Significance Tests in Climate Science

Peter Ukkonen[1]

[1]Affiliation not available

March 6, 2019

Mockup conference abstract based on Ambaum (2010)

Significance tests are frequently misused in climate science. In a recent, randomly selected issue of the Journal of Climate, roughly three-quarters of the articles were found to use significance tests in an erroneous or misleading way. In a randomly selected issue 10 years prior, misuse of significance tests occurred in about half of the articles. While typically being only a small part of the evidence presented, and not invalidating results from such papers, it is nonetheless a misleading and rarely useful piece of information.

Significance tests are often performed to test for a "significant correlation" or a "significant trend" without understanding that it only quantifies the likelihood of an observation, given a (null) hypothesis, and not the likelihood of a hypothesis being true, given an observation. While these two probabilities are related by Bayes' theorem, they are different. This error of the *transposed conditional* is one of two major types of errors commonly found in the climate literature. The other is what's called a *category error* in philosophy, and in climate science includes statements like "the two time series are significantly correlated at the 95% level". In this case, a p value is ascribed to the *related* time series, where actually it is a property of *unrelated* time series.

More formally, the use of significance tests for this purpose can described as follows. An experiment produces two time series, which are found to be correlated. The hypothesis is that the time series are related and the correlation r_0 is a measure of the relation (note that correlation is a statistical property, while *relation* implies a physical dependence). The null hypothesis in this case is that the two time series are not related and that the observed correlation is a fluke. This may be tested if some synthetic time series can be produced which have similar properties to the original time series but with the hypothesized relationship explicitly switched off. The probability of finding a correlation at least as large as r_0 between a pair of unrelated time series is given by the p value. We can define a threshold correlation r_p which corresponds the given p value for the null hypothesis that the two time series are unrelated, e.g. $p = 5\%$. For repeated experiments, the proportion of unrelated synthetic time series showing a correlation larger than r_p should then approach 5%, assuming a correct experiment design. For repeated experiments where the time series are related by construct (the hypothesized relationship is turned on), we expect a fairly large fraction to produce a high correlation ($r > r_p$), e.g. 60%. It is clear that the p value of 5% is a property of the unrelated time series, and says nothing about the related time series.

To answer the question of whether the relationship is real, given a measured correlation r_0, we cannot use the p value alone, but need to invoke Bayes theorem. To obtain $p(H|r > r_p)$ where H is the hypothesized relationship, also need the prior of the null (opposite) relationship $p(\bar{H})$ and conditional probability $p(r > r_0)$:

$$p(H|r > r_p) = 1 - p(r > r_p|\bar{H})\frac{p(\bar{H})}{p(r > r_0)},$$

where $p(r > r_p|\bar{H})$ is the p-value. In practice, it is often impossible to retrieve the prior and the conditional probability.

To conclude, misuses of significance tests are very prevalent in climate science. Low p values are often interpreted as evidence of a hypothesis, where they only indicate that the measurement is unlikely if the null hypothesis were true. Likewise, high p values merely mean that the null hypothesis cannot be rejected. Significance tests are often used as quantitative evidence of a physical relation, but this is incorrect. Similarly, whether an upward trend is "significant" can easily be misunderstood; a low p-value as often reported does not actually mean that the null hypothesis (there is no upward trend, just natural variability) is unlikely, only that the observation (of an upward trend) is unlikely under the null hypothesis.

References

Maarten H. P. Ambaum. Significance Tests in Climate Science. *Journal of Climate*, 23(22):5927–5932, nov 2010. doi: 10.1175/2010jcli3746.1. URL https://doi.org/10.1175%2F2010jcli3746.1.