Journal of the Korean Physical Society, Vol. 72, No. 4, February 2018, pp. 545∼550

Effect of Outliers and Non-consecutive Data Points on the Detrended Cross-Correlation Analysis

Gyuchang Lim and Seungsik Min*

Department of Natural Sciences, Korea Naval Academy, Changwon 51704, Korea

(Received 7 December 2017, in final form 1 January 2018)

In this paper, we investigate the robustness of the well-known DCCA (detrended cross-correlation analysis) methodology and give a qualitative analysis result. Due to the non-stationarity inherent in most observational data sets, the results of DCCA and its variants may be spurious. In particular, oceanographic data sets contaminated with measurement errors are subject to unusual records, making it difficult to trust DCCA results. To ensure the validity of the DCCA methodology for the oceanographic time series, we conduct simulation studies based on the ARFIMA process and perform statistical tests using surrogate methods. First, so-called outliers due to measurement error lead to the spurious results of DCCA methods while discontinuities in the time series have been found to have little effect on the results. This means that the cross-correlation structure is robust to the discontinuity of time series. Second, statistical significance of cross-correlation was obtained through a surrogate statistical test for the oceanographic time series.

PACS numbers: 89.20.−a, 02.50.−r, 05.40.−a
Keywords: Cross-correlation, Oceanographic data, DCCA methodology, Surrogate methods
DOI: 10.3938/jkps.72.545

I. INTRODUCTION

In recent years, the DCCA (detrended cross-correlation analysis) methodology and its variants have been proposed as new approaches to analyzing power-law cross-correlations between various time series from complex systems to understand the underlying dynamics [1–4], and have become a cornerstone in a variety of fields - finance [5–7], meteorology [8], seismology [9], neuroscience [10], and others. However, in the application of the DCCA methodology, the detrending method is very important because non-stationarities inherent in most experimental / empirical time series lead to the spurious estimation of cross-correlations. To this end, a higher order detrending approach was proposed and turned out to be valid when analyzing non-stationary time series with periodic trends [11].

Outliers found in oceanographic / meteorological time series are often due to the measurement equipment errors. In this case, the usual detrending method used in the so-called DCCA approach is not suitable for direct application, so some data points of the time series should be selectively removed. In this process, we are forced to lose the requirement that all data points be recorded consecutively. This is the critical problem we face when we analyze the cross-correlation of those time series using the DCCA methodology. Figure 1 clearly illustrates this problem: the raw data of 5 items are plotted and outliers are obviously found in 4 items. Performing a cross-correlation analysis on these obtained data without any data preprocessing can lead to a spurious result.

For this reason, we use ARFIMA processes to conduct simulation studies on the validity of the DCCA methodology. First, we investigate the effect of outliers on DCCA results by inserting artificial outliers into the time series generated in the ARFIMA process. The strength of an outlier is controlled by a multiple of the max value of the ARFIMA series. Second, we deliberately remove parts of the ARFIMA series to investigate the impact on DCCA results. Lastly, based on the above two simulation results, we perform a statistical test for the cross-correlation structure of oceanographic / meteorological time series using the surrogate method [4,12].

The organization of this paper are as follows. In the next section, we give a brief description of oceanographic / meteorological time series. In Sec. III, we perform the simulation studies and present the results. And, in Sec. IV, we apply the DCCA method to the preprocessed data set and give a statistical test for the DCCA results. The last section is closed with summary and discussions.

II. DATA DESCRIPTION

We obtained a set of oceanographic / meteorological data from the Korea Hydrographic and Oceanographic

*E-mail: fieldsmin@gmail.com

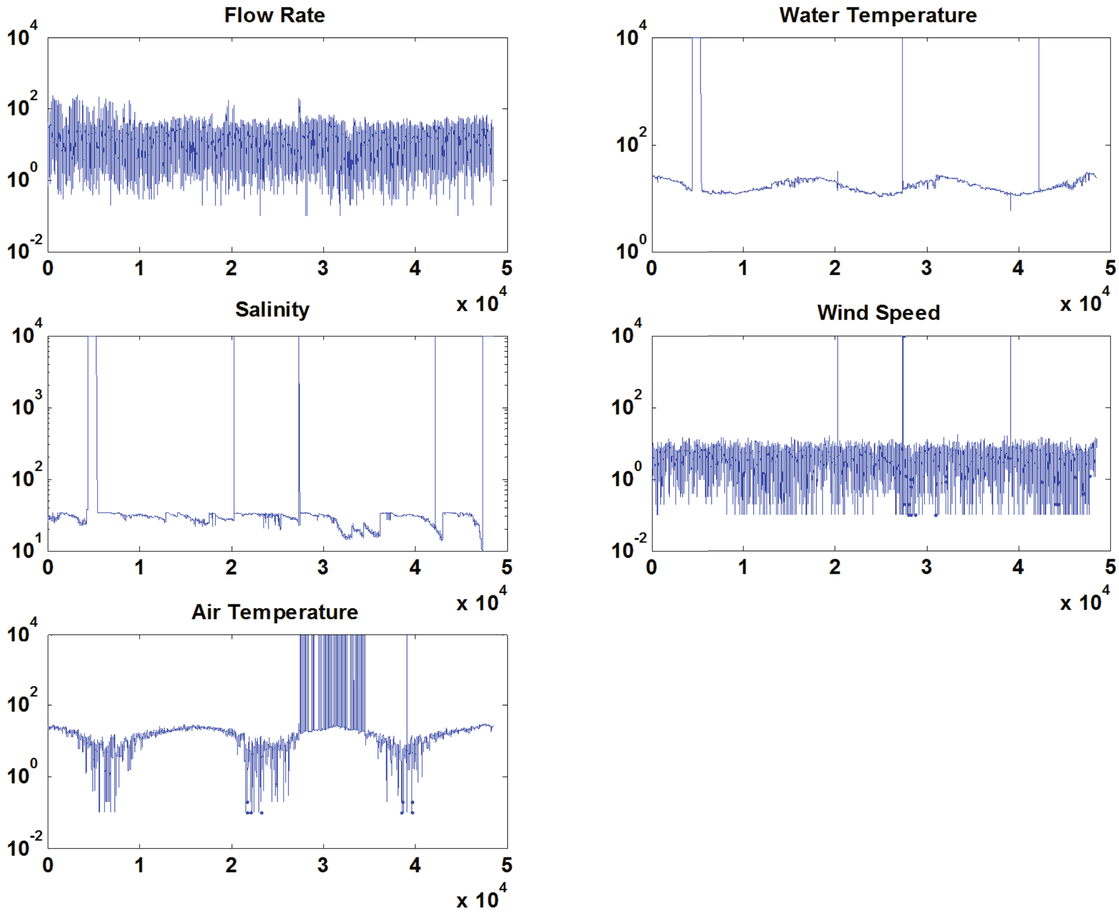

Fig. 1. (Color online) The 5 raw time series are plotted. These are simultaneously recorded and equipment errors are obviously found in water temperature, salinity, wind speed and air temperature.

Agency [15] , which are measured hourly with floating buoy equipment from Sep. 2013 to Aug. 2016. Since the raw data have a lot of outliers due to measurement errors as shown in Fig. 1, we deliberately removed those outliers by visual selection and so-called processed time series are presented in Fig. 2. To maintain the concurrency of the multivariate time series, we cut out all the data points of the multivariate measurements of the same time. This preprocessing can make a devastating effect on the results of the cross-correlation analysis. And we still have a validity issue with data manipulation because there is no criterion. In next section, we deal with this problem via simulation studies.

III. METHODOLOGY

1. Detrended cross-correlation analysis (DCCA)

We first consider two time series $\{x_i\}$ and $\{y_i\}$, both consisting of N measurements. Without loss of generality, we assume that these time series have zero means, and we build the profile to be $X_k = \sum_{i=1}^{k} x_i$ and $Y_k = \sum_{i=1}^{k} y_i$ with $k = 1, \cdots, N$. And we divide the profile into $N - n$ overlapping segments, each containing n values. For both time series, in each segment that starts at i and ends at $i + n - 1$, we define the local trends, $\widetilde{X}_{k,i}$ and $\widetilde{Y}_{k,i}$, to be the ordinate of a nonlinear least squares fit; we take the polynomial detrending procedure with order 16. Lastly, we define the detrended walk as the difference between the original walk and the local trend, that is, $X_k - \widetilde{X}_{k,i}$ and $Y_k - \widetilde{Y}_{k,i}$. Next, we calculate the variance and covariance of the residuals in each segment [1,13] defined as

$$f^2_{\mathrm{DFA}}(n,i) \equiv \frac{1}{n} \sum_{k=1}^{n+i-1} (X_k - \widetilde{X}_{k,i})^2 \qquad (1)$$

$$f^2_{\mathrm{DCCA}}(n,i) \equiv \frac{1}{n} \sum_{k=1}^{n+i-1} (X_k - \widetilde{X}_{k,i})(Y_k - \widetilde{Y}_{k,i}), \qquad (2)$$

where i denotes the box index.

Finally, we can obtain the detrended variance/covariance function by summing over all overlap-

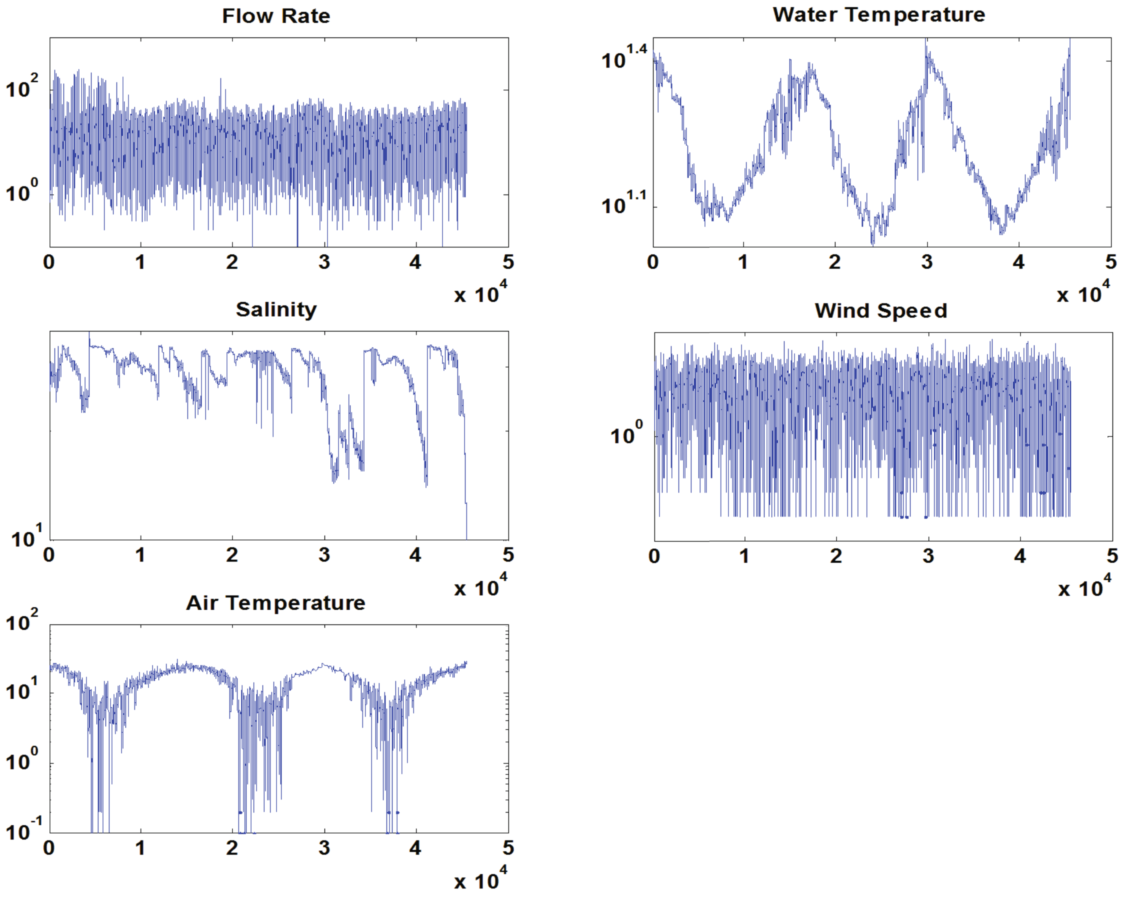

Fig. 2. (Color online) The processed time series, by removing outliers, are plotted. Since the length of the time series is reduced by about 6%, the periodic trend becomes clearer.

ping $N-n$ segments of size n:

$$F^2_{\mathrm{DFA}}(n) \equiv \frac{1}{(N-n)}\sum_{i=1}^{N-n} f^2_{\mathrm{DFA}}(n,i) \qquad (3)$$

$$F^2_{\mathrm{DCCA}}(n) \equiv \frac{1}{(N-n)}\sum_{i=1}^{N-n} f^2_{\mathrm{DCCA}}(n,i). \qquad (4)$$

If the both series are power-law cross-correlated, the following scaling relation is fulfilled:

$$f_{\mathrm{DCCA}}(n) \sim n^{\lambda}, \qquad (5)$$

where the scaling exponent λ is related to the cross correlation exponents γ_x, which satisfies the following relations:

$$X(n) \sim n^{-\gamma_x}, \qquad (6)$$

where $X(n) \equiv \overline{(X_k-\mu_x)(y_{k+n}-\mu_y)}/\sigma_x\sigma_y$.

Otherwise, we used the DCCA coefficient [3] defined as

$$\rho_{\mathrm{DCCA}}(n) \equiv \frac{F^2_{\mathrm{DCCA}}(n)}{\sqrt{F_{\mathrm{DFA}\{x_i\}}(n)F_{\mathrm{DFA}\{y_i\}}(n)}}. \qquad (7)$$

This coefficient outperforms the Pearson correlation coefficient r on detecting both the memory effect and the real cross-correlations for non-stationary series [14].

2. Simulation study

In order to examine the robustness of the DCCA method, we simulate the ARFIMA process generating long-range power-law correlated time series. We here use a periodic two-component ARFIMA:

$$Z_i = \left[\sum_{n=1}^{\infty} a_n(\rho_1)Z_{i-n}\right] + A_1 \sin\left(\frac{2\pi}{T_1}i\right) + \eta_i \qquad (8)$$

$$Z'_i = \left[\sum_{n=1}^{\infty} a_n(\rho_2)Z'_{i-n}\right] + A_2 \sin\left(\frac{2\pi}{T_2}i\right) + \eta_i \qquad (9)$$

Here, η_t is shared between z_i and z'_i in order to enable cross-correlations, $T_1(T_2)$ is the sinusoidal period, A_1 and A_2 are two sinusoidal amplitudes, and $a_n(\rho)$ is a statistical weight defined by $a_n(\rho) = \Gamma(n-\rho)/(\Gamma(-\rho)\Gamma(1+n))$, where Γ denotes the Gamma function and ρ is a parameter ranging from -0.5 to 0.5. According to the work

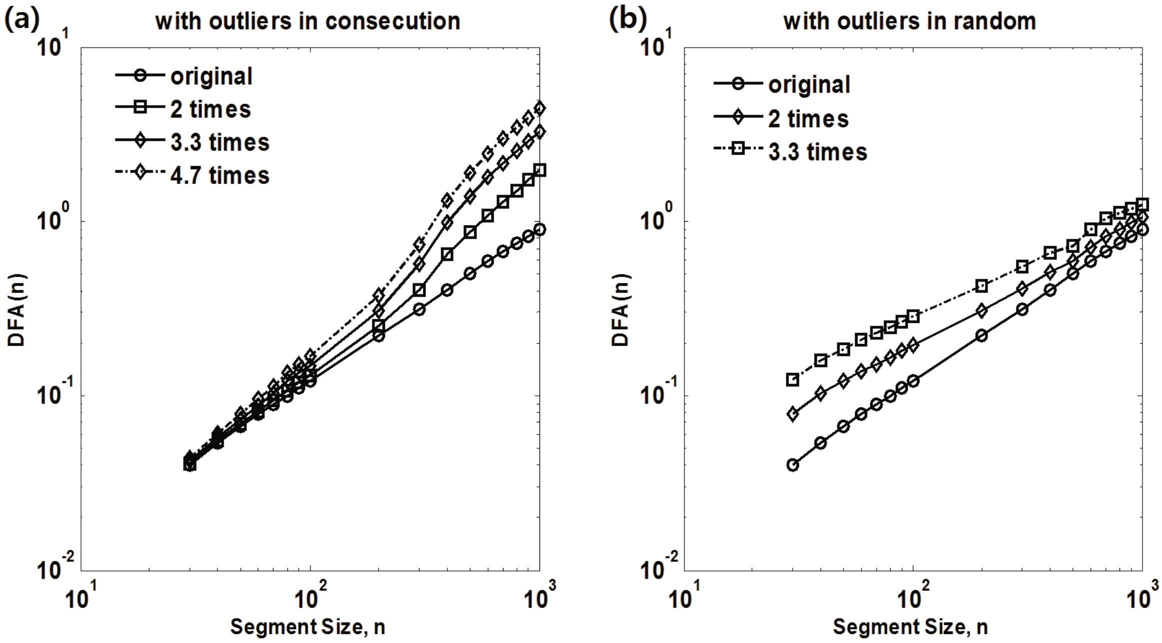

Fig. 3. (a) 30 outliers inserted consecutively : The crossover is clearly detected and the scaling exponent gets larger. (b) 30 outliers inserted randomly : the scaling exponent gets smaller with growing strength, that is, it is getting closer to the random noise.

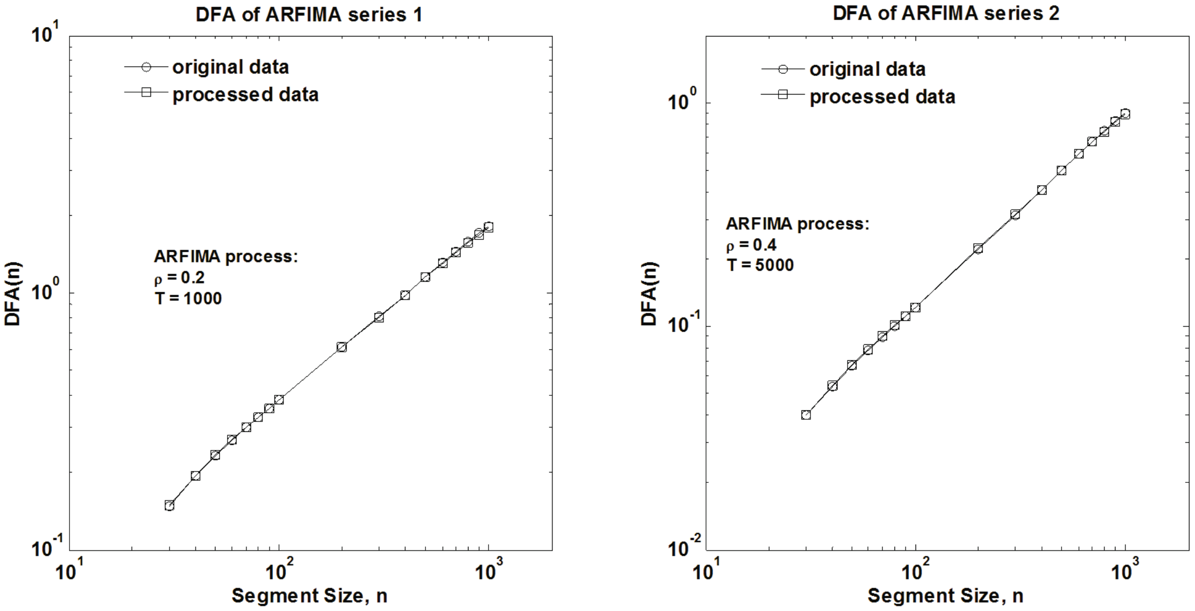

Fig. 4. DCCA results such as the power-law scaling and the DCCA coefficient are also robust to the removed data points.

of D. Horvatic *et al.* [4], the detrending method with varying order of the polynomial well eliminates the periodic trend and preserves the inherent cross-correlation between two signals. Importantly, the polynomial order l is increasing with the segment size n. In our work, we generated two ARFIMA series of the same length as oceanographic time series of 45,000 recordings and the parameters are assigned as follows for each series:

$$\begin{aligned} &\rho_1 = 0.2, && \rho_2 = 0.4 \\ &T_1 = 1000, && T_2 = 5000 \\ &A_1 = A_2 = 1 \end{aligned} \tag{10}$$

And we adopted the detrending polynomial of order 16.

Applying the higher order of fitting polynomial leads to a spurious crossover in small box size n, so we set the smallest box size to 30.

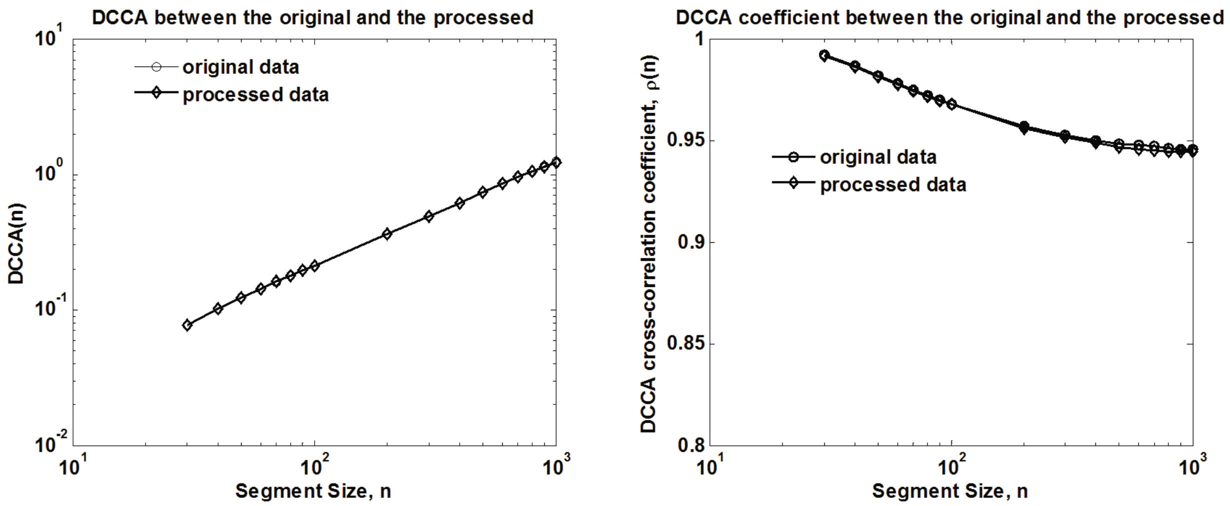

Fig. 5. DCCA results such as the power-law scaling and the DCCA coefficient are also robust to the removed data points.

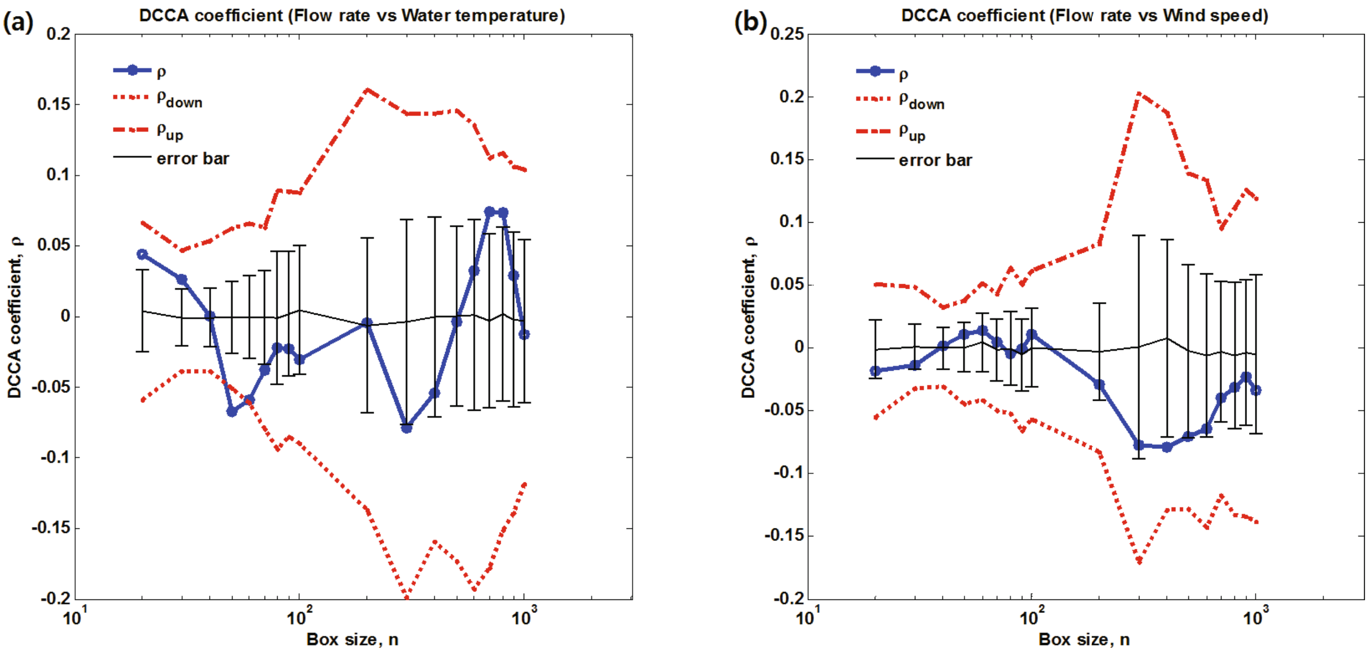

Fig. 6. (Color online) (a) DCCA coefficients between flow rate and water temperature are plotted *w.r.t.* box size, n. No clear evidence of a genuine cross-correlation is found. (b) Likewise, no evidence of a cross-correlation between flow rate and wind speed.

First, we examined the effect of outliers on the detrended variance / covariance analysis results by inserting artificial outliers into the ARFIMA series. For ease of comparison, we adjusted the strength of the outliers by a multiple of the maximum value of the original data, and change the number of outliers. For the number of artificial outliers, we raised one by one and found the crossover at 30. In this paper, we illustrate the simulation result with 30 artificial outliers. As shown in Fig. 3, increasing the strength of outliers leads to a dramatic change of the slope in a log-log plot of DFA. However, a distinctive difference is revealed by the placement of outliers: (a) a dramatic crossover is clearly detected when 30 outliers are placed in one segment, while (b) no crossover is detected when the outliers are placed randomly. This means that the effect of the outliers is intensified by getting the outliers closer to each other. Our findings imply that the strength of outliers and their placement play a conflicting role in estimating the power-law scaling exponent.

This finding means that the strength of outliers and their placement play a conflicting role in estimating the power-law scaling exponent. Random placement weakens the inherent scaling structure in the time series while gathering with each other strengthens the correlation structure. Therefore, a good tool of analysis can sometimes result in spurious results by the data itself.

Second, we investigate the effect of removing data points on the DFA / DCCA results. In the following subsection, we discuss the simulation results. Figure 4 shows that the scaling exponent of DFA makes no difference between the original data and the processed data, where about 6% of the whole length are randomly removed because we removed about 6% outliers of the oceanographic time series. Likewise, the DCCA results show no difference in both series as shown in Fig. 5. This finding implies that the removed data points do not affect the correlation structure between the time series.

IV. APPLICATION

So far, we have conducted the simulation study and confirmed two facts : one is that the effect of outliers is critical in estimating the power-law correlation scaling exponent, and the other is that removing data points has no effect on the DFA / DCCA results within at least about 6%. Based on the results of simulation studies, we investigated the cross-correlation structure of oceanographic / meteorological time series using the DFA / DCCA methodology after data preprocessing. There was no clear evidence for a power-law cross-correlation structure, so we compute the DCCA coefficient [3].

And we perform statistical tests for the DCCA coefficients to see whether the cross-correlations are significant or not. We first determine the null hypothesis as follows : each series must be a power-law correlated with the same DFA exponent as the corresponding oceanographic or meteorological data. So we generate 100 surrogate series for each real time series by applying the Fourier phase randomization [12], and compute the DCCA coefficients between two surrogate series. Figure 6 shows the estimated DCCA coefficients *w.r.t.* the box size for the 95% confidence level under the assumption abovementioned. Without loss of generality, we present two analysis results because the others are not much different, that is, we could not find any evidence for the existence of a genuine cross-correlation between the time series.

V. CONCLUSION

In summary, we examined the feasibility of the DFA / DCCA methodology on empirical or experimental time series that are defective in the data itself. By looking at two cases that occur in the process of improving defects in time series, we found that outliers have much effect on the DFA / DCCA results while removal of defected data points has less impact on it. Additionally, the effect of outliers appeared in two aspects: one is the strength of an outlier, and the other is the way in which outliers are placed in the time series. For the latter, a random placement weakens the correlation structure while a consecutive placement strengthens the correlation structure. Lastly, by performing a surrogate statistical test, we have found that no genuine cross-correlation exists between ocean and weather

ACKNOWLEDGMENTS

This work was supported by the Korea Naval Academy Maritime Institute in 2017, and the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT under Contract 2017R1C1B5017491.

REFERENCES

[1] B. Podobnik and H. E. Stanley, Phys. Rev. Lett. **100**, 084102 (2008).
[2] W-X. Zhou, Phys. Rev. E **77**, 066211 (2008).
[3] G. Zebende, Physica A **390**, 614 (2011).
[4] B. Podobnik, Z-Q. Jiang, W-X. Zhou and H. E. Stanley, Phys. Rev. E **84**, 066118 (2011).
[5] B. Podobnik, D. Horvatic, A. Petersen and H. E. Stanley, P. Natl. Acad. Sci, USA **106**, 22079 (2009).
[6] L-Y. He and S-P. Chen, Physica A **390**, 3806 (2011).
[7] G. Cao, L. Xu and J. Cao, Physica A **391**, 4855 (2012).
[8] G. Lim, K. Kim, J-K. Park and K-H. Chang, J. Korean Phys. Soc. **62**, 193 (2013).
[9] S. Shadkhoo and G. Jafari, Eur. Phys. J. B **72**, 679 (2009).
[10] W. Jun and Z. Da-Qing, Chin. Phys. B **21**, 028703 (2012).
[11] D. Horvatic, H. E. Stanley and B. Podobnik, Europhys. Lett. **94**, 18007 (2011).
[12] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian and J. D. Farmer, Physica D **58**, 77 (1992).
[13] C. Peng, S. Buldyrev, S. Havlin, M. Simons, H. Stanley and A. Goldberger, Phys. Rev. E **49**, 1685 (1994).
[14] L. Kristoufek, Physica A **402**, 291 (2014).
[15] Website: http://www.khoa.go.kr/.