Bringing physical reasoning into statistical practice in climate-change science

Climatic Change (2021) 169:2
https://doi.org/10.1007/s10584-021-03226-6

Bringing physical reasoning into statistical practice in climate-change science

Theodore G. Shepherd[1]

Received: 7 June 2021 / Accepted: 16 September 2021

Abstract
The treatment of uncertainty in climate-change science is dominated by the far-reaching influence of the 'frequentist' tradition in statistics, which interprets uncertainty in terms of sampling statistics and emphasizes *p*-values and statistical significance. This is the normative standard in the journals where most climate-change science is published. Yet a sampling distribution is not always meaningful (there is only one planet Earth). Moreover, scientific statements about climate change are hypotheses, and the frequentist tradition has no way of expressing the uncertainty of a hypothesis. As a result, in climate-change science, there is generally a disconnect between physical reasoning and statistical practice. This paper explores how the frequentist statistical methods used in climate-change science can be embedded within the more general framework of probability theory, which is based on very simple logical principles. In this way, the physical reasoning represented in scientific hypotheses, which underpins climate-change science, can be brought into statistical practice in a transparent and logically rigorous way. The principles are illustrated through three examples of controversial scientific topics: the alleged global warming hiatus, Arctic-midlatitude linkages, and extreme event attribution. These examples show how the principles can be applied, in order to develop better scientific practice.

"La théorie des probabilités n'est que le bon sens reduit au calcul." (Pierre-Simon Laplace, *Essai Philosophiques sur les Probabilités*, 1819).
"It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude." (Harold Jeffreys, *Theory of Probability*, 1st edition, 1939).

Keywords Climate change · Statistics · Uncertainty · Inference · Bayes factor · Bayes theorem

This article belongs to the topical collection "*Perspectives on the quality of climate information for adaptation decision support*", edited by Marina Baldissera Pacchetti, Suraje Dessai, David A. Stainforth, Erica Thompson, and James Risbey.

✉ Theodore G. Shepherd
theodore.shepherd@reading.ac.uk

[1] Department of Meteorology, University of Reading, Reading RG6 6BB, UK

1 Introduction

As climate change becomes increasingly evident, not only in global indicators but at the local scale and in extreme events, the challenge of developing climate information for decision-making becomes more urgent. It is taken as given that such information should be based on sound science. However, it is far from obvious what that means. As with many other natural sciences, controlled experiments on the real climate system cannot be performed, and climate change is by definition statistically non-stationary. Together this means that scientific hypotheses cannot be tested using traditional scientific methods such as repeated experimentation. (Experiments can be performed on climate simulation models, but the models differ from the real world in important respects, and often disagree with each other.) On the global scale, it is nevertheless possible to make scientific statements with high confidence, and to speak of what can be considered to be effectively climate change 'facts' (e.g. the anthropogenic greenhouse effect, the need to go to net-zero greenhouse gas emissions in order to stabilize climate), which are sufficient to justify action on mitigation. This is because the process of spatial aggregation tends to reduce the relevant physical principles to energetic and thermodynamic ones which are anchored in fundamental theory (Shepherd 2019), and to beat down much of the climate noise so that the signals of change emerge clearly in the observed record (Sippel et al. 2020).

Yet for many aspects of climate-change science, there is no consensus on what constitutes fundamental theory, the signals of change are not unambiguously evident in the observed record, and climate models provide conflicting results. This situation occurs with so-called climate 'tipping points', due to uncertainties in particular climate feedbacks (Lenton et al. 2008). It also occurs on the space and time scales relevant for climate adaptation, where atmospheric circulation strongly determines climatic conditions, yet there is very little confidence in its response to climate change (Shepherd 2014). These uncertainties compound in the adaptation domain, where human and natural systems play a key role (Wilby and Dessai 2010).

This situation of ambiguous possible outcomes is illustrated by Fig. 1, which shows the precipitation response to climate change across the CMIP5 climate models as presented by IPCC (2013). Stippling indicates where the multi-model mean change is large compared to internal variability, and 90% of the models agree on the sign of change, whilst hatching indicates where the multi-model mean change is small compared to internal variability. These metrics embody the concept of 'statistical significance', which underpins the usual approach to uncertainty in climate-change science. Yet they are seen to be agnostic over many populated land regions, including most of the Global South, which are neither hatched nor stippled. Zappa et al. (2021) have shown that in those regions, the models suggest precipitation responses that are potentially large but are non-robust (i.e. uncertain in sign), and that the same situation holds with the CMIP6 models.

It follows that if climate-change science is to be informative for decision-making, it must be able to adequately reflect the considerable uncertainty that can exist in the information. The traditional language of science is usually framed in terms of findings, which for climate change might be explanations of past behaviour (attribution), or predictions of future behaviour (known as 'projections' when made conditional on the future climate forcing). To give just one example, the title of Sippel et al. (2020) is "Climate change now detectable from any single day of weather at global scale". In the peer-reviewed literature, these findings are generally presented in a definitive, unconditional manner (op. cit., where it is indeed justified); some journals even insist that the titles of their articles are worded that

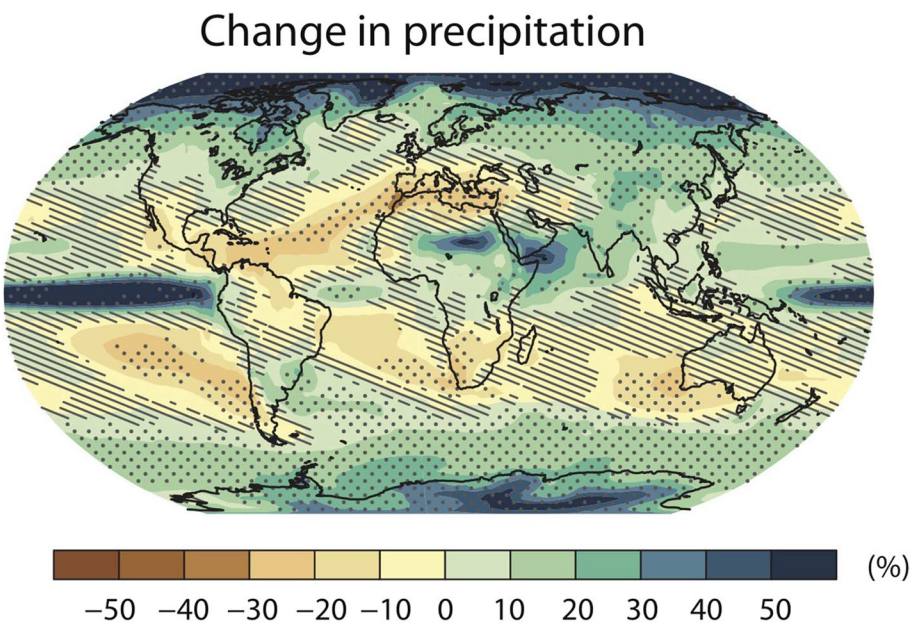

Fig. 1 Projected changes in precipitation (in %) over the twenty-first century from the CMIP5 models under a high climate forcing scenario (RCP8.5). Stippling indicates where the multi-model mean change is large compared with natural internal variability in 20-year means (greater than two standard deviations) and where at least 90% of models agree on the sign of change. Hatching indicates where the multi-model mean change is small compared with internal variability (less than one standard deviation). From the Summary for Policymakers of IPCC (2013)

way. Caveats on the findings are invariably provided, but it is not straightforward to convert those to levels of confidence in the finding. When made quantitative, the uncertainties are represented through some kind of error bar, usually around a best estimate. As Stirling (2010) has argued, such 'singular, definitive' representations of knowledge are inappropriate and potentially highly misleading when the state of knowledge is better described as 'plural, conditional', as for mean precipitation changes in the unmarked regions in Fig. 1. There are many methods available for dealing with 'plural, conditional' knowledge within a decision framework (Weaver et al. 2013; Rosner et al. 2014), so there is certainly no requirement for climate information to be expressed in a 'singular, definitive' manner in order to be useable.

There are many reasons for this situation, some of which are non-scientific (e.g. the reward system, both for authors and for journals). My goal here is to focus on one of the scientific reasons, namely the statistical practice that characterizes most climate-change science, which is still dominated by procedures that originate from the so-called 'frequentist' tradition in statistics. This tradition interprets uncertainty in terms of sampling statistics of a hypothetical population, and places a strong emphasis on *p*-values and statistical significance. It does not provide a language for expressing the probability of a hypothesis being true, nor does it provide a home for the concept of causality. Yet scientific reasoning is about hypotheses (including the 'findings' mentioned earlier), and reasoning under uncertainty is simply a form of extended logic, generalizing the true/false dichotomy of Aristotelian logic to situations where a hypothesis has a probability of being true that lies

between 0 and 1 (Jeffreys 1961; Jaynes 2003). Moreover, the concept of causality is central to physical science, as well as to decision-making since otherwise there is no connection between decisions and consequences, and causality has a logical formulation as well (Pearl and Mackenzie 2018; Fenton and Neil 2019). Those elements of physical reasoning are part of scientific practice in climate-change science, but are not connected to statistical practice in an explicit way. Thus, it seems crucial to bring these elements into the treatment of uncertainty.

In lay terms, probability is the extent to which something is likely to happen or to be the case. This includes frequency-based (or long-run) probabilities — the frequentist paradigm — as a special case, but it applies to single-outcome situations as well, such as a scientific hypothesis concerning climate change, where probability is interpreted as degree of belief. (For scientists, the word "belief" may cause some discomfort, but we can interpret belief as expert judgement, which is a widely accepted concept in climate-change science, including by the IPCC (Mastrandrea et al. 2011).) The two concepts of uncertainty are quite distinct, yet are commonly confused, even by practicing climate scientists. Even the use of frequency-based probabilities requires a degree of belief that they may be appropriately used for the purpose at hand, which is a highly non-trivial point when one is making statements about the real world. Jeffreys (1961) and Jaynes (2003) both argue that whilst the frequentist methods generally produce acceptable outcomes in the situations for which they were developed (e.g. agricultural trials, quality control in industry), which are characterized by an abundance of data and little in the way of prior knowledge, they are not founded in rigorous principles of probability (the 'extended logic' mentioned above, which is so founded (e.g. Cox 1946)), and are not appropriate for the opposite situation of an abundance of prior knowledge and little in the way of data. For climate-change science, especially (although not exclusively) in the adaptation context, we are arguably in the latter situation: we have extensive physical knowledge of the workings of the climate system and of the mechanisms involved in climate impacts, and very little data that measures what we are actually trying to predict, let alone under controlled conditions. This motivates a reappraisal of the practice of statistics in climate-change science. In this I draw particularly heavily on Jeffreys (1961), since he was a geophysicist and thus was grappling with scientific problems that have some commonality with our own.

This paper is aimed at climate scientists. Its goal is to convince them that the frequentist statistical methods that are standard in climate-change science should be embedded within a broader logical framework that can connect physical reasoning to statistical practice in a transparent way. Not only can this help avoid logical errors, it also provides a scientific language for representing physical knowledge even under conditions of deep uncertainty, thereby expanding the set of available scientific tools. In this respect, making explicit and salient the conditionality of any scientific statement is a crucial benefit, especially for adaptation where a variety of societal values come into play (Hulme et al. 2011). Note that I am not arguing for the wholesale adoption of Bayesian statistical methods, although these may have their place for particular problems (see further discussion in Sect. 4). Rather, I am simply arguing that we should follow Laplace's dictum and embed our statistical calculations in common sense, so as to combine them with physical reasoning. Section 2 starts by reprising the pitfalls of 'null hypothesis significance testing' (NHST); although the pitfalls have been repeatedly pointed out, NHST continues to be widespread in climate-change science, and its dichotomous misinterpretation reinforces the 'singular, definitive' representation of knowledge. Section 2 goes on to discuss how the concept of frequency fits within the broader concepts of probability and inference. Section 3 examines a spectrum of case studies: the alleged global warming hiatus, Arctic-midlatitude linkages, and extreme event

attribution. Together these illustrate how the principles discussed in Sect. 2 can be applied, in order to improve statistical practice. The paper concludes with a discussion in Sect. 4.

2 Back to basics

The ubiquitous use of NHST has been widely criticized in the published literature (e.g. Amrhein et al. 2019). To quote from the abstract of the psychologist Gerd Gigerenzer's provocatively titled paper 'Mindless statistics' (2004):

> Statistical rituals largely eliminate statistical thinking in the social sciences. Rituals are indispensable for identification with social groups, but they should be the subject rather than the procedure of science. What I call the 'null ritual' consists of three steps: (1) set up a statistical null hypothesis, but do not specify your own hypothesis nor any alternative hypothesis, (2) use the 5% significance level for rejecting the null and accepting your hypothesis, and (3) always perform this procedure.

Gigerenzer refers to the social sciences, but is it actually any different in climate science[1]? Nicholls (2000) and Ambaum (2010) both provide detailed assessments showing the widespread use of NHST in climate publications. This practice does not appear to have declined since the publication of those papers; indeed, my impression is that it has only increased, exacerbated by the growing dominance of the so-called 'high-impact' journals which enforce the statistical rituals with particular vigour, supposedly in an effort to achieve a high level of scientific rigour. Ambaum (2010) suggests that the practice may have been facilitated by the ready availability of online packages that offer significance tests as a 'black box' exercise, even though no serious statistician would argue that the practice of statistics should become a 'black box' exercise. I would add that Gigerenzer's insightful comment about "identification with social groups" may also apply to climate scientists, in that statistical rituals become a working paradigm for certain journals and reviewers. I suspect I am not alone in admitting that most of the statistical tests in my own papers are performed in order to satisfy these rituals, rather than as part of the scientific discovery process itself.

Gigerenzer (2004) shows that NHST, as described above, is a bastardized hybrid of Fischer's null hypothesis testing and Neyman–Pearson decision theory, and has no basis even in orthodox frequentist statistics. According to Fischer, a null hypothesis test should only be performed in the absence of any prior knowledge, and before one has even looked at the data, neither of which applies to the typical applications in climate science. Violation of these conditions leads to the problem known as 'multiple testing'. Moreover, failure to reject the null hypothesis does not prove the null hypothesis, nor does rejection of the null hypothesis prove an alternative hypothesis. Yet, these inferences are routinely made in climate science, and the oxymoronic phrase "statistically significant trend" is commonplace.

Amrhein et al. (2019) argue that the main problem lies in the dichotomous interpretation of the result of a NHST — i.e. as the hypothesis being either true or false depending on the p-value — and they argue that the concept of statistical significance should be dropped entirely. (Their comment gathered more than 800 signatories from

[1] I sometimes use the term "climate science", because the points I make are applicable to climate science in general, but use "climate-change science" when I wish to emphasize that particular aspect of climate science.

researchers with statistical expertise.) Instead, they argue that all values of a sampling statistic that are compatible with the data should be considered as plausible; in particular, two studies are not necessarily inconsistent simply because one found a statistically significant effect and the other did not (which, again, is a common misinterpretation in climate science). This aligns with Stirling's (2010) warning, mentioned earlier, against 'singular, definitive' representations of knowledge when the reality is more complex, and all I can do in this respect is urge climate scientists to become aware of the sweeping revolution against NHST in other areas of science. Instead, I wish to focus here on bringing physical reasoning into statistical practice, which is of particular relevance to climate-change science for the reasons discussed earlier.

Misinterpretation of NHST is rooted in the so-called 'prosecutor's fallacy', which is the transposition of the conditional. The p-value quantifies the probability of observing the data D under the null hypothesis H that the apparent effect occurred by chance. This is written $P(D|H)$, sometimes called the likelihood function, and is a frequentist calculation based on a specified probability model for the null hypothesis, which could be either theoretical or empirical. (As noted earlier, the specification of an appropriate probability model is itself a scientific hypothesis, but let us set that matter aside for the time being.) However, one is actually interested in the probability that the apparent effect occurred by chance, which is $P(H|D)$. The two quantities are not the same, but are related by Bayes' theorem:

$$P(H|D) = \frac{P(H)}{P(D)} P(D|H). \tag{1}$$

To illustrate the issue, consider the case where H is not the null hypothesis but is rather the hypothesis that one has a rare illness, having tested positive for the illness (data D). Even if the detection power of the test is perfect, i.e. $P(D|H) = 1$, a positive test result may nevertheless indicate only a small probability of having the illness, i.e. $P(H|D)$ being very small, if the illness is sufficiently rare and there is a non-negligible false alarm rate, such that $P(H) \ll P(D)$. This shows the error that can be incurred from the transposition of the conditional if one does not take proper account of prior probabilities. In psychology, it is known as 'base rate neglect' (Gigerenzer and Hoffrage 1995).

The example considered above of the medical test is expressed entirely in frequentist language, because the probability of the test subject having the illness (given no other information, and before taking the test), $P(H)$, is equated to the base rate of the illness within the general population, which is a frequency. However, this interpretation is not applicable to scientific hypotheses, for which the concept of a 'long run' frequency is nonsensical. To consider this situation, we return to the case of H being the null hypothesis and write Bayes' theorem instead as

$$P(H|D) = \frac{P(D|H)}{P(D)} P(H). \tag{2}$$

Equation (2) is mathematically equivalent to Eq. (1) but has a different interpretation. Now the probability of the apparent effect having occurred by chance, $P(H|D)$, is seen to be the prior probability of there being no real effect, $P(H)$, multiplied by the factor $P(D|H)/P(D)$. The use of Bayes' theorem in this way is often criticized for being sensitive to the prior $P(H)$. However, expert (prior) knowledge is also used in the formulation (1) to determine how to control for confounding factors and for other aspects of the statistical analysis, and it is widely used in climate-change science to determine how much weight to

place on different pieces of evidence. It is thus a strength, rather than a weakness, of Bayes' theorem that it makes this aspect of the physical reasoning explicit.

The factor $P(D|H)/P(D)$ in Eq. (2) represents the power of the data for adjusting one's belief in the null hypothesis. But whilst $P(D|H)$ is the p-value, we have another factor, $P(D)$; how to determine it? This can only be done by considering the alternative hypotheses that could also account for the data D. We write $\neg H$ as the complement of H, so that $P(\neg H) = 1 - P(H)$. (In practice, $\neg H$ should be enumerated over all the plausible alternative hypotheses.) From the fundamental rules of probability,

$$P(D) = P(D|H)P(H) + P(D|\neg H)P(\neg H), \tag{3}$$

which can be substituted into Eq. (2). Thus, we can eliminate $P(D)$, but only at the cost of having to determine $P(D|\neg H)$. In that case, it is simpler to divide Eq. (2) by the same expression with H replaced by $\neg H$, which eliminates $P(D)$ and yields the 'odds' version of Bayes' theorem:

$$\frac{P(H|D)}{P(\neg H|D)} = \frac{P(D|H)}{P(D|\neg H)} \frac{P(H)}{P(\neg H)}. \tag{4}$$

This states that the odds on the data occurring by chance — the left-hand side of Eq. (4) — equal the prior odds of the null hypothesis multiplied by the first term on the right-hand side of Eq. (4), which is known as the Bayes factor (Kass and Raftery 1995) and was heavily used by Jeffreys (1961). The deviation of the Bayes factor from unity represents the power of the data for discriminating between the null hypothesis and its complement. (Note that Eq. (4) holds for any two hypotheses, but its interpretation is simpler when the two hypotheses are mutually exclusive and exhaustive, as here.) One of the attractive features of the Bayes factor is that it does not depend on the prior odds, and is amenable to frequentist calculation when the alternative hypothesis can be precisely specified.

The formulation (4) represents in a clear way the aphorism that 'strong claims require strong evidence': if the prior odds of the null hypothesis are very high, then it requires a very small Bayes factor to reject the null hypothesis. But Eq. (4) makes equally clear that the power of the data is represented not in the p-value $P(D|H)$ but rather in the Bayes factor, and that failure to consider the Bayes factor is a serious error in inference. To quote Jeffreys (1961, p. 58):

> We get no evidence for a hypothesis by merely working out its consequences and showing that they agree with some observations, because it may happen that a wide range of other hypotheses would agree with those observations equally well. To get evidence for it we must also examine its various contradictories and show that they do not fit the observations.

Thus, for both reasons, the p-value $P(D|H)$ on its own is useless for inferring the probability of the effect occurring by chance, and thus for rejecting the null hypothesis, even though this is standard practice in climate science. Rather, we need to consider both the prior odds of the null hypothesis, and the p-value for the alternative hypothesis, $P(D|\neg H)$. We will discuss the implications of this in more detail in Sect. 3 in the context of specific climate-science examples. Here, we continue with general considerations. With regard to the difference between the p-value $P(D|H)$ and the Bayes factor, Bayesian statisticians have ways of estimating $P(D|\neg H)$ in general, and the outcome is quite shocking. Nuzzo (2014), for example, estimates that a p-value of 0.05 generally corresponds to a Bayes factor of only 0.4 or so, almost 10 times larger. The reason why the p-value can differ so much

from the Bayes factor is because the latter penalizes imprecise alternative hypotheses, which are prone to overfitting. The difference between the two is called the 'Ockham factor' by Jaynes (2003, Chapter 20), in acknowledgement of Ockham's razor in favour of parsimony: "The onus of proof is always on the advocate of the more complicated hypothesis" (Jeffreys 1961, p. 343). The fact that such a well-established principle of logic is absent from frequentist statistics is already telling us that the latter is an incomplete language for describing uncertainty.

It follows that in order for a p-value of 0.05 to imply a 5% likelihood of a false alarm (i.e. no real effect) — which is the common misinterpretation — the alternative hypothesis must already be a good bet. For example, Nuzzo (2014) estimates a 4% likelihood of a false alarm when $P(H) = 0.1$, i.e. the null hypothesis is already considered to be highly improbable. For a toss-up with $P(H) = 0.5$, the likelihood of a false alarm given a p-value of 0.05 rises to nearly 30%, and it rises to almost 90% for a long-shot alternative hypothesis with $P(H) = 0.95$. Yet despite this enormous sensitivity of the inferential power of a p-value to the prior odds of the null hypothesis, nowhere in any climate science publication have I ever seen a discussion of prior odds (or probabilities) entering the statistical interpretation of a p-value.

In fact, in much published climate science, the alternative hypothesis to the null is already a good bet, having been given plausibility by previous research or by physical arguments advanced within the study itself. In other words, the statistical analysis is only confirmatory, and the p-value calculation performed merely as a sanity check. However, it is important to understand the prior knowledge and assumptions that go into this inference. For transparency, they should be made explicit, and a small p-value should in no way be regarded as a 'proof' of the result.

There is an exception to the above, when the data does strongly decide between the two hypotheses. This occurs in the case of detection and attribution of anthropogenic global warming, where the observed warming over the instrumental record can be shown to be inconsistent with natural factors, and fully explainable by anthropogenic forcing (e.g. IPCC 2013). In that case, the Bayes factor is very small, and a strong inference is obtained without strong prior assumptions (mainly that all potential explanatory factors have been considered). However, this 'single, definitive' situation is generally restricted to thermodynamic aspects of climate change on sufficiently coarse spatial and temporal scales (Shepherd 2014).

A similar issue arises with confidence intervals. The frequentist confidence interval represents the probability distribution of a sampling statistic of a population parameter. However, it does not represent the likely range of the population parameter, known as the 'credible interval' (Spiegelhalter 2018). To equate the two, as is commonly done in climate science publications, is to commit the error of the transposed conditional. In particular, it is common to assess whether the confidence interval around a parameter estimate excludes the null hypothesis value for that parameter, as a basis for rejecting the null hypothesis. This too is an inferential error. However, the confidence interval can approximately correspond to the credible interval if a wide range of prior values are considered equally likely (Fenton and Neil 2019, Chap. 12), which is effectively assuming that the null hypothesis value (which is only one such value) is highly unlikely. Thus, once again, provided we are prepared to acknowledge that we are assuming the null hypothesis to be highly unlikely, the use of a frequentist confidence interval may be acceptable.

There is one more general point that is worth raising here before we go on to the examples. In most climate science, the use of 'statistical rituals' means that particular statistical metrics (such as the p-value) are used without question. However, statisticians well

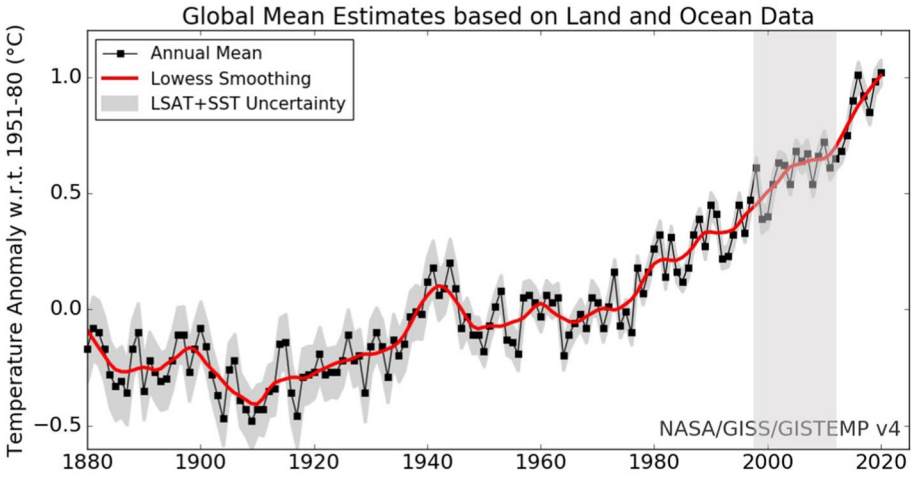

Fig. 2 NASA GISTEMP time series of estimated observed annual global mean surface air temperature expressed as anomalies relative to the 1951–1980 reference period. The red line is a smoothed version of the time series. The grey band (added here) indicates the time period 1998–2012, which was defined as the hiatus period in Box TS.3 of IPCC (2013). From https://data.giss.nasa.gov/gistemp/, downloaded 31 May 2021

appreciate that every statistical metric involves a trade-off, and that the choice will depend on the decision context. For example, in forecasts, there is a trade-off between discrimination and reliability, and in parameter estimation, there is a trade-off between efficiency and bias. There is no objective basis for how to make those trade-offs. Part of better statistical practice in climate-change science is to recognize these trade-offs and acknowledge them in the presentation of the results.

3 Examples

3.1 The alleged global warming hiatus

The alleged global warming 'hiatus' was the apparent slowdown in global-mean warming in the early part of the twenty-first century. Looking at it now (Fig. 2), it is hard to see why it attracted so much attention. Yet it was the focus of much media interest, and a major challenge for the IPCC during the completion of the AR5 WGI report, in 2013. There are good scientific reasons to try to understand the mechanisms behind natural variability in climate, and there was also a question at the time of whether the climate models were overestimating the warming response to greenhouse gases. However, the media attention (and the challenge for the IPCC) was mostly focused on whether climate change had weakened, or even stopped — as many climate sceptics claimed (Lewandowsky et al. 2016). We focus here on that specific question.

Given that there is high confidence in the basic physics of global warming, a reasonable null hypothesis would have been that the hiatus was just the result of natural variability. Then the logical thing to have done would have been to determine the Bayes factor comparing the hypothesis of continued climate change to that of a cessation to climate change. If the Bayes factor was of order unity, then the data would not have differentiated

between the two hypotheses; in other words, the hiatus would have been entirely consistent with natural variability together with continued long-term climate change, and there would have been no need to adjust the prior hypothesis (which would have to have been given an extremely high likelihood, given previous IPCC reports).

Yet such an approach was not taken. Instead, there were many published studies examining the statistical significance of the observed trends, and much attention in the technical summary of the AR5 WGI report (Box TS.3 of IPCC 2013) was devoted to the hiatus, which was defined by IPCC to be the period 1998–2012 (grey shading in Fig. 2). The fact that small adjustments to the data sets could make the difference between statistical significance or not (Cowtan and Way 2014) should have raised alarm bells that this frequentist-based approach to the data analysis, with multiple testing together with transposing the conditional to make inferences about physical hypotheses, was ill-founded. Completely ignoring all the knowledge from previous IPCC reports in the statistical assessment was also somewhat perverse, given that our confidence in the physics of anthropogenic global warming does not rest on observed warming alone, let alone warming over a 14-year period.

Rahmstorf et al. (2017) revisited the hiatus controversy, and deconstructed many of the published analyses of the hiatus, showing how they fell into many of the pitfalls discussed in Sect. 2. A particularly egregious one is the selection bias that arises from selectively focusing on a particular period and ignoring the others (also known as the 'multiple testing problem'), which is apparent by eye from Fig. 2. They also showed that a more hypothesis-driven approach to the data analysis would have deduced that there was nothing unusual about the hiatus, which is equivalent to saying that the Bayes factor would have been close to unity. (That is even before bringing prior odds into the picture.) An independent and very interesting confirmation of this result is the study of Lewandowsky et al. (2016), which took the observed global-mean temperature time series (up to 2010), relabelled it as "World Agricultural Output", and asked non-specialists whether they saw any weakening of the trend. The answer was a resounding no. This appears to show the power of the human brain for separating signal from noise, much more reliably than frequentist-based analysis methods.

I cannot resist pointing out that Fig. 10.6 of the IPCC AR5 WGI report showed clearly that the hiatus was entirely explainable from a combination of ENSO variability and the decline in solar forcing, even in the presence of continued anthropogenic warming. To this day, I still cannot understand why the IPCC chose to ignore this piece of evidence in its discussion of the hiatus, relying instead on purely statistical analyses without the incorporation of the huge amount of knowledge within the WGI report itself. When I asked someone about this, the answer I got was that Fig. 10.6 did not "prove" the case. But that's not the point. Given all the knowledge that existed, it was surely sufficient to show that no other explanation was needed. To again draw on Jeffreys (1961, p. 342):

> Variation is random until a contrary is shown; and new parameters in laws, when they are suggested, must be tested one at a time unless there is specific reason to the contrary.

3.2 Arctic-midlatitude connections

The Arctic amplification of global warming is a robust aspect of climate change, and the observed decline in Arctic sea-ice extent is its poster-child. The sea-ice decline is largest in the summer season, but the additional warmth in that season is absorbed by the colder

ocean and released back to the atmosphere during winter, when the atmosphere is colder. Hence, Arctic amplification is mainly manifest in the winter season. Based on observed trends, Francis and Vavrus (2012) made the claim that Arctic amplification led to a wavier jet stream, causing more extreme winter weather at midlatitudes, including (somewhat counterintuitively in a warming climate) more cold spells. This claim has subsequently generated heated debate within the scientific community, and is an extremely active area of research (e.g. Screen et al. 2018; Cohen et al. 2020).

In contrast to the example of the hiatus, here, the prior knowledge is not very strong. Much of the evidence that is cited in favour of the claim of Arctic-to-midlatitude influence is from congruent trends in observational time series. However, the waviness of the jet stream will itself induce Arctic winter warming through enhanced poleward heat transport (see Shepherd 2016a), so any attempt to isolate the causal influence of Arctic warming on midlatitude weather must control for this opposing causal influence. Kretschmer et al. (2016) used the time lags of the various hypothesized physical processes to do this from observations, using a causal network framework, and inferred a causal role for sea ice loss in the Barents–Kara seas inducing midlatitude atmospheric circulation changes that are conducive to cold spells. This approach builds in prior knowledge to constrain the statistical analysis. Mostly, however, researchers have used frequentist-based methods applied to the change in long-term trends since 1990, when Arctic warming appeared to accelerate. This places a lot of weight on what from a climate perspective are relatively short time series (which *is* similar to the hiatus situation). Moreover, climate change affects both midlatitude conditions and the Arctic, representing a common driver and thus a confounding factor for any statistical analysis (Kretschmer et al. 2021). The theoretical arguments for a wavier jet stream are heuristic, and more dynamically based considerations are inconclusive (Hoskins and Woollings 2015). Climate models provide inconsistent responses, and there are certainly good reasons to question the fidelity of climate models to capture the phenomenon, given the fact they are known to struggle with the representation of persistent circulation anomalies such as blocking. Overall, there are certainly sufficient grounds to develop plausible physical hypotheses of Arctic-midlatitude linkages, even if not through the Francis and Vavrus (2012) mechanism. Indeed, several large funding programmes have been established to explore the question.

Yet with all this uncertainty, it is difficult to understand how the published claims can be so strong, on both sides. Whilst the whiplash of conflicting claims may help generate media attention, it must be very confusing for those who want to follow the science on this issue. Adopting a more 'plural, conditional' perspective would surely be helpful, and much more representative of the current state of knowledge. Kretschmer et al. (2020) examined the previously hypothesized link between Barents–Kara sea-ice loss (where the changes are most dramatic) and changes in the strength of the stratospheric polar vortex — known to be a causal factor in midlatitude cold spells (Kretschmer et al. 2018) and a major driver of uncertainty in some key wintertime European climate risks (Zappa and Shepherd 2017) — across the CMIP5 models. They found that the link in the models was so weak as to be undetectable in the year-to-year variability, which means that it will be difficult to find Bayes factors between the hypothesis of a causal influence and the null hypothesis of no such influence that are very informative. Yet even such a weak link had major implications for the forced changes, given the large extent of projected Barents–Kara sea-ice loss (essentially 100%) compared to other changes in the climate system. Returning to Eq. (4), the weakness of the link may help explain why the scientific findings in this subject seem to be so closely linked to scientists' prior beliefs. There would be nothing wrong with that so long as those beliefs were made explicit, which would happen naturally if scientists also

considered all plausible alternative hypotheses, as Eq. (4) obliges them to do. (The quote given earlier from Jeffreys (1961, p. 58) is relevant here.) Alternatively, one can present the different hypotheses in conditional form, as storylines, allowing the user of the information to impose their own beliefs (Shepherd 2019). This is useful for decision-making since within probability theory, beliefs can incorporate consequences (Lindley 2014). Once again, there is a relevant quote from Jeffreys (1961, p. 397):

> There are cases where there is no positive evidence for a new parameter, but important consequences might follow if it was not zero, and we must remember that [a Bayes factor] > 1 does not prove that it is zero, but merely that it is more likely to be zero than not. Then it is worth while to examine the alternative [hypothesis] further and see what limits can be set to the new parameter, and thence to the consequences of introducing it.

3.3 Extreme event attribution

Since weather and climate extremes have significant societal impacts, it is no surprise that many of the most severe impacts of climate change are expected to occur through changes in extreme events. If climate is understood as the distribution of all possible meteorological states, then the effect of climate change on extreme events is manifest in the changes in that distribution. This is the subject of a large literature. Over the last 20 years, the different topic of extreme event attribution has emerged, which seeks to answer the question of whether, or how, a particular extreme event can be attributed to climate change. In contrast to the two previous examples, which concerned clear climate-science questions, here, it is far from obvious how to even pose the question within a climate-science framework, since every extreme event is unique (NAS 2016). This 'framing' question of how to define the event raises its own set of issues for statistical practice and scientific reasoning.

The most popular approach, first implemented by Stott et al. (2004) for the 2003 European heat wave, has been to estimate the probability of an event at least as extreme as the observed one occurring (quantified in a return period), under both present-day and pre-industrial conditions, and attributing the change in probability to climate change. This is done by defining an 'event class' (typically univariate, and at most bivariate) which is sufficiently sharp to relate to the event in question, but sufficiently broad to allow a frequency-based calculation of probability. Clearly, there is a trade-off involved here, which will depend on a variety of pragmatic factors. For example, in Stott et al. (2004), the event was defined by the average temperature over a very large region encompassing Southern Europe, over the entire summer period (June through August), for which the observed extreme was only 2.3 °C relative to preindustrial conditions, and around 1.5 °C relative to the expected temperature in 2003. Such an anomaly was very rare for that highly aggregated statistic, but clearly nobody dies from temperatures that are only 1.5 °C above average. Given that this 'probabilistic event attribution' (PEA) is based on a frequentist definition of probability, along with related concepts such as statistical significance, it is worth asking how a widening of the perspective of probability and reasoning under uncertainty, along the lines described in this paper, might enlarge the set of scientific tools that are available to address this important scientific topic.

The first point to make is that from the more general perspective of probability theory discussed in Sect. 2, there is no imperative to adopt a frequentist interpretation of probability. As Jeffreys says (1961, p. 401), 'No probability….is simply a frequency'. A frequency is at best a useful mathematical model of unexplained variability. The analogy that is often made of increased risk from climate change is that of loaded dice. But if a die turns up 6,

whether loaded or unloaded, it is still a 6. On the other hand, if an extreme temperature threshold is exceeded only very rarely in pre-industrial climate vs quite often in present-day climate, the nature of these exceedances will be different. One is in the extreme tail of the distribution, and the other is not, so they correspond to very different meteorological situations and will be associated with very different temporal persistence, correlation with other fields, and so on. Since pretty much every extreme weather or climate event is a compound event in one way or another, this seems like quite a fundamental point. It is perfectly sensible to talk about the probability of a singular event, so we should not feel obliged to abandon that concept.

The fact is that climate change changes *everything*; the scientific question is not whether, but how and by how much. When the null hypothesis is logically false, as here, use of NHST is especially dangerous. Following Ockham's razor, the more relevant question is whether a particular working hypothesis (which would then be the null hypothesis) is enough to provide a satisfactory answer to the question at hand. As noted, most extreme weather and climate events are associated with unusual dynamical conditions conducive to that event, which we denote generically by N. The event itself we denote by E. An example event might be a heat wave, for which N could be atmospheric blocking conditions; or a drought, for which N could be the phase of ENSO. The effect of climate change can then be represented as the change in the joint probability $P(E,N)$ between present-day, or factual (subscript f) conditions, and the same conditions without climate change, which are a counter-factual (subscript c), expressed as a risk ratio. From NAS (2016),

$$\frac{P_f(E,N)}{P_c(E,N)} = \frac{P_f(E|N)}{P_c(E|N)} \frac{P_f(N)}{P_c(N)}. \tag{5}$$

This simple equation, which is based on the fundamental laws of probability theory, shows that the risk ratio factorizes into the product of two terms. The first is a ratio of conditional probabilities, namely the change in probability for a given dynamical conditioning factor N. The second expresses how the probability of the conditioning factor might itself change.

The scientific challenge here is that for pretty much any relevant dynamical conditioning factor for extreme events, there is very little confidence in how it will change under climate change (Shepherd 2019). This lack of strong prior knowledge arises from a combination of small signal-to-noise ratio in observations, inconsistent projections from climate models, and the lack of any consensus theory. If one insists on a frequentist interpretation of this second factor, as in PEA, then this can easily lead to inconclusive results, and that is indeed what tends to happen for extreme events that are not closely tied to global-mean warming (NAS 2016). But there is an alternative. We can instead interpret the second factor on the right-hand side of (5) as a degree of belief — which is far from inappropriate, given that the uncertainty here is mainly epistemic — and consider various hypotheses, or storylines (Shepherd 2016b). The simplest hypothesis is that the second factor is unity, which can be considered a reasonable null hypothesis. One should of course be open to the possibility that the second factor differs from unity, but in the absence of strong prior knowledge in that respect, that uncertainty would be represented by a prior distribution centred around unity. The advantage of this partitioning is that the first term on the right-hand side of Eq. (5) is generally much more amenable to frequentist quantification than is the second, and if the dynamical conditioning is sufficiently strong, it tends to focus the calculation on the thermodynamic aspects of climate change about which there is comparatively much greater confidence.

It seems worth noting that this approach is very much in line with the IPCC's guidance on the treatment of uncertainty (Mastrandrea et al. 2011), which only allows a probabilistic quantification of uncertainty when the confidence levels are high.

This approach is actually used implicitly in much PEA. For example, in the analogue method (e.g. Cattiaux et al. 2010), atmospheric circulation regimes are used for N, and when using large-ensemble atmosphere-only models (as in Stott et al. 2004), sea-surface temperature anomalies are used for N. Both methods have been considered as perfectly acceptable within the PEA framework (Stott et al. 2016), despite effectively assuming that the second factor in Eq. (5) is unity. This assumption is very often not even discussed, and if it is, the argument is typically made that there is no strong evidence in favour of a value other than unity (see e.g. van Oldenborgh et al. 2021 for a recent example). Yet for some reason, when exactly the same approach was proposed for the detailed dynamical situation of a highly unusual meteorological configuration (Trenberth et al. 2015), it was argued by the PEA community that it was invalid scientific reasoning. For example, Stott et al. (2016, p. 33) say:

> By always finding a role for human-induced effects, attribution assessments that only consider thermodynamics could overstate the role of anthropogenic climate change, when its role may be small in comparison with that of natural variability, and do not say anything about how the risk of such events has changed.

There is a lack of logical consistency here. First, since climate change has changed everything, at least to some degree, there is nothing logically wrong with "always finding a role for human-induced effects". Second, this approach is not biased towards overstating the role of anthropogenic climate change, as it could equally well understate it. As Lloyd and Oreskes (2018) have argued, whether one is more concerned about possible overstatement or understatement of an effect is not a scientific matter, but one of values and decision context. Third, "small compared with natural variability" can be consistent with an effect of anthropogenic climate change. For example, in van Garderen et al. (2021), global spectral nudging was used to apply the storyline approach to the 2003 European and 2010 Russian heat waves. The study clearly showed that the anthropogenic warming was small in magnitude compared to the natural variability that induced the heat waves, but the high signal-to-noise ratio achieved through the approach provided a quantitative attribution at very fine temporal and spatial scales, potentially allowing for reliable impact studies (and avoiding the need to choose an arbitrary 'event class', which blurs out the event). Finally, whilst it is true that the approach does not say anything about how the risk of such events has changed, I am not aware of a single PEA study that has a definitive attribution of changes in the conditioning factor N leading to the event, so they are subject to exactly the same criticism. Instead, the attribution in PEA studies is invariably explained in terms of well-understood thermodynamic processes. That seems like a pretty good justification for the storyline approach. In this way, the two approaches can be very complementary (see Table 2 of van Garderen et al. 2021). And if there are strong grounds for considering changes in dynamical conditions (as in Schaller et al. 2016, where the change in flood risk in the Thames Valley changed sign depending on the modelled circulation changes), then probability theory, as in Eq. (5), provides the appropriate logical framework for considering this question in a hypothesis-driven manner, through storylines of circulation change (Zappa and Shepherd 2017). In such cases, a 'plural, conditional' perspective is called for.

Yet again, there is a relevant quote from Jeffreys (1961, p. 302):

> In induction there is no harm in being occasionally wrong; it is inevitable that we shall be. But there is harm in stating results in such a form that they do not represent

the evidence available at the time when they are stated, or make it impossible for future workers to make the best use of that evidence.

4 Discussion

In an application where there is little in the way of prior knowledge, and a lot of data, the Bayes factor rapidly overpowers the influence of the prior knowledge, and the result is largely insensitive to the prior. However, many aspects of climate-change science, especially (although not exclusively) in the adaptation context, are in the opposite situation of having a large amount of prior knowledge, and being comparatively data-poor (in terms of data matching what we are actually trying to predict). In particular, the observed record provides only a very limited sample of what is possible, and is moreover affected by sources of non-stationarity, many of which may be unknown. Larger data sets can be generated from simulations using climate models, but those models have many failings, and it is far from clear which aspects of model simulations contain useful information, and which do not. Physical reasoning is therefore needed at every step. In such a situation, using statistical methods that eschew physical reasoning and prior knowledge — "letting the data speak for itself", some might say — is a recipe for disaster. Statistical practice in climate-change science simply has to change.

A statistician might at this point argue that the answer is to use Bayesian statistics. Indeed, Bayesian methods are used in particular specialized areas of climate science, such as inverse methods for atmospheric sounding (Rodgers 2000) including pollution-source identification (Palmer et al. 2003), sea-level and ice-volume variations on palaeoclimate timescales (Lambeck et al. 2014), and climate sensitivity (Sherwood et al. 2020). Mostly, this involves introducing prior probability distributions on the estimated parameters, but Sherwood et al. (2020) discuss the constraints on climate sensitivity in terms of the confidence that can be placed in physical hypotheses. There have been brave attempts to employ Bayesian methods more widely, e.g. in the UK climate projections (Sexton et al. 2012). The difficulty is that Bayesian calibration for climate-change projections requires knowing the relationship between model bias in present-day climate (which is measurable) and the spread in a particular aspect of model projections. Such a relationship is known as an 'emergent constraint' (Hall et al. 2019), and it has been recognized from the outset that in order to be predictive, it must be causal. Given the huge number of potential relationships, data mining can easily lead to spurious but apparently statistically significant relationships (Caldwell et al. 2014), and correlations can also reflect common drivers. Indeed, several published emergent constraints have subsequently been debunked (by Pithan and Mauritsen 2013; Simpson and Polvani 2016; Caldwell et al. 2018), and the field is something of a Wild West. Hall et al. (2019) emphasize the crucial importance of anchoring emergent constraints in physical mechanisms, and argue that emergent constraints are most likely to be found when those mechanisms are direct and linear. This may help explain why it has been so challenging to find emergent constraints for circulation aspects of climate change (relevant for adaptation), since there is no consensus on the relevant mechanisms and the circulation responses appear to involve multiple interacting factors, and potential nonlinearity.

For climate information to be useable, its uncertainties must be comprehensible and salient, especially in the face of apparently conflicting sources of information, and the connection between statistical analysis and physical reasoning must be explicit rather than implicit. This argues for bringing the Bayesian spirit of hypothesis testing more explicitly

into our scientific reasoning, forgoing the 'mindless' performance of statistical rituals as a substitute for reasoning, resisting true/false dichotomization, and being ever vigilant for logical errors such as multiple testing and the transposed conditional. As a recent Nature editorial states (Anonymous 2019), "Looking beyond a much used and abused measure [statistical significance] would make science harder, but better." Yet we can still use familiar statistical tools, such as *p*-values and confidence intervals, so long as we remember what they do and do not mean. They are useful heuristics, which researchers have some experience interpreting. And we need to make sure that we are not chasing phantoms.

Neuroscience has shown that human decision-making cannot proceed from facts alone but involves an emotional element, which provides a narrative within which the facts obtain meaning (Damasio 1994). Narratives are causal accounts, which in the scientific context can be regarded as hypotheses. To connect physical reasoning and statistical practice, these narratives need to run through the entire scientific analysis, not simply be a 'translation' device bolted on at the end. To return to the quote from Jeffreys at the beginning of this piece, we need to recognize that data does not speak on its own; there is no answer without a question, and the answer depends not only on the question but also on how it is posed.

Acknowledgements The author acknowledges the support provided through the Grantham Chair in Climate Science at the University of Reading. He is grateful to Michaela Hegglin, Marlene Kretschmer, Michael McIntyre, and Marina Baldisserra Pacchetti, as well as the two reviewers, for comments on an earlier version of this manuscript.

Author contribution Not applicable.

Availability of data and materials Not applicable.

Code availability Not applicable.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The author declares no competing interests.

References

Ambaum MHP (2010) Significance tests in climate science. J Clim 23:5927–5932
Amrhein V, Greenland S, McShane B (2019) Retire statistical significance. Nature 567:305–307

Anonymous (2019) It's time to talk about ditching statistical significance. Nature 567:283 (online version)
Caldwell PM, Bretherton CS, Zelinka MD, Klein SA, Santer BD, Sanderson BM (2014) Statistical significance of climate sensitivity predictors obtained by data mining. Geophys Res Lett 41:1803–1808
Caldwell PM, Zelinka MD, Klein SA (2018) Evaluating emergent constraints on equilibrium climate sensitivity. J Clim 31:3921–3942
Cattiaux J, Vautard R, Cassou C, Yiou P, Masson-Delmotte V, Codron F (2010) Winter 2010 in Europe: a cold extreme in a warming climate. Geophys Res Lett 37:L20704
Cohen J, Zhang X, Francis J, Jung T, Kwok R, Overland J, Ballinger TJ, Bhatt US, Chen HW, Coumou D, Feldstein S, Gu H, Handorf D, Henderson G, Ionita M, Kretschmer M, Laliberte F, Lee S, Linderholm HW, Maslowski W, Peings Y, Pfeiffer K, Rigor I, Semmler T, Stroeve J, Taylor PC, Vavrus S, Vihma T, Wang S, Wendisch M, Wu Y, Yoon J (2020) Divergent consensuses on Arctic amplification influence on midlatitude severe winter weather. Nature Clim Chang 10:20–29
Cowtan K, Way RG (2014) Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. Quart J Roy Meteor Soc 140:1935–1944
Cox RT (1946) Probability, frequency and reasonable expectation. Amer J Phys 14:1–13
Damasio A (1994) Descartes' error. G. P. Putnam's Sons
Fenton N, Neil M (2019) Risk assessment and decision analysis with Bayesian networks, 2nd edn. CRC Press
Francis JA, Vavrus SJ (2012) Evidence linking Arctic amplification to extreme weather in mid-latitudes. Geophys Res Lett 39:L06801
Gigerenzer G (2004) Mindless statistics. J Socio-Econom 33:587–606
Gigerenzer G, Hoffrage U (1995) How to improve Bayesian reasoning without instructions: frequency formats. Psychol Rev 102:684–704
Hall A, Cox P, Huntingford C, Klein S (2019) Progressing emergent constraints on future climate change. Nature Clim Chang 9:269–278
Hoskins B, Woollings T (2015) Persistent extratropical regimes and climate extremes. Curr Clim Chang Rep 1:115–124
Hulme M, O'Neill SJ, Dessai S (2011) Is weather event attribution necessary for adaptation funding? Science 334:764–765
IPCC (Intergovernmental Panel on Climate Change) (2013) Climate change 2013: the physical basis. Stocker TF et al. (eds) Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK
Jaynes ET (2003) Probability theory: the logic of science. Cambridge University Press
Jeffreys H (1961) The theory of probability, 3rd edn. Oxford University Press
Kass RE, Raftery AE (1995) Bayes factors. J Amer Stat Assoc 90:773–795
Kretschmer M, Coumou D, Donges JF, Runge J (2016) Using causal effect networks to analyze different Arctic drivers of midlatitude winter circulation. J Clim 29:4069–4081
Kretschmer M, Coumou D, Agel L, Barlow M, Tziperman E, Cohen J (2018) More-persistent weak stratospheric polar vortex states linked to cold extremes. Bull Amer Meteor Soc 99:49–60
Kretschmer M, Zappa G, Shepherd TG (2020) The role of Barents-Kara sea ice loss in projected polar vortex changes. Wea Clim Dyn 1:715–730
Kretschmer M, Adams SV, Arribas A, Prudden R, Robinson N, Saggioro E, Shepherd TG (2021) Quantifying causal pathways of teleconnections. Bull Amer Meteor Soc, in press. https://doi.org/10.1175/BAMS-D-20-0117.1
Lambeck K, Rouby H, Purcell A, Sun Y, Sambridge M (2014) Sea level and global ice volumes from the Last Glacial Maximum to the Holocene. Proc Natl Acad Sci USA 111:15296–15303
Lenton TM, Held H, Kriegler E, Hall JW, Lucht W, Rahmstorf S, Shellnhuber HJ (2008) Tipping elements in the Earth's climate system. Proc Natl Acad Sci USA 105:1786–1793
Lewandowsky S, Risbey JS, Oreskes N (2016) The 'pause' in global warming: turning a routine fluctuation into a problem for science. Bull Amer Meteor Soc 97:723–733
Lindley DV (2014) Understanding uncertainty, revised edition. Wiley
Lloyd EA, Oreskes N (2018) Climate change attribution: when is it appropriate to accept new methods? Earth's Future 6:311–325
Mastrandrea MD, Mach KJ, Plattner GK et al (2011) The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups. Clim Chang 108:675–691
NAS (National Academies of Sciences, Engineering and Medicine) (2016) Attribution of extreme weather events in the context of climate change. The National Academies Press, Washington, DC. https://doi.org/10.17226/21852

Nicholls N (2000) The insignificance of significance testing. Bull Amer Meteor Soc 82:981–986
Nuzzo R (2014) Statistical errors. Nature 506:150–152
Palmer PI, Jacob DJ, Jones DBA, Heald CL, Yantosca RM, Logan JA, Sachse GW, Streets DG (2003) Inverting for emissions of carbon monoxide from Asia using aircraft observations over the western Pacific. J Geophys Res 108:8828
Pearl J, Mackenzie D (2018) The book of why. Penguin Random House
Pithan F, Mauritsen T (2013) Comments on "Current GCMs' unrealistic negative feedback in the Arctic". J Clim 26:7783–7788
Rahmstorf S, Foster G, Cahill N (2017) Global temperature evolution: recent trends and some pitfalls. Environ Res Lett 12:1–7
Rodgers CD (2000) Inverse methods for atmospheric sounding: theory and practice. World Scientific
Rosner A, Vogel RM, Kirshen PH (2014) A risk-based approach to flood management decisions in a nonstationary world. Water Resources Res 50:1928–1942
Schaller N, Kay AL, Lamb R, Massey NR, van Oldenborgh GJ, Otto FEL, Sparrow SN, Vautard R, Yiou P, Ashpole I, Bowery A, Crooks SM, Haustein K, Huntingford C, Ingram WJ, Jones RG, Legg T, Miller J, Skeggs J, Wallom D, Weisheimer A, Wilson S, Stott PA, Allen MR (2016) Human influence on climate in the 2014 southern England winter floods and their impacts. Nature Clim Chang 6:627–634
Screen JA, Deser C, Smith DM, Zhang X, Blackport R, Kushner PJ, Oudar T, McCusker KE, Sun L (2018) Consistency and discrepancy in the atmospheric response to Arctic sea-ice loss across climate models. Nature Geosci 11:155–163
Sexton DMH, Murphy JM, Collins M, Webb MJ (2012) Multivariate probabilistic projections using imperfect climate models. Part I: Outline of methodology. Clim Dyn 38:2513–2542
Shepherd TG (2014) Atmospheric circulation as a source of uncertainty in climate change projections. Nat Geosci 7:703–708
Shepherd TG (2016a) Effects of a warming Arctic. Science 353:989–990
Shepherd TG (2016b) A common framework for approaches to extreme event attribution. Curr Clim Chang Rep 2:28–38
Shepherd TG (2019) Storyline approach to the construction of regional climate change information. Proc R Soc A 475:20190013
Sherwood S, Webb MJ, Annan JD, Armour KC, Forster PM, Hargreaves JC, Hegerl G, Klein SA, Marvel KD, Rohling EJ, Watanabe M, Andrews T, Braconnot P, Bretherton CS, Foster GL, Hausfather Z, von der Heydt AS, Knutti R, Mauritsen T, JNorris JR, Proistosescu C, Rugenstein M, Schmidt GA, Tokarska KB, Zelinka MD (2020) An assessment of Earth's climate sensitivity using multiple lines of evidence. Rev Geophys 58:e2019RG000678
Simpson IR, Polvani L (2016) Revisiting the relationship between jet position, forced response, and annular mode variability in the southern midlatitudes. Geophys Res Lett 43:2896–2903
Sippel S, Meinshausen N, Fischer EM, Székely E, Knutti R (2020) Climate change now detectable from any single day of weather at global scale. Nature Clim Chang 10:35–41
Spiegelhalter D (2018) The art of statistics: learning from data. Pelican Books
Stirling A (2010) Keep it complex. Nature 468:1029–1031
Stott PA, Stone DA, Allen MR (2004) Human contribution to the European heatwave of 2003. Nature 432:610–614
Stott PA, Christidis N, Otto FEL, Sun Y, Vanderlinden JP, van Oldenborgh GJ, Vautard R, von Storch H, Walton P, Yiou P, Zwiers FW (2016) Attribution of extreme weather and climate-related events. WIREs Clim Chang 7:23–41
Trenberth KE, Fasullo JT, Shepherd TG (2015) Attribution of climate extreme events. Nature Clim Chang 5:725–730
van Garderen L, Feser F, Shepherd TG (2021) A methodology for attributing the role of climate change in extreme events: a global spectrally nudged storyline. Nat Hazards Earth Syst Sci 21:171–186
van Oldenborgh GJ, Krikken F, Lewis S, Leach NJ, Lehner F, Saunders KR, van Weele M, Haustein K, Li S, Wallom D, Sparrow S, Arrighi J, Singh RK, van Aalst MK, Philip SY, Vautard R, Otto FEL (2021) Attribution of the Australian bushfire risk to anthropogenic climate change. Nat Hazards Earth Syst Sci 21:941–960
Weaver CP, Lempert RJ, Brown C, Hall JA, Revell D, Sarewitz D (2013) Improving the contribution of climate model information to decision making: the value and demands of robust decision frameworks. WIREs Clim Chang 4:39–60
Wilby RL, Dessai S (2010) Robust adaptation to climate change. Weather 65:180–185

Zappa G, Shepherd TG (2017) Storylines of atmospheric circulation change for European regional climate impact assessment. J Clim 30:6561–6577
Zappa G, Bevacqua E, Shepherd TG (2021) Communicating potentially large but non-robust changes in multi-model projections of future climate. Int J Clim 41:3657–3669