Comment on 'How unusual is the recent series of warm years?' by Zorita, Stocker, and von Storch

Gerd Bürger[*]

[*Zorita et al.*, 2008], shortly ZSS, introduce a strategy aimed at detecting unusual (anthropogenic) influences on the recent series of warm years, with no reliance on climate models. From the null hypothesis that the global temperature record is a stationary stochastic process, they derive a probability distribution after which the likelihood of the actual record is somewhere between 10^{-3} and 10^{-5}, depending on the assumed memory characteristics of the process. Hence the risk to erroneously reject the null hypothesis, as ZSS suggest, is correspondingly low.

The study contains two fallacies, one that might be called the 'Mexican Hat fallacy' and one that is known as prosecutor's fallacy.

Firstly, let me describe what is known as the 'Mexican Hat' argument. The Mexican Hat is a curious stone formation in Arizona that looks very much like what its name suggests. [*von Storch and Zwiers*, 1999], henceforth SZ, use it as a pedagogical example, or rather a counterexample, of statistical methodology applied to empirical science. It deals with the question: Is the Mexican Hat of natural origin or is it man-made? By collecting enough natural stones from the area and comparing them to the Mexican Hat, one would surely find that the null hypothesis 'the stone is natural' is quite unlikely, and it must be rejected in favor of a human influence. In view of this obvious absurdity SZ conclude: 'The problem with these null hypotheses is that they were derived from the same data used to

[*] Institut für Meteorologie, Freie Universität Berlin, Berlin, Germany

conduct the test. We already know from previous exploration that the Mexican Hat is unique, and its rarity leads us to conjecture that it is unnatural.' A statistical test of this kind 'can not be viewed as an objective and unbiased judge of the null hypothesis.'

Despite of that, Mexican Hat reasoning is found frequently in climate research, as SZ purport. It was encountered, for example, when a strong solar-terrestrial correlation was discovered by [*Labitzke and Van Loon*, 1988]. For that publication, to bolster the significance of the findings the reviewers had demanded a corresponding statistical test. Labitzke and Van Loon reluctantly gave in and rendered a test that was bound to be of Mexican Hat type. SZ conclude that 'truly independent confirmatory analysis can only be performed with observations in the future because we can collect the necessary independent information in the future.' In short, the definition of a statistical test must *precede* any observations that undergo the test.

Methodologically, the finding of the solar-terrestrial correlation by [Labitzke and Van Loon, 1988] and the clustering of warm years found by ZSS are comparable. They both represent events that are quite unlikely to appear by chance, but both came into existence as observations and not by being predicted *ex ante*. Therefore, if the statistical significance of [Labitzke and Van Loon, 1988] cannot be established, the same is true of ZSS. Consequently, ZSS have not detected that the recent unusual series of warm years is significantly non-random. Without independent evidence, either from future observations or, for example, 'deductively' from climate models no such detection is possible.

Secondly, regardless of the above argument ZSS slip, probably unknowingly, into what is known as prosecutor's fallacy [*Thompson and Schumann*, 1987]. They conclude that 'the risk of erroneously rejecting the null hypothesis strongly depends on the assumption about

47 the character of the memory'. However, in their analysis only the *p-value* is memory-de-
48 pendent, and that value is different from the probability of the null-hypothesis being true.
49 According to Bayes Theorem

$$P(H_0|d)=\frac{P(H_0)}{P(d)}\cdot P(d|H_0) \quad , \tag{1}$$

50 the posterior probability of the null-hypothesis H_0 given the data d, $P(H_0|\,d)$, and the
51 p-value, $P(d\,|\,H_0)$, are related via the ratio $P(H_0)/P(d)$ that can take on almost any value.
52 The low probabilities reported by ZSS, hence, say nothing about the probability of the
53 temperature data being natural.

54 It should be noted that both fallacies can be avoided under a Bayesian framework.

Reference

55 Labitzke, K., and H. Van Loon (1988), Associations between the 11-year solar cycle, the
56 QBO and the atmosphere. I: the troposphere and stratosphere in the northern
57 hemisphere in winter, *Journal of atmospheric and terrestrial physics*, *50*(3), 197-206.

58

59 von Storch, H., and F. W. Zwiers (1999), *Statistical analysis in climate research*, Cambridge
60 University Press.

61

62 Thompson, W. C., and E. L. Schumann (1987), Interpretation of statistical evidence in crim-
63 inal trials, *Law and Human Behavior*, *11*(3), 167-187, doi:10.1007/BF01044641.

64

Zorita, E., T. F. Stocker, and H. von Storch (2008), How unusual is the recent series of warm years?, *Geophysical Research Letters*, *35*(24).

Response to the comment by G. Bürger on 'How unusual is the recent clustering of warm years?'

E. Zorita

Institute for Coastal Research, GKSS-Research Centre, Geesthacht, Germany

T. F. Stocker

Physics Institute and Oeschger Centre for Climate Change Research,

University of Bern, Switzerland

H. v. Storch

Institute for Coastal Research, GKSS-Research Centre, Geesthacht, Germany

---

Eduardo Zorita, Institute for Coastal Research, GKSS-Research Centre, Max-Planck-Straße 1 D-21502 Geesthacht, Germany.

Thomas Stocker Physics Institute. University of Bern Sidlerstrasse 5 3012 Bern, Switzerland

Hans von Storch, Institute for Coastal Research, GKSS-Research Centre, Max-Planck-Straße 1 D-21502 Geesthacht, Germany.

We thank Gerd Bürger for raising two interesting points about our analysis of the recent clustering of record warm years. These two points reappear frequently in studies, as ours, that aim at establishing whether or not an observed rare event is compatible with a 'usual' or 'natural' probability distribution. Although in a general setting the objections raised by *Bürger* [2009] are correct, its application to our study is more subtle than it may appear at first sight. The first issue has been known in the literature as the Mexican hat fallacy [*Storch and Zwiers*, 1999], perfectly described in the comment by *Bürger* [2009]. This Mexican hat fallacy arises in the context of statistical hypothesis testing when a null hypothesis is put forward based on some empirical data set and this very same data set is subsequently used to empirically estimate the probability distribution of the test statistics under the null hypothesis. Our study would have fallen into this category if we had performed the analysis as follows: 1) it has been observed that the 13 warmest years in the observational record are included within the last 16 years (the Mexican-hat-type event). 2) This or a similar clustering of warm years is not observed at any other stage of the observational record -in other words, the 'natural' probability distribution would have been empirically constructed as a histogram based on the previously observed record. 3) it would have been established that the observed rare event falls outside the most extreme quantile of the observational probability distribution, and 4) we have concluded that the observed clustering is 'unnatural'. This chain of reasoning is indeed a tautology, as one knows beforehand that the Mexican-hat event lies outside the most extreme quantile of the empirical distribution, and therefore the result of the test is already determined from the very beginning.

Our study, however, deviates in two essentials points from this line of reasoning. One is that the rise in global temperatures and from this, the clustering of warmest years *at the end* of the observation record, was already predicted more than 100 years ago [*Arrhenius*, 1896]. Our 'Mexican hat' is therefore placed not in a random position in the desert but in the place predicted by a physically-based theory. The second point derives from the fact that we do not derive that 'natural' probability distribution directly from an histogram of the previous observations. Instead, we assume – and this assumption belongs also to the null hypothesis – that the observations are the outcome of a certain stochastic model. These models are not an ad-hoc construct for this particular analysis, but are widely used to represent a large range of other observational timeseries. They contain just one free parameter. We then derive the probability distribution of the test statistics conditional on the underlying stochastic process, the parameters of which are indeed derived from the observational record. In our analysis we assumed that the statistical properties of several temperature records (global mean, regional means and individual station records) could be represented by a simple autoregressive processes or, alternatively, by a more complex Long Term Persistence processes [*Hosking*, 1981]. Contrary to the typical case of the Mexican hat fallacy, it is very difficult to know *a priori* how the probability distribution of clustering of warm years may look like under these two scenarios without performing the Monte Carlo calculations described in our paper [*Zorita et al.*, 2009]. The main goal of our study was precisely to calculate some quantiles of these probability distributions. Indeed, contrary to the empirically-derived distribution of the test statistics in the Mexican-hat fallacy,

both stochastic models do produce infrequent realizations of the rare event, indicating that they are not a simple reflection of the available observations.

Furthermore, our conclusions do not critically hinge on particular values of the parameters of these stochastic models, the value of the lag-one autocorrelation or the value of the fractional differencing parameter d. This would have contained a tautological character, as these parameters are estimated the observations. On the contrary, to estimate the values of these parameters we could include or exclude the part of the record containing the rare event, without having to change the main conclusions. Furthermore, we discuss a range of possible values of these parameters, compatible with, but not limited to, the observational record , in an attempt to be reasonably conservative in the rejection of the null hypothesis, e.g. by considering values of the lag-one autocorrelation or of the fractional differencing parameter d higher than the best fit obtained from observations (see Figure 1 in [*Zorita et al.* 2009].

The use of a stochastic model of the temperature records to derive the 'natural' probability distribution of the clustering of warm years also changes the formulation of the null hypothesis compared to the Mexican hat case. Instead of formulating the null-hypothesis as 'the rare event is natural', the null hypothesis has to be reformulated as 'the observed rare event is the outcome of a natural autoregressive stochastic processes (or of a natural Long Term Persistence process)'.The conclusion of our study was that it is very unlikely that either type of stochastic process may produce a clustering of record warm years as in the observations. Therefore, from the rejection of the null hypothesis it may be concluded

that the observed clustering is unnatural and/or that these two stochastic processes are not adequate to describe the statistical properties of the observed temperature records.

To summarize, a "Mexican" rendition of our actual analysis would have been an observer that is aware of the Mexican-hat formation, uses a pre-existing model for the sand compaction and erosion processes, considers a range of plausible parameters for this model for the case of the North American desert, and then concludes that a Mexican hat formation is still very unlikely. The difference to the original fallacy is subtle but essential.

The prosecutor fallacy is related to the *attribution* of an observed rare event to a particular agent. It arises when it is overseen that the observed rare event could have also been caused, with unknown probability, by other agents. In other words, it is unknown how *specific* the rare event is as a marker for the particular agent. *Berry* [2008] has illustrated this fallacy in the context of doping tests in professional sports. In our case, *Bürger* [2009] basically argues that we attribute the clustering of warm years at the end of the observational record to anthropogenic CO_2 forcing simply because this clustering is observed simultaneously with the increase of atmospheric CO_2; he would further argue that we are ignorant about the probability of occurrence of similar clustering in other centuries, when the increase in CO_2 forcing was not important.

Although our paper states that the recent clustering of warm years is indeed completely consistent with the recent increase of atmospheric CO_2 and other anthropogenic forcings, we do not formally attribute it to this type of forcings. It would have been difficult to do so without considering explicitly the evolution of all other external forcings in the 20th century. Alternatively, records of reconstructed or simulated global temperatures in the

past centuries could have been included in the analysis to statistically isolate the rising CO_2 concentrations as a causing agent. As *Bürger* [2009] and *Berry* [2008] indicate, additional evidence is required for the attribution problem. This aspect was, however, not dealt with in our paper but it could certainly be considered in future analyses.

Acknowledgments. T.F.Stocker acknowledges funding by the Swiss National Science Foundation and by the NCCR Climate.

References

Arrhenius S. (1896) On the influence of carbonic acid in the air upon the temperature on the ground.*Phil. Magazine 41*, 237-276.

Berry, D. (2008) The science of doping. *Nature 454*, 692-693.

Bürger, G (2009) Comment on 'How unsual is the recent series of warm years?' , This issue

von Storch, H., and F.W. Zwiers (1999) *Statistical Analysis in Climate Research*, Cambridge University Press, 499 pp.

Hoskin, J.R.M. (1981) Fractional differencing. *Biometrika 68*, 165-176.

Zorita, E., T.F. Stocker, and H.v. Storch (2009) How unusual is the recent series of warm years? *Geophys. Res. Lett.35*, L24706, doi: 10.1029/2008GL036228.