Robust detection of statistically significant correlations in geophysical timeseries: A Monte Carlo method accounting for serial dependence and sampling uncertainty

M.S. Armstrong[a], **A.S Kiem**[a] **and G. Kuczera**[b]

[a]Centre for Water, Climate and Land, University of Newcastle, Australia
[b]School of Engineering, University of Newcastle, Australia
Email: matthew.armstrong@uon.edu.au

Abstract: Two geophysical timeseries may share a common low-frequency signal that is distorted by high-frequency noise. As such, these timeseries are often filtered to remove the high-frequency noise prior to performing statistical analysis. However, this filtering artificially increases the serial dependence of the timeseries, meaning that the assumption of independent data underlying most standard correlation tests (e.g. Pearson's correlation) is violated. Monte Carlo methods that account for serial dependence when comparing serially dependent data are typically focused on either (a) calculating the p-value of the observed correlation with respect to an empirically derived null distribution, which is derived by calculating the correlation between independently generated replicates of the observed data or (b) estimating sampling uncertainty in the observed statistic by performing a block bootstrap, with block size proportional to the serial dependence in the timeseries. In this study, we present a Monte Carlo test that combines these two approaches and, in doing so, explicitly accounts for serial dependence and sampling uncertainty when comparing two timeseries. A case study is presented that demonstrates the ability of the proposed method to detect statistically insignificant correlations when performed on filtered white noise timeseries. Crucially, existing methods accounting for serial dependence detected a statistically significant, spurious correlation. This demonstrates that the proposed method is suitable for use when performing statistical analysis on filtered timeseries.

Keywords: *Monte Carlo, Bootstrap, Computational statistics*

1. INTRODUCTION

Bivariate geophysical statistical analysis is complicated by (a) serially dependent processes; (b) sampling uncertainty due to short observational records; and (c) confounding high-frequency noise that can distort a common low-frequency signal shared by two timeseries. Prior to comparing two geophysical timeseries, a filter can be applied to extract this common low-frequency signal. However, filtering increases the serial dependence of a timeseries and many conventional bivariate statistical tests (e.g. Pearson's correlation) assume serial independence – serially dependent data violates this assumption and reduces the number of independent observations (Yule, 1926). Given that geophysical timeseries are often serially dependent, and this serial dependence is increased if the timeseries is filtered, this assumption of independence increases the likelihood of making a Type 1 error (e.g., detecting a statistically significant correlation between two timeseries when there is none).

Monte Carlo methods can be used to detect statistically significant correlations in the presence of serial dependence. These methods derive an empirical null distribution by generating independent, persistence preserving replicates of each time series from which the correlation is calculated (Ebisuzaki, 1997; Macias-Fauria *et al.*, 2012). The p-value of the observed correlation is then calculated with respect to this null distribution. However, these methods do not consider uncertainty in the observed correlation (i.e. sampling uncertainty) - the observed correlation is treated as a fixed quantity (Figure 1b).

There are numerous methods of estimating sampling uncertainty from serially dependent data. For example, Mudelsee (2003) accounts for serial dependence when generating a correlation sampling distribution using the stationary bootstrap of DiCiccio and Efron (1996). However, this method does not compare the resultant sampling distribution to a reference null distribution. This may be an issue because, for different statistics, different types of serial dependence may result in different thresholds of statistical significance. Therefore, it is necessary to compare a sampling distribution with an empirically derived null distribution when determining statistical significance in the presence of sampling uncertainty (Figure 1c).

Both the Wilcoxon-Mann-Whitney and Kolmogorov-Smirnov tests could be used to identify a statistically significant difference between the null and sampling distributions. However, given the large sample size of the null/sampling distributions (often several thousand) and that the different sampling methods used to generate respective distributions can result in distributions with a slightly different shape, these tests will have a high Type 1 error rate. For example, sampling distributions of statistics calculated from serially dependent data also tend to be skewed (Politis and Romano, 1994) and may have reduced variance relative to the true null (Mudelsee, 2003). The Wilcoxon-Mann-Whitney and Kolmogorov-Smirnov tests are sensitive to such differences in skew/variance (Fagerland and Sandvik, 2009), meaning that they are unsuited for determining statistically significant differences between corresponding null and sampling distributions.

When assessing the statistical significance of a correlation, given that geophysical timeseries are often (a) short, meaning they have large sampling uncertainty, and (b) filtered prior to statistical analysis, which increases timeseries serial dependence, there is a need to account for sampling uncertainty and serial dependence. These factors can be accounted for by comparing a correlation sampling distribution with an empirically derived null distribution. However, standard statistical tests used for comparing differences between two distributions (e.g. Mann-Whitney U) are not suitable for comparing these distributions. As such, in this study we propose a new, Monte Carlo based statistical test that can detect statistically significant correlations in the presence of serial dependence and sampling uncertainty.

2. METHOD

As existing methods comparing two distributions (e.g. Mann-Whitney U) are sensitive to the differences in skew/variance evident between corresponding null/sampling distributions, in this analysis we present a non-parametric method of detecting statistical significance that accounts for serial dependence and sampling uncertainty. This method extends the Monte Carlo concepts underpinning that of Ebisuzaki (1997) to a statistic that gives a measure of similarity between two distributions: the overlapping coefficient.

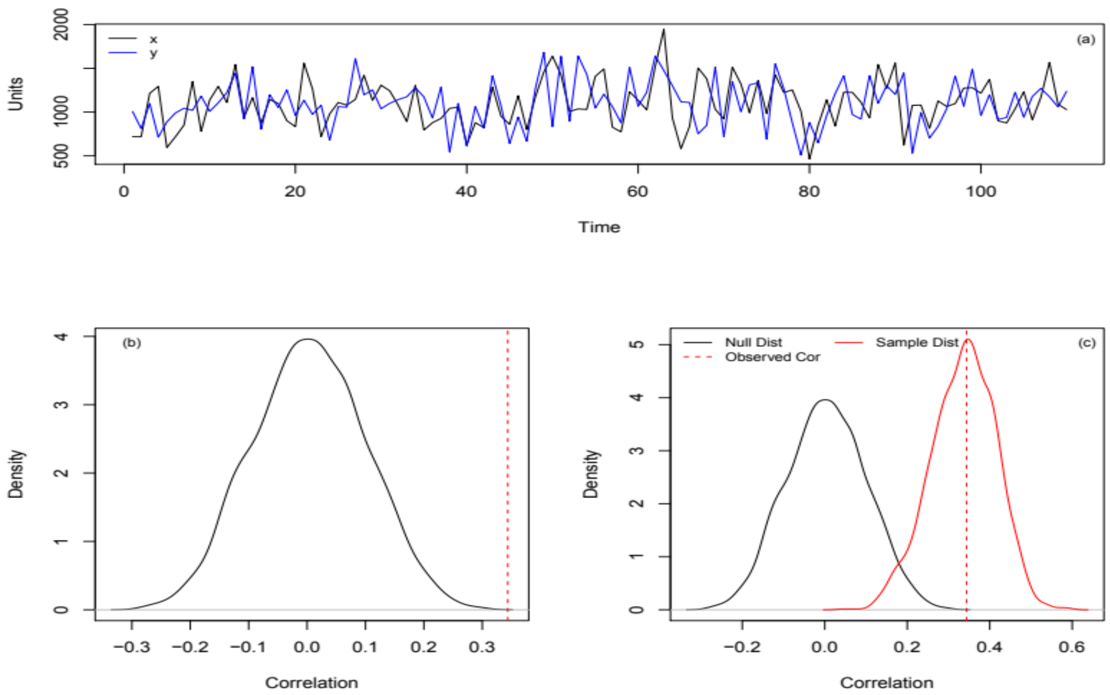

Figure 1. (a) Synthetic, serially dependent x and y time series; (b) empirically derived null distribution for the correlation between x and y assuming no association - the red dashed line is the observed correlation; (c) empirical null and sampling distributions for the correlation between x and y. Sampling distribution was derived using the method of Mudelsee (2003).

The overlapping coefficient (also referred to as the proportion of overlap) refers to the shared area of two (possibly) intersecting probability distributions. The definition of the overlapping coefficient is shown below. This equation can be solved using various computational methods, which typically involve kernel density estimation of the respective distributions, followed by numerical integration (Schmid and Schmidt, 2006). Calculation method aside, the key point to note is that an overlapping coefficient of 1 indicates that the two distributions are identical, whereas a value of 0 indicates complete divergence.

$$(1) \qquad Overlap = \int \min\{f(x_1), f(x_2)\}dx$$

The overlapping coefficient is used as a diagnostic metric in various fields. For example, ecologists use the overlapping coefficient to compare the relative activity of different animals (Ridout and Linkie, 2009). In behavioural science, the overlapping coefficient is used to when comparing the relative differences between control and treatment groups (Cohen, 1988). The Perkins Skill Score, a metric commonly used in climatology to evaluate the performance of climate models, is also a calculation of overlap between modelled and observed climate variable distributions (Perkins *et al.*, 2007).

The proposed method, which calculates the statistical significance of the overlap between an empirically derived null distribution and an observed sampling distribution, involves the following steps:

- For two time series x and y of length n, identify some bivariate statistic of interest (e.g. correlation).
- As with the method of Ebisuzaki (1997), derive an empirical null distribution for the statistic of interest. This involves generating replicate time series that have the same serial dependence and length as x and y

but with different sequencing. This distribution reflects the values we can expect from independent time series with the same persistence structure as the observed values.

- Generate random, paired replicates of the time series x and y of length n. For each paired replicate, calculate the corresponding sampling distribution of the statistic of interest.
    - Various methods that account for serial dependence can be used to generate the sampling distribution. In this study, we use the bootstrap method of Mudelsee (2003), which is based on the stationary block bootstrap of DiCiccio and Efron (1996). However, other methods – such as the maximum entropy bootstrap of Vinod and Lopez-de-Lacalle (2009) – could also be used.
- For each sampling distribution, calculate the proportion of overlap with the empirical null distribution. This is the “null overlapping coefficient” distribution - the values of the overlapping coefficient we could expect between the empirical null distribution and sampling distributions given the time series x and y are independent.
- Using the observed time series x and y, calculate the observed sampling distribution for the statistic of interest. The sampling distribution must be calculated using the same method as the paired replicates.
- Calculate the overlap between the observed sampling distribution and the empirical null distribution.
- Calculate the percentile of the observed overlapping coefficient with respect to the null overlapping coefficient distribution. This percentile is the p-value of the observed overlapping coefficient. The statistical significance of this p-value can then be determined using standard thresholds (e.g., a p-value less than 0.05).

3. RESULTS

The ability of the method to detect significant/insignificant statistics in situations where standard statistical tests do not work well is demonstrated using an example that involves filtered timeseries. Geophysical time series are inherently noisy, so in many cases a filter is applied to these time series in order to extract an underlying signal (Mann, 2008; Thompson *et al.*, 2018; Sanchez-Morales *et al.*, 2019). However, filtering introduces artificial serial dependence, which can lead to the emergence of spurious correlations between filtered time series.

In this analysis, we used two randomly generated white noise timeseries (Figure 2) and applied a smoothing spline (span of 0.7) to each (Figure 3a). The statistical significance of the correlation between the filtered timeseries was then calculated using (a) two existing methods typically used for serially dependent timeseries - the methods of Ebisuzaki (1997) and Mudelsee (2003) - and (b) the proposed method.

When comparing the two timeseries, the methods of Ebisuzaki (1997) and Mudelsee (2003) both returned statistically significant p-values (0.008 and 0.002 respectively, Figure 3b and Figure 3c). Given that the original timeseries are not correlated (Figure 2), these are spurious results artificially introduced by filtering. In contrast, Figure 3d shows that, by accounting for serial dependence and sampling uncertainty, the proposed method correctly returned a statistically insignificant p-value (0.27). This highlights the potential utility of the proposed method – determining the significance of a correlation between filtered timeseries.

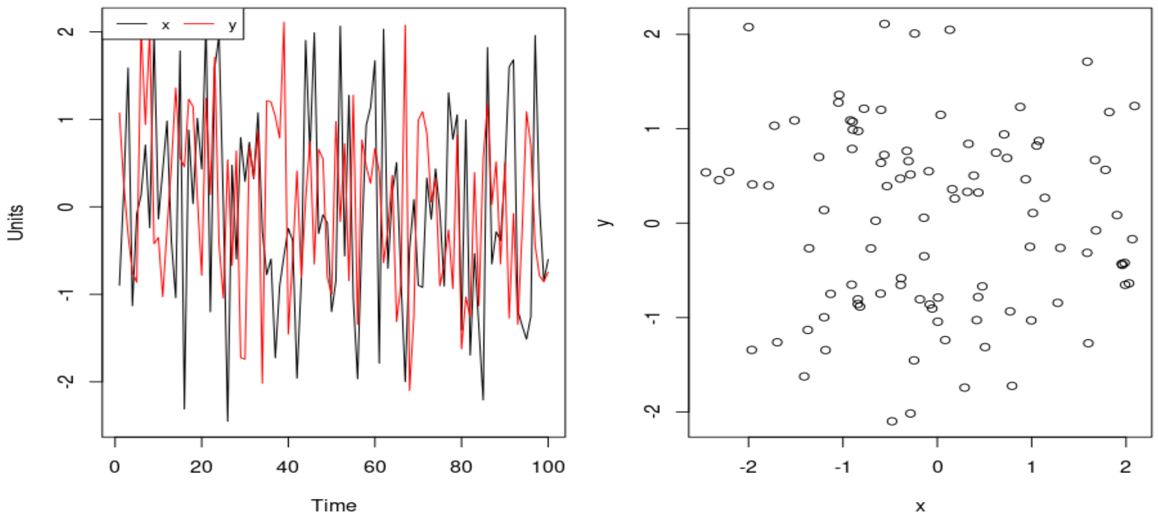

Figure 2. Left - two randomly generated white noise timeseries. Right - association between the two randomly generated white noise timeseries.

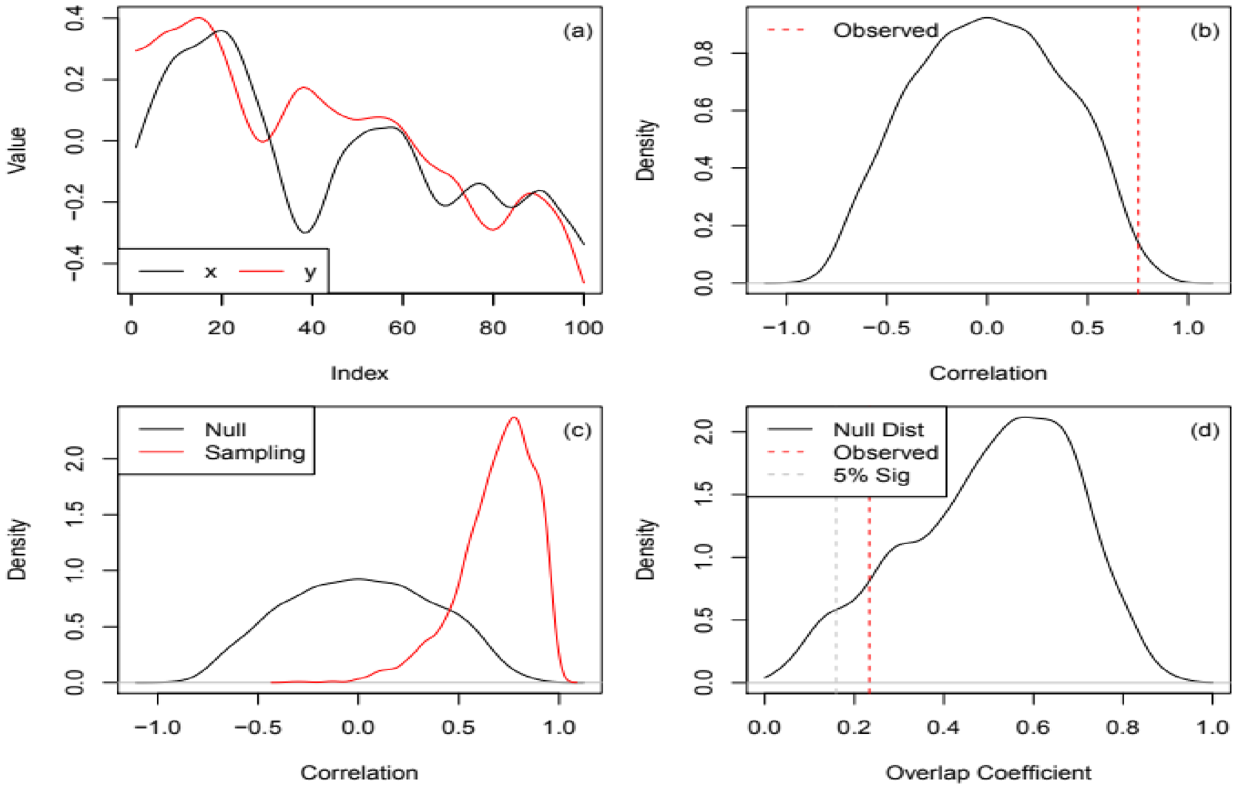

Figure 3. (a) Timeseries in Figure 3 after the application of a smoothing spline (span of 0.7). b) Results from the Ebisuzaki (1997) method applied to the timeseries in (a), which identified a statistically significant correlation. c) Comparing null and sampling distributions for the timeseries in (a) – the sampling distribution was derived using the method of Mudelsee (2003), with.the 95% confidence intervals not containing 0. d) Results from the proposed method determining the significance of the overlap between null and sampling distributions. Note that in (b) and (c) respective methods returned statistically significant correlations, whereas the proposed method (d) returned a statistically insignificant correlation.

4. DISCUSSION AND CONCLUSION

The proposed method has widespread potential applications in geophysical timeseries analysis, which often filters data prior to performing statistical analysis. For example, the Interdecadal Pacific Oscillation (IPO) is a filtered index of basin wide Pacific Ocean sea surface temperature variability which is associated with increased/reduced rainfall for various regions around the world (Power *et al.*, 1999) – the significance of these identified associations may be misrepresented by this filtering. Furthermore, palaeoclimate proxy data, used to infer climate conditions in periods without instrumental measurements, are also often filtered prior to statistical analysis (Sanchez-Morales *et al.*, 2019) – once again, this filtering may misrepresent the significance of any associations identified using the filtered data.

Although filtering of geophysical timeseries prior to statistical analysis can result in the identification of spurious correlations, this filtering still serves an important purpose. Given that geophysical timeseries are inherently noisy, and that this noise can distort common low-frequency signal, removing this noise via a filter is a necessary step when examining associations/relationships between geophysical timeseries. However, in doing so, it is important to use the appropriate statistical tools. In this study we've shown that, under certain conditions (i.e., the application of a smoothing spline with span 0.7 to two white noise timeseries of length 100), conventional methods that account for serial dependence/sampling uncertainty when identifying correlations between timeseries - the methods of Ebisuzaki (1997) and Mudelsee (2003) – can produce spurious results. In contrast, the proposed method correctly identified that the filtered timeseries were independent, highlighting its potential application in identifying/not identifying correlations between filtered geophysical timeseries.

In summary, in this study we propose a new Monte Carlo method that offers potential for statistical analysis of filtered geophysical timeseries. The proposed method builds on existing methods by explicitly accounting for how serial dependence can (a) increase the threshold required for a statistically significant result and (b) increase sampling uncertainty. Although the method demonstrates promise for statistical analysis of filtered timeseries, more work is needed to verify the conditions under which the proposed method provides added benefit over existing methods, and the likelihood of such conditions arising during geophysical timeseries statistical analysis.

REFERENCES

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge. doi: doi.org/10.4324/9780203771587.

DiCiccio, T. J. and Efron, B. (1996) 'Bootstrap confidence intervals', *Statist. Sci.*, 11(3), pp. 189–228. doi: 10.1214/ss/1032280214.

Ebisuzaki, W. (1997) 'A Method to Estimate the Statistical Significance of a Correlation When the Data Are Serially Correlated', *Journal of Climate*, 10(9), pp. 2147–2153. doi: 10.1175/1520-0442(1997)010<2147:AMTETS>2.0.CO;2.

Fagerland, M. W. and Sandvik, L. (2009) 'The Wilcoxon–Mann–Whitney test under scrutiny', *Statistics in Medicine*, 28(10), pp. 1487–1497. doi: 10.1002/sim.3561.

Macias-Fauria, M. *et al.* (2012) 'Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series', *Dendrochronologia*, 30(2), pp. 179–187. doi: 10.1016/j.dendro.2011.08.003.

Mann, M. E. (2008) 'Smoothing of climate time series revisited', *Geophysical Research Letters*, 35(16). doi: https://doi.org/10.1029/2008GL034716.

Mudelsee, M. (2003) 'Estimating Pearson's Correlation Coefficient with Bootstrap Confidence Interval from Serially Dependent Time Series', *Mathematical Geosciences*, 35(6), pp. 651–665. doi: 10.1023/B:MATG.0000002982.52104.02.

Perkins, S. E. *et al.* (2007) 'Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions', *Journal of Climate*, 20(17), pp. 4356–4376. doi: 10.1175/JCLI4253.1.

Politis, D. N. and Romano, J. P. (1994) ‘The Stationary Bootstrap’, *Journal of the American Statistical Association*, 89(428), pp. 1303–1313. doi: 10.1080/01621459.1994.10476870.

Power, S. *et al.* (1999) ‘Inter-decadal modulation of the impact of ENSO on Australia’, *Climate Dynamics*, 15(5), pp. 319–324. doi: 10.1007/s003820050284.

Ridout, M. S. and Linkie, M. (2009) ‘Estimating overlap of daily activity patterns from camera trap data’, *Journal of Agricultural, Biological, and Environmental Statistics*, 14(3), pp. 322–337. doi: 10.1198/jabes.2009.08038.

Sanchez-Morales, J. *et al.* (2019) ‘A new method for reconstructing past-climate trends using tree-ring data and kernel smoothing’, *Dendrochronologia*, 55, pp. 1–15. doi: https://doi.org/10.1016/j.dendro.2019.03.002.

Schmid, F. and Schmidt, A. (2006) ‘Nonparametric estimation of the coefficient of overlapping—theory and empirical application’, *Computational Statistics and Data Analysis*, 50(6), pp. 1583–1596. doi: https://doi.org/10.1016/j.csda.2005.01.014.

Thompson, L. G. *et al.* (2018) ‘Ice core records of climate variability on the Third Pole with emphasis on the Guliya ice cap, western Kunlun Mountains’, *Quaternary Science Reviews*, 188, pp. 1–14. doi: https://doi.org/10.1016/j.quascirev.2018.03.003.

Vinod, H. D. and Lopez-de-Lacalle, J. (2009) ‘Maximum Entropy Bootstrap for Time Series: The meboot R Package’, *Journal of Statistical Software*, 29(i05). doi: http://hdl.handle.net/10.

Yule, G. U. (1926) ‘Why do we Sometimes get Nonsense-Correlations between Time-Series?--A Study in Sampling and the Nature of Time-Series’, *Journal of the Royal Statistical Society*, 89(1), pp. 1–63. doi: 10.2307/2341482.