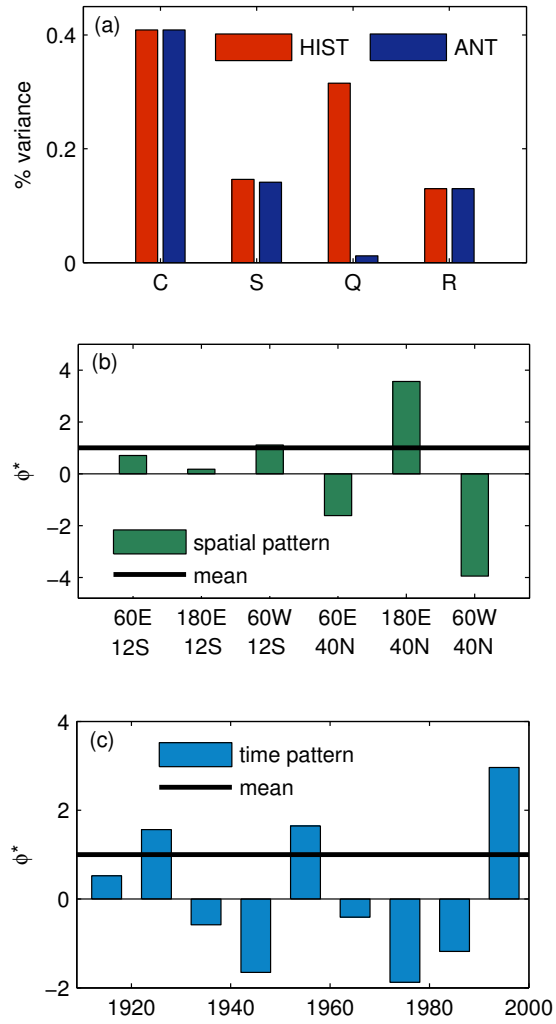


FIG. 8. Same as Figure 3 for the mapping ϕ^+ projected onto the leading eigenvectors of \mathbf{C} .