Bayesian approaches to detection and attribution

J. D. Annan
RIGC/JAMSTEC, Yokohama, Japan

February 19, 2010

Abstract

We consider the Bayesian alternative to the classical frequentist approach to detection and attribution of climate change. Some of the notable advantages of the Bayesian paradigm include a more consistent approach to competing hypotheses, a coherent interpretation of all available data, and an intuitively natural interpretation of the results.

Detection and attribution (D&A) of climate change is a powerful tool to demonstrate the reality of the anthropogenic impact on the climate system. This process consists of demonstrating that specific climate changes that have occurred are unlikely (at some level of probability) to occur in an unforced system (the "detection" component), and furthermore are consistent with results from climate models which include the major anthropogenic and natural forcings ("attribution"). Perhaps the most important result in this field has been the conclusion that the natural variability of the climate system is "very unlikely" (meaning a probability of less than 5%) to result in as much warming as has been observed over the 20th century, and moreover that the changes are consistent with estimates generated by climate models driven by historical forcings [Solomon et al., 2007, Ch 9]. Typically, the attribution methodology is also used to calculate best guesses and confidence intervals for factors by which the model results should be scaled to best match reality, which enables the individual contributions of different forcing factors to be estimated. For a recent review of D&A in climate science, the interested reader is referred to Solomon et al. [2007, Ch 9].

This comment aims to present some of the advantages of a Bayesian approach to evaluating observations of climate change. In particular, Bayesian methods allow a rather more natural and intuitive approach to probabilistic estimation, and also readily enable a more flexible approach to the hypotheses to be tested. The frequentist approach cannot in principle generate probabilistic estimates that are credible in the Bayesian sense, and the desire for such credible estimates has lead to some rather inconsistent statements being made. We argue that the Bayesian paradigm provides a far more natural approach to understanding and predicting climate change. Rougier [2009] explores some of the technical aspects which vary between frequentist and Bayesian approaches. Here we focus more on the foundational differences.

Bayesian versus frequentist probability

In order to consider the core issues concerning Bayesian approaches (and alternatives) to D&A, it is necessary to ensure first that the distinction between frequentist and Bayesian approaches to probability is clearly understood. In the common frequentist interpretation, the probability of an event is defined as its relative frequency as the number of trials increases. In contrast, the Bayesian interpretation admits the use of probability as a means of describing the degree of belief in a hypothesis [Bernardo and Smith, 1994].

The detection component of D&A is essentially the frequentist approach of null hypothesis significance testing. Here one calculates the classical Fisher p-value, which is the probability of obtaining data "more extreme than observed" (for some predefined notion of "extreme"), if the null hypothesis is true. In this application, the null hypothesis is generally that there is no external influence on the climate system and all changes are merely the result of natural internal variability. It was known to Fisher [1959, p35] that this p-value cannot not represent the "probability that the null hypothesis is true", since the hypothesis is not a sample from a random distribution and therefore the probability of it being true or not is simply not a valid concept in the frequentist sense. However, this misinterpretation of the p-value (the fallacy of the transposed conditional, or Prosecutor's Fallacy [Thompson and Schumann, 1987]) has been so commonly applied that this point has needed to be repeatedly emphasised over subsequent years [eg Cohen, 1994, Nicholls, 2001].

This misinterpretation may be taken as an indication of the limitations of the frequentist approach: the question that people are generally interested in is how strongly they should believe a hypothesis is true or false, not how likely we would be to obtain data as extreme as observed if a specific hypothesis were true. It is therefore inconvenient that the frequentist approach cannot in principle address the former question. One practical impact of the misinterpretation of the p-value as the probability of the null being true is that it routinely exaggerates the weight of evidence against the null hypothesis that would be apparent in any realistic Bayesian analysis [Sellke et al., 2001, Goodman, 2001]. Conversely, it may also trick the unwary into accepting the truth of a null hypothesis that actually has zero prior probability ('nil hypothesis', Cohen [1994]).

A similar but more subtle error, which is perhaps even more commonly made, applies to the interpretation of the results of interval estimation. For a classical frequentist calculation, the correct interpretation of a 95% confidence interval for an unknown but fixed parameter is that, were the experiment to be repeated a large number of times, and a new confidence interval calculated according to the same (predefined) method, then 95% of the intervals so calculated would contain the true value of the parameter. This is *not* equivalent to the (Bayesian) belief that the parameter lies in the specific interval obtained, with probability 95%. While there are situations in which this common confusion is relatively harmless and thus this distinction may seem somewhat petty, it can have significant implications in some circumstances, a point to which we return below. We emphasise that the concept of Bayesian probability is fundamentally a subjective one in which the probabilities are a property of the researcher, rather than of the system being studied.

Bayesian approaches to D&A

The Bayesian calculation requires a prior distribution of probabilities for hypotheses H_i, $i = 1, \ldots, n$ which are updated in the light of observational evidence O to the posterior belief via Bayes' Theorem:

$$P(H_i|O) = P(O|H_i)P(H_i)/P(O).$$

This approach allows the researcher a great deal of freedom as to how to partition the problem into a set of competing hypotheses.

Although there has been relatively little research into Bayesian approaches for detection and attribution compared to the classical frequentist method, what literature there is already indicates a wide range of approaches to the problem. Perhaps the most natural starting point is the direct analogue of the standard frequentist approach, in which one calculates scaling factors for model estimates of the forced responses. Leroy [1998] analyses this problem, and considers how the results depend on the prior knowledge (implemented as Gaussian estimates of the scaling factors). He notes that in his framework, the results reduce to the frequentist confidence intervals as the prior variance on the factors increases without limit. This result has been widely used to justify the Bayesian re-interpretation of essentially frequentist analyses [Houghton et al., 2001, p 701]. However, it should be noted that a uniform prior is not always a reasonable choice [Annan and Hargreaves, 2009], so such an interpretation is not automatically valid. Leroy also notes that the hypotheses do not form a strict partition, since none of the models is actually correct in the strictest sense, and thus it would usually be more appropriate to present results in terms of relative probabilities $P(H_i|O)/P(H_j|O)$.

Berliner et al. [2000] and Lee et al. [2005] also consider the problem in terms of estimating linear scaling factors, but they explicitly present two competing hypotheses, which are that anthropogenic forcing either has virtually no effect, or that the modelled response is approximately correct, respectively. They both use as their prior a mixture of two Gaussians for the scaling factors, centred at 0 and 1. This approach allows them to directly consider the probabilities of competing hypotheses in a more symmetrical approach, in contrast to the two-step process of frequentist detection and attribution, in which a signal must rise above the threshold of natural variability before it is considered further. These scaling factors have independent value beyond the formal process of D&A, since they can also be used to rescale future climate change projections [Kettleborough et al., 2007]. Therefore, it might be useful to consider a more comprehensive analysis in which scaling factors are individually estimated for significant forcing factor. However, this approach does not appear to have been pursued further.

Schnur and Hasselmann [2005] increase the number of competing hypotheses to three, representing two different combinations of forcings and a "null" of no response. However, they do not consider scaling factors, but instead directly evaluate the point hypotheses. They show how the effect of model error can be explicitly included in the Bayesian analysis, in contrast to the frequentist approach. One fact we can be very sure of is that models are, and always will be, imperfect, so this is a welcome improvement that has the potential to increase confidence in the results. They also note that the Bayesian analysis has no need of a metric for climate changes to define "more extreme than observed", but simply uses the data that are available. Thus the filtering they used is determined purely in terms of maximising the impact of the data on the posterior probabilities. Schnur and Hasselmann also prefer to report their results as Bayes factors $P(O|H_i)/P(O|H_j)$ which indicate the extent to which the data prefer one hypothesis over another, without any need to prescribe a prior [Kass and Raftery, 1995]. It seems that Bayes Factors could be attractive to those who are squeamish about subjective priors, but it must also be recognised that numerous subjective decisions are inevitably made even in a frequentist analysis of any complex data set, and in both cases the robustness of results to the underlying assumptions can (and indeed should) be checked with sensitivity tests.

Min and Hense [2006] suggest another alternative to the reporting of probabilities, explicitly treating the issue as a decision problem in which the expected loss is to be minimised and thus emphasising the close link between Bayesian probability and decision theory. The companion paper Min and Hense [2007] considers the issue of D&A on a regional and seasonal basis. Uncertainties are relatively higher at smaller scales, and moreover it is on a local basis that climate change will actually impact the environment. Therefore, this area of research is likely to remain important long after the main questions of climate change on the global scale are considered settled.

Hasselmann [1998] discusses the difference between frequentist and Bayesian approaches to probability in some detail. He particularly emphasises the flexibility of the Bayesian approach - and perhaps this is its biggest advantage over frequentist methods - in allowing for the integration of diverse sources of information. It is well known that additional information is generally expected to reduce uncertainty, therefore the use of prior knowledge or additional sources of data will generate more precise results. This point has been frequently acknowledged in subsequent reviews of D&A [eg Barnett et al., 1999] and is also mentioned in Solomon et al. [2007, Ch 9] in an informal manner, but has rarely been considered quantitatively outside of a handful of illustrative calculations. Hasselmann [1998] suggests that numerous different climate indicators can be considered statistically independent and that their joint likelihood can therefore be calculated as the product of their individual likelihood functions, and this approach was also adopted by Schnur and Hasselmann [2005] for several time series. As Schnur and Hasselmann mention, however, this may be too optimistic, as it seems likely that many indicators will exhibit significant covariation in their natural variability. On the other hand, this same covariance should then greatly strengthen our belief in anthropogenic influence on indicators that have not yet been sufficiently well observed for their own signal to rise so clearly above the floor of natural variability.

For example, although the IPCC authors claim that the surface warming is "very likely" to have been caused by greenhouse gas emissions, it is only "likely" that these emissions have contributed to warming of the upper ocean [Solomon et al., 2007, Table 9.4]. The paradoxical conclusion we are forced to draw is that the IPCC authors assign significant probability for strong surface warming to be associated with ocean cooling, even though there is no physically plausible mechanism for this, and moreover the oceans have in fact been observed to warm in a similar manner to model predictions [Barnett et al., 2005]. The main source of this apparent inconsistency is that the observations of ocean temperature are limited in time and space. The relatively sparse and short-term nature of these observations, together with model uncertainties, make it hard to determine to what extent the observations lie outside the bounds of natural variability. Therefore, a frequentist analysis which only considers the ocean data alone will struggle to strongly reject the null hypothesis of no anthropogenic influence. A more explicitly Bayesian approach, which considers both the prior probability of a cooling response and the corroborating evidence of surface warming, would surely not have generated such an implausible statement.

Conclusion

While classical D&A has proved to be an interesting and useful way of testing our models and understanding of climate change, it is in principle incapable of generating probabilistic predictions (an implicitly Bayesian concept) and therefore cannot directly address questions of policy relevance, whether directed in the mitigation or adaptation sphere. Bayesian methods are capable of mimicking the D&A approach, but can also cover a rather more diverse set of approaches based on different sets of hypotheses. They also enable model error to be explicitly considered. Furthermore, their results are directly and intuitively interpretable in terms of (relative) probabilities on hypotheses, rather than the relatively obscure and widely misinterpreted p-values and confidence intervals. A further, potentially large but as yet largely untapped, advantage is their ability to consider numerous diverse data sources, either as prior information or directly in the likelihood function. Inclusion of this extra information can be expected to generate a more coherent and accurate depiction of the anthropogenic influence on the climate.

References

J. D. Annan and J. C. Hargreaves. On the generation and interpretation of probabilistic estimates of climate sensitivity. *Climatic Change*, 2009. DOI: 10.1007/s10584-009-9715-y.

TP Barnett, K Hasselmann, M Chelliah, T Delworth, G Hegerl, P Jones, E Rasmusson, E Roeckner, C Ropelewski, and B Santer. Detection and attribution of recent climate change: A status report. *Bull Amer Met Soc*, 80(12):2631–2659, 1999.

TP Barnett, DW Pierce, KM AchutaRao, PJ Gleckler, BD Santer, JM Gregory, and WM Washington. Penetration of human-induced warming into the world's oceans. *Science*, 309(5732):284–287, 2005.

L.M. Berliner, R.A. Levine, and D.J. Shea. Bayesian climate change assessment. *Journal of Climate*, 13(21):3805–3820, 2000.

J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, Chichester, UK, 1994.

J. Cohen. The earth is round ($p < 0.05$). *American Psychologist*, 49(12): 997–1003, 1994.

R.A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, 1959.

S.N. Goodman. Of P-Values and Bayes: A Modest Proposal. *Epidemiology*, 12(3):295, 2001.

K. Hasselmann. Conventional and Bayesian approach to climate-change detection and attribution. *Quarterly Journal of the Royal Meteorological Society*, 124(552), 1998.

J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. Van Der Linden, X. Dai, K. Maskell, and C. A. Johnson. *Climate Change 2001: Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2001.

R.E. Kass and A.E. Raftery. Bayes factors and model uncertainty. *Journal of the American Statistical Association*, 90(773):95, 1995.

JA Kettleborough, BBB Booth, PA Stott, and MR Allen. Estimates of uncertainty in predictions of global mean surface temperature. *Journal of Climate*, 20(5):843–855, 2007.

T.C.K. Lee, F.W. Zwiers, G.C. Hegerl, X. Zhang, and M. Tsao. A Bayesian Climate Change Detection and Attribution Assessment. *Journal of Climate*, 18(13):2429–2440, 2005.

S.S. Leroy. Detecting Climate Signals: Some Bayesian Aspects. *Journal of Climate*, 11(4):640–651, 1998.

S.K. Min and A. Hense. A Bayesian assessment of climate change using multimodel ensembles. Part I: Global mean surface temperature. *Journal of Climate*, 19(13):3237–3256, 2006.

S.K. Min and A. Hense. A Bayesian assessment of climate change using multimodel ensembles. Part II: Regional and seasonal mean surface temperatures. *Journal of Climate*, 20(12):2769–2790, 2007.

N. Nicholls. Commentary and analysis: The insignificance of significance testing. *Bulletin of the American Meteorological Society*, 82(5):981–986, 2001.

J C Rougier. Climate change detection and attribution. *ISBA Bulletin*, pages 1–4, Jan 2009.

R. Schnur and K. Hasselmann. Optimal filtering for Bayesian detection and attribution of climate change. *Climate Dynamics*, 24(1):45–55, 2005.

T. Sellke, MJ Bayarri, and J.O. Berger. Calibration of ρ Values for Testing Precise Null Hypotheses. *The American Statistician*, 55(1):62–71, 2001.

S. Solomon, D. Qin, M. Manning, Z. Chen, et al. Climate change 2007: The physical science basis. Contribution of the Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007.

W.C. Thompson and E.L. Schumann. Interpretation of statistical evidence in criminal trials. *Law and Human Behavior*, 11(3):167–187, 1987.