METEOROLOGY

University of Reading

BRINGING PHYSICAL REASONING INTO STATISTICAL PRACTICE IN CLIMATE-CHANGE SCIENCE

Ted Shepherd, Grantham Chair of Climate Science, Department of Meteorology, University of Reading

Acknowledgements: (1) The opportunity to teach MT2SWC; (2) Unfunded research

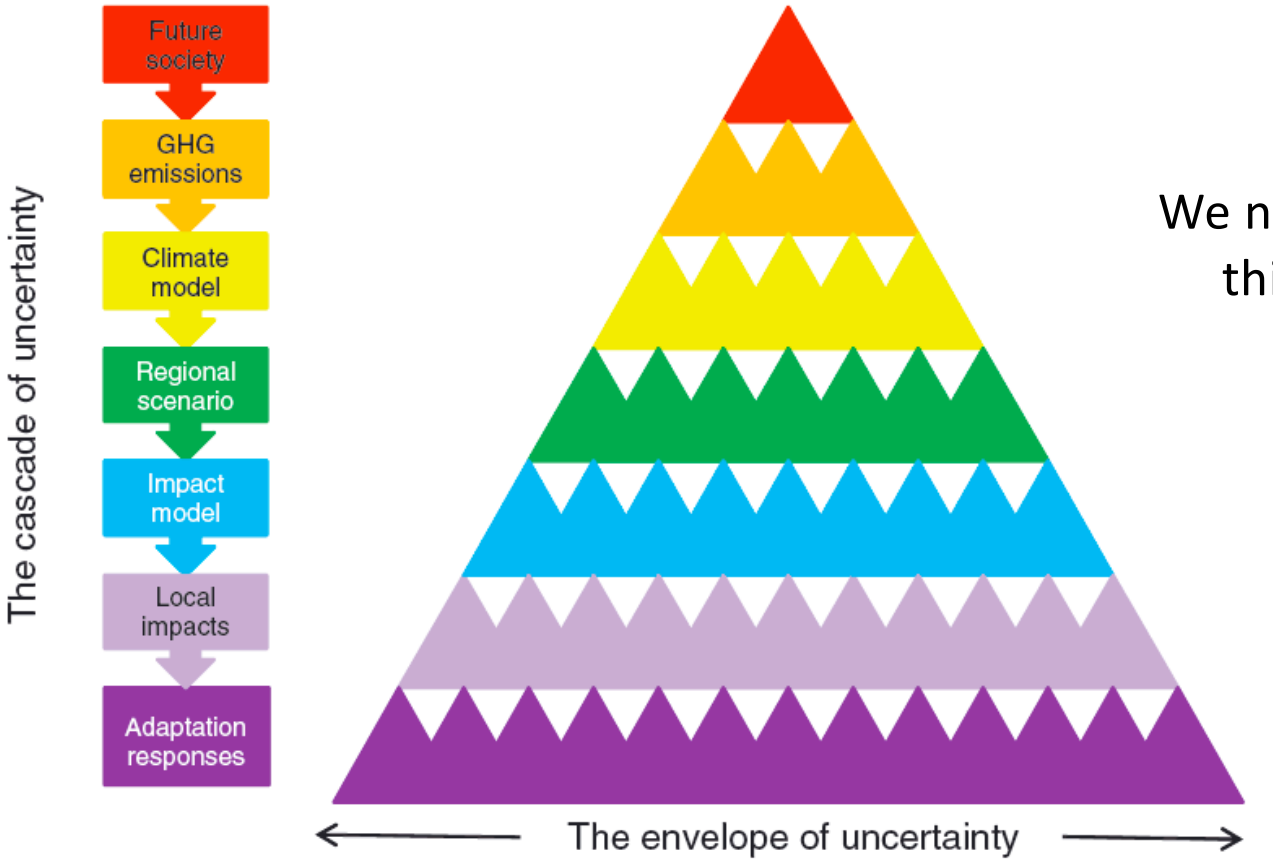
The heart of the matter

- Climate-change science is **anchored in physical understanding**
  - Controlled experiments on the real climate system cannot be performed
  - We have very little data measuring what we are trying to predict
- The **treatment of uncertainty** in published climate-change science is dominated by the far-reaching influence of the 'frequentist' tradition in statistics
  - In the frequentist tradition, uncertainty is interpreted in terms of sampling statistics, and there is an emphasis on p-values and **statistical significance**
    - But for climate-change science, a sampling distribution is not always meaningful: there is only one planet Earth!
  - There is **no way of expressing the uncertainty of a scientific hypothesis**
  - There is **no room for the concept of causality** (cause-effect relationships)
- All this creates a **disconnect between physical reasoning and statistical practice** in climate-change science

→ See my recent paper in *Climatic Change* (2021)

- **Example:** In a very important paper, Zelinka et al. (2020 GRL) show that the CMIP6 models have a higher climate sensitivity than the CMIP5 models due to stronger positive cloud feedbacks from decreasing extratropical low cloud coverage and albedo
  - These changes in the representation of cloud processes are believed to be realistic
  - However, they say that the higher climate sensitivity is "not statistically significant"
    - What on earth does this mean? Does it mean we should ignore the result?
- The significance test has to assume hypothetical populations of CMIP5-type models and of CMIP6-type models; what do we mean by that?
  - When you count everything, the concept of statistical significance is irrelevant
  - Users definitely need to know that climate sensitivity in the CMIP6 models is higher!
  - A p-value is effectively being used as a descriptive statistic, so why not just report the difference that way (e.g. relative to ensemble spread)?
- The significance test also has to assume that within each hypothetical population, there is a true CMIPn-type climate sensitivity, and that deviations from it result only from chance (and not from physically informed climate model development!)

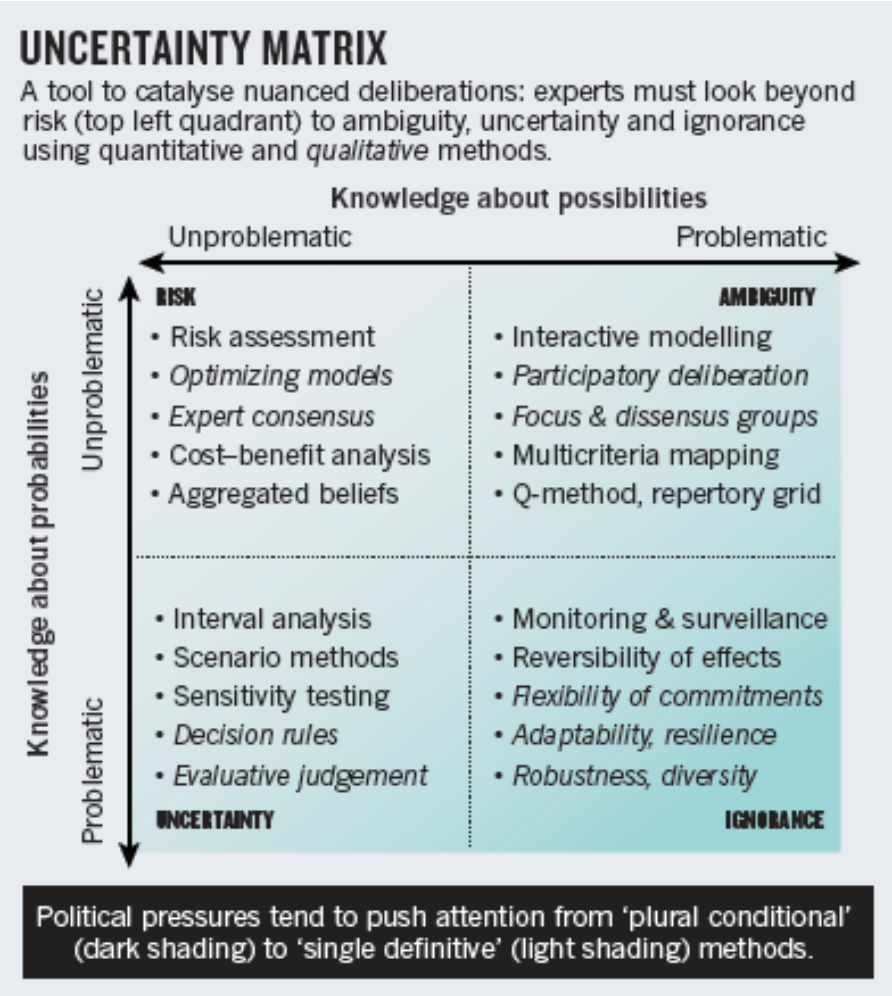
- Consideration of all the uncertainties in climate change in the traditional way can lead to a "**cascade of uncertainty**" which obscures the climate information content



We need to navigate this somehow!

Wilby & Dessai (2010 Weather)

Methods for decision-making under uncertainty



Stirling (2010 Nature)

- Scientists are pressured to issue 'single, definitive' statements
- In consensus mode, can lead to reliable but rather uninformative statements, e.g.

  "...there is low confidence in projected changes in the North Atlantic storm tracks" (IPCC AR6 WGI SPM 2021)
- We need a language for expressing a 'plural, conditional' state of knowledge:
  - **Multiple, mutually exclusive hypotheses**
  - **Logic for structured scientific reasoning**
- Frequentist statistics *has no such language, or any logic behind it*

nature

EDITORIAL · 20 MARCH 2019

It's time to talk about ditching statistical significance

Looking beyond a much used and abused measure would make science harder, but better.

- An obvious point is that Null Hypothesis Significance Testing (NHST) and $p < 0.05$ should not be interpreted dichotomously (as True/False), but the issue runs much deeper than this

The Journal of Socio-Economics 33 (2004) 587–606

Mindless statistics

Gerd Gigerenzer*

Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Abstract

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. I report evidence of the resulting collective confusion and fears about sanctions on the part of students and teachers, researchers and editors, as well as textbook writers.

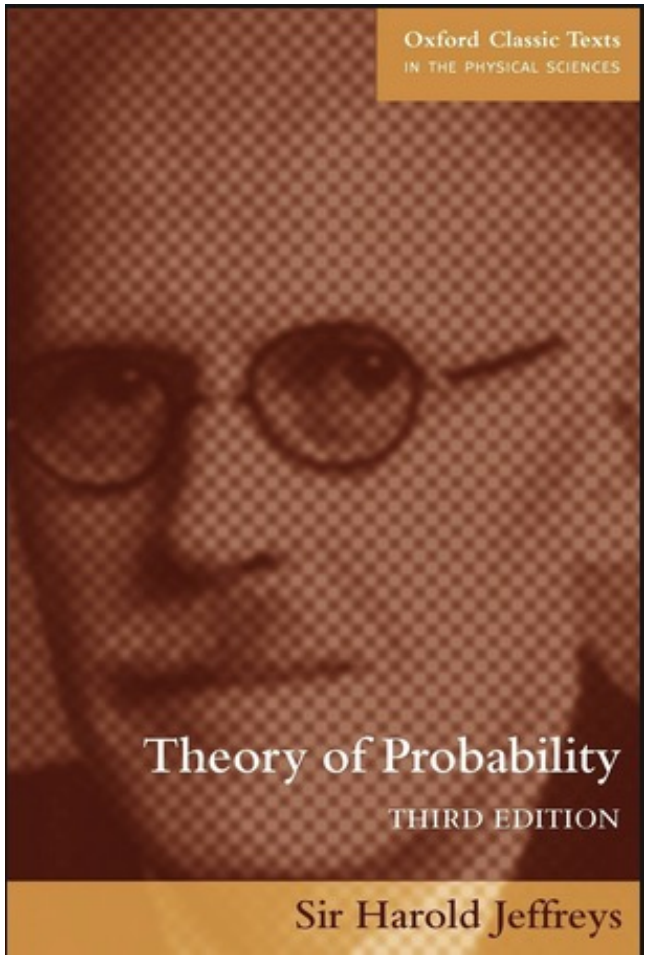
Gigerenzer refers to the social sciences, but is it really any different in climate science?

- Gigerenzer (2004) explains that **null hypothesis significance testing** is a bastardization of Fisher's null hypothesis testing (which should only be done when you have not looked at the data, and in the absence of prior information) and Neyman-Pearson decision theory, and was condemned by Fisher himself

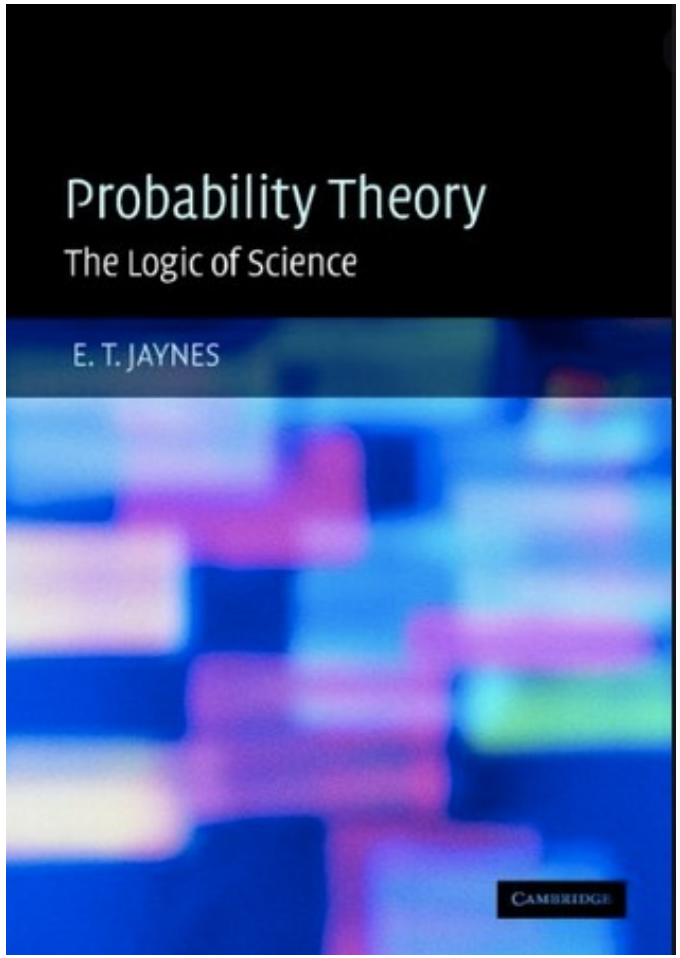
> . . . no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.
>
> Sir Ronald A. Fisher (1956)

- If you have already looked at the data, then you are prone to the 'multiple testing problem' (recognised already by Gilbert Walker in 1914)
- See Nicholls (2000) and Ambaum (2010) on the misuse of NHST in climate science
  - I think it has only got worse since those papers were published, abetted by the ready access of online statistical 'black boxes', and the obsessive emphasis on NHST by the so-called 'high-impact' journals (in spite of their own editorials!)

Some background.....



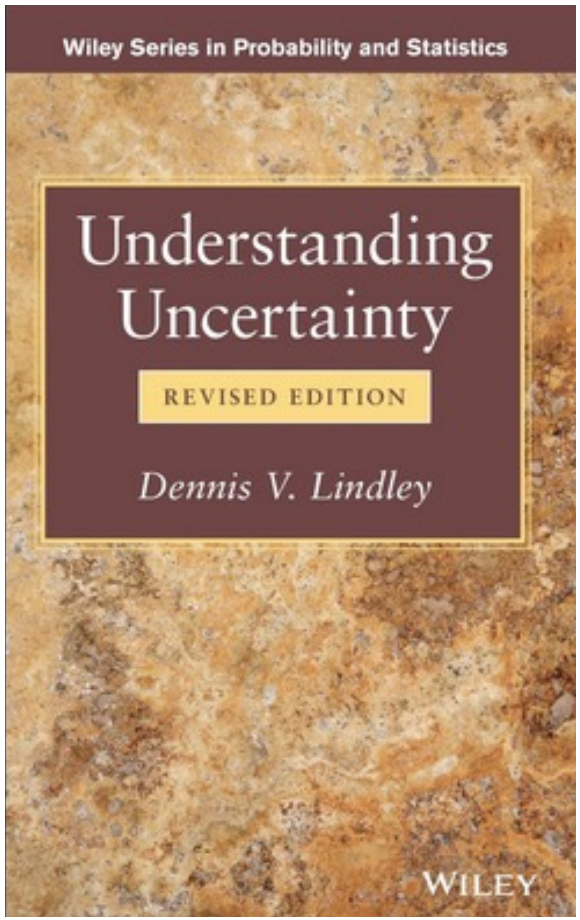
First published in 1939; 3rd edition 1961

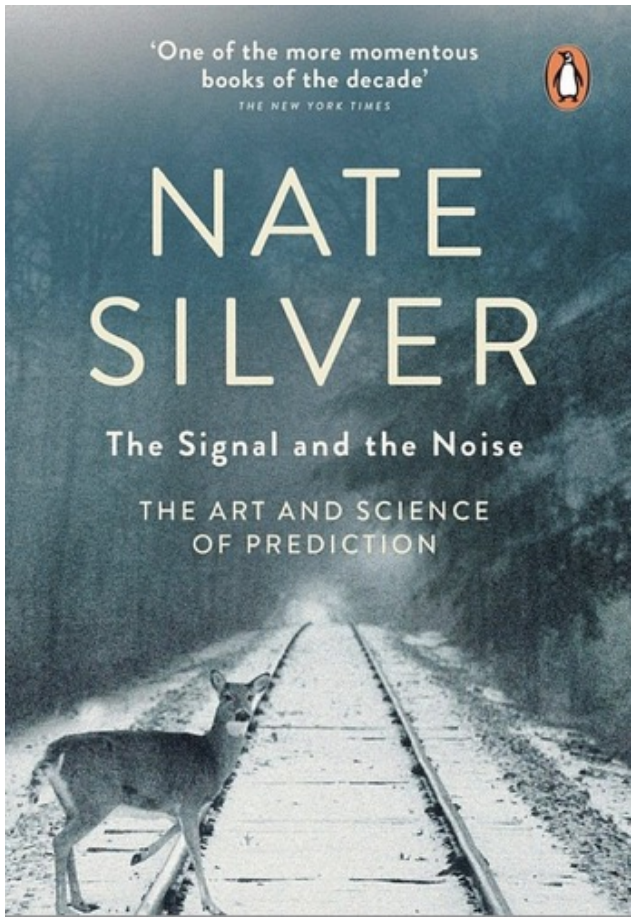


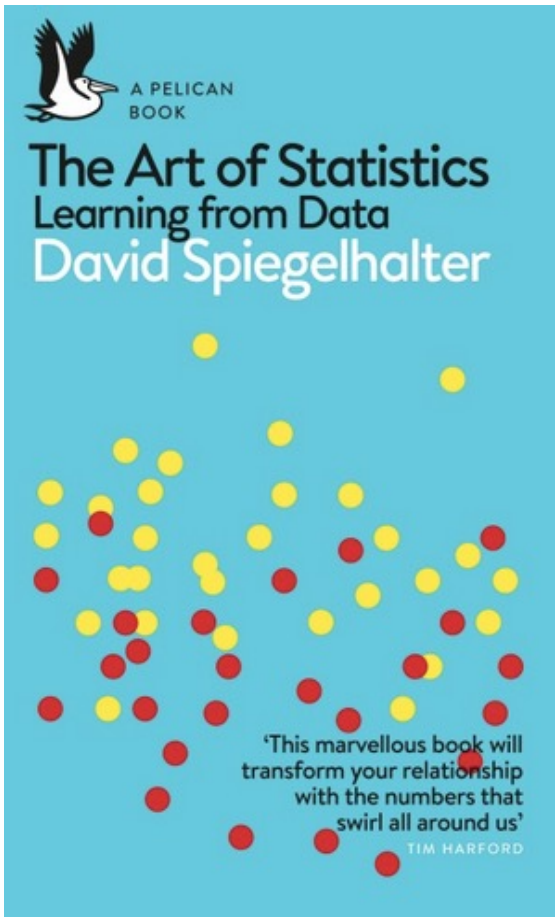
"*Dedicated to the memory of Sir Harold Jeffreys, who saw the truth and preserved it*"

Published posthumously in 2003

Or if you want something lighter....





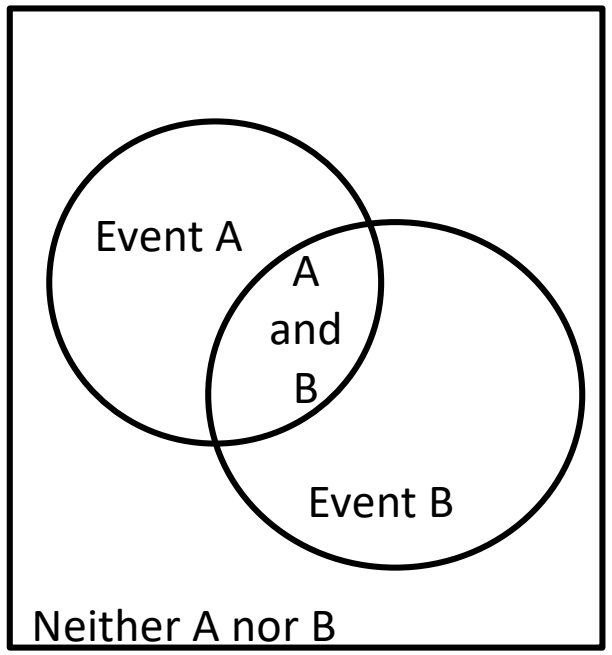


- **Two ways of thinking about probability** (Cox 1946 Amer. J. Phys.)
  - Likely outcome from a chance process → frequentist (sampling) statistics
  - Reasonable expectation → Bayesian (belief) statistics
- Hypotheses in climate science correspond to beliefs, not to frequencies
  - There is only one planet Earth
- Yet our statistical tools are mainly frequentist!

- This introduces a **mismatch between physical and statistical concepts**
- Frequentist methods were developed for situations with an abundance of data and little prior information (e.g. agricultural trials, quality control in industry)
- In climate-change science, we are in the opposite situation of an abundance of prior information and little in the way of data (given the size of the 'phase space')
  - Here I consider climate models to be prior information, not data
- Motivates a **reappraisal of the practice of statistics in climate-change science**

The fundamental principles of probability

- Generalization of Aristotelian (or Boolean) logic to case where $0 < P(\cdot) < 1$; reduces to it when $P(\cdot)$ is either 0 (false) or 1 (true)
  - Visualizable in Venn diagrams
- Product rule: $P(AB) = P(A|B)\,P(B) = P(B|A)\,P(A)$
  - A product means "AND". From this follows Bayes' theorem
- Sum rule: $P(A) + P(\neg A) = 1$, where $\neg A$ is the complement of A
  - From these follow: $P(A+B) = P(A) + P(B) - P(AB)$
  - A sum means "OR"
- Principle of indifference: if the hypotheses $H_1,\ldots H_N$ are mutually exclusive and exhaustive, and none are favoured over the others, then $P(H_i)=1/N$ $(1 \le i \le N)$
- **And that is (basically) it!**

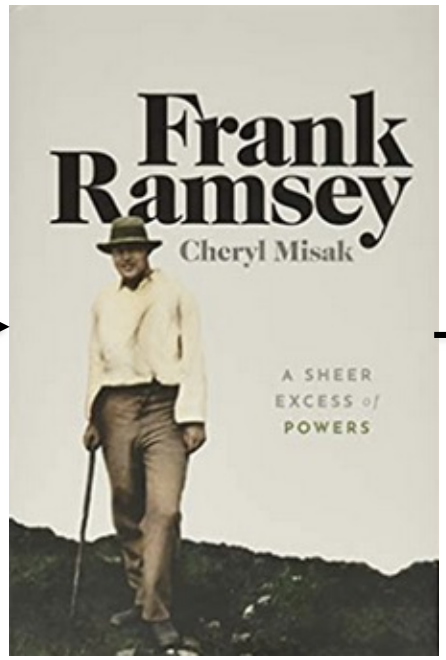


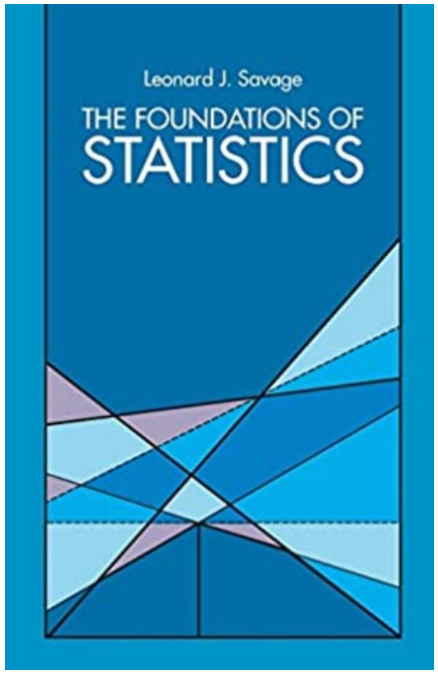
"La théorie des probabilités n'est que le bon sens reduit au calcul."
(Pierre-Simon Laplace, 1819)

- One can alternatively define probability in terms of **preferences** (or **proclivity to action**)
- Savage (1954) used the "sure thing principle" as the basis of his formulation of probability theory (this leads to the same rules of probability)

If action A is preferred over action B when C is true, and is also preferred when C is false, then it is preferred when C is uncertain

C.S. Peirce



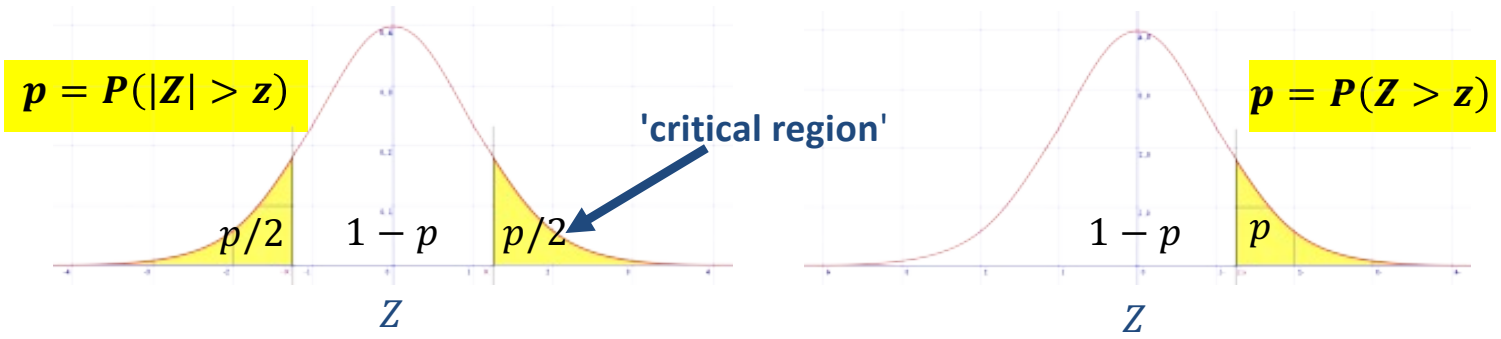


The relevance to decision-making under uncertainty is obvious!

It's also the basis of Judea Pearl's causality calculus

Misak (145): "Peirce argued that a belief is in part a habit which cashes out in behaviour."

- Let H be the **null hypothesis** (that nothing special is happening, i.e. that the data D occurred by chance)
- If one has a precise definition of what is meant by 'chance', then one can compute the **likelihood function**, $P(D|H)$
  - The '**p-value**' is the likelihood that something **at least as extreme as D** could happen under the null hypothesis H
  - Can be based on either a **two-sided** or a **one-sided** calculation (also known as **two-tailed** or **one-tailed**), depending on the alternative hypothesis



- The NHST: if $p < 0.05$, then your finding is "statistically significant" (at the 5% level)
  - You reject the null hypothesis (and accept your own hypothesis) **THIS IS WRONG**

- We must avoid **the error of the transposed conditional**: in general, $P(H|D) \neq P(D|H)$
- **Bayes' theorem** tells us: $P(H|D) = \frac{P(D|H)}{P(D)} P(H)$
- $P(H)$ reflects the relevance of prior knowledge: "strong claims require strong evidence"
- $P(D)$ requires consideration of **all** possible explanations for the data: if $\neg H$ is the negation (or complement) of H (possibly including several explanations),

$$P(D) = P(D|H)P(H) + P(D|\neg H)P(\neg H)$$

- Yet **nowhere** in any climate science publication have I seen any explicit consideration of these two factors, which strongly affect the inference that can be obtained from a p-value!

"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." (Jeffreys 1961)

'Sherlock Holmes' principle

- It is convenient to work with the '**odds**' form of Bayes' theorem

$$\frac{P(H|D)}{P(\neg H|D)} = \frac{P(D|H)}{P(D|\neg H)} \times \frac{P(H)}{P(\neg H)}$$

($\neg H$ = negation of H)

Bayes factor

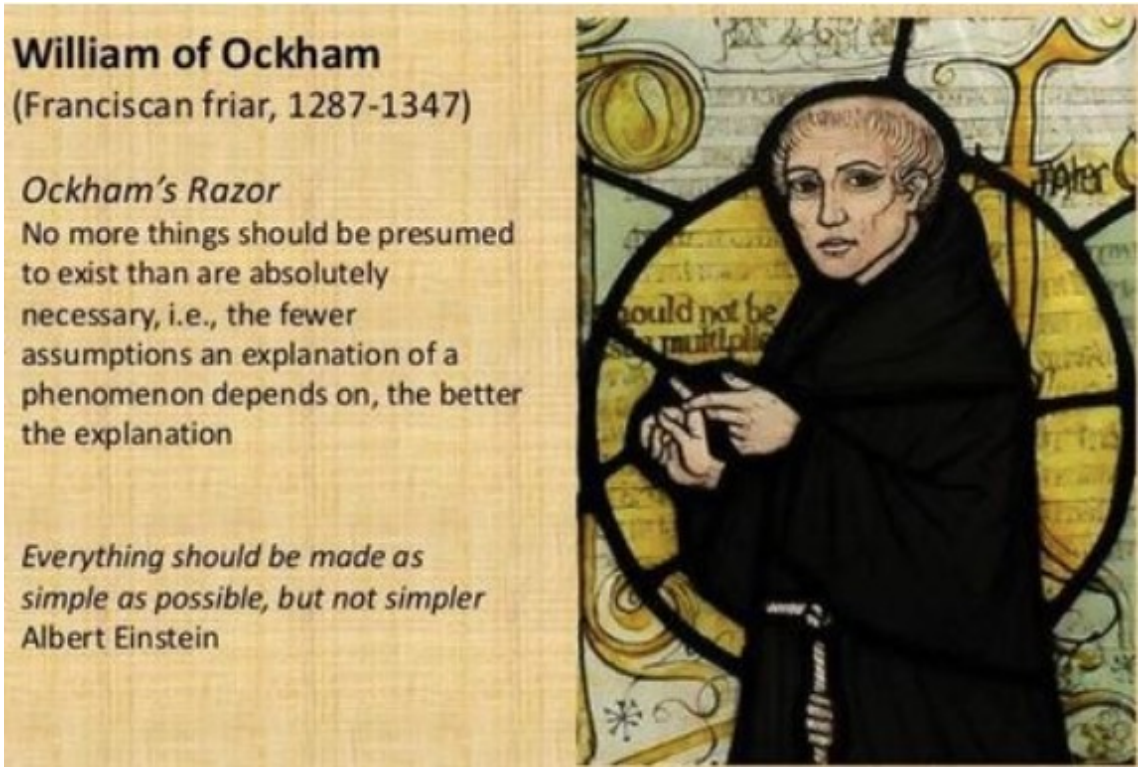
- Thus: in order to interpret a p-value $P(D|H)$, **we need a well-defined alternative hypothesis** $\neg H$ whose likelihood function $P(D|\neg H)$ can also be calculated
  - We cannot just 'go fishing' with a vaguely specified alternative hypothesis
- We also need to consider the **prior odds on the null hypothesis**
  - To do otherwise is to throw out all the background information we have
  - **If the prior odds are even, then the posterior odds equal the Bayes factor**
- If the alternative hypothesis is only vaguely specified, then Bayesian statisticians have ways of modelling $P(D|\neg H)$, using uninformative priors
  - These penalize imprecise alternative hypotheses, which are prone to overfitting

- We should not be given much credit for explaining the data when our hypothesis for doing so is only vaguely specified
  - Another form of the 'multiple testing problem'

nysora.com

- The p-value has to be downweighted by the fraction of the prior distribution for the alternative hypothesis that fits the data well (see e.g. Cousins 2017 Synthese)

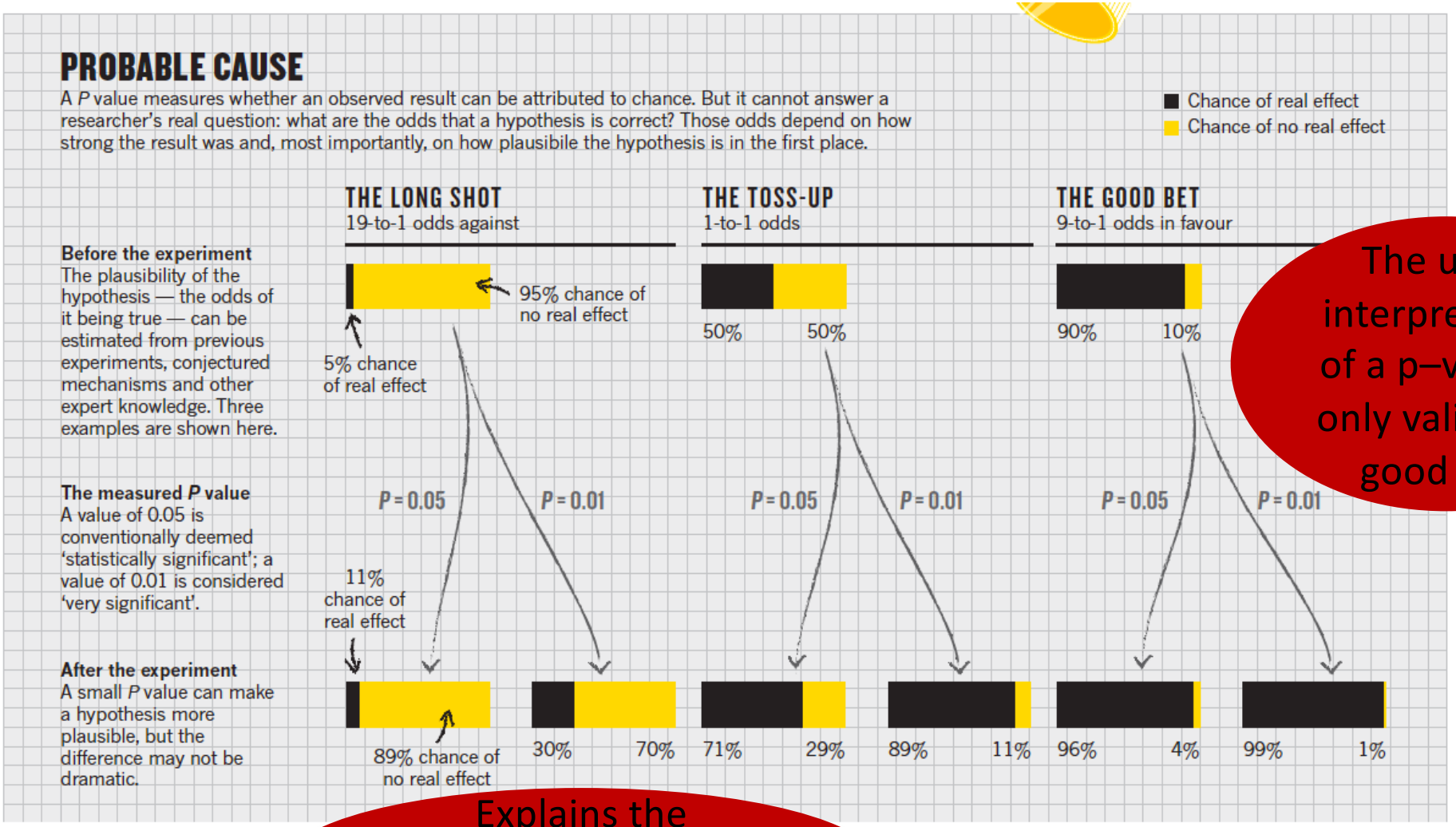
- According to the widely used parameterisation of Sellke et al. (2001 Am. Stat.), a p-value of 0.05 corresponds to a Bayes factor of only 0.4 or so, almost 10 times larger!
- The ratio of the two is called the 'Ockham factor'



www.zen-tools.net

- 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

R. NUZZO; SOURCE: T. SELLKE *ET AL. AM. STAT.* **55**, 62–71 (2001)



The usual interpretation of a p–value is only valid for a good bet!

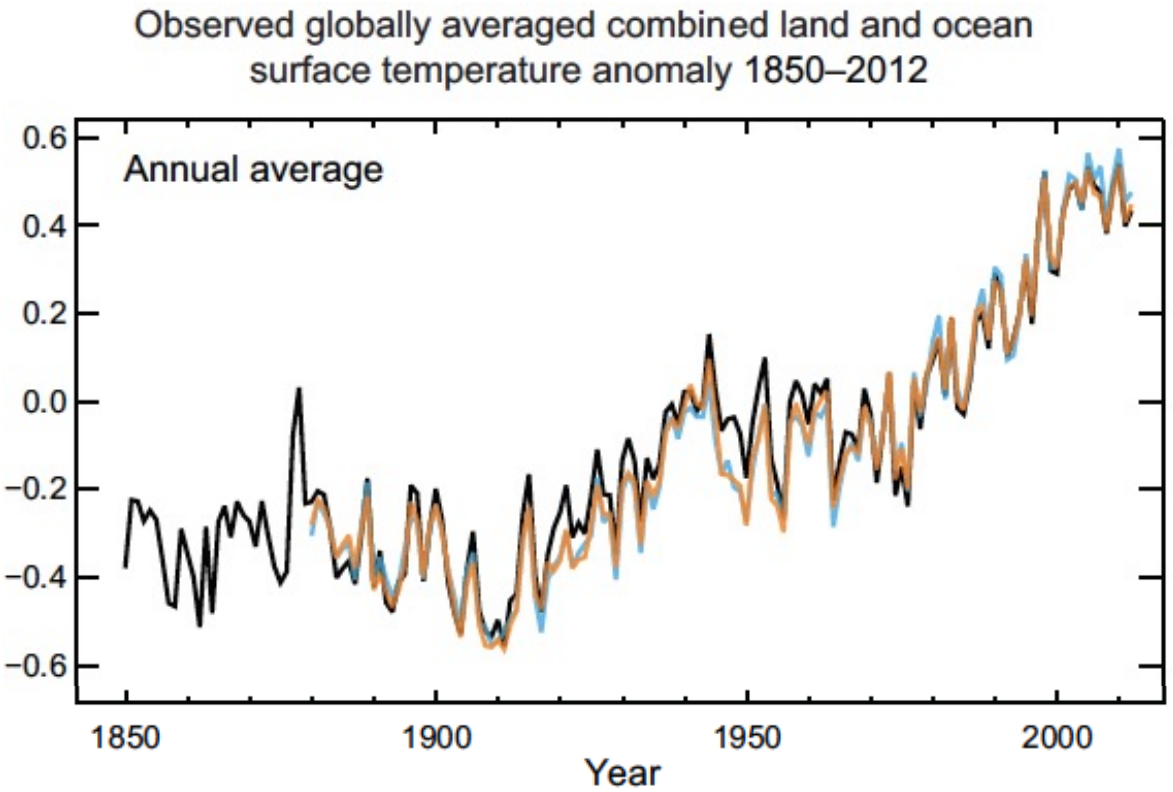
Explains the 'reproducibility crisis' in data-driven fields

Nuzzo (2014 Nature)

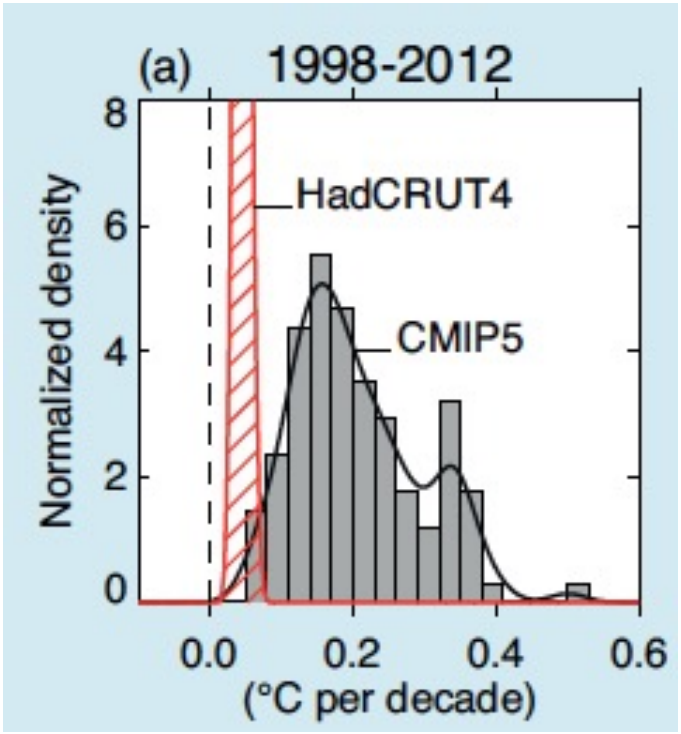
- **Example:** The development of the ozone hole over the last decades of the 20th century led to a delay in the late-spring breakdown of the stratospheric polar vortex
  - Was the main driver of observed trends in the SH summertime midlatitude jet
  - There is high confidence (e.g. in IPCC) that after the ozone hole stopped worsening (c. 2000), this trend in the delay of the polar vortex breakdown would not continue
- Zambri et al. (2021 Nature Geosci.) recently showed that the changes in the observed trends before and after 2000 are consistent with what climate models predict
- However, they also say that the trend differences "are statistically significant ($p < 0.05$)"
- Again, we can ask what this means
  - The null hypothesis was that the previous trends would continue: $P(H) \ll 1$
  - The alternative hypothesis was only vaguely specified: $\text{Bayes Factor} \gg P(D|H)$
  - With an implausible null hypothesis and a vaguely specified alternative hypothesis, the evidentiary power of a small p-value is very small indeed
- As with most published studies in climate science, the scientific claim rests on prior knowledge and physical reasoning, and the statistical test is only a sanity check

Example: the alleged global warming hiatus

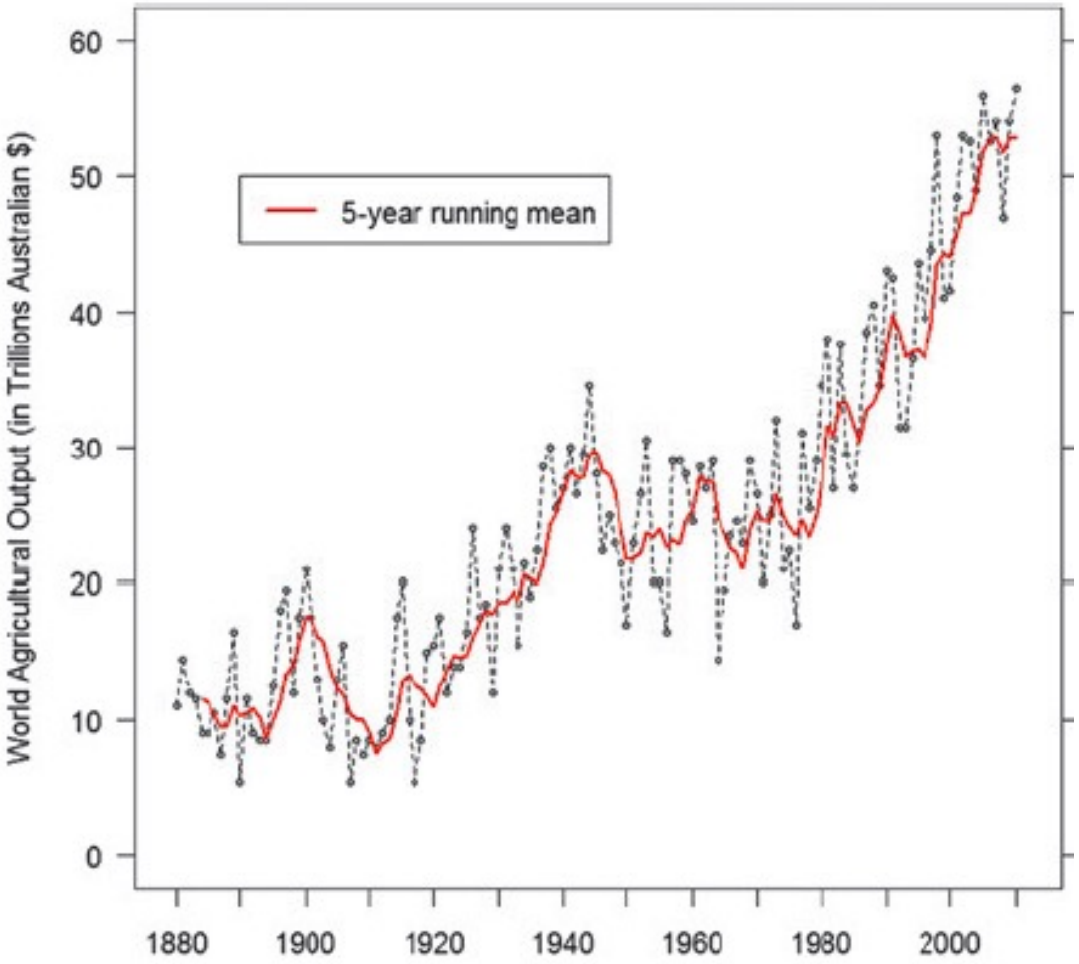
- Even very reputable climate scientists fell afoul of the 'multiple testing problem' (right panel); in fact, the behaviour was not at all anomalous (Rahmstorf et al. 2017 ERL)



IPCC AR5 SPM



IPCC AR5 TS

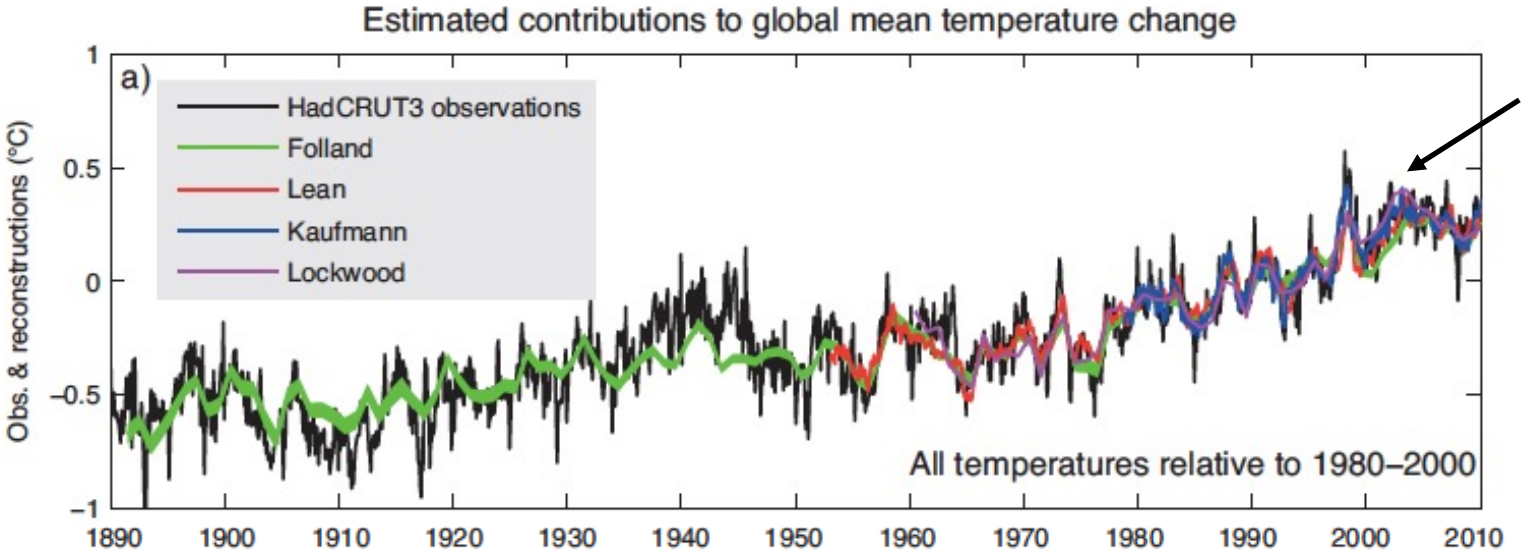


- Removed from its context, no experts in statistical analysis saw a pause — they all predicted future increases
- They also judged statements such as "the increase has stopped" to be disingenuous
- The eye is a very good detector of signal and noise!

Lewandowsky et al. (2016 Bull. Amer. Meteor. Soc.)

- Within the AR5 report itself there was all the evidence that was needed to show that the hiatus was plausibly explainable from known processes (see figure below)
- Given that the prior on anthropogenic global warming being true was extremely high, it should have taken a very small Bayes Factor to have raised any doubt

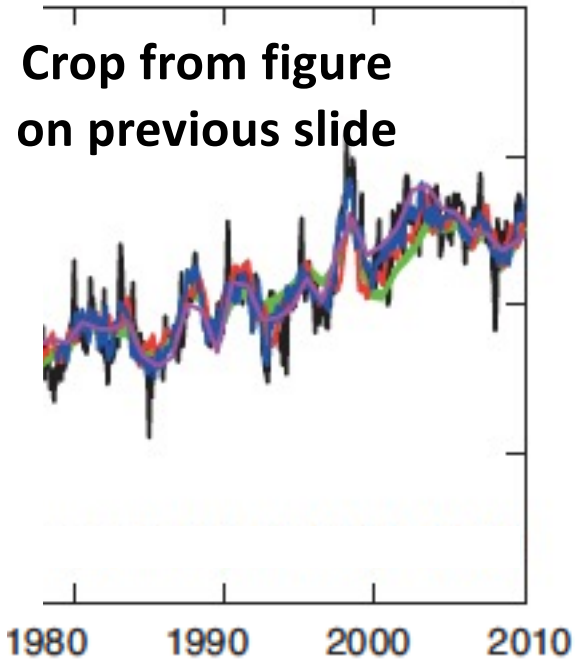
"The onus of proof is always on the advocate of the more complicated hypothesis.... There is no point in rejecting the null hypothesis until there is something to put in its place...Variation is random until a contrary is shown" (Jeffreys 1961)

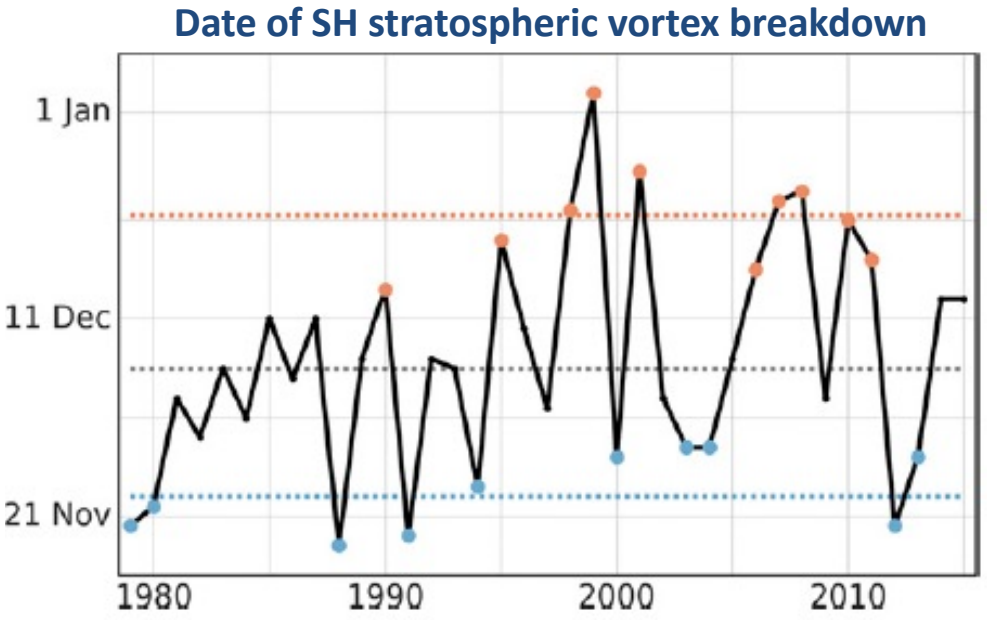


Hiatus mainly because of ENSO, with a small contribution from declining solar forcing

IPCC AR5 Ch 10

- It is instructive to compare the hiatus (left) with the change in stratospheric vortex breakdown dates before and after 2000 (right), discussed earlier
- Statistically, the changes in trends do not look so different, but our scientific conclusions are completely different — in fact, opposite — in the two cases
  - This is because we did not doubt global warming, and we did not expect a continuation of the pre-2000 vortex breakdown trend





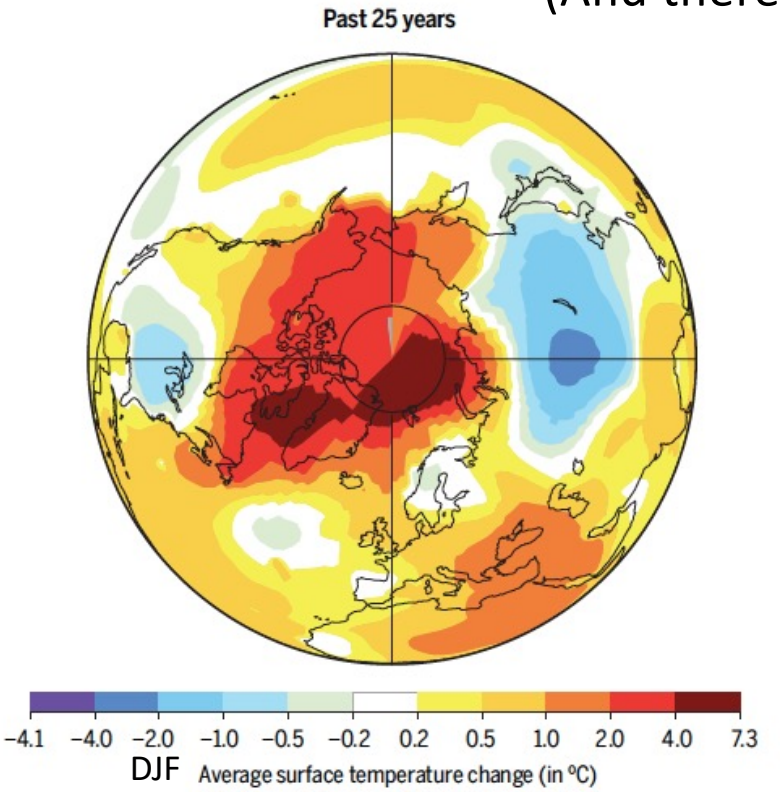
Byrne et al. (2017 J. Clim.)

- Data does not speak for itself!
- **So why do we use statistical tools that assume this?**

Example: Arctic-to-midlatitude connections

- Highly controversial topic, with many papers having quite unconditional titles
  - Yet absence of evidence is not evidence of absence! (inversion of the conditional)

(And there are a lot of large funding programmes on this very topic)



Shepherd (2016 Science)

Insignificant effect of Arctic amplification on the amplitude of midlatitude atmospheric waves

Minimal influence of reduced Arctic sea ice on coincident cold winters in mid-latitudes

Twenty-five winters of unexpected Eurasian cooling unlikely due to Arctic sea-ice loss

Midlatitudes unaffected by sea ice loss

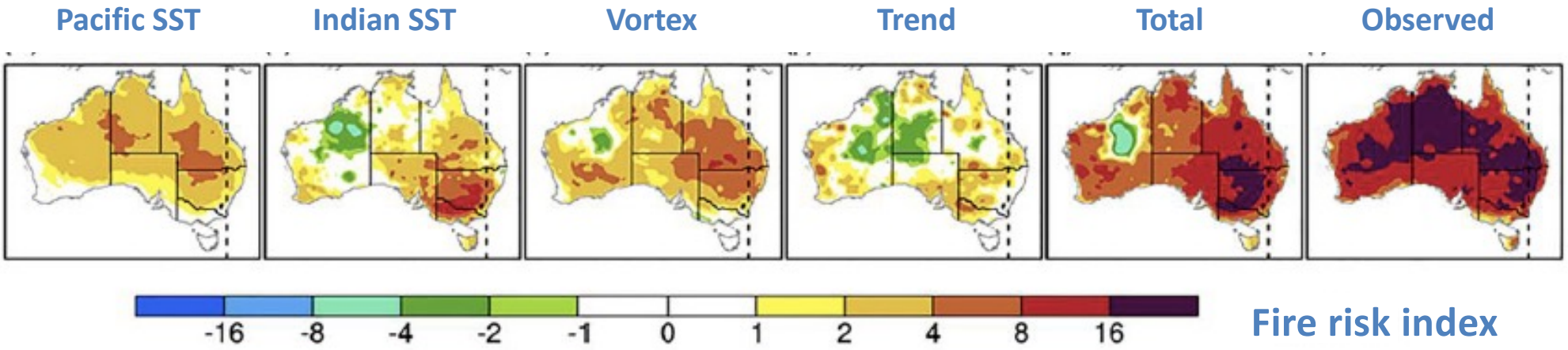
- Challenges (see Shepherd 2016 Science)
  - Background knowledge in this area is not very strong
  - Climate models have acknowledged deficiencies
  - The observational record is short
  - Causality is difficult to disentangle, since midlatitudes certainly affect the Arctic
- Essentially, **the Bayes Factor is close to unity**, meaning that 'believers' can publish papers in favour of the connection, and 'skeptics' can publish papers against it
- It is important to be explicit about the 'belief' (in the form of a scientific hypothesis)
  - See Kretschmer, Zappa & Shepherd (2020 Wea. Clim. Dyn.) for an explicit example

"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." (Jeffreys 1961)

Example: extreme event attribution

- Detection and attribution of changes in the statistics of weather and climate extremes has long been a subject of scientific study (e.g. IPCC SREX 2012)
- More recently, the question has emerged of the attribution of *single* extreme events to climate change
  - It is far from obvious how to even pose the question within a climate-science framework (see e.g. the 2016 US National Academies report)
- The most popular methodology (Stott et al. 2004 Nature, et seq.) takes a **frequentist** approach to the question, and frames the answer in a **single, definitive** manner
- Requires defining an 'event class' in order to create a 'population' of events
- There are **several issues with this approach** (Shepherd 2016 Curr. Clim. Change Rep.)
  - Definition of extreme needs to be sufficiently weak to allow enough events
  - The same extreme in a warming world will be very different meteorologically
  - Extreme impact is not equivalent to extreme hazard (van der Wiel et al. 2020 ERL)

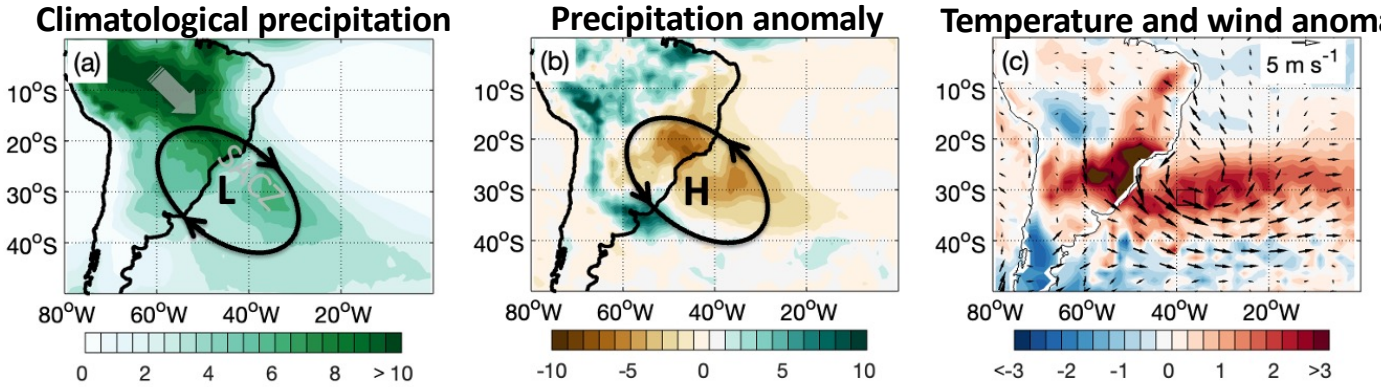
- Also, in most extreme events, the role of **unusual dynamical conditions** is generally a very important causal factor
  - **How those dynamical conditions could change** represents a major source of uncertainty in climate information for adaptation
- For the 2019 Australian wildfires, long-term warming ("Trend") was actually only a minor contributor to increased fire risk, which mainly arose from **drying associated with unusual dynamical states (atmospheric circulation)**

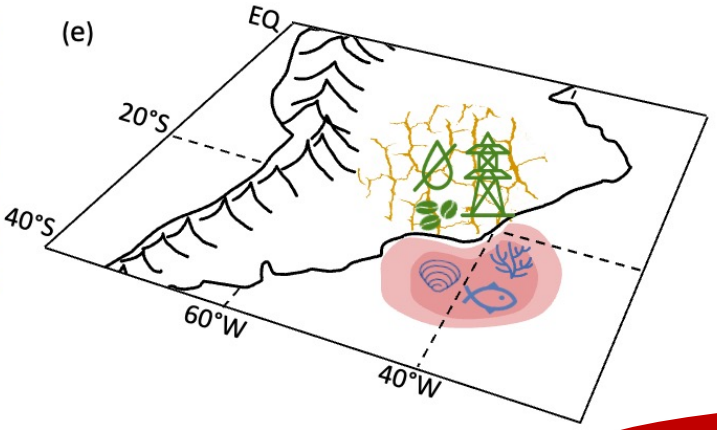


Lim et al. (2021 BAMS)

Example: a compound extreme event in southeast Brazil

- Anomalous anti-cyclonic circulation led to failure of 2013/14 South American monsoon
- Caused drought and heatwaves, affected food-water-energy nexus: correlated risk

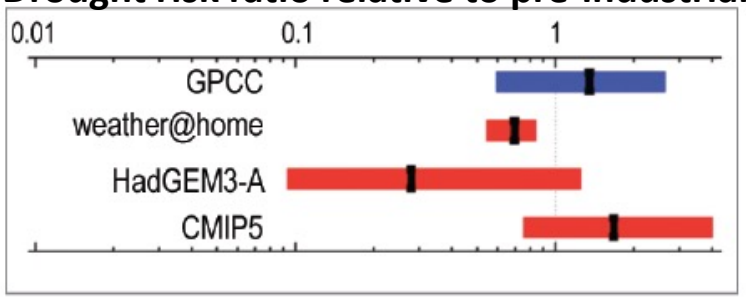




Rodrigues & Shepherd (2022 PNAS Nexus)

- A probabilistic attribution study of the event found "insufficient evidence" that climate change increased drought risk

Drought risk ratio relative to pre-industrial



Martins et al. (2017 BAMS)

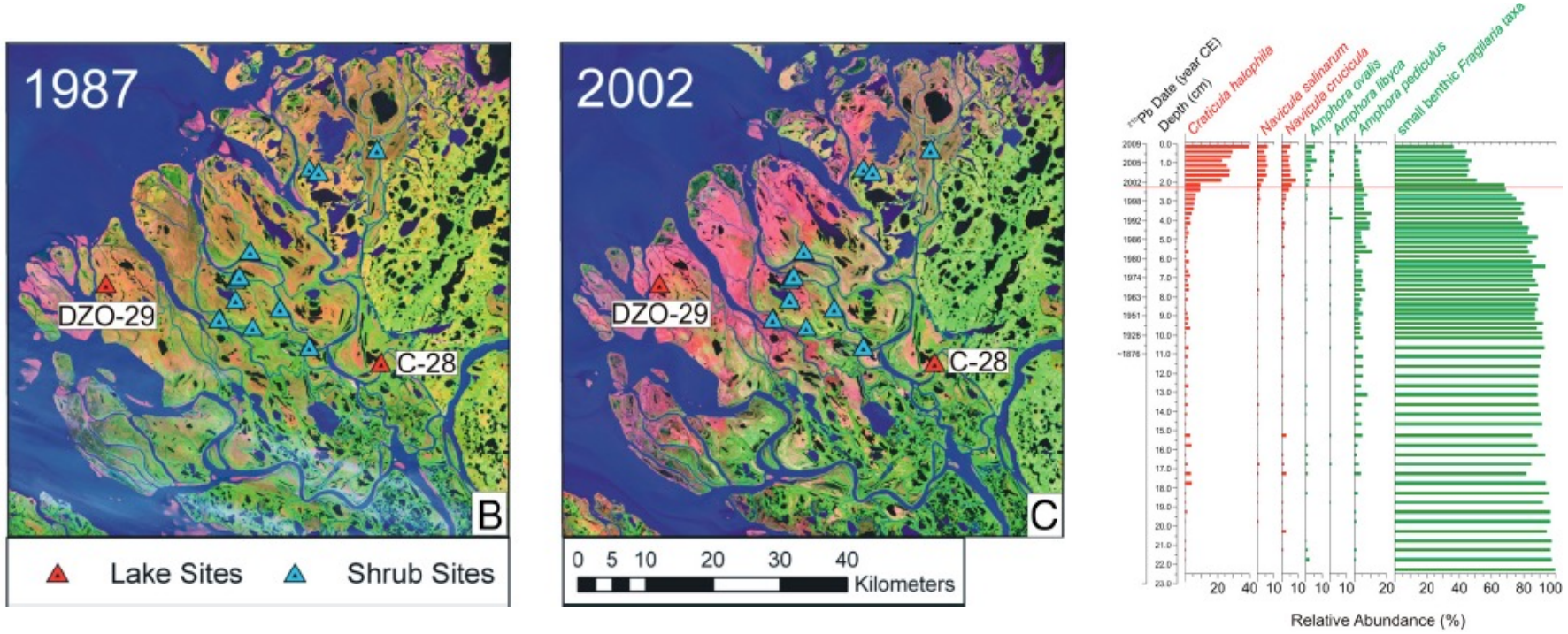
We can ask: **insufficient for whom?**

- The uncertainty can be partitioned using the fundamental rules of probability

$$\frac{p_1(E,C)}{p_0(E,C)} = \frac{p_1(E\,|\,C)}{p_0(E\,|\,C)} \times \frac{p_1(C)}{p_0(C)} \qquad \text{(US NAS 2016)}$$

  - p_1 is future, p_0 is present-day (for example)
  - E is the event of interest, C is the circulation regime conducive to that event
- The ratio of conditional probabilities represents the effects of climate change *for a given circulation regime*
  - **Builds in what we know with confidence about climate change**
  - Sometimes called the 'thermodynamic' component of change; can be defined in various ways (is not a precise distinction, but is very useful)
- The second ratio, representing the 'dynamical' component of change, **should be treated separately, e.g. via storylines** (Shepherd 2016 CCCR; 2019 Proc. R. Soc. A)
  - The uncertainty space is represented **discretely**, in a **plural, conditional** manner
  - Builds in **self-consistency**, which is essential for consideration of correlated risk

- **Example of an event storyline: Arctic ecosystem collapse**
  - A saltwater storm surge in the Mackenzie Delta (Canadian Arctic coast) in late September of 1999 led to irreversible changes from freshwater (green) to brackish (red) species, unmatched in over 1000 years (right)
  - Such a **singular event** is best described through a **storyline, or narrative**



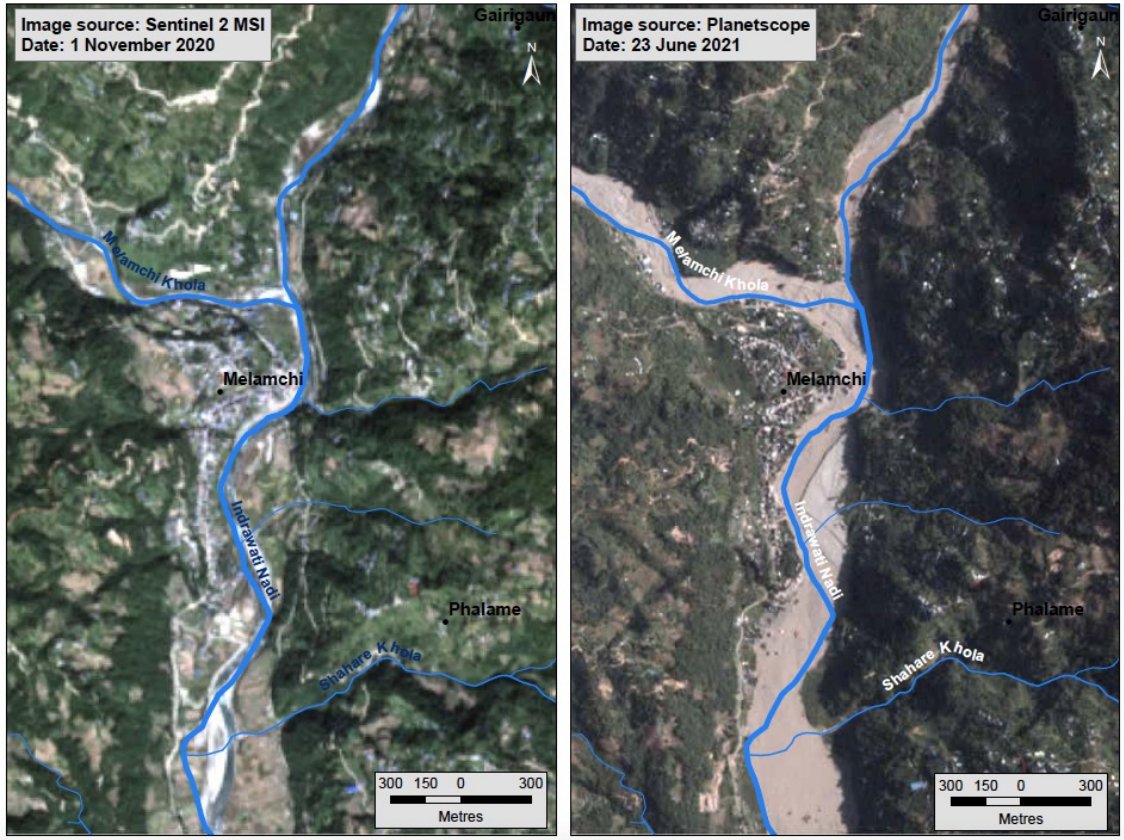
Pisaric et al. (2011 Proc. Natl. Acad. Sci. USA)

Narrative in science

"Natural historians have too often been apologetic, but most emphatically should not be in supporting a plurality of legitimately scientific modes, including a narrative or historical style that explicitly links the explanation of outcomes *not only to spatiotemporally invariant laws of nature, but also, if not primarily, to the specific contingencies of antecedent states*, which, if constituted differently, could not have generated the observed result." [emphasis added)

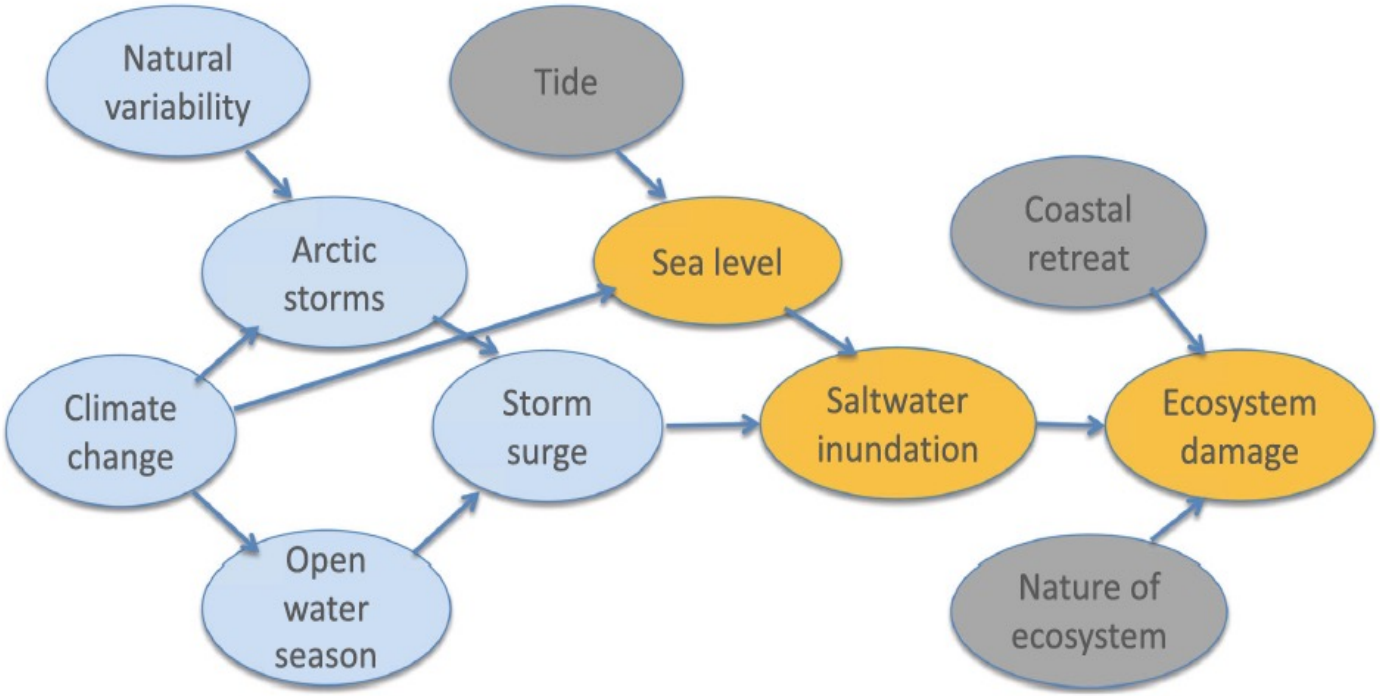
Stephen Jay Gould, *The Structure of Evolutionary Theory* (2002)

- So **why not climate scientists too?** (see Shepherd & Lloyd 2021 Climatic Change)



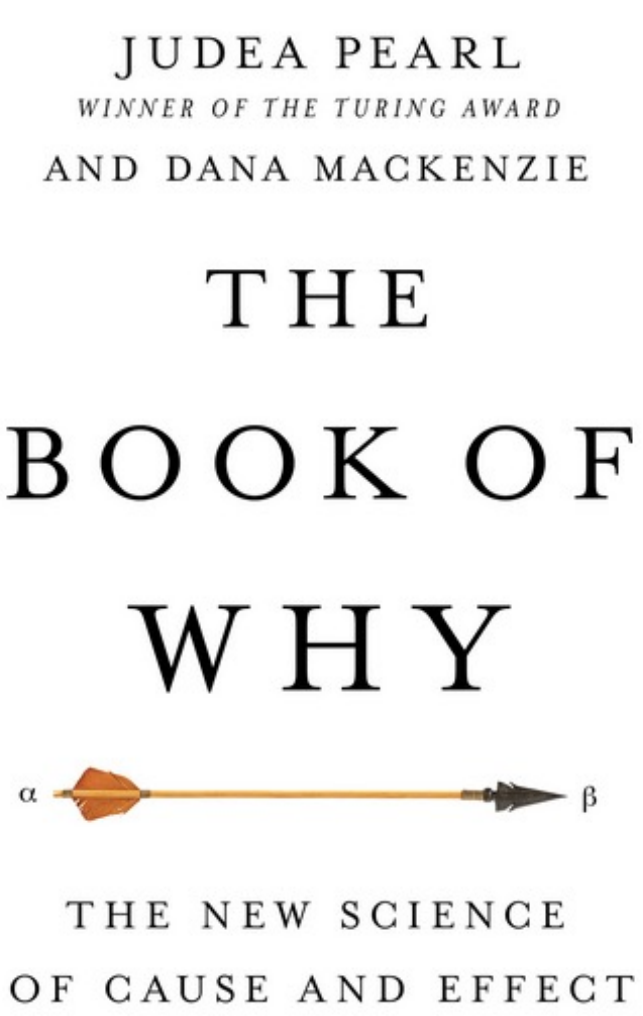
Debris outflow from the Melamchi (Nepal) flood disaster of 15 June 2021 (ICIMOD 2021)

- Storylines provide **conditional explanations**, and can be represented in **causal networks**
- Pisaric et al. (2011) discuss all the factors below, and conclude that the only essential ones were the longer open-water season from climate change, and the Arctic storm
  - There is no assessment of 'statistical significance', or of likelihood



- The storyline approach aligns well with the forensic approach to attribution in the ecosystem literature
- It also aligns well with liability under tort law (Lloyd & Shepherd 2021 Climatic Change)

Lloyd & Shepherd (2020 Ann. NY Acad. Sci.)

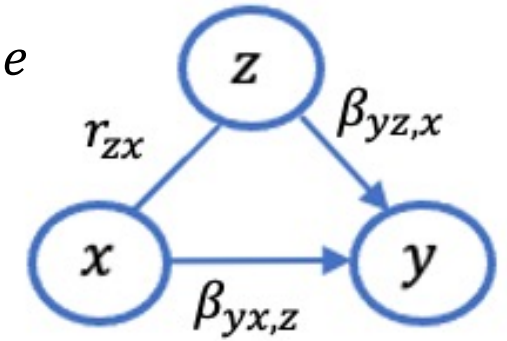


- **Causality** is not usually discussed in statistics textbooks
- However, understanding the causality involved in a particular situation is crucial for setting up the statistical analysis, and for interpreting the results
  - The mathematics is agnostic about causality, but the physical interpretation is not!
  - e.g. in an observed correlation between x and y, whether z is a **confounder** or a **mediator** depends on the direction of causation between x and z

$$y_i = \beta_{yx,z} x_i + \beta_{yz,x} z_i + noise$$

$$\Longrightarrow \quad r_{yx} = \underbrace{\beta_{yx,z}}_{\text{Direct}} + \underbrace{\beta_{yz,x} r_{zx}}_{\text{Indirect}}$$

(special case of the **path-tracing rule**)



Kretschmer et al. (2021 Bull. Amer. Meteor. Soc.)

- **Example:** Variability of Southern Hemisphere midlatitude jet in early austral summer (OND) is correlated with ENSO: $r_{JE} = -0.14$
- During this season the SH midlatitude jet is also known to be affected by interannual variability in the seasonal breakdown of the stratospheric polar vortex
  - Based on the NCEP reanalysis over 1949–2019, MLR gives

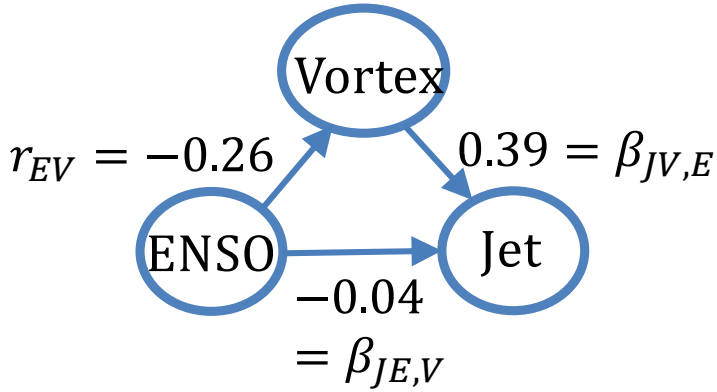
$$\text{Jet} = -0.04\ \text{ENSO} + 0.39\ \text{Vortex} + \text{noise}$$

- Most of the influence of ENSO on Jet is via the indirect stratospheric pathway

$$r_{JE} = -0.14 \approx -0.04 + 0.39 \times (-0.26)$$

$$r_{yx} = \beta_{yx,z} + \beta_{yz,x} r_{zx}$$

- MLR provides a built–in narrative
- Generalizes naturally to conditional probabilities
- Can be used to construct storylines



Kretschmer et al. (2021 BAMS)

- How can we ensure storylines are not **"just so stories"**?
  - And if we abandon NHST, does all hell break loose?
- Such a reaction (which I invariably get) seems reminiscent of the scientific community's reaction to **Thomas Kuhn**
- Kuhn responsed to his critics in his paper "Objectivity, value judgment, and theory choice" (1973)
  - Bottom line: there is no place to hide
- However, scientists also need constructive guidance
  - The answer here surely lies (in part) in probability theory, and the logic of Bayesian reasoning
- Kuhn (1962) put it thus: "If we can learn to substitute evolution-from-what-we-know for evolution-toward-what-we-wish-to-know, a number of vexing problems may vanish in the process"

Conclusions

- Scientific reasoning in climate science involves a combination of physically-based logic and statistical calculation (since we cannot do controlled experiments)

  "It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude." (Jeffreys 1961)

- In climate-science publications, the physically-based logic is generally stated in words, and the statistics in terms of 'rituals', but they are not brought together in a systematic way, **and statistical inferences tend to be treated as true/false statements**
  - Physical (causal) reasoning has been divorced from statistical practice: history!
  - Statistical practice should be embedded within structured logical reasoning (e.g. in the form of causal networks), which will help avoid the errors of inference that can easily arise when the statistical analysis is treated as an end in itself
- I am **not** arguing for full-blown Bayesian analysis, which can quickly become opaque
- I **am** arguing for following some of the very basic logical principles of reasoning under uncertainty: **be explicit about your assumptions, and consider alternative explanations**