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/

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):

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

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

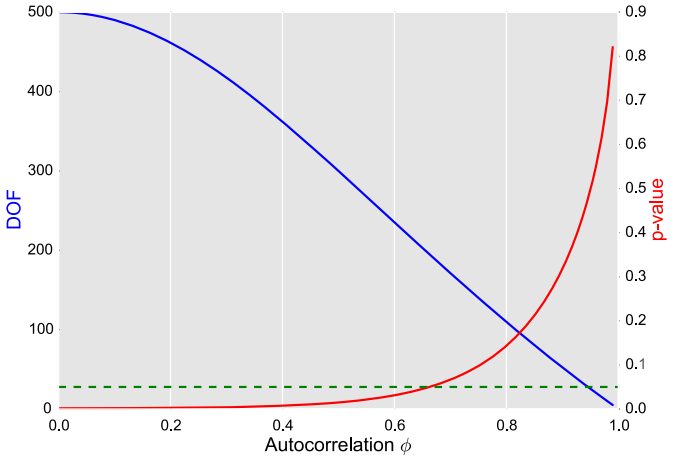

Fig. 1. The p-value and numbers of degrees of freedom (DOF) of the correlation (0.13) between two AR(1) time series (500 samples each) with the changing autocorrelation ϕ. The green dashed line is the 5% criteria for 5% level significance test.

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).

3. Challenge #2: test multiplicity

3.1. Theory

When assessing correlations with a field, multiple tests are carried out at different locations simultaneously. For example, if a correlation test is applied to 1000 locations with a significance level of 5%, one would expect about 50 hypotheses to be falsely rejected (in this case, 50 correlations would be deemed significant when in fact they are not), which is unacceptably high. The fundamental problem is that the test level α (the probability of false positives, i.e. the probability of falsely rejecting the null hypothesis of zero correlation) applies to pairwise comparisons, but not to multiple such comparisons. The finer the grid, the more such tests are simultaneously carried out, and the higher the risk of identifying spurious correlations as "significant" when in fact they are not. This test multiplicity problem is well known in the statistical literature and solutions exist (Benjamini and Hochberg, 1995; Storey, 2002).

In particular, the False Discovery Rate (FDR) procedure (Benjamini and Hochberg, 1995) has been widely applied (>34,000 Google Scholar citations at the time of writing) to control the proportion of falsely rejected null hypotheses out of all rejected null hypotheses, thus offering a level of scientific rigor that naive correlation testing does not afford. The term "false discovery" here is synonymous with "falsely identified significant correlations". Instead of restricting the occurrence of falsely rejected null hypotheses, the FDR procedure controls the proportion of erroneously rejected null hypotheses. Setting $q = 5\%$ in the FDR procedure, guarantees that 5% or fewer of the locations where the null hypothesis is rejected are false detections on average, so the proportion of false rejections is controlled. The FDR procedure of Benjamini and Hochberg (1995) proceeds as follows:

1. Carry out the test (calculate p-values) at all m locations;
2. Rank the p-values $p_{(i)}$ in increasing order: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$;
3. Define k as the largest i for which $p_{(i)} \leq q\frac{i}{m}$;
4. Reject hypotheses at locations $i = 1, 2, \ldots, k$.

In this procedure, the FDR treats locations with low p-values as most significant, ranking p-values from high to low. If the largest p-value is less than its corresponding threshold, then all tests are regarded as significant. If the largest p-value is greater than its corresponding threshold, then it is compared to the second largest p-value with a more restricted threshold – and so on. These thresholds guarantee that the expected rate of falsely positive hypotheses are smaller than α. Through this procedure, the number of rejected hypotheses is k, and the expected number of falsely rejected hypotheses is smaller than $mp_{(k)}$, such that the fraction of falsely rejected hypotheses is smaller than $mp_{(k)}/k$. The third step in the FDR procedure limits the fraction of falsely rejected hypotheses to be smaller than q, which is the threshold for the fraction of falsely rejected hypotheses. Thus the FDR procedure ensures that the fraction of erroneously detected relationships is smaller than a specified threshold.

This procedure is graphically illustrated by Fig. 2, adapted from Ventura et al. (2004). In order to clearly show the difference between the traditional and the FDR procedure, q and α were both set to 20%. At each location ($p_{(i)}$), p-values were ranked in increasing order and plotted against i/m, where m is the total number of locations. The blue dashed line is the traditional significance threshold. All dots below the blue dashed line are significant by the traditional procedure. With the FDR procedure, only those p-values under the red line (green dots) are significant. Therefore,

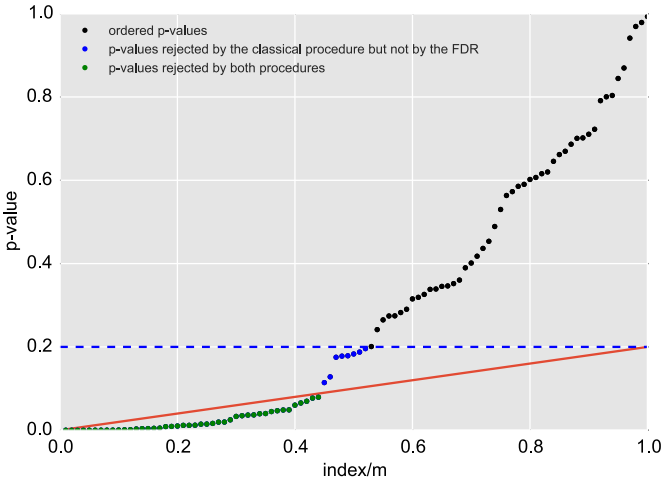

Fig. 2. Illustration of the traditional significance test procedure and the FDR procedure on an illustrative example, with $q = \alpha = 20\%$. p-values at each grid point ($p_{(i)}$) are ranked in increasing order, plotted against i/m, where m is the total number of grid points. The blue dashed line is the traditional α threshold for the p-value. Green dots indicate they are significant by both traditional and FDR procedure, and blue dots indicate they are only significant by the traditional procedure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

some p-values are deemed significant by the traditional procedure but not by the FDR procedure. Applying this methodology lowers the likelihood of identifying spurious correlations as significant, hence making correlation tests more stringent. We should note that the absence of significant correlations does not imply that the correlation is inexistent – only that the data do not provide enough evidence to reject the null hypothesis of zero correlation.

3.2. Application

Ventura et al. (2004) proposed a simple implementation of the FDR procedure for climate fields, and showed that its assumptions are robust to the spatial correlation levels typical of climate fields. Here we use the code of Benjamini and Hochberg (1995), which is equivalent. To show how to apply this procedure to field correlations, we use the study of Zhu et al. (2012) as an example. The authors generated a cellulose δ^{18}O record of Merkus pines for the past 140 years in Kririrom National Park, southern Cambodia (KRPM15B, 11.29°N; 104.25°E; 675 m). This record was dated by ring-counting. The authors assert that this cellulose δ^{18}O record is dominantly controlled by convection over the Indo-Pacific Warm Pool (IPWP), an interpretation buttressed by high correlations of cellulose δ^{18}O with instrumental precipitation and outgoing longwave radiation over the IPWP (Fig. 7 in Zhu et al., 2012). This explanation is reasonable because the cellulose δ^{18}O is mainly controlled by the δ^{18}O in precipitation during the rainy season, which is often depleted when the rainfall increases (the so-called "amount effect"; Dansgaard, 1964). During El Niño events, the precipitation in Southeast Asia is usually suppressed, which would lead to higher δ^{18}O values.

Here we consider the false discovery rate in this spatial correlation (Fig. 7 in Zhu et al., 2012). When the FDR is considered, none of the correlations are found significant, even at the relatively permissive 10% level chosen by authors (Fig. 3). This calls into question the proposed relationship between cellulose δ^{18}O and interannual changes in tropical convection at this site. However, we did find that the correlation between the cellulose δ^{18}O and (unsmoothed) NINO4 index is significant at the 5% level ($r = 0.45$, p-value $= 2.3 \times 10^{-6}$), considering autocorrelation as above. This

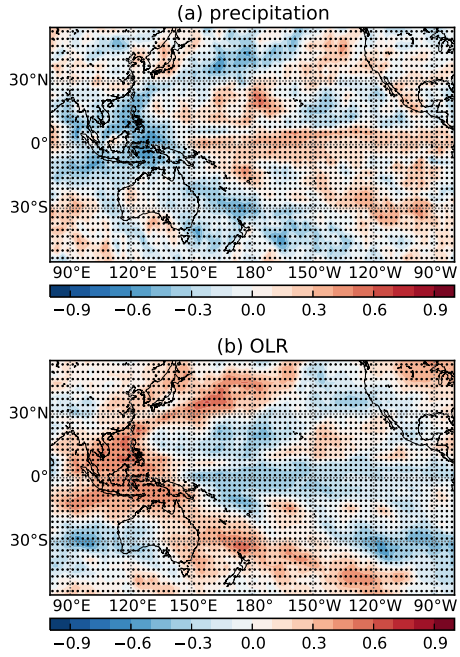

Fig. 3. Spatial correlation of October Kirirom cellulose δ^{18}O values with the October–November–December mean of (a) CMAP precipitation, and (b) NOAA interpolated OLR. Black dots indicate that the correlation does not pass the significance test at the 90% level.

indicates that cellulose δ^{18}O, or at least rainfall δ^{18}O at the site, may be connected with large scale changes due to the El Niño-Southern Oscillation. However, the explanation that local precipitation changes during El Niño events dominate the cellulose δ^{18}O may need further examination, for instance via forward modeling (Evans, 2007).

4. Challenge #3: age uncertainties

4.1. Theory

Age uncertainties have long been known to affect inferences made from paleoclimate data (e.g. Clark and Thompson, 1979), including correlation analysis (see Mudelsee, 2001; Rehfeld and Kurths, 2014, for recent examples). Indeed, Wunsch (2003) showed that even two randomly generated, unrelated time series could appear to correlate well with each other by adjusting the chronology within age uncertainties. Thus, quantifying age uncertainties is critical to the analysis of paleoclimate records. Many methods have been developed to do so (e.g. Haslett and Parnell, 2008; Blaauw and Christen, 2011). Additionally, various methods have been devised to deal with correlation (or covariance) under age uncertainties (Haam and Huybers, 2010). In the next section, we illustrate how age uncertainties may influence correlation analysis, and how they may compound the effects of autocorrelation and test multiplicity by drawing an example from a high-profile publication (McCabe-Glynn et al., 2013).

4.2. Compound challenges in one: the case of Crystal Cave

Recently, McCabe-Glynn et al. (2013) applied time-based calibrations to δ^{18}O data from a stalagmite from Crystal Cave, southern California. They interpreted the record as a proxy for sea surface temperature (SST) in the Kuroshio Extension region, arguing that warm SST in the area may generate southwesterly wind anomalies over southern California, bringing isotopically-enriched moisture to the cave. They also found a strong 22-year periodicity, which they linked to solar cycles, and found that some southwestern North American droughts coincided with episodes of warm Kuroshio Extension SSTs, as reconstructed from the Crystal Cave stalagmite. Some of these conclusions are based on a correlation analysis that encounters the three challenges of interest in this study: serial correlation, test multiplicity and age uncertainties. Here we will show one way to properly address these challenges.

4.3. Effect of serial correlation

We reuse the δ^{18}O data from stalagmite CRC-3, collected from Crystal Cave in Sequoia National Park, California (36.59°N; 118.82°W; 1386 m), which was interpolated at annual scale and archived online.[1] As in the original study, the record is correlated to SST anomaly data from the Kaplan SST v2 dataset (Kaplan et al., 1998) (1856–2007).

The correlation is shown in Fig. 4b, to be compared with Supplementary Fig. S6a in McCabe-Glynn et al. (2013). Because both the δ^{18}O series and SST field are intrinsically autocorrelated (the autocorrelation of δ^{18}O series is 0.95, and the autocorrelation of SST is 0.11–0.76, depending on location), the effective number of degrees of freedom is much lower than its theoretical value ($N - 2 = 149$). Indeed ν_{eff} is less than 60 in the Kuroshio Extension region (Fig. 4a). Hence, when this effect is considered (Fig. 4c), far fewer correlations pass the significance test (Fig. 4b, c). While McCabe-Glynn et al. (2013) had considered the effect of serial correlation using the method of Macias-Fauria et al. (2012), they did not graphically represent these results, giving little indication of where the relationship might be reliable. Nonetheless, our result is consistent with theirs, in that correlations over the Kuroshio Extension region pass the significance test with both approaches.

4.4. Effect of test multiplicity

Since the correlation between the δ^{18}O record and SST is also a field correlation, we need to consider test multiplicity. The result is shown in Fig. 4d, e. If autocorrelation is ignored, the FDR procedure results in fewer correlations passing the significance test (Fig. 4d vs. Fig. 4b). When both autocorrelation and multiplicity are considered, no correlation passes the significance test (Fig. 4e). This result suggests that the correlation between δ^{18}O record and instrumental SST may not be used as a basis for the record's interpretation.

4.5. Effect of age uncertainties

Another problem compounding serial correlation stems from the fact that the age model for the speleothem carries uncertainties of years to decades (Cheng et al., 2013), and these uncertainties may propagate to other inferences made from the proxy. The age uncertainties were quantified in McCabe-Glynn et al. (2013) using the StalAge algorithm (Scholz and Hoffmann, 2011). While the StalAge code exports 95%-confidence limits for the corresponding ages, it does not export possible age ensembles, which are essential for propagating uncertainty to other inferences. Here we leverage the power of ensembles to quantifying age uncertainties in correlation analysis.

4.5.1. Age model

We chose to model age uncertainties using Bchron (Haslett and Parnell, 2008), a Bayesian probability model allowing for random variations in accumulation rate between tie points. Bchron is capable of dealing with outliers and hiatuses (Parnell et al., 2011),

[1] ftp://ftp.ncdc.noaa.gov/pub/data/paleo/speleothem/northamerica/usa/california/crystal2013.txt.

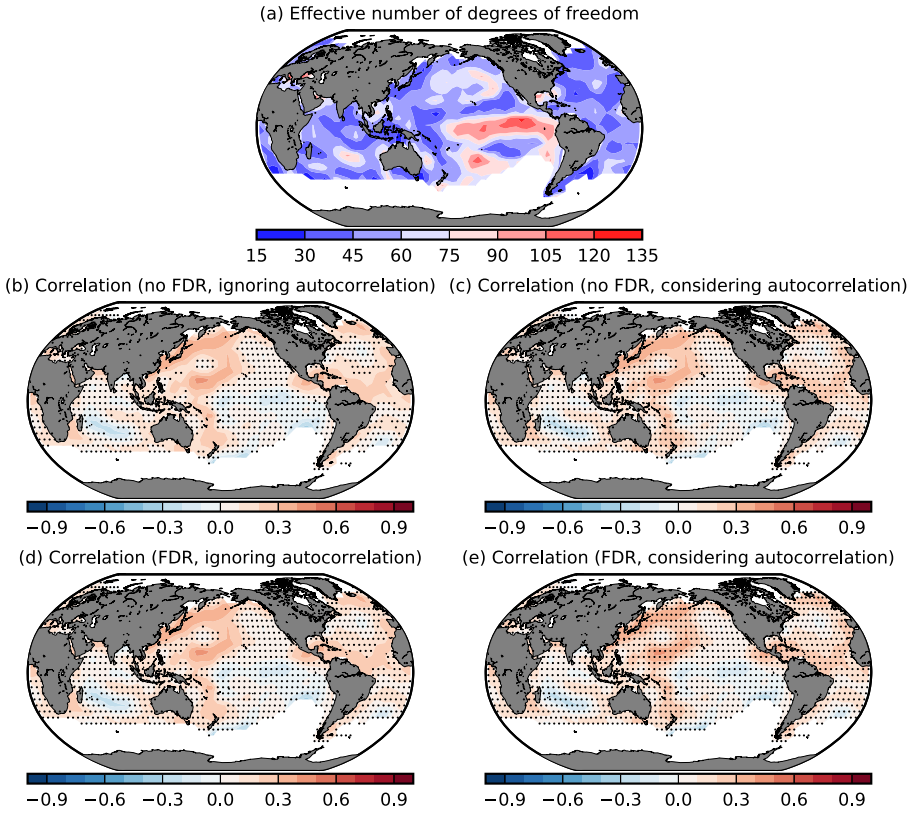

Fig. 4. (a) The effective number of degrees of freedom of the correlation between Crystal Cave δ^{18}O and SST from 1856–2007; The correlation between these series, considering the test multiplicity problem (False Discovery Rate) (d, e), or not (b, c); considering autocorrelation (c, e) or not (b, d). Black dots indicate that the correlation does not pass the significance test at the 5% level.

and naturally produces more hiatuses than StalAge. Bchron proceeds as follows: for each dated sample, the algorithm randomly selects a calendar date consistent with the age information (measured age and age error) and monotonicity (deeper sample with older age). It then inserts several points (at depths we want to know ages) between dated samples consistent with monotonicity, and then linearly interpolates between those points. Finally, it repeats this process many times until enough realizations fit the measured ages. The main advantage of Bchron here is the ability to extract an ensemble of age models all consistent with the posterior distribution of ages. For a review of different approaches to age modeling, see Scholz et al. (2012).

The age models are compared in Fig. 5, and one can see that they are quite close. The choice of age model (blue vs. black curve) introduces relatively small differences in the median age model, but the inclusion of a full ensemble of 1000 plausible age realizations makes a great difference indeed. Each of these age realizations corresponds to a different δ^{18}O time series: Fig. 6 (left) shows three of them, corresponding to the lower (2.5%), median (50%) and upper (97.5%) quantiles of the age distribution. One can see that many of the major features of the δ^{18}O timeseries can shift by about 50 years, making a correlation to instrumental data fraught with uncertainty.

This may be seen in more detail in Fig. 5. While the median age models from the two methods (blue and black curves) do appear quite close, there are large offsets between δ^{18}O timeseries (Fig. 6, right), especially before AD 1960. For instance, the large peak ca. 1890 in McCabe-Glynn et al. (2013), is centered around 1900 in the median Bchron age model. Such differences are especially significant if one tries to correlate them to other climatic timeseries, as we now do.

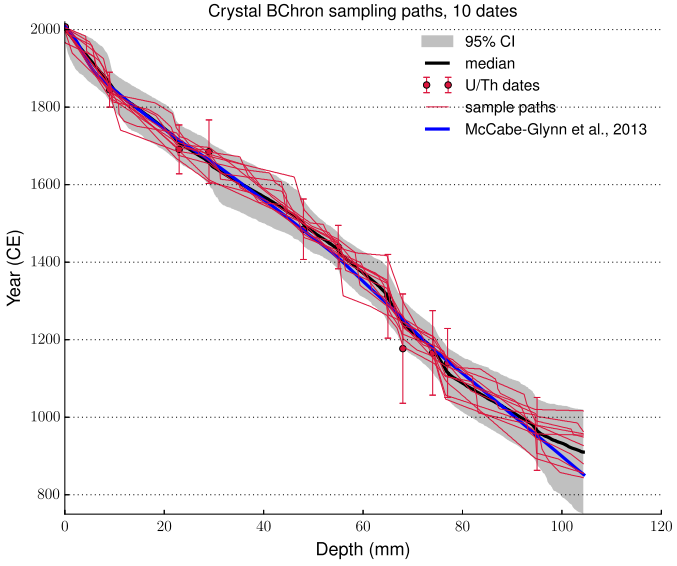

Fig. 5. The age modeling results of the Crystal Cave δ^{18}O record using a Bchron age model. The gray area is the 95% confidence interval of the age at each depth. The red lines show 10 random paths out of the 1000 age models generated, and the blue curve shows the StalAge-generated model used by McCabe-Glynn et al. (2013). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.5.2. Correlations considering age uncertainty and autocorrelation

We assess the impact of chronological uncertainties by repeating the previous analysis on each of the 1000 realizations of the δ^{18}O time series, similarly interpolated to an annual scale to facilitate correlations to SST. In the following analysis, we will also take autocorrelation into account.

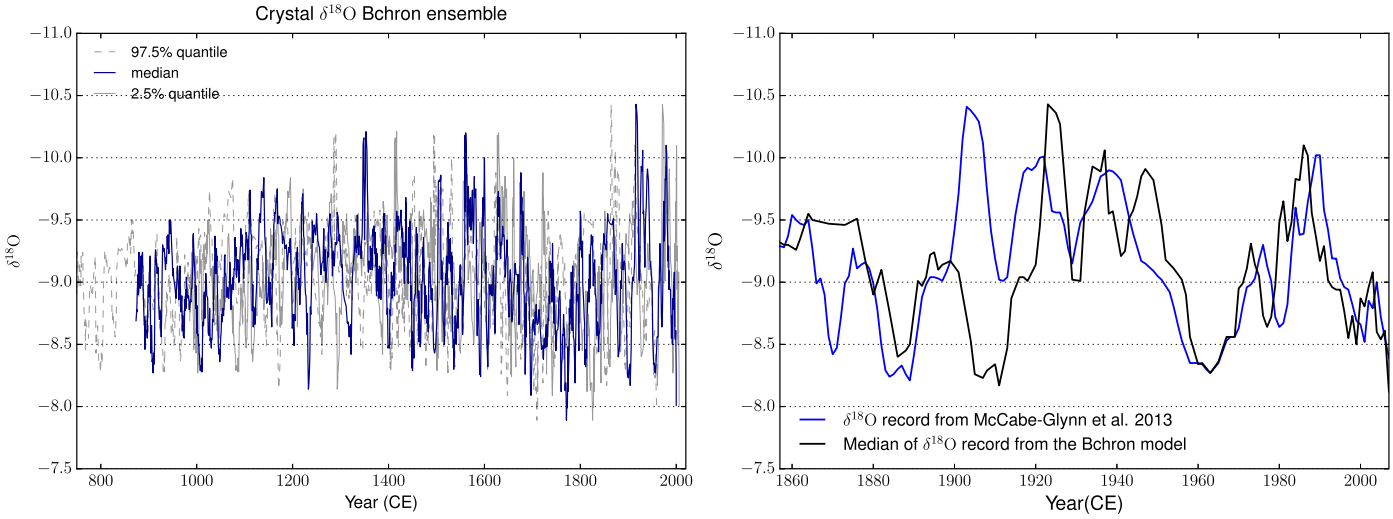

Fig. 6. Left panel: The median, 97.5% quantile and 2.5% quantile of the ensemble of 1000 Crystal δ^{18}O time generated from the Bchron model. Right panel: The time series of δ^{18}O record from McCabe-Glynn et al. (2013) (blue) and the median of δ^{18}O record from the Bchron model (black). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

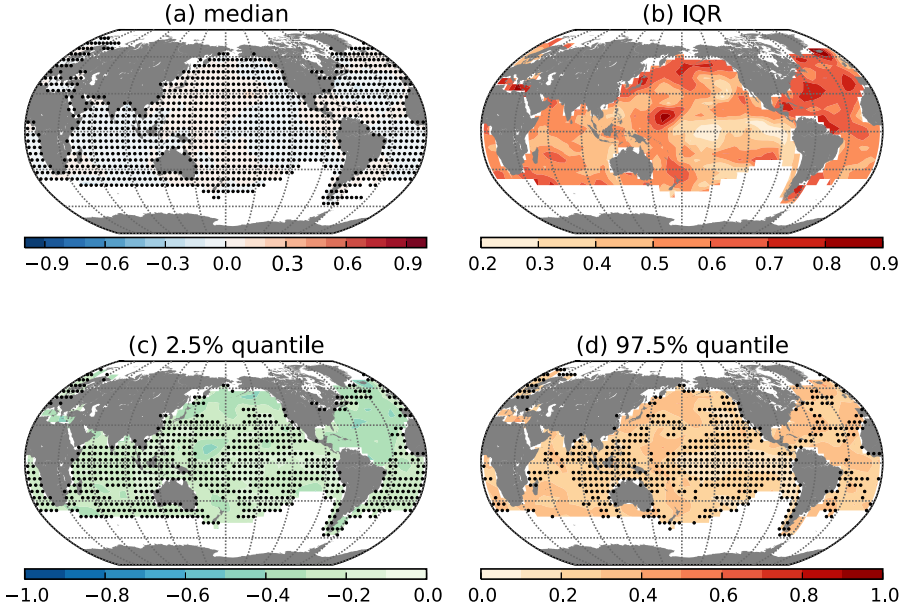

Fig. 7. Correlations considering age uncertainty. The median (a), interquartile range (b), the 2.5% (c) and the 97.5% (d) percentile of the correlation between the δ^{18}O age ensemble and SST. Black dots indicate that the correlation does not pass the significance test at the 5% level, accounting for serial correlation. Note that the interquartile range here is a measure of distributional spread, and has no measure of significance attached to it.

Using the full age ensemble, one obtains 1000 δ^{18}O-SST correlations for each grid point, from which one may infer an empirical distribution, whose median, 2.5% and 97.5% percentiles, and interquartile range (IQR) are reported in Fig. 7. The median is the aspect of the distribution that most studies use exclusively (for Gaussians, the median, mean and mode coincide). Due to the construction of our ensemble, the 2.5% quantile gathers some of the strongest negative correlations, and the 97.5% quantile gathers some of the strongest positive correlations. The IQR measures the spread between the 25% and 75% quantiles (the width of the distribution), and is therefore an indication of the spread of correlations due to age uncertainties alone.

The pattern of correlations for the median age model is similar to Fig. 4b, but the absolute values of correlations are much smaller, and the positive center in the North Pacific is shifted southward. Also, none of the median correlations pass the significance test at the 5% level. Since McCabe-Glynn et al. (2013) use the StalAge model, this suggests that the correlation between the δ^{18}O record and instrumental SST may be dependent on the age model. However, Fig. 5 clearly shows that the StalAge model is within the 95% confidence bounds of the Bchron age model, which underlines that age uncertainties are generally quite large compared to the timescale of variability in the SST record, and should therefore be accounted for in the analysis of correlations.

The pattern of the range of the correlation (IQR, Fig. 7b) is quite similar to Fig. 4b, which indicates that the regions of highest correlation in McCabe-Glynn et al. (2013) correspond to the regions of largest uncertainties in the age models. The 97.5% quantile and the 2.5% quantiles (Fig. 7c, d) also correspond to the regions of positive/negative correlation in Fig. 4b, and the correlation in some regions passes the significance test. However, the corresponding pattern differs from age ensemble to age ensemble, and from the published result, indicating a lack of robust relationship on which to build a reliable interpretation.

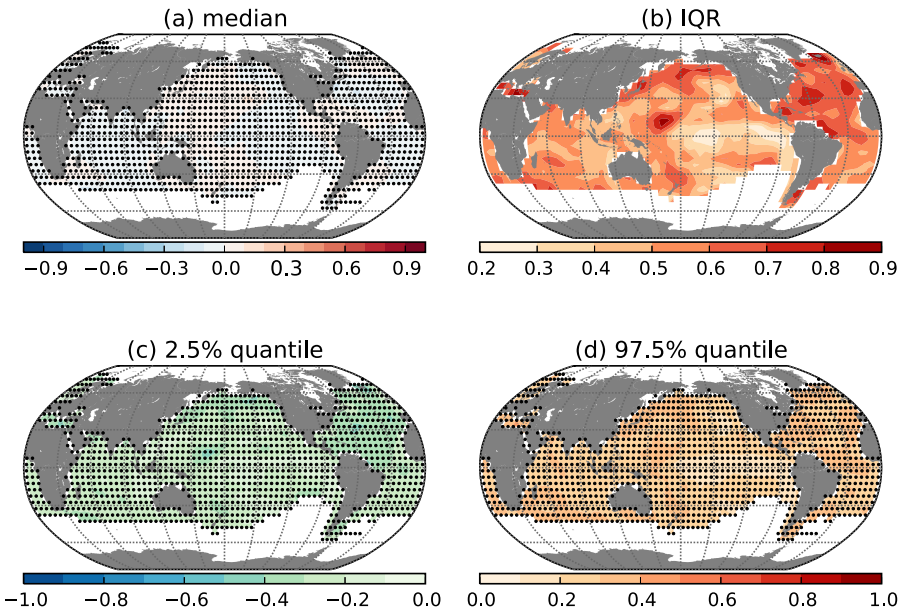

Fig. 8. Same as Fig. 7, but testing for field correlations while controlling for the False Discovery Rate.

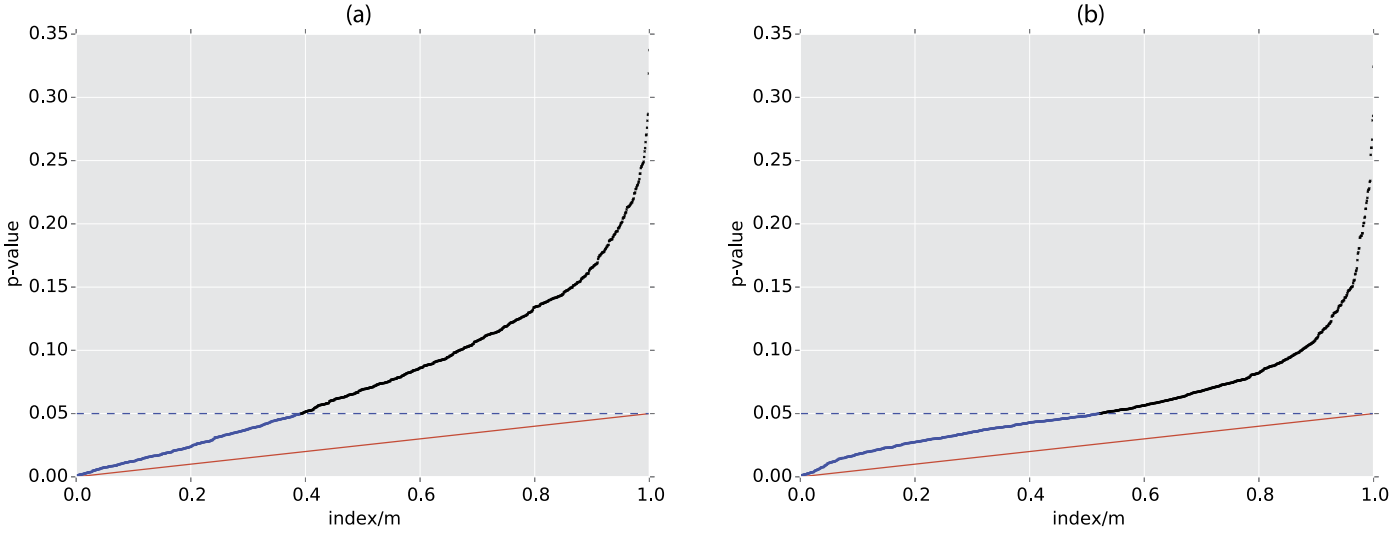

Fig. 9. Illustration of the FDR procedure on the 2.5% quantile (a) and the 97.5% quantile (b) correlations shown in Fig. 8, using the false discovery rate $q = 5\%$ (red line). p-values at each grid point ($p_{(i)}$) are ranked in increasing order, plotted against i/m, where m is the total number of grid points. The blue dashed line is the traditional 5% threshold for the p-value. Dots with p-values below this threshold are shown in blue. In this example, many dots fall below the nominal threshold (5%), but none fall below the red line, which means that they are not significant according to the FDR-controlling procedure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.6. Correlations considering all three challenges

Adding to these challenges, when the multiple-test problem is considered, none of the correlations for the 97.5% or 2.5% quantiles of the age ensemble pass the significance test (Fig. 8). This is because, while some p-values fall below the nominal threshold (Fig. 9, blue dashed lines), none drop under the FDR threshold (Fig. 9, red line), implying that these correlations are not significant at the 5% level. Thus the interpretation of Crystal Cave δ^{18}O as a proxy for SST in the Kuroshio Extension (or SST anywhere) should be revisited.

4.7. Time-uncertainty spectral analysis

Finally, we assess the influence of chronological uncertainties on inferences made in the spectral domain. We use the multi-taper method (Thomson, 1982), which achieves an optimal tradeoff between leakage and resolution. We use a half-bandwidth parameter of 4, which favors statistical significance over resolution, and is therefore the most conservative setting. The multi-taper spectra of the δ^{18}O record (855–2007 AD) from the two age models are presented in Fig. 10. It shows that the δ^{18}O record derived from the median of the Bchron ensemble has a dominant period ca. 18 years while the published data exhibit a dominant period of 21 years. This is clearly within age uncertainties, despite McCabe-Glynn et al. (2013) having used REDFIT, a variant of the Lomb–Scargle periodogram (Schulz and Mudelsee, 2002). Therefore, the multi-taper spectral method is suitable for the comparison. Considering that much of the uncertainty lies above the red line and gray area in Fig. 10, it is hard to distinguish significant periodicities from red noise. This suggests that there may not be any notable harmonic cycles in the δ^{18}O record at any scale less than 1000 years. Thus more evidence would be needed to draw a connection to the 22-year Hale solar cycle, as done in McCabe-Glynn et al. (2013).

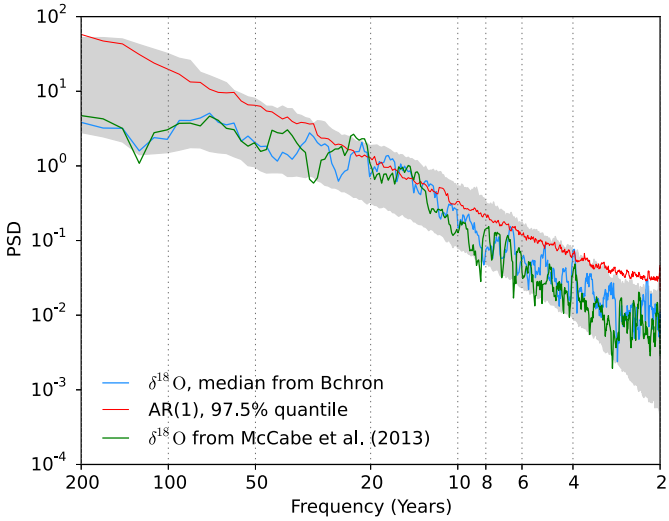

Fig. 10. The MTM-estimated spectra of the δ^{18}O record from the Bchron age model (blue line and gray shaded area) compared to the spectrum of the published record, together with a simulated AR(1) benchmark. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We showed how to address the three main challenges of correlation analysis and how these challenges weaken the conclusions of the original study. The example illustrates how each challenge in isolation would be enough to question the main conclusions, and their combination even more so. In closing, we note that although we do not find significant correlations between the Crystal Cave record and sea-surface temperature, we cannot rule out the existence of such correlations. The short calibration period and relatively large age uncertainties simply preclude the establishment of significant correlations with the current datasets.

Finally, we stress that the interpretation of the oxygen isotope composition of speleothem calcite is complex (LeGrande and Schmidt, 2009), and that other factors may influence the δ^{18}O record in Crystal Cave. The variation of δ^{18}O of precipitation in California can be affected by changes in condensation height (Buenning et al., 2013), and δ^{18}O in precipitation is known to be influenced by source-water composition due to shifts in storm-track location (Oster et al., 2015). Thus, the published interpretation of Crystal Cave δ^{18}O needs to be further investigated.

5. Discussion

In a provocative paper, Ioannidis (2005) concluded that "most published research findings are false". Central to this problem is a widespread tendency to hunt for statistical effects (often, correlations) that pass a significance threshold of 95% (i.e. a 5% probability of false positive), a practice often dubbed "p-hacking". While the paper focused on biomedical research, some of its conclusions transfer to paleoclimatology, and the Earth sciences at large. In particular, the existence of important challenges in correlation analysis (autocorrelation, false-discovery rate and age uncertainties) should be recognized by all practicing paleoclimatologists. In this article we illustrate these challenges using three examples, showing how published interpretations may be changed by one challenge alone, or by a combination of them.

We began with showing how autocorrelation and non-normality affect correlation analysis by using the example from Proctor et al. (2000), who reconstructed the NAO index using layer thickness in a stalagmite from Scotland. We found that properly taking those into account did not change the interpretation, though the latter would have been more robust had the authors taken these challenges into consideration. Autocorrelation is commonplace in paleoclimate proxies because many processes (e.g. bioturbation in sediment cores, groundwater mixing in speleothems, and firn diffusion in ice cores) smooth out climate signals. Autocorrelation should always be a concern unless it can be demonstrated otherwise.

The test multiplicity problem is another serious challenge of correlation analysis, as we showed in the example of Zhu et al. (2012). The spatial correlations between their cellulose δ^{18}O record and instrumental precipitation/outgoing longwave radiation over the Indo-Pacific Warm Pool in this paper are not significant after considering the false discovery rate. However, we find correlations with the NINO4 index significant, with the index explaining about 25% of the record's variance.

Age uncertainties also challenge the robustness of correlation analysis and we show that they may also combine with other challenges in the example of McCabe-Glynn et al. (2013), who claimed, on the basis of correlations to the SST field, to identify a relationship between the δ^{18}O record of a speleothem from southern California and the SST in the Kuroshio Extension region. We show, by considering all three challenges, that no correlation survives the test.

These three examples lead us to draw attention to the importance of:

- using established statistical procedures to guard against the misleading impacts of spurious correlations. Several books (Wilks, 2011; Mudelsee, 2013; Emile-Geay, 2016) address these challenges in more detail.
- using the rich output of age modeling software (not just the median age) to appraise the effect of age uncertainties on a study's conclusions.
- establishing a mechanistic understanding for proxy signals, and only relying on statistical approaches when there are sufficient numbers of degrees of freedom to unequivocally reject chance correlations.

While we do not dispute that the records scrutinized here contain potentially valuable climatic information, our main message is that it is often impossible to establish so by a purely statistical approach. Instead, we encourage detailed process studies to elucidate the climatic and/or hydrological controls on δ^{18}O, or other proxies, in various archives.

For speleothems, possible strategies involve the forward modeling of cave processes (such as Baker et al., 2012; Partin et al., 2013, and others), and/or cave instrumentation (such as Spötl et al., 2005; Partin et al., 2012, and others) to better ascertain the processes that control the recorded oxygen isotope signal. For many speleothem studies, age modeling considerations will be of secondary importance: in studies of glacial-interglacial cycles, for instance, age offsets of a few decades are immaterial. However, the interpretation of δ^{18}O in climate proxies (whether in terms of rainfall, temperature, or other factors) is usually complex, and therefore should be backed by isotope-enabled models (such as LeGrande and Schmidt, 2009; Pausata et al., 2011, and others). We note that proxy system modeling is a burgeoning field with applications to all paleoclimate archives, not just speleothems (e.g. Schmidt, 1999; Evans, 2007; Dee et al., 2015).

For the data collecting and measuring stage, high resolution sampling is suggested; von Gunten et al. (2012) suggested collecting 80–100 data points over the calibration period for time-based calibrations to achieve a sufficient effective sample size. If the smoothing scale is known, the record with smaller smoothing scale should be used in the time-based calibration. Also we should note that age uncertainties set the limit of the usage of proxy data. For example, proxy data with decadal age uncertainties cannot be

used in interannual-scale research questions, but may be used in centennial-scale research questions (Birks et al., 2012).

Our goal in presenting these results is not to indict a particular set of authors, as the unsophisticated data-analytical practices of the original studies are unfortunately rather common in the paleoclimate literature. Instead, we wish to draw attention to under-appreciated statistical issues, with the hope of lessening the occurrences of proxy interpretations based on spurious correlations, and to improve the robustness of future paleoclimate studies. In this spirit, we are making the Python code associated with this study freely available at https://github.com/ClimateTools/Correlation_EPSL in order to disseminate best practices.

Acknowledgements

We thank Chris Paciorek for his False Discovery Rate code, and thank Andy Baker for providing the weather record of Assynt in Scotland. This work was supported by Grant 1347213 from the US National Science Foundation and by a Dornsife Merit Fellowship from the University of Southern California.

References

Baker, A., Bradley, C., Phipps, S., Fischer, M., Fairchild, I., Fuller, L., Spötl, C., Azcurra, C., 2012. Millennial-length forward models and pseudoproxies of stalagmite δ^{18}O: an example from NW Scotland. Clim. Past 8, 1153–1167. http://dx.doi.org/10.5194/cp-8-1153-2012.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc., Ser. B 57, 289–300. http://dx.doi.org/10.2307/2346101.

Birks, J.B., Lotter, A.F., Juggins, S., Smol, J.P., 2012. Tracking Environmental Change Using Lake Sediments: Data Handling and Numerical Techniques, vol. 5. Springer, Netherlands.

Blaauw, M., Christen, J.A., 2011. Flexible paleoclimate age-depth models using an autoregressive gamma process. Bayesian Anal. 6, 457–474. http://dx.doi.org/10.1214/ba/1339616472.

Black, B.A., Griffin, D., van der Sleen, P., Wanamaker, A.D., Speer, J.H., Frank, D.C., Stahle, D.W., Pederson, N., Copenheaver, C.A., Trouet, V., et al., 2016. The value of crossdating to retain high-frequency variability, climate signals, and extreme events in environmental proxies. Glob. Change Biol. http://dx.doi.org/10.1111/gcb.13256.

Buenning, N.H., Stott, L., Kanner, L., Yoshimura, K., 2013. Diagnosing atmospheric influences on the interannual ^{18}O/^{16}O variations in Western US precipitation. Water 5, 1116–1140. http://dx.doi.org/10.3390/w5031116.

Cheng, H., Edwards, R.L., Shen, C.C., Polyak, V.J., Asmerom, Y., Woodhead, J., Hellstrom, J., Wang, Y., Kong, X., Spötl, C., et al., 2013. Improvements in ^{230}Th dating, ^{230}Th and ^{234}U half-life values, and U–Th isotopic measurements by multi-collector inductively coupled plasma mass spectrometry. Earth Planet. Sci. Lett. 371, 82–91. http://dx.doi.org/10.1016/j.epsl.2013.04.006.

Clark, R., Thompson, R., 1979. A new statistical approach to the alignment of time series. Geophys. J. Int. 58, 593–607. http://dx.doi.org/10.1111/j.1365-246X.1979.tb04796.x.

Crowley, T.J., 1999. Correlating high-frequency climate variations. Paleoceanography 14, 271–272. http://dx.doi.org/10.1029/1999PA900003.

Dansgaard, W., 1964. Stable isotopes in precipitation. Tellus 16, 436–468. http://dx.doi.org/10.1111/j.2153-3490.1964.tb00181.x.

Dawdy, D., Matalas, N., 1964. Statistical and Probability Analysis of Hydrologic Data, Part III: Analysis of Variance, Covariance and Time Series. McGraw-Hill.

Dee, S., Emile-Geay, J., Evans, M.N., Allam, A., Steig, E.J., Thompson, D.M., 2015. PRYSM: an open-source framework for PRoxY System Modeling, with applications to oxygen-isotope systems. J. Adv. Model. Earth Syst. 7, 1220–1247. http://dx.doi.org/10.1002/2015MS000447.

Emile-Geay, J., 2016. Data Analysis in the Earth & Environmental Sciences, second ed. FigShare. http://dx.doi.org/10.6084/m9.figshare.1014336.

Emile-Geay, J., Tingley, M., 2016. Inferring climate variability from nonlinear proxies: application to palaeo-ENSO studies. Clim. Past 12, 31–50. http://dx.doi.org/10.5194/cp-12-31-2016.

Evans, M.N., 2007. Toward forward modeling for paleoclimatic proxy signal calibration: a case study with oxygen isotopic composition of tropical woods. Geochem. Geophys. Geosyst. 8, Q07008. http://dx.doi.org/10.1029/2006GC001406.

Ghil, M., Allen, M., Dettinger, M., Ide, K., Kondrashov, D., Mann, M., Robertson, A.W., Saunders, A., Tian, Y., Varadi, F., et al., 2002. Advanced spectral methods for climatic time series. Rev. Geophys. 40. http://dx.doi.org/10.1029/2001RG000092.

Haam, E., Huybers, P., 2010. A test for the presence of covariance between time-uncertain series of data with application to the Dongge Cave speleothem and atmospheric radiocarbon records. Paleoceanography 25. http://dx.doi.org/10.1029/2008PA001713.

Haslett, J., Parnell, A., 2008. A simple monotone process with application to radiocarbon-dated depth chronologies. J. R. Stat. Soc., Ser. C, Appl. Stat. 57, 399–418. http://dx.doi.org/10.1111/j.1467-9876.2008.00623.x.

Ioannidis, J., 2005. Why most published research findings are false. PLoS Med. 2, e124. http://dx.doi.org/10.1371/journal.pmed.0020124.

Jones, P., Briffa, K., Osborn, T., Lough, J., van Ommen, T., Vinther, B., Luterbacher, J., Wahl, E., Zwiers, F., Mann, M., Schmidt, G., Ammann, C., Buckley, B., Cobb, K., Esper, J., Goosse, H., Graham, N., Jansen, E., Kiefer, T., Kull, C., Kuttel, M., Mosley-Thompson, E., Overpeck, J., Riedwyl, N., Schulz, M., Tudhope, A., Villalba, R., Wanner, H., Wolff, E., Xoplaki, E., 2009. High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. Holocene 19, 3–49. http://dx.doi.org/10.1177/0959683608098952.

Kaplan, A., Cane, M.A., Kushnir, Y., Clement, A.C., Blumenthal, M.B., Rajagopalan, B., 1998. Analyses of global sea surface temperature 1856–1991. J. Geophys. Res., Oceans 103, 18567–18589. http://dx.doi.org/10.1029/97JC01736.

Khider, D., Huerta, G., Jackson, C., Stott, L., Emile-Geay, J., 2015. A Bayesian, multivariate calibration for Globigerinoides ruber Mg/Ca. Geochem. Geophys. Geosyst. 16, 2916–2932. http://dx.doi.org/10.1002/2015GC005844.

LeGrande, A., Schmidt, G., 2009. Sources of Holocene variability of oxygen isotopes in paleoclimate archives. Clim. Past 5, 441–455. http://dx.doi.org/10.5194/cp-5-441-2009.

Macias-Fauria, M., Grinsted, A., Helama, S., Holopainen, J., 2012. Persistence matters: estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia 30, 179–187. http://dx.doi.org/10.1016/j.dendro.2011.08.003.

McCabe-Glynn, S., Johnson, K.R., Strong, C., Berkelhammer, M., Sinha, A., Cheng, H., Edwards, R.L., 2013. Variable North Pacific influence on drought in southwestern North America since AD 854. Nat. Geosci. 6, 617–621. http://dx.doi.org/10.1038/NGEO1862.

Mudelsee, M., 2001. The phase relations among atmospheric CO_2 content, temperature and global ice volume over the past 420 ka. Quat. Sci. Rev. 20, 583–589. http://dx.doi.org/10.1016/S0277-3791(00)00167-0.

Mudelsee, M., 2013. Climate Time Series Analysis. Springer.

Oster, J.L., Montañez, I.P., Santare, L.R., Sharp, W.D., Wong, C., Cooper, K.M., 2015. Stalagmite records of hydroclimate in central California during termination 1. Quat. Sci. Rev. 127, 199–214. http://dx.doi.org/10.1016/j.quascirev.2015.07.027.

Parnell, A.C., Buck, C.E., Doan, T.K., 2011. A review of statistical chronology models for high-resolution, proxy-based Holocene palaeoenvironmental reconstruction. Quat. Sci. Rev. 30, 2948–2960. http://dx.doi.org/10.1016/j.quascirev.2011.07.024.

Partin, J., Quinn, T., Shen, C.C., Emile-Geay, J., Taylor, F., Maupin, C., Lin, K., Jackson, C., Banner, J., Sinclair, D., Huh, C.A., 2013. Multidecadal rainfall variability in South Pacific Convergence Zone as revealed by stalagmite geochemistry. Geology. http://dx.doi.org/10.1130/G34718.1.

Partin, J.W., Jenson, J.W., Banner, J.L., Quinn, T.M., Taylor, F.W., Sinclair, D., Hardt, B., Lander, M.A., Bell, T., Miklavič, B., et al., 2012. Relationship between modern rainfall variability, cave dripwater, and stalagmite geochemistry in Guam, USA. Geochem. Geophys. Geosyst. 13. http://dx.doi.org/10.1029/2011GC003930.

Pausata, F.S., Battisti, D.S., Nisancioglu, K.H., Bitz, C.M., 2011. Chinese stalagmite δ^{18}O controlled by changes in the Indian monsoon during a simulated Heinrich event. Nat. Geosci. 4, 474–480. http://dx.doi.org/10.1038/ngeo1169.

Proctor, C., Baker, A., Barnes, W., Gilmour, M., 2000. A thousand year speleothem proxy record of North Atlantic climate from Scotland. Clim. Dyn. 16, 815–820. http://dx.doi.org/10.1007/s003820000077.

Rehfeld, K., Kurths, J., 2014. Similarity estimators for irregular and age-uncertain time series. Clim. Past 10, 107–122. http://dx.doi.org/10.5194/cp-10-107-2014.

Schmidt, G.A., 1999. Forward modeling of carbonate proxy data from planktonic foraminifera using oxygen isotope tracers in a global ocean model. Paleoceanography 14, 482–497. http://dx.doi.org/10.1029/1999PA900025.

Scholz, D., Hoffmann, D.L., 2011. StalAge – an algorithm designed for construction of speleothem age models. Quat. Geochronol. 6, 369–382. http://dx.doi.org/10.1016/j.quageo.2011.02.002.

Scholz, D., Hoffmann, D.L., Hellstrom, J., Ramsey, C.B., 2012. A comparison of different methods for speleothem age modelling. Quat. Geochronol. 14, 94–104. http://dx.doi.org/10.1016/j.quageo.2012.03.015.

Schulz, M., Mudelsee, M., 2002. REDFIT: estimating red-noise spectra directly from unevenly spaced paleoclimatic time series. Comput. Geosci. 28, 421–426.

Spötl, C., Fairchild, I.J., Tooth, A.F., 2005. Cave air control on dripwater geochemistry, Obir Caves (Austria): implications for speleothem deposition in dynamically ventilated caves. Geochim. Cosmochim. Acta 69, 2451–2468. http://dx.doi.org/10.1016/j.gca.2004.12.009.

Storey, J.D., 2002. A direct approach to false discovery rates. J. R. Stat. Soc., Ser. B, Stat. Methodol. 64, 479–498. http://dx.doi.org/10.1111/1467-9868.00346.

Thomson, D.J., 1982. Spectrum estimation and harmonic-analysis. Proc. IEEE 70, 1055–1096. http://dx.doi.org/10.1109/PROC.1982.12433.

Tierney, J.E., Tingley, M.P., 2014. A Bayesian, spatially-varying calibration model for the TEX86 proxy. Geochim. Cosmochim. Acta 127, 83–106. http://dx.doi.org/10.1016/j.gca.2013.11.026.

Tingley, M.P., Craigmile, P.F., Haran, M., Li, B., Mannshardt, E., Rajaratnam, B., 2012. Piecing together the past: statistical insights into paleoclimatic reconstructions. Quat. Sci. Rev. 35, 1–22. http://dx.doi.org/10.1016/j.quascirev.2012.01.012.

Trouet, V., Esper, J., Graham, N.E., Baker, A., Scourse, J.D., Frank, D.C., 2009. Persistent positive North Atlantic Oscillation mode dominated the medieval climate anomaly. Science 324, 78–80. http://dx.doi.org/10.1126/science.1166349.

Ventura, V., Paciorek, C.J., Risbey, J.S., 2004. Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. J. Climate 17, 4343–4356. http://dx.doi.org/10.1175/3199.1.

von Gunten, L., Grosjean, M., Kamenik, C., Fujak, M., Urrutia, R., 2012. Calibrating biogeochemical and physical climate proxies from non-varved lake sediments with meteorological data: methods and case studies. J. Paleolimnol. 47, 583–600. http://dx.doi.org/10.1007/s10933-012-9582-9.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences: An Introduction. Academic Press, San Diego.

Wunsch, C., 2003. Greenland–Antarctic phase relations and millennial time-scale climate fluctuations in the Greenland ice-cores. Quat. Sci. Rev. 22, 1631–1646. http://dx.doi.org/10.1016/S0277-3791(03)00152-5.

Yule, G.U., 1926. Why do we sometimes get nonsense-correlations between time-series – a study in sampling and the nature of time-series. J. R. Stat. Soc. 89, 1–63. http://dx.doi.org/10.2307/2341482.

Zhu, M., Stott, L., Buckley, B., Yoshimura, K., Ra, K., 2012. Indo-Pacific Warm Pool convection and ENSO since 1867 derived from Cambodian pine tree cellulose oxygen isotopes. J. Geophys. Res., Atmos. 117. http://dx.doi.org/10.1029/2011JD017198.