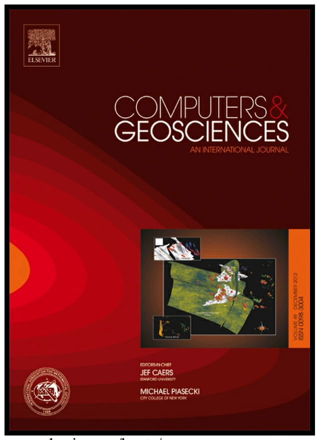
Author's Accepted Manuscript



www.elsevier.com/locate/cageo

Correlation confidence limits for unevenly sampled data

Jason Roberts, Mark Curran, Samuel Poynter, Andrew Moy, Tas van Ommen, Tessa Vance, Carly Tozer, Felicity Graham, Duncan Young, Christopher Plummer, Joel Pedro, Don Blankenship, Martin Siegert

PII: S0098-3004(16)30445-9
DOI: http://dx.doi.org/10.1016/j.cageo.2016.09.011
Reference: CAGEO3842

To appear in: *Computers and Geosciences*

Received date: 3 December 2015
Revised date: 18 July 2016
Accepted date: 25 September 2016

Correlation confidence limits for unevenly sampled data

Jason Roberts[a,b,*], Mark Curran[a,b], Samuel Poynter[c], Andrew Moy[a,b], Tas van Ommen[a,b], Tessa Vance[b], Carly Tozer[b,d], Felicity Graham[e], Duncan Young[f], Christopher Plummer[b,e], Joel Pedro[b,g], Don Blankenship[f], Martin Siegert[h]

[a]*Department of the Environment, Australian Antarctic Division, Kingston, Tasmania 7050, Australia*
[b]*Antarctic Climate and Ecosystems Cooperative Research Centre, University of Tasmania, Private Bag 80, Hobart, Tasmania 7001, Australia*
[c]*Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Burnley, Victoria 3121, Australia*
[d]*University of Newcastle, Callaghan, NSW, Australia*
[e]*Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania 7001, Australia*
[f]*Jackson School of Geosciences, University of Texas at Austin, Austin, Texas, USA*
[g]*Centre for Ice and Climate, Niels Bohr Institute, University of Copenhagen, Denmark*
[h]*Grantham Institute and Department of Earth Science and Engineering, Imperial College London, London, UK*

Abstract

Estimation of correlation with appropriate uncertainty limits for scientific data that are potentially serially correlated is a common problem made seriously challenging especially when data are sampled unevenly in space and/or time. Here we present a new, robust method for estimating correlation with uncertainty limits between autocorrelated series that does not require either resampling or interpolation. The technique employs the Gaussian kernel method with a bootstrapping resampling approach to derive the probability density function and resulting uncertainties. The method is validated using an example from radar geophysics. Autocorrelation and error bounds are estimated for an airborne radio-echo profile of ice sheet thickness. The

*corresponding author Jason.Roberts@aad.gov.au

computed limits are robust when withholding 10%, 20%, and 50% of data. As a further example, the method is applied to two time-series of methanesulphonic acid in Antarctic ice cores from different sites. We show how the method allows evaluation of the significance of correlation where the signal-to-noise ratio is low and reveals that the two ice cores exhibit a significant common signal.

Keywords: Unevenly sampled data, autocorrelation, bootstrapping, Gaussian Kernel Method, confidence limits

1. Introduction

Sparse data correlation techniques, and the confidence limits associated with them, are a keystone of quantitative analysis in geoscience. However, uneven sampling of data is a common feature in many fields, and our inability to prescribe appropriate interpolations between data may hinder the statistical application of results. In many cases, this may come about as an inherent sampling non-uniformity. In the case of ice cores, for example, the relationship between the spatial and temporal distribution of a sample material varies with depth such that uniform spatial sampling generates non-uniform sampling on a temporal scale. Further difficulty arises from missing data or data gaps, which may be caused by physical sample size constraints, damage, or loss of samples due to contamination or analytical problems. Where numerical methods require evenly sampled data, interpolation is necessary, but must be used cautiously to avoid signal artifacts. The use of common software tools to interpolate between data points often comes at the expense of robustness, as bias may be introduced.

Rehfeld and Kurths (2014) investigated this issue in detail, bench-marking a variety of techniques to overcome the challenges introduced by irregularly-sampled time series. The use of a Gaussian kernel method gave a reliable and robust estimation in comparison to commonly-used interpolation approaches such as resampling onto a common uniform independent grid. Complications arise for irregularly-sampled data with inherent autocorrelations, however, as the estimation of a confidence interval, or some other measure of significance, requires explicit and quantitative consideration of the autocorrelations (Mudelsee, 2003; Ólafsdóttir and Mudelsee, 2014). Several methods exist for the assessment of significance, for evenly sampled data, in the presence of autocorrelation. Such methods include the effective spatial degrees of freedom method of Bretherton et al. (1999) which uses classical tests with a reduced number of degrees of freedom to account for autocorrelations in the data, and data surrogates such as bootstrapping and Fourier space methods. These latter methods make no assumptions on the distribution of the data (Mudelsee, 2003), so may be more appropriate for many real-world datasets. Compared to standard bootstrapping techniques, Fourier space methods have the advantage of preserving linear correlations, but lose many of their computational advantages for irregularly-sampled data.

Here, we report the development of the Gaussian kernel method, extended to provide confidence interval information, with application to airborne glacier geophysical data. An evenly-sampled, highly autocorrelated dataset of Antarctic ice thicknesses from the ICECAP (Investigating Cryospheric Evolution through Collaborative Aerogeophysical Profiling) project (see Fig. 1 for location) provides a suitable test data set to validate the approach. The correlation and confidence interval distribution is compared to a recently published method (Ólafsdóttir and Mudelsee, 2014). Data were randomly removed to simulate the effect of uneven data spacing and the resulting autocorrelation distributions compared.

As a second independent demonstration of the strength of the technique we compute the correlation between time series of methanesulphonic acid (MSA) concentration in two Antarctic ice cores (see Fig. 1 for location). MSA has been used as a proxy for Antarctic sea ice extent (Curran et al., 2003), based on the production of MSA from sea ice-associated phytoplankton which are known to be a dominant sulphur source from the sea-ice edge in Antarctica (Vance et al., 2013). Confirming that a statistically significant (at a 95% confidence interval) relationship exists between the two MSA records supports the hypothesis that the records preserve a common environmental signal.

While the Mudelsee (2003); Ólafsdóttir and Mudelsee (2014) method can be used on unevenly spaced climate time series data, in cases where the data are both unevenly spaced and on a different time base their method requires interpolation or resampling. Our Gaussian Kernel-based method removes the need for such resampling, making it well suited to computing correlations between paleoclimate records from different locations and different archives, in which different time bases are ubiquitous.

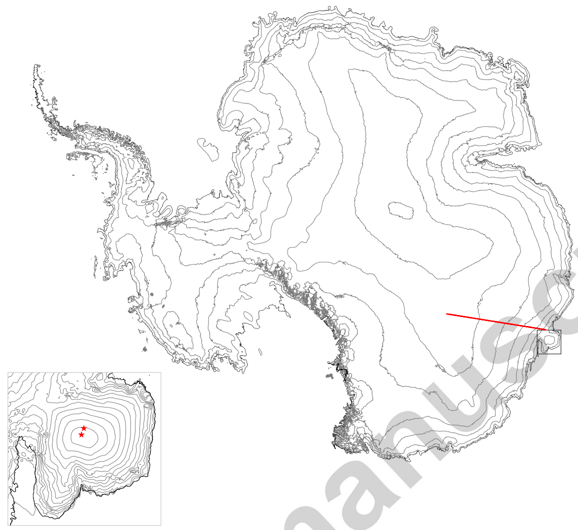


Figure 1: Location of an airborne radar transect yielding ice thickness data (red line). Elevation contours at 500 m are from Bamber et al. (2009) (grey lines) and the ice sheet grounding line is from Bindschadler et al. (2011). Inset shows the Law Dome region of East Antarctica and the sites of the two ice cores (red stars), with DSS97 being closer to the dome summit and W10K being close to the 1300 m elevation contour.

2. Method

2.1. Correlation

Correlations (C_{xy}) between unevenly and differently sampled series (x_i and y_j) are calculated using the Gaussian kernel correlation slotting (Rehfeld et al., 2011).

$$C_{xy} = \frac{1}{\sigma_x \sigma_y} \frac{\sum_{i=1}^{n_x} \sum_{j=i}^{n_y} \left(x_i - \overline{x}\right)\left(y_j - \overline{y}\right) K\left(d_{x_i} - d_{y_j}\right)}{\sum_{i=1}^{n_x} \sum_{j=i}^{n_y} K\left(d_{x_i} - d_{y_j}\right)} \tag{1}$$

where the average of the two series x_i and y_j (of length n_x and n_y) is $\overline{x}$ and $\overline{y}$, respectively, and d_x and d_y are the independent variables (typically time or distance) for x and y respectively, and may differ from each other. The Gaussian kernel $K(d) = \frac{1}{\sqrt{2\pi h}} \exp\left(-d^2/2h^2\right)$ uses a width parameter (h) of one quarter of the larger of the average spacing of the two data series. Unlike Rehfeld et al. (2011) who normalise the signals to have zero mean and unit variance, we use the original signals and correct for the mean and estimate the standard deviations (σ_x and σ_y) using the same weighted summation Gaussian kernel ($K(d)$) as used in Eq. 1.

2.2. Bootstrapping

Confidence intervals (95%) are estimated using a stationary bootstrapping technique (Politis and Romano, 1994). This method accounts for persistence (serial correlation) and the associated reduction in the effective degrees of freedom in the data (Wilks, 2006) by generating resampled data series based on blocks of data from the original series. The block lengths vary randomly, but the average block length is a function of the persistence of the data (Mudelsee, 2003). We estimate the persistence of the data from the offset required for the autocorrelation to fall to $\frac{1}{e}$, and account for the uneven time intervals between data samples.

For each member of the dataset of bootstrap resamples, in which $n =$ 2000, the correlation is estimated. The 95% confidence interval for the correlation coefficient is then calculated using the bias-corrected and accelerated (BCa) bootstrap method (Efron, 1987), which adjusts the result to account for bias in the resampled set compared to the original point estimate of the correlation coefficient from the original data.

Finally, in order to improve the robustness of the estimations, an additional 24 bootstrap resample sets of 2000 members each are generated and separate estimates of the 95% confidence intervals made. The final estimate of the lower (and upper) 95% confidence interval bounds is the median of the 25 BCa bootstrap estimates of the lower (upper) bound. The use of the median of 25 bootstrap resample sets was found to produce robust results for the test datasets used here; additional resample sets may be required for particularly problematic data.

The cumulative probability density function and its inverse, both used for the BCa calculations, are based on algorithms from Abramowitz and Stegun (1968) and Wichura (1988) respectively.

Bootstrapping methods can produce accurate estimates for relatively short data series without significant autocorrelation. However, the advantages of no underlying assumptions on the distribution of data comes at a cost: longer data series are required when the data has significant autocorrelation. Mudelsee (2003) investigated this and concluded that several hundred or more data points may be required for highly autocorrelated data. Additionally, for highly skewed data, the BCa method may predict 95% confidence bounds that do not include the correlation estimate for the original data.

3. Validation

To validate the method, a uniformly sampled data set of 951 points was used. This allowed for comparison with existing methods for uniformly sampled data (Ólafsdóttir and Mudelsee, 2014). In addition, unevenly sampled data can be simulated by removing a portion of the data on a random basis. Three unevenly sampled datasets were generated, withholding 10%, 20%, and 50% of the data respectively.

The validation data set chosen was 950 line km of ice thickness data (see Fig. 2) over Eastern Antarctica (Roberts et al., 2011; Young et al., 2011) (ICECAP flight ASB/JKB1a/R10Eb). This dataset covers different subglacial terrains, including flat sedimentary basins (Wright et al., 2012), steep-sided glacial valleys (Young et al., 2011) and rough highlands (Roberts et al., 2011). This dataset has a high degree of autocorrelation at large displacements (see Fig. 3), indicative of long characteristic length scales. The high autocorrelations over relatively large spatial scales shown in Fig. 3 is consistent with the results of Smith et al. (2007) who found that for East Antarctica, variograms of topography could be well represented by exponential models with length scales up to 700 km. Since the ice sheet has a relatively smooth surface (compared with the underlying topography), bedrock topography is strongly related to ice thickness. Therefore it is not unexpected that the ice thickness may show a long characteristic scale distribution as terrain frequently exhibits long characteristic scales especially when preserved

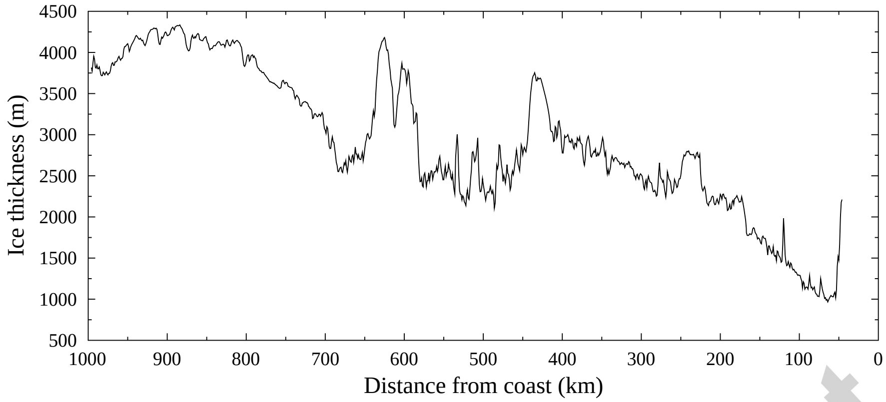


Figure 2: ICECAP radar data from flight ASB/JKB1a/R10Eb, revealing a major deep subglacial sedimentary basin, between 1000 and 700 km, in which the bed is remarkably flat; steep sided 1500 m-deep valleys between 700 and 600 km, and at 400 km; rough highland between 600 and 500 km; and flat subglacial terrain between 400 and 0 km. The profile illustrates a range of topographies measured by airborne radar profiling. The data are often hampered by losses at great ice depth, over very steep sided topography, or as a consequence of scattering at the ice surface and bed.

beneath ice sheets.

The large autocorrelation in the datasets can influence the performance of the bootstrapping estimates of the confidence interval, as the autocorrelation results in long sequences of resampled data which are very similar to the original data and, hence, artificially narrow 95% confidence intervals (e.g., the collapsing of the 95% confidence intervals at displacements around 80 and 95 km in Fig. 3). This is consistent with the finding of Mudelsee (2003) that long data series may be needed for strongly autocorrelated data.

Including multiple BCa bootstrap estimates and the resulting median estimates for the 95% confidence interval bounds produces more robust (and more smoothly varying with displacement) estimates than the Ólafsdóttir and Mudelsee (2014) method.

The estimate is also robust to both the reduction in the amount of data and unevenly sampled data, as shown in Fig. 3b and c. Results are consistent even with 50% data removal. However, it should be noted that this dataset is highly autocorrelated and, as such, contains much redundant information (at least in terms of correlation estimates). The persistently high autocorrelation at increasing displacements (Fig. 3b) is due to the non-stationary nature of the ice thickness data: when the low frequency trend is removed by high-pass filtering (using a Gaussian filter with an equivalent half-power width of 10 km) the correlations rapidly fall towards zero (Fig. 3c), although the results still remain consistent with up to 50% data removal.

4. MSA

The MSA records from two distinct ice core locations (DSS97 and W10k) 10 km apart were compared to establish if a significant common signal exists, which would support the use of MSA as a sea ice proxy in this region. For broad consistency with the work of Curran et al. (2003), we low-pass filter the two MSA records (Fig. 4) using a Gaussian filter with width $\sigma = 0.2994$ (equivalent half power width of 1 year), maintaining their unevenly sampled and different time-bases. The correlation between the two MSA records is 0.394 [0.231–0.625], and as the confidence interval does not cross zero, we can conclude that this correlation is significant at the 95% level. The results here demonstrate coherence between two ice core records which experience differ-

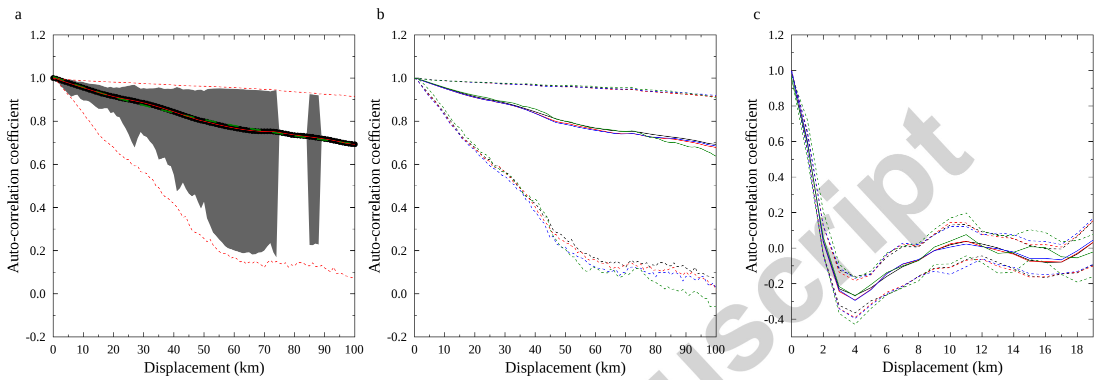


Figure 3: Autocorrelation coefficient for ICECAP ice thickness data as a function of offset displacement. a) Uniformly sampled data, red line Gaussian kernel correlation and associated 95% confidence limits (dashed red). Black line is existing method correlation calculation (Ólafsdóttir and Mudelsee, 2014) and associated 95% confidence limits (grey banding). Exponential data fit to correlation coefficient (green dashed). b) Unevenly sampled data, 0% missing (black), 10% missing (red), 20% missing (blue) and 50% missing (green). c) As per b) except for high-pass filtered ICECAP ice thickness data and different axis limits.

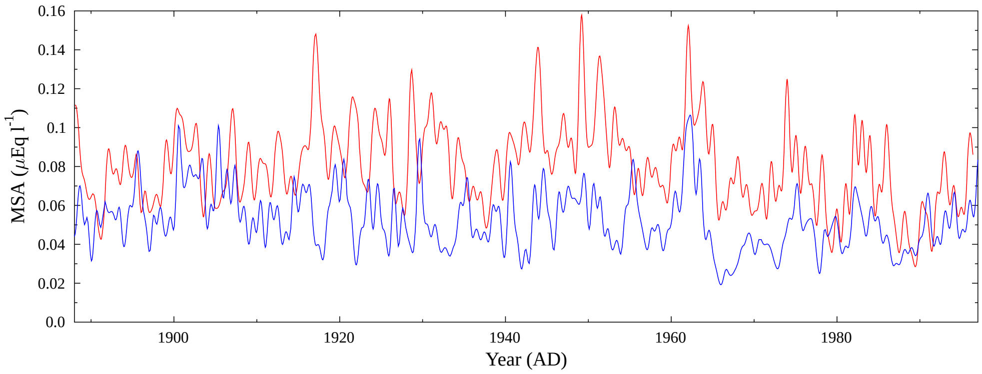


Figure 4: MSA records for two East Antarctic ice cores, DSS97 (red) and W10k (blue).

ent snow accumulation rates. DSS97 has a mean annual accumulation rate of 0.69 metres of ice equivalent per year (Roberts et al., 2015) and W10k of around 0.5 metres of ice equivalent per year. The different mean snow accumulation rates and associated snow densification rates result in different diffusion properties of MSA within the ice cores (Abram et al., 2008; Roberts et al., 2009), and different losses of MSA to the atmosphere, resulting in reduced coherence. Additionally, timing noise arises from variable snow deposition and surface relief. Despite these sources of difference, the coherent signal between the two sites provides evidence that the co-variation seen between MSA in the Law Dome region and sea-ice extent represents a regional signal that is distinct from local noise processes. This supports the interpretation of MSA as a proxy for sea-ice extent in this region.

5. Software performance and concluding remarks

We provide Fortran 90 source code (Intel 12.1.0, PGI 15.10-0 and gfortran 4.8.4 compatible) as well as a Windows executable and MATLAB (R2015a)/Octave (3.8.1) source code in the supplementary material.

The bootstrap confidence interval calculation requires the calculation of 50 000 (25 replicates of 2000 members) Gaussian kernel method correlations which, for larger datasets, can be quite time consuming. For example, an execution time of around 11 minutes on an Intel Core i7-4712HQ CPU for the MSA example where the sample sizes are $n_x = 1879$ and $n_y = 1491$. The majority of the time is in the nested summation ($n_x \times n_y$ repetitions) of the exponential function in Eq. 1. To decrease the execution time, the code has OpenMP compiler directives to parallelise the calculation of this nested exponential function. The speed-up obtained on the test machine is approximately 4.2 on 8 CPUs. This appears to be hardware limited, as the simultaneous execution of eight single-CPU versions of the code obtains a very similar speed-up. An OpenACC version of the code is also provided for execution on GPU or other accelerators if available.

The assertion that the nested exponential calculations is the most time intensive part of the calculation is confirmed by the replacement of the exponential with an approximation. Using the relation

$$\exp(x) = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n \tag{2}$$

results in the following code

```
d=ty(j)-tx(i)
b=exp(-d**2/(2*h**2))/sqrt(2*pi*h)
```

being replaced by the approximation

```
d=-(ty(j)-tx(i))**2/(2.0*h**2)
! if-then-else is approximation for b=exp(d)
if (d<-20) then
  b=0.0
else
  b=1.0+d/1024
  b=b*b
  b=b*b
  b=b*b
  b=b*b
  b=b*b
  b=b*b
  b=b*b
  b=b*b
  b=b*b
  b=b*b/sqrt(2*pi*h)
endif
```

This version of the code runs around 4.6 times faster than the version using the exponential function. Note that this version of the code is quicker (by around 25%) compared to replacing all the `b=b*b` with a single `b=b**1024`. This approximate version is included in the source code. For the MSA example the difference in the calculated correlation value is -0.005%, with larger differences in the confidence interval of around 0.8%. Since the confidence interval is only an estimate, this difference may be acceptable for some applications with very large datasets and associated long execution times.

The method presented here combined with freely available software provides a new and valuable tool for evaluation of correlation significance in common circumstances of unevenly and differently sampled, autocorrelated data series.

6. Acknowledgements

Funding for this work was provided by the UK Natural Environment Research Council grant NE/D003733/1, NSF grant ANT-0733025, the Jackson School of Geosciences and the G. Unger Vetlesen Foundation. The Australian Antarctic Division provided funding and logistical support (AAS 757, 1172, 2198, 3025, 3103, 4061, 4077 and 4346). This work was supported by the Australian Government's Cooperative Research Centres Programme through the Antarctic Climate and Ecosystems Cooperative Research Centre (ACE CRC). This research was supported under Australian Research Council's Special Research Initiative for Antarctic Gateway Partnership (Project ID SR140300001).

Abram, N., Curran, M., Mulvaney, R., Vance, T., 2008. The preservation of methanesulphonic acid in frozen ice-core samples. Journal of Glaciology 54, 680–684.

Abramowitz, M., Stegun, I. (Eds.), 1968. Handbook of Mathematical Functions. United States Department of Commerce, Washington.

Bamber, J. L., Gomez-Dans, J. L., Griggs, J. A., 2009. A new 1 km digital elevation model of the Antarctic derived from combined satellite radar and laser data Part 1 : Data and methods. The Cryosphere 3, 101–111.

Bindschadler, R., Choi, H., Wichlacz, a., Bingham, R., Bohlander, J., Brunt, K., Corr, H., Drews, R., Fricker, H., Hall, M., Hindmarsh, R., Kohler, J., Padman, L., Rack, W., Rotschky, G., Urbini, S., Vornberger, P., Young, N., Jul. 2011. Getting around Antarctica: new high-resolution mappings of the grounded and freely-floating boundaries of the Antarctic ice sheet created for the International Polar Year. The Cryosphere 5 (3), 569–588.

Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M., Blade, I., 1999. Effective Number of Spatial Degrees of freedom of a Time-Varying Field. Journal of Climate 12 (1969), 1990–2009.

Curran, M. A. J., Ommen, T. D. V., Morgan, V. I., Phillips, K. L., Palmer, A. S., 2003. Ice Core Evidence for Antarctic Sea Ice Decline Since the 1950s. Science 302 (5648), 1203–1206.

Efron, B., 1987. Better bootstrap confidence intervals. Journal of the American Statistical Association 82 (397), 171–185.

Mudelsee, M., Aug. 2003. Estimating Pearson's Correlation Coefficient with Bootstrap Confidence Interval from Serially Dependent Time Series. Mathematical Geology 35 (6), 651–665.

Ólafsdóttir, K. B., Mudelsee, M., Feb. 2014. More accurate, calibrated bootstrap confidence intervals for estimating the correlation between two time series. Mathematical Geosciences 46 (4), 411–427.

Politis, D., Romano, J., 1994. The stationary bootstrap. Journal of the American Statistical Association 89 (428), 1303–1313.

Rehfeld, K., Kurths, J., 2014. Similarity estimators for irregular and age-uncertain time series. Climate of the Past 10 (1), 107–122.

Rehfeld, K., Marwan, N., Heitzig, J., Kurths, J., Jun. 2011. Comparison of correlation analysis techniques for irregularly sampled time series. Nonlinear Processes in Geophysics 18 (3), 389–404.

Roberts, J., Plummer, C., Vance, T., van Ommen, T., Moy, A., Poynter, S., Treverrow, A., Curran, M., George, S., 2015. A two thousand year annual record of snow accumulation rates for Law Dome, East Antarctica. Climate of the Past 11, 697–707.

Roberts, J. L., van Ommen, T. D., Curran, M. A. J., Vance, T. R., 2009. Methanesulphonic acid loss during ice-core storage: recommendations based on a new diffusion coefficient. Journal of Glaciology 55 (193), 784–788.

Roberts, J. L., Warner, R. C., Young, D., Wright, A., van Ommen, T. D., Blankenship, D. D., Siegert, M., Young, N., Tabacco, I. E., Forieri, A., Passerini, A., Zirizzotti, A., Frezzotti, M., 2011. Refined broad-scale subglacial morphology of Aurora Subglacial Basin, East Antarctica derived by an ice-dynamics-based interpolation scheme. The Cryosphere 5, 551–560.

Smith, A., Bell, R., Velicogna, I., Studinger, M., 2007. Methods for determining topography in data sparse regions of East Antarctica. Antarctica:

A Keystone in a Changing WorldOnline Proceedings for the 10th International Symposium on Antarctic Earth Sciences, edited by AK Cooper et al., US Geol. Surv. Open File Rep 1047.

Vance, T., Davidson, A., Thomson, P., Levasseur, M., Lizotte, M., Curran, M., Jones, G., 2013. Rapid DMSP production by an Antarctic phytoplankton community exposed to natural surface irradiances in late spring. Aquatic Microbial Ecology 71, 117–129.

Wichura, M., 1988. The Percentage Points of the Normal Distribution. Journal of the Royal Statistical Society, Series C 37 (3), 477–484.

Wilks, D. S., 2006. Statistical Methods in the Atmospheric Sciences. International Geophysics Series. Academic Press, London, UK.

Wright, A. P., Young, D. A., Roberts, J. L., Schroeder, D. M., Bamber, J. L., Dowdeswell, J. A., Young, N. W., Le Brocq, A. M., Warner, R. C., Payne, A. J., Blankenship, D. D., van Ommen, T. D., Siegert, M. J., Mar. 2012. Evidence of a hydrological connection between the ice divide and ice sheet margin in the Aurora Subglacial Basin, East Antarctica. Journal of Geophysical Research 117 (F1), 1–15.

Young, D., Wright, A., Roberts, J., Warner, R., Young, N., Greenbaum, J., Schroeder, D., Holt, J., Sugden, D., Blankenship, D., van Ommen, T., Siegert, M., 2011. A dynamic early East Antarctic Ice Sheet suggested by ice-covered fjord landscapes. Nature 474, 72–75.