Earth and Planetary Science Letters 459 (2017) 362–371

Contents lists available at ScienceDirect

Earth and Planetary Science Letters

www.elsevier.com/locate/epsl

Correlation-based interpretations of paleoclimate data – where statistics meet past climates

Jun Hu [a], Julien Emile-Geay [a,*], Judson Partin [b]

[a] Department of Earth Sciences, University of Southern California, Los Angeles, CA 90089, USA
[b] Institute for Geophysics, University of Texas at Austin, Austin, TX 78758, USA

ARTICLE INFO

Article history:
Received 27 June 2016
Received in revised form 5 November 2016
Accepted 24 November 2016
Available online 12 December 2016
Editor: H. Stoll

Keywords:
speleothems
calibration
age-uncertainties
false discovery rate
spurious significance

ABSTRACT

Correlation analysis is omnipresent in paleoclimatology, and often serves to support the proposed climatic interpretation of a given proxy record. However, this analysis presents several statistical challenges, each of which is sufficient to nullify the interpretation: the loss of degrees of freedom due to serial correlation, the test multiplicity problem in connection with a climate field, and the presence of age uncertainties. While these issues have long been known to statisticians, they are not widely appreciated by the wider paleoclimate community; yet they can have a first-order impact on scientific conclusions. Here we use three examples from the recent paleoclimate literature to highlight how spurious correlations affect the published interpretations of paleoclimate proxies, and suggest that future studies should address these issues to strengthen their conclusions. In some cases, correlations that were previously claimed to be significant are found insignificant, thereby challenging published interpretations. In other cases, minor adjustments can be made to safeguard against these concerns. Because such problems arise so commonly with paleoclimate data, we provide open-source code to address them. Ultimately, we conclude that statistics alone cannot ground-truth a proxy, and recommend establishing a mechanistic understanding of a proxy signal as a sounder basis for interpretation.

1. Introduction

Inferring past climate conditions from proxy archives is a central tenet of paleoclimatology. The calibration of paleoclimate proxies is accomplished in two main ways: space-based calibrations and time-based calibrations (defined below). In space-based calibrations, the values of a proxy at different locations are calibrated to measured climate indicators at the same locations, as exemplified by the calibration of paleothermometers in the core-top of marine sediments (e.g. Tierney and Tingley, 2014; Khider et al., 2015). This approach is relatively forgiving of time uncertainties, as long as core-top values are broadly contemporaneous, in relation to the question being asked of the cores. In time-based calibrations, on the other hand, proxy timeseries overlapping with the instrumental era are calibrated against an instrumental target (e.g. Jones et al., 2009; Tingley et al., 2012), via correlation analysis or the closely-related linear regression.

Thus "ground-truthing" a proxy record often involves establishing that its correlation to an instrumental climate variable (whether local, regional, or global) is significant in some way. Significance of correlations is most commonly assessed via a t-test, which assumes that samples are independent, identically-distributed, and Gaussian. However, these criteria may not be fulfilled in paleoclimate timeseries due to their intrinsic properties (Ghil et al., 2002).

Indeed, the loss of degrees of freedom due to autocorrelation has long been known to challenge the assumption of independence (Yule, 1926), though workarounds are known (e.g. Dawdy and Matalas, 1964). Non-Gaussianity may also prove an issue, especially for precipitation timeseries, though relatively simple transformations may alleviate it (Emile-Geay and Tingley, 2016).

Additionally, correlating proxies with instrumental climate fields is a common way of establishing the ability of a proxy to capture large-scale climate information. Unfortunately when implemented as a mining exercise using a large, spatially gridded dataset, test multiplicity becomes a problem. We will review how this problem may be successfully circumvented using simple statistical approaches (Benjamini and Hochberg, 1995; Storey, 2002).

Finally, the presence of age uncertainties may bring substantial uncertainties to time-based correlations between records (e.g. Crowley, 1999; Wunsch, 2003; Black et al., 2016). We will show

* Corresponding author.
E-mail address: julieneg@usc.edu (J. Emile-Geay).

http://dx.doi.org/10.1016/j.epsl.2016.11.048
0012-821X/© 2016 Published by Elsevier B.V.

a robust approach to quantifying age uncertainties and how they propagate to correlation and other analyses.

The article is structured as follows. In Section 2 we show the importance of considering autocorrelation in cross-correlation analyses. In Section 3, we briefly introduce the "test multiplicity" problem and the false discovery rate and show how it affects correlations with a climate field. In Section 4, we introduce the effects of age uncertainties, how they influence the interpretation of a speleothem record, and how this compounds with the other two challenges. We finish with a discussion of the significance of these results, and propose strategies to mitigate these statistical issues going forward.

2. Challenge #1: serial correlation

2.1. Theory

The most common way to determine the significance of Pearson's product-moment correlation involves a t-test. Student's t distribution is fully determined by the number of degrees of freedom available in the sample (ν). For N independent samples, $\nu = N - 2$, but it may be considerably lower when this assumption is violated, leading to overconfident assessments of significance.

As an example, consider correlations between two timeseries $x(t)$ and $y(t)$ generated by autoregressive processes of order 1 (a common timeseries model for serially correlated data; e.g. Emile-Geay, 2016, Chapter 8). Each process is evenly sampled 500 times and their correlation coefficient is 0.13, which is significant at the 5% level assuming independence (hence, $\nu = 498$). However, the lag-1 autocorrelation of each time series (ϕ) is 0.8, which is common for climate variables like temperature, as well as for many paleoclimate records, which tend to have a red spectrum (Ghil et al., 2002). This means that neighboring samples are highly dependent, so the effective numbers of degrees of freedom, ν_{eff}, is much lower. This number may be estimated via the following relation (Dawdy and Matalas, 1964):

$$\nu_{\text{eff}} = N \frac{1 - \phi_x \cdot \phi_y}{1 + \phi_x \cdot \phi_y} \quad (1)$$

where ϕ_x, ϕ_y are the lag-1 autocorrelation coefficients of two time series x, y respectively.

Based on equation (1), when either lag-1 autocorrelation coefficient increases, the effective number of degrees of freedom decreases, and the p-value of the test increases. In this case, the effective number of degrees of freedom decreases from 498 to 99 after considering the autocorrelation, and the p-value rises to 0.19, suggesting the correlation is no longer significant at the 5% level. Fig. 1 shows how the p-value and the degrees of freedom change for a time series of 500 samples and a fixed correlation of 0.13 just by changing the autocorrelation coefficients ($\phi_x = \phi_y = \phi$ for simplicity). As the autocorrelation increases, the p-values increases, and the degrees of freedom decrease. When all samples are independent ($\phi_x = \phi_y = 0$), the p-value is far smaller than 5%. When the autocorrelation increases to about 0.65, the p-value becomes larger than 5%, making the correlation insignificant at this level. The problem only worsens as ϕ increases, and as we shall see in this article, values above 0.8 are quite typical of paleoclimate timeseries.

Autocorrelation is thus a very serious challenge, which alone can substantially raise the bar of a significance test; if ignored, it may lead to overconfident assessments of significance.

2.2. Application

To see this effect at work in the real world, consider the example of Proctor et al. (2000), who used the band width in a

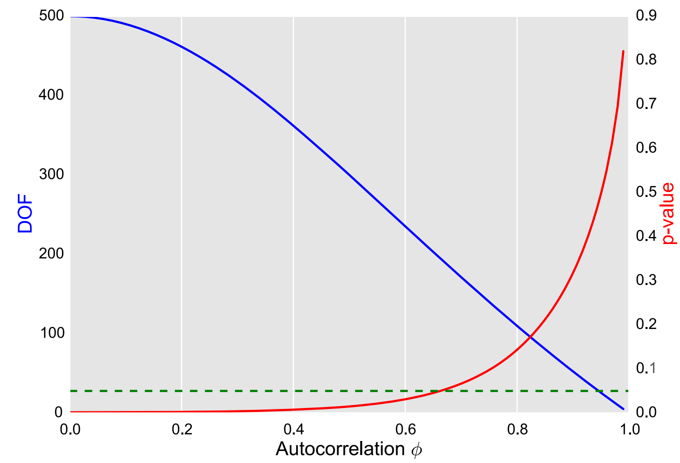

Fig. 1. The p-value and numbers of degrees of freedom (DOF) of the correlation (0.13) between two AR(1) time series (500 samples each) with the changing autocorrelation ϕ. The green dashed line is the 5% criteria for 5% level significance test.

stalagmite (SU-96-7) from Uamh an Tartair (northwest Scotland) to reconstruct the North Atlantic Oscillation (NAO). The record was dated by counting annual bands, with only 17 bands as double annual bands, implying a counting error less than 20 years. When compared to the whole length of the entire 1087-year-long record, this amounts to only 2%. Therefore, the influence of age uncertainties can be neglected to first order.

The climatic interpretation of the stalagmite was based on the high correlation between the band width and the temperature/precipitation ratio ($r = 0.80$) as well as the correlation between band width and the winter NAO index ($r = -0.70$) by using decadally-smoothed data. Here we apply the effective degrees of freedom in testing the significance of correlation, since the correlation significance may be biased by autocorrelation due to the effect of smoothing. Also, inherent aspects of these records leads to complications using statistics based on normally distributed populations, as the band width distribution of the stalagmite record is bimodal instead of normal. The t-test for correlation significance assumes that both time series are normally distributed, negating its use as a statistical tool unless appropriate transformations are made.

Considering the autocorrelation of the smoothed data, the high correlation between the band width of stalagmites and the temperature/precipitation ratio (T/P) in the instrumental period is not significant at 5% significance level (the adjusted p-value is 0.44). The correlation between the band width of stalagmites and winter NAO is also not significant, because of high autocorrelations of the smoothed time series of the band width ($\phi = 0.99$), T/P ($\phi = 0.99$) and winter NAO ($\phi = 0.95$). However, this result is based on an assumption of normality, and as discussed above, the distribution of the band width in this speleothem is bimodal, hence non-normal (not shown). Thus, transforming the non-normal series to normality (Emile-Geay and Tingley, 2016) is necessary. After this transformation, the correlations pass the significance test at the 5% level: for the correlation between the band width and T/P, ν_{eff} is 93 ($N = 115$), and the p-value is 3×10^{-3}; for the correlation between the band width and winter NAO, ν_{eff} is 95 ($N = 126$), and the p-value is 4×10^{-2}, just under the 5% threshold. While we conclude that the original interpretation is supported by our analysis, the authors reached this conclusion thanks to error compensation, potentially undermining their point.

We note, however, that the decrease of DOF due to smoothing was considered when this reconstruction was used for studying the long-term variability of the NAO in the high-profile study of Trouet et al. (2009).