Statistica Sinica Preprint No: SS-2023-0379	
Title	A Warped Self-normalized Two-sample Test for Time Series with Staggered Observation Periods
Manuscript ID	SS-2023-0379
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0379
Complete List of Authors	Weiliang Wang and Ting Zhang
Corresponding Authors	Ting Zhang
E-mails	tingzhang@uga.edu
Notice: Accepted version subject to English editing.	

A WARPED SELF-NORMALIZED TWO-SAMPLE TEST FOR TIME SERIES WITH STAGGERED OBSERVATION PERIODS

Weiliang Wang and Ting Zhang

Boston University and University of Georgia

Abstract: We consider the problem of two-sample testing for time series with staggered observation periods, where the two time series can have different starting and ending observation times and can be of different lengths. In addition, we allow the two time series to depend on each other in a general way, which makes the staggered observation periods nontrivial to deal with as it now requires accommodating the joint dependence in the presence of overlapping and nonoverlapping segments when designing a valid inference protocol. This also makes existing self-normalization methods inapplicable to the current problem. To address this, we propose a warped self-normalized two-sample test, which uses warped self-normalized subsamples to provide uncertainty quantification of the global two-sample statistic. The method can be readily applied to compare quantities beyond the mean such as the variance or quantiles, and the associated asymptotic theory has been established. Numerical experiments including a simulation study and a real data analysis are also provided to further illustrate the proposed method.

Key words and phrases: self-normalization; staggered time series; subsampling; two-sample test.

1. Introduction

Applications from various scientific problems often require the comparison of data from two populations, which can be, for example, clinical trial data from the treatment and control groups, temperature record data from different countries or regions, electricity usage data from different industries, among many others. The problem is often phrased as a two-sample test in statistical analysis, which has been widely studied in the literature; see for example Hotelling (1931), Cressie and Whitford (1986), Hall and Martin (1988), Bai and Saranadasa (1996), Keselman et al. (2004), Chen and Qin (2010), Chen et al. (2013), Cai et al. (2014), Gregory et al. (2015), Xu et al. (2016), Städler and Mukherjee (2017), Chen et al. (2019), Zhang et al. (2020), and references therein. The aforementioned works mostly concerned the situation when the data can be viewed as independent samples from an underlying distribution. For time series data, Dette and Weißbach (2009) considered testing the difference between regression functions of two stationary conditional heteroskedastic autoregressive processes. Politis and Romano (2010) considered the use of subsampling in two-sample or multi-

sample problems of time series data. Horváth et al. (2013) considered testing the equality of means in two functional time series by using a normal approximation for the functional sample mean and estimating the long-run covariance kernel. Dette et al. (2020) considered the problem of testing relevant hypotheses in two functional time series. In the aforementioned works, however, the two time series being compared are often assumed to be independent of each other. When dependence exists between the two time series to be compared, Zhang and Shao (2015) proposed the use of self-normalization to pivotalize the test statistic to accommodate the joint dependence. Their method is based on the key assumption that the sample size of the two time series are the same, namely the balanced two-sample case. The extension to the unbalanced case can be highly nontrivial, as Shao (2015) demonstrated with counterexamples that self-normalization can work in two-sample problems only when the two time series are independent or when they are of the same length.

In many applications, however, the two time series to be compared are not necessarily of the same length and can have different starting and ending observation times. For example, in the CRUTEM data hosted by the Met Office Hadley Centre that contains hemispheric and global mean time series of land air temperature anomalies, the northern hemisphere has anomaly

values started in 1850 while the southern hemisphere only has data started in 1857 mainly due to the poor land data coverage in the southern hemisphere before 1857. Also, for the United Kingdom (UK) precipitation data studied in Section 4.2, the monthly precipitation record started in January 1873 for the Northwest England and Wales (NwEW) region and January 1931 for the Northern Ireland (NI) region. More generally in meteorological science and climate science, it is often the case that one time series has a longer history of record than the other, and monitoring stations can be built or demolished at different times in different regions making their collected data on different time periods. In finance, portfolios and index funds can be created and abandoned at different times making their recording periods different. In economics, it is often the case that the economic variable of interest has published data available on different time horizons for different countries. In infectious diseases research such as the most recent COVID-19, different countries often start and end the data collection at different times making their observation periods different. When the two time series to be compared do not share the same observation period, it then becomes difficult to accommodate their joint dependence. In this case, using independent random block subsamples from the two time series as in Politis and Romano (2010) may no longer preserve the underlying de-

pendence structure in the target global two-sample statistic. On the other hand, the limiting distribution of the self-normalized statistic as in Shao (2015) will be affected by the underlying joint dependence in this case and is no longer pivotalized unless the two time series are of the same length. We shall here fill the gap and propose a new approach that can survive when the two time series to be compared do not share the same observation period, have possibly different lengths, and exhibit nonnegligible joint dependence.

To handle staggered observation periods in the presence of joint dependence, we propose a warped self-normalized two-sample test which incorporates a time domain warping based on how the two time series are staggered to recover and match the original dependence structure. The test can be applied to handle settings with different staggering patterns, for example when the two time series share the same starting time but have different ending times, when one time series is observed on a time subset of the other, or when the two time series are observed on time intervals that have overlapping and nonoverlapping segments. A distinguishable feature of the proposed method is that it allows the length of nonoverlapping segments to increase with the sample size so that the data sample is not necessarily dominated by the overlapping segment. In addition, it can be readily

applied to compare quantities beyond the mean such as the variance or quantiles in a unified manner, and the associated asymptotic guarantee is also established. In Section 2, we revisit the subsampling approach of Politis and Romano (2010) and the self-normalization approach of Shao (2015) to understand what caused their inapplicability to the current setting. Although combining the strength of subsampling and self-normalization has been demonstrated to be useful in certain problems (Bai et al., 2016), its direct application still cannot handle the joint dependence under staggered observation periods. For this, we in Section 3 propose a new ingredient, called the time domain warping, which leads to a warped self-normalized two-sample test that is able to address the problem. Numerical experiments including a Monte Carlo simulation and a real data analysis are provided in Section 4 to further illustrate the proposed method. Technical proofs are deferred to the supplementary material.

2. Subsampling and Self-Normalization: A Revisit

In this section, we revisit the subsampling approach of Politis and Romano (2010) and the self-normalization approach of Shao (2015) to illustrate, and more importantly to understand, why such popular approaches can become inapplicable in the current setting. For this, suppose we observe

$X_k, \ldots, X_m$ and $Y_l, \ldots, Y_n$ from a stationary bivariate time series (X_i, Y_i), where the observation periods $[k, m]$ and $[l, n]$ for the two time series are not necessarily the same and we denote the associated sample sizes by $N_x = m-k+1$ and $N_y = n-l+1$ respectively. As commented by Politis and Romano (2010), the literature on comparing time series of different lengths seems scarce, and Politis and Romano (2010) proposed to address the problem using subsampling. To illustrate, we consider the case of the mean, and we denote $\bar{X}_{k,m} = (m-k+1)^{-1} \sum_{i=k}^{m} X_i$ and $\bar{Y}_{l,n} = (n-l+1)^{-1} \sum_{j=l}^{n} Y_j$, then the method of Politis and Romano (2010) approximates the distribution of $\bar{X}_{k,m} - \bar{Y}_{l,n}$ by that of subsamples after a suitable scale adjustment. In particular, let $X_i, \ldots, X_{i+B_x-1}$ and $Y_j, \ldots, Y_{j+B_y-1}$ be subsamples of lengths B_x and B_y from the two time series, and we denote the underlying means and long-run variances by

$$\mu_x = \mathrm{E}(X_0), \quad \mu_y = \mathrm{E}(Y_0), \quad g_x = \sum_{k\in\mathbb{Z}} \mathrm{cov}(X_0, X_k), \quad g_y = \sum_{k\in\mathbb{Z}} \mathrm{cov}(Y_0, Y_k),$$

then by Politis and Romano (2010) the distribution of

$$Z_{\mathrm{PR10}} = (N_x^{-1} g_x + N_y^{-1} g_y)^{-1/2} (\bar{X}_{k,m} - \bar{Y}_{l,n})$$

under the null hypothesis of equal mean can be approximated by the empirical distribution $\hat{F}_{\mathrm{PR10}}(u)$, $u \in \mathbb{R}$, of the subsamples

$$H_{i,j} = (B_x^{-1} g_x + B_y^{-1} g_y)^{-1/2} (\bar{X}_{i,i+B_x-1} - \bar{Y}_{j,j+B_y-1}),$$

where $1 \leq i \leq q_x = N_x - B_x + 1$ and $1 \leq j \leq q_y = N_y - B_y + 1$. In particular, let $\mathrm{I}(\cdot)$ be the indicator function, then we can write

$$\hat{F}_{\mathrm{PR10}}(u) = q_x^{-1} q_y^{-1} \sum_{i=1}^{q_x} \sum_{j=1}^{q_y} \mathrm{I}(H_{i,j} \leq u).$$

Assuming independence between (X_i) and (Y_j), Politis and Romano (2010) provided the theoretical guarantee of such a two-sample subsampling approach under strong mixing. When the two time series are not independent of each other, however, the following theorem suggests that the subsampling approach of Politis and Romano (2010) may become inapplicable even for the simple case of d-dependent processes that share the same observation period.

Theorem 1. *Assume that (X_i, Y_i) is a d-dependent stationary process with finite second moment and we observe $X_k, \ldots, X_m$ and $Y_l, \ldots, Y_n$ with $k = l$ and $m = n$. If the long-run variances g_x and g_y are both bounded away from zero and the subsample lengths satisfy $1/B_x + B_x/N_x \to 0$ and $1/B_y + B_y/N_y \to 0$, then under the null hypothesis of $\mu_x = \mu_y$, as $N_x = N_y \to \infty$ we have*

$$\hat{F}_{\mathrm{PR10}}(u) \to_p \mathrm{pr}(Z_{\mathrm{PR10}} \leq c_{xy} u),$$

where $\to_p$ denotes the convergence in probability and

$$c_{xy} = \left(\frac{g_x + g_y - a_{xy}}{g_x + g_y} \right)^{1/2}, \quad a_{xy} = \sum_{k \in \mathbb{Z}} \{ \mathrm{cov}(X_0, Y_k) + \mathrm{cov}(Y_0, X_k) \}.$$

When the two time series (X_i) and (Y_j) are not independent as in Politis and Romano (2010), the cross term a_{xy} is generally nonzero which makes the constant $c_{xy} \neq 1$. As a result, the subsampling distribution $\hat{F}_{\mathrm{PR10}}$ in this case will become a distorted approximation to the distribution of $Z_{\mathrm{PR10}} = (N_x^{-1} g_x + N_y^{-1} g_y)^{-1/2} (\bar{X}_{k,m} - \bar{Y}_{l,n})$, which makes the method of Politis and Romano (2010) inapplicable to the current setting. Note that implementing the above discussed subsampling approach also requires estimating the long-run variances g_x and g_y, which can itself be a nontrivial problem and may involve the selection of additional tuning parameters. This motivates the use of self-normalization in two-sample problems as considered in Shao (2015) and Zhang and Shao (2015).

The idea of self-normalization is to use a sequence of recursive estimators to pivotalize the asymptotic distribution of the two-sample statistic. Following Shao (2015), we let $N = N_x + N_y$ and consider the case when $k = l = 1$ so that the two time series (X_i) and (Y_j) share the same starting time. In this case, the recursive two-sample differences can be constructed as

$$D_{i,N} = \bar{X}_{1,\lfloor iN_x/N \rfloor} - \bar{Y}_{1,\lfloor iN_y/N \rfloor}, \quad i = 1, \ldots, N, \tag{2.1}$$

and the self-normalized two-sample statistic of Shao (2015) takes the form

$$Z_{\mathrm{S15}} = \frac{N D_{N,N}^2}{N^{-2}\sum_{i=1}^{N} i^2 (D_{i,N} - D_{N,N})^2}. \tag{2.2}$$

To derive the asymptotic distribution of Z_{S15}, we follow Shao (2015) and make the following assumption.

(IP) There exist $a, b \neq 0$ and c such that for any $M \to \infty$,

$$M^{-1/2} \sum_{i=1}^{\lfloor Mt \rfloor} \begin{pmatrix} X_i - \mu_x \\ Y_i - \mu_y \end{pmatrix} \Rightarrow \begin{pmatrix} a & 0 \\ -c & -b \end{pmatrix} \begin{Bmatrix} W_1(t) \\ W_2(t) \end{Bmatrix},$$

where $\Rightarrow$ denotes the weak convergence in the Skorokhod space and $W_1(t)$ and $W_2(t)$ are two independent standard Brownian motions.

Assumption (IP) is generally referred to as the invariance principle, which has been widely studied under various short-range dependence conditions; see for example Hannan (1979), Herrndorf (1984), Wu (2007), Berkes et al. (2014) and references therein. We also refer to Shao (2010), Zhang and Lavitas (2018) and Zhang et al. (2019) for additional references on the use of the invariance principle in self-normalization. Let $\to_d$ denotes the convergence in distribution, then by assumption (IP) and the continuous mapping argument as in Shao (2015) and Zhang and Lavitas (2018), one can show that as $N_x/N \to p_x$ and $N_y/N \to p_y$ for some $p_x, p_y \in (0, 1)$ the

self-normalized two sample statistic

$$Z_{\mathrm{S15}} \to_d \frac{\{V_{p_x,p_y}(1;a,b,c)\}^2}{\int_0^1 \{V_{p_x,p_y}(t;a,b,c) - tV_{p_x,p_y}(1;a,b,c)\}^2 dt},$$

where

$$V_{p_x,p_y}(t;a,b,c) = \frac{a}{p_x}W_1(p_x t) + \frac{c}{p_y}W_1(p_y t) + \frac{b}{p_y}W_2(p_y t).$$

In the case when $p_x = p_y$ for which the two time series share the same length, we can write

$$V_{p_x,p_y}(t;a,b,c) = \frac{a+c}{p_x}W_1(p_x t) + \frac{b}{p_x}W_2(p_x t),$$

which then has the same distribution as the process $p_x^{-1/2}\{(a+c)^2+b^2\}^{1/2}W_1(t)$ on $t \in [0,1]$. As a result, one can show that in this case,

$$Z_{\mathrm{S15}} \to_d \frac{\{W_1(1)\}^2}{\int_0^1 \{W_1(t) - tW(1)\}^2 dt},$$

where the asymptotic distribution is pivotal. When the two time series are not of the same length, however, the process $V_{p_x,p_y}(t;a,b,c)$, $t \in [0,1]$, is not necessarily a Brownian motion and the asymptotic distribution of Z_{S15} can then depend on the underlying unknown dependence structure; see also the discussion in Shao (2015). Note that even if the two time series are of the same length, when one is started recording earlier than the other, the aforementioned self-normalization of Shao (2015) can become inapplicable

as well. To illustrate, we consider the case when $N_x = N_y$ but $k \neq l$, and we can generalize the recursive two-sample difference sequence in (2.1) to consider

$$D^{\star}_{i,N} = \bar{X}_{k,k+\lfloor iN_x/N \rfloor - 1} - \bar{Y}_{l,l+\lfloor iN_y/N \rfloor - 1}, \quad i = 1, \ldots, N. \tag{2.3}$$

Similar to (2.1), the quantity in (2.3) represents the recursive two-sample difference obtained by using the first i/N proportion of the data from both time series, where $D^{\star}_{N,N} = \bar{X}_{k,m} - \bar{Y}_{l,n}$ continues to represent the global two-sample difference. The self-normalized two-sample statistic can then be constructed as

$$Z^{\star}_{\mathrm{S15}} = \frac{N(D^{\star}_{N,N})^2}{N^{-2}\sum_{i=1}^{N} i^2 (D^{\star}_{i,N} - D^{\star}_{N,N})^2},$$

and its asymptotic distribution is given in Theorem 2.

Theorem 2. *Assume condition (IP) and $N_x = N_y$. If $k = 1$, $N \to \infty$, and $l/N \to \ell$ for some $\ell \in (0, 1/2)$, then under the null hypothesis of $\mu_x = \mu_y$ we have*

$$Z^{\star}_{\mathrm{S15}} \to_d \frac{\{V^{\star}_{\ell}(1; a, b, c)\}^2}{\int_0^1 \{V^{\star}_{\ell}(t; a, b, c) - tV^{\star}_{\ell}(1; a, b, c)\}^2 dt},$$

where

$$V^{\star}_{\ell}(t; a, b, c) = 2aW_1(t/2) + 2c\{W_1(\ell + t/2) - W_1(\ell)\} + 2b\{W_2(\ell + t/2) - W_2(\ell)\}.$$

By Theorem 2, the asymptotic distribution of the self-normalized statistic $Z^{\star}_{\text{S15}}$ in this case can be affected by the recording lag in time represented by ℓ and is generally not pivotal unless $\ell = 0$. Therefore, a direct application of self-normalization as in Shao (2015) seems to be able to address the joint dependence only in the case when the two time series are observed during the same period, and such a method can become inapplicable for time series with staggered observation periods. We shall in the following propose a new method that can handle staggered observation periods in the presence of joint dependence.

3. Warped Self-Normalized Subsampling

3.1 The Mean Case: Illustration of the Idea

We first illustrate the idea by considering the mean case. Following the setting in Section 2, suppose we observe $X_k, \ldots, X_m$ and $Y_l, \ldots, Y_n$ from a stationary bivariate time series (X_i, Y_i), where the observation periods $[k, m]$ and $[l, n]$ for the two time series are not necessarily the same and we denote the associated sample sizes by $N_x = m-k+1$ and $N_y = n-l+1$ respectively. Write $N = N_x + N_y$, and without loss of generality, we assume that $1 = k \leq l$. Our idea is to use self-normalized statistics calculated from windows with appropriately warped times between the two time series to precisely

capture the underlying dependence as in the global two-sample statistic. Let B_N be a sequence of nonnegative real numbers, we first construct a pair of index sets that represent the desired time warping as

$$\mathcal{I}_{x,i,B_N} = \{s \in \mathbb{Z} : \ l + i - 1 - \lfloor B_N(l-1)/N \rfloor \leq s \leq l - \lfloor B_N(l-1)/N \rfloor + \lfloor B_N N_x/N \rfloor - 1 + i - 1\}$$

and

$$\mathcal{I}_{y,i,B_N} = \{s \in \mathbb{Z} : \ l + i - 1 \leq s \leq l + \lfloor B_N N_y/N \rfloor - 1 + i - 1\}.$$

The time warping association between index sets of the two time series as proposed above is carefully constructed to preserve the dependence structure, and serves as a key component in self-normalized inference when the two time series have staggered observation periods. Intuitively, $\mathcal{I}_{x,i,j}$ and $\mathcal{I}_{y,i,j}$ are constructed to mimic the staggering pattern of the original data, so that the joint dependence between the original two time series with full length can be well approximated by that of the recursive pairs using the warped index sets when constructing the self-normalizer. To illustrate, we consider the simple example when $N_x = N_y = 100$ with $k = 1$ and $l = 51$. In this case, $N = N_x + N_y = 200$ and we observe $X_1, \ldots, X_{100}$ and $Y_{51}, \ldots, Y_{150}$, which have an overlap of length 50 equaling to half of their own length. If we apply the proposed time warping,

then it can be seen that the recursive time-warped index pairs are $\mathcal{I}_{x,1,j} = \{50-\lfloor j/4\rfloor+1,\ldots,50-\lfloor j/4\rfloor+\lfloor j/2\rfloor\}$ and $\mathcal{I}_{y,1,j} = \{50+1,\ldots,50+\lfloor j/2\rfloor\}$, which continue to have an overlap equaling to approximately half of their own length. The idea is then to use $\mathcal{I}_{x,1,j}$ and $\mathcal{I}_{y,1,j}$ recursively over j to construct the time-warped self-normalizer, so that the associated recursive means will have asymptotically the same joint dependence as the original $X_1,\ldots,X_{100}$ and $Y_{51},\ldots,Y_{150}$. However, if we ignore the staggered observation periods and use $Z^{\star}_{\mathrm{S15}}$ in Section 2, then it amounts to the use of $X_1,\ldots,X_{\lfloor j/2\rfloor}$ and $Y_{50+1},\ldots,Y_{50+\lfloor j/2\rfloor}$ when constructing the self-normalizer, which can have a different dependence structure when compared to the full data $X_1,\ldots,X_{100}$ and $Y_{51},\ldots,Y_{150}$. Therefore, an appropriate time warping as carefully constructed in the current article plays an important role to preserve the dependence structure for valid self-normalized inference.

For any nonempty index set $\mathcal{I}$, we use $|\mathcal{I}|$ to denote its cardinality, and we define the averages $\bar{X}_{\mathcal{I}} = |\mathcal{I}|^{-1}\sum_{i\in\mathcal{I}} X_i$ and $\bar{Y}_{\mathcal{I}} = |\mathcal{I}|^{-1}\sum_{i\in\mathcal{I}} Y_i$. Then, based on the suitably constructed warped index sets $\mathcal{I}_{x,i,B_N}$ and $\mathcal{I}_{y,i,B_N}$ as proposed above, we can form the time-warped self-normalized statistic as

$$T_{i,B_N}(\Delta) = \frac{B_N(\bar{X}_{\mathcal{I}_{x,i,B_N}} - \bar{Y}_{\mathcal{I}_{y,i,B_N}} - \Delta)^2}{B_N^{-2}\sum_{j=1}^{B_N} j^2\{(\bar{X}_{\mathcal{I}_{x,i,j}} - \bar{Y}_{\mathcal{I}_{y,i,j}}) - (\bar{X}_{\mathcal{I}_{x,i,B_N}} - \bar{Y}_{\mathcal{I}_{y,i,B_N}})\}^2}. \quad (3.4)$$

The presence of Δ allows the statistic to be used for assessing a more

general null value of the difference $\mu_x - \mu_y$, and we write $T_{i,B_N} = T_{i,B_N}(0)$ for the null hypothesis of $\mu_x = \mu_y$. When $B_N = N$ and $i = 1$, we have $\mathcal{I}_{x,1,N} = \{j \in \mathbb{Z} : \ 1 \le j \le m\}$ and $\mathcal{I}_{y,1,N} = \{j \in \mathbb{Z} : \ l \le j \le n\}$, in which case $T_{1,N}$ now represents the global time-warped self-normalized two-sample statistic.

We shall here provide a brief discussion on the connection and comparison with the conventional self-normalized statistic reviewed in Section 2. In particular, when the two time series share the same observation period with the same length for which $N_x = N_y$ and $k = l = 1$, then by Shao (2015) the self-normalized statistic $Z^{\star}_{\mathrm{S15}}$ in Section 2 will continue to work. In this case, there is no need to perform the additional time warping, and it can be shown that the proposed time-warped self-normalized statistic $T_{1,N}$ will automatically reduce to $Z^{\star}_{\mathrm{S15}}$, which can be an attractive feature. On the other hand, when the two time series to be compared do not share the same observation period or when they are of different lengths, then the new self-normalized statistic $T_{1,N}$ can now be different from $Z^{\star}_{\mathrm{S15}}$ in Section 2 due to the new self-normalizer used in the construction of $T_{1,N}$ that incorporates an additional time warping between the two time series to preserve the underlying dependence structure. With the same time warping, the distribution of $T_{1,N}$ can then be well approximated by that of the time-warped

self-normalized subsamples T_{i,B_N}.

We shall in the following present the detailed algorithm that describes and implements the proposed warped self-normalized two-sample test for time series with staggered observation periods.

(i) Given the observed time series $X_k, \ldots, X_m$ and $Y_l, \ldots, Y_n$ with $1 = k \leq l$, use (3.4) to compute the global warped self-normalized two-sample statistic $T_{1,N}$.

(ii) For a preselected B_N, use (3.4) to compute the warped self-normalized statistics for all subsamples of the designated size with Δ replaced by $\hat{\Delta} = \bar{X}_{k,m} - \bar{Y}_{l,n}$ to neutralize the effect of the null for the subsamples, and denote them by $T_{i,B_N}(\hat{\Delta})$, $1 \leq i \leq M_N$, where

$$M_N = \min(m-l+2+\lfloor B_N(l-1)/N \rfloor - \lfloor B_N N_x/N \rfloor, n-l+2-\lfloor B_N N_y/N \rfloor).$$

(iii) Compute the $(1-\alpha)$-th quantile $\hat{q}_{T,1-\alpha}$ of $T_{i,B_N}(\hat{\Delta})$, $1 \leq i \leq M_N$, and reject the null hypothesis of $\mu_x = \mu_y$ at level α if $T_{1,N} > \hat{q}_{T,1-\alpha}$.

To provide the theoretical justification of the proposed warped self-normalized method, we use the strong mixing framework of Rosenblatt (1956). In particular, let $\mathcal{F}_{a,b}$ be the σ-field generated by $(X_a, Y_a), \ldots, (X_b, Y_b)$ for

$-\infty \le a \le b \le \infty$, then the strong mixing coefficient is defined as

$$\alpha(k) = \sup_{A \in \mathcal{F}_{-\infty,0},\ B \in \mathcal{F}_{k,\infty}} |\mathrm{pr}(A \cap B) - \mathrm{pr}(A)\mathrm{pr}(B)|.$$

Let Υ_k be the covariance matrix between the two vectors (X_i, Y_i) and (X_{i+k}, Y_{i+k}), then the long-run covariance matrix of the process (X_i, Y_i) is given by $\Upsilon = \sum_{k\in\mathbb{Z}} \Upsilon_k$. We make the following assumptions.

(S1) The process (X_i, Y_i) is strong mixing with

$$\mathrm{E}(|X_i|^\iota) < \infty,\ \mathrm{E}(|Y_i|^\iota) < \infty,\ \sum_{k=1}^{\infty} \{\alpha(k)\}^{1-2/\iota} < \infty$$

for some $\iota > 2$.

(S2) The long-run covariance matrix Υ is positive definite.

Assumptions (S1) and (S2) are standard primitive conditions for the invariance principle (IP) in Section 2; see for example Phillips and Durlauf (1986). The strong mixing condition has been widely used to study limit theorems of dependent processes, and we refer to the survey paper by Bradley (2005) and the book by Bradley (2007) for a detailed review and additional references. Theorem 3 provides the asymptotic justification of the proposed warped self-normalized subsampling method.

Theorem 3. *Assume (S1), (S2),* $N_x/N \to p_x \in (0,1)$, $N_y/N \to p_y \in (0,1)$ *and* $l/N \to \ell \in [0,1)$ *as* $N \to \infty$. *If* $B_N \to \infty$ *and* $B_N/N \to 0$, *then for*

any $\alpha \in (0,1)$ *we have (i) under the null hypothesis of* $\mu_x = \mu_y$,

$$\mathrm{pr}(T_{1,N} > \hat{q}_{T,1-\alpha}) \to \alpha;$$

and (ii) under the alternative when $N^{1/2}|\mu_x - \mu_y| \to \infty$,

$$\mathrm{pr}(T_{1,N} > \hat{q}_{T,1-\alpha}) \to 1.$$

By Theorem 3, the asymptotic size of the proposed warped self-normalized two-sample test is guaranteed to converge to its nominal level under the null, and the power of the proposed test will converge to one under the alternative when $N^{1/2}|\mu_x - \mu_y| \to \infty$. The proposed warped self-normalized two-sample test provides a unified solution to different staggering schemes of the two time series, for example when the two time series share the same starting time but have different ending times, when one time series is observed on a time subset of the other, or when the two time series are observed on time intervals that have overlapping and nonoverlapping segments. This makes it convenient to use in practice as one does not need to perform a case-by-case study and can simply apply the proposed method to obtain a meaningful p-value justified by a solid statistical theory. We shall in the following section consider extending the proposed test to quantities beyond the mean, such as the variance or quantiles.

3.2 The General Case: Quantities Beyond the Mean

Due to the incorporation of self-normalization, the proposed method can be easily extended to cases beyond the mean. For this, suppose one is interested in comparing the quantity $\theta_x = Q(F_{x,d})$ with $\theta_y = Q(F_{y,d})$, where Q is a functional that maps the d-dimensional marginal distributions $F_{x,d}$ and $F_{y,d}$ of the two time series to their respective quantities of interest. As illustrated in Shao (2010) and Zhang and Lavitas (2018), such a framework covers many commonly used quantities of interest as special cases. For example, if we set $d = 1$ and use $F_{x,1}$ to denote the marginal distribution, then taking $Q(F_{x,1}) = \int_{\mathbb{R}} uF_{x,1}(u)du$ leads us to the mean case considered in Section 3.1. As additional examples, we can set $Q(F_{x,1}) = \int_{\mathbb{R}} u^2F_{x,1}(u)du - \{\int_{\mathbb{R}} uF_{x,1}(u)du\}^2$ to focus on the variance case and $Q(F_{x,1}) = F^{-1}_{x,1}(q)$ for some $q \in (0, 1)$ to focus on the quantile case; see also Shao (2010) and Zhang and Lavitas (2018) for further discussions. For index sets $\mathcal{I}_x$ and $\mathcal{I}_y$, let $\hat{F}_{x,d,\mathcal{I}_x}$ and $\hat{F}_{y,d,\mathcal{I}_y}$ be the associated empirical distributions calculated from $(X_i)_{i\in\mathcal{I}_x}$ and $(Y_i)_{i\in\mathcal{I}_y}$ respectively, then $\hat{\theta}_{x,\mathcal{I}_x} = Q(\hat{F}_{x,d,\mathcal{I}_x})$ and $\hat{\theta}_{y,\mathcal{I}_y} = Q(\hat{F}_{y,d,\mathcal{I}_y})$ represent the parameter estimators calculated from $(X_i)_{i\in\mathcal{I}_x}$ and $(Y_i)_{i\in\mathcal{I}_y}$ respectively. We can now form the time-warped self-normalized

statistic as

$$T^{\theta}_{i,B_N}(\Delta) = \frac{B_N(\hat{\theta}_{x,\mathcal{I}_{x,i,B_N}} - \hat{\theta}_{y,\mathcal{I}_{y,i,B_N}} - \Delta)^2}{B_N^{-2}\sum_{j=1}^{B_N} j^2\{(\hat{\theta}_{x,\mathcal{I}_{x,i,j}} - \hat{\theta}_{y,\mathcal{I}_{y,i,j}}) - (\hat{\theta}_{x,\mathcal{I}_{x,i,B_N}} - \hat{\theta}_{y,\mathcal{I}_{y,i,B_N}})\}^2}, \tag{3.5}$$

which extends (3.4) from the mean case to more general quantities. Similarly, we write $T^{\theta}_{i,B_N} = T^{\theta}_{i,B_N}(0)$ for the null hypothesis of $\theta_x = \theta_y$. By substituting this generalized statistic (3.5) into the algorithm described in Section 3.1, we can similarly compute the global time-warped self-normalized statistic $T^{\theta}_{1,N}$ along with its subsample version $T^{\theta}_{i,B_N}(\hat{\Delta}^{\theta})$ for $1 \le i \le M_N$ where $\hat{\Delta}^{\theta} = \hat{\theta}_{x,\mathcal{I}_{x,1,N}} - \hat{\theta}_{y,\mathcal{I}_{y,1,N}}$. Let $\hat{q}_{T^{\theta},1-\alpha}$ be the $(1-\alpha)$-th quantile of $T^{\theta}_{i,B_N}(\hat{\Delta}^{\theta})$, $1 \le i \le M_N$, we can then reject the null hypothesis of $\theta_x = \theta_y$ at level α if $T^{\theta}_{1,N} > \hat{q}_{T^{\theta},1-\alpha}$. To provide the theoretical justification under this more general setting, we introduce the notion of influence function (Hampel et al., 1986) defined as

$$\mathrm{IF}_Q(z, F_d) = \lim_{\varepsilon \downarrow 0} \frac{Q\{(1-\varepsilon)F_d + \varepsilon\delta_z\} - Q(F_d)}{\varepsilon},$$

where $z \in \mathbb{R}^d$ and δ_z denotes the point mass at z. Then for any index set $\mathcal{I}$, we can follow Shao (2010) and apply the expansion

$$\hat{\theta}_{x,\mathcal{I}} = Q(\hat{F}_{x,d,\mathcal{I}}) = Q(F_{x,d}) + \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathrm{IF}_Q(X_i, F_{x,d}) + R_{x,\mathcal{I}}$$

and

$$\hat{\theta}_{y,\mathcal{I}} = Q(F_{y,d}) + \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathrm{IF}_Q(Y_i, F_{y,d}) + R_{y,\mathcal{I}},$$

where $\mathrm{E}\{\mathrm{IF}_Q(X_i, F_{x,d})\} = \mathrm{E}\{\mathrm{IF}_Q(Y_i, F_{y,d})\} = 0$ and $R_{\mathcal{I}} = (R_{x,\mathcal{I}}, R_{y,\mathcal{I}})^\top$ with $^\top$ being the transpose represents the remainder term. We make the following assumptions.

(A1) The process (X_i, Y_i) is strong mixing with

$$\mathrm{E}\{|\mathrm{IF}_Q(X_i, F_{x,d})|^\iota\} < \infty, \quad \mathrm{E}\{|\mathrm{IF}_Q(Y_i, F_{y,d})|^\iota\} < \infty,$$
$$\sum_{k=1}^{\infty} \{\alpha(k)\}^{1-2/\iota} < \infty$$

for some $\iota > 2$.

(A2) The long-run covariance matrix of the influence function process $\{\mathrm{IF}_Q(X_i, F_{x,d}), \mathrm{IF}_Q(Y_i, F_{y,d})\}^\top$ is positive definite.

(A3) The remainder term $|\mathcal{I}|R_{\mathcal{I}} = o_p(N^{1/2})$ uniformly over $\mathcal{I} = \{i \in \mathbb{Z} : Ns \le i \le Nt\}$ for $0 \le s \le t \le 1$ for all large N.

Assumptions (A1)–(A3) are standard in self-normalized inference; see for example Shao (2010) and the discussions therein. Intuitively, Assumptions (A1) and (A2) imply an invariance principle for the influence functions which generalizes the mean invariance principle (IP) in Section 2. Assumption (A3) is a negligibility condition on the remainder term, which is also mild; see for example the discussion in Shao (2010) and its verification for the class of smooth function models. We also remark that it is possible to

replace or alleviate Assumptions (A1)–(A3) by, for example, the functional delta approach as in Volgushev and Shao (2014). Theorem 4 provides the asymptotic justification of the proposed warped self-normalized subsampling method for quantities beyond the mean.

Theorem 4. *Assume (A1)–(A3),* $N_x/N \to p_x \in (0,1)$, $N_y/N \to p_y \in (0,1)$ *and* $l/N \to \ell \in [0,1)$ *as* $N \to \infty$. *If* $B_N \to \infty$ *and* $B_N/N \to 0$, *then for any* $\alpha \in (0,1)$ *we have (i) under the null hypothesis of* $\theta_x = \theta_y$,

$$\mathrm{pr}(T^{\theta}_{1,N} > \hat{q}_{T^{\theta},1-\alpha}) \to \alpha;$$

and (ii) under the alternative when $N^{1/2}|\theta_x - \theta_y| \to \infty$ *we have*

$$\mathrm{pr}(T^{\theta}_{1,N} > \hat{q}_{T^{\theta},1-\alpha}) \to 1.$$

Although in practice one is often interested in testing the null hypothesis of $\theta_x = \theta_y$ or equivalently $\theta_x - \theta_y = 0$, it is possible to extend the proposed time-warped self-normalized test to handle the more general null hypothesis of $\theta_x - \theta_y = \Delta$ for some prespecified $\Delta \in \mathbb{R}$. For this, we use the global time-warped self-normalized statistic $T^{\theta}_{1,N}(\Delta)$ and follow the algorithm described in Section 3.1 to obtain its subsample version $T^{\theta}_{i,B_N}(\hat{\Delta}^{\theta})$ for $1 \le i \le M_N$ where $\hat{\Delta}^{\theta} = \hat{\theta}_{x,\mathcal{I}_{x,1,N}} - \hat{\theta}_{y,\mathcal{I}_{y,1,N}}$. Note that the subsample statistics $T^{\theta}_{i,B_N}(\hat{\Delta}^{\theta})$, $1 \le i \le M_N$, use the estimate $\hat{\Delta}^{\theta}$ in their construction, and as a result the associated subsampling distribution will not be affected by whether the null

$\theta_x - \theta_y = \Delta$ is true or not. Corollary 1 provides the theoretical justification of the proposed time-warped self-normalized test for handling such a more general null hypothesis.

Corollary 1. *Assume (A1)–(A3),* $N_x/N \to p_x \in (0,1)$, $N_y/N \to p_y \in (0,1)$ *and* $l/N \to \ell \in [0,1)$ *as* $N \to \infty$. *If* $B_N \to \infty$ *and* $B_N/N \to 0$, *then for any* $\alpha \in (0,1)$ *we have (i) under the null hypothesis of* $\theta_x - \theta_y = \Delta$,

$$\mathrm{pr}(T^{\theta}_{1,N}(\Delta) > \hat{q}_{T^{\theta},1-\alpha}) \to \alpha;$$

and (ii) under the alternative when $N^{1/2}|(\theta_x - \theta_y) - \Delta| \to \infty$ *we have*

$$\mathrm{pr}(T^{\theta}_{1,N}(\Delta) > \hat{q}_{T^{\theta},1-\alpha}) \to 1.$$

The proposed warped self-normalized subsampling method provides a unified approach for two-sample testing of mean and other quantities when the two time series to be compared can depend on each other, be collected on staggered time periods, and have different lengths. In contrast, existing results are often developed for the setting when the two time series are either independent or collected on the same period, and may no longer work in the current more general setting; see for example the discussion in Section 2.

4. Numerical Experiments

4.1 Simulation Results

We shall here provide a simulation study to examine the finite-sample performance of the proposed warped self-normalized subsampling method, denoted by WSNS hereafter. We also make a comparison with the two-sample subsampling approach of Politis and Romano (2010), denoted by PR10 hereafter, and the self-normalized two-sample inference of Shao (2015), denoted by S15 hereafter. For this, let (ζ_i) and (ξ_i) be independent autoregressive processes generated as

$$\begin{pmatrix} \zeta_i \\ \xi_i \end{pmatrix} = \rho \begin{pmatrix} \zeta_{i-1} \\ \xi_{i-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix},$$

where ϵ_{1i} and ϵ_{2i}, $i \in \mathbb{Z}$, are independent standard normal random variables. Additional simulation results for other innovation types are provided in the supplementary material. We consider the joint process

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ r & \sqrt{1-r^2} \end{pmatrix} \begin{pmatrix} \zeta_i \\ \xi_i \end{pmatrix},$$

where r controls the degree of dependence between the two time series (X_i) and (Y_i). The observations are taken as $X_k, \ldots, X_m$ and $Y_l, \ldots, Y_n$, for which we consider three scenarios.

- Scenario 1: $k = l = 1$, $m = n = 500s$, $s \in \{1, 2, 3\}$, which represents the situation when the two time series share the same observation period.

- Scenario 2: $k = 1$, $l = 100s + 1$, $m = 600s$, $n = 700s$, $s \in \{1, 2, 3\}$, which represents the situation when the two time series are recorded on shifted observation periods.

- Scenario 3: $k = 1$, $l = 250s + 1$, $m = n = 750s$, $s \in \{1, 2, 3\}$, which represents the situation when the observation period of one time series is a subset of the other.

For all the three scenarios, the overlapping length between (X_i) and (Y_i) is set as $N_c = 500s$, $s \in \{1, 2, 3\}$. For each realization, we apply the proposed WSNS method, the PR10 method of Politis and Romano (2010) and the S15 method of Shao (2015) to perform two-sample tests on the mean, median and variance of the two time series. Let $\rho = 0.3$, $N_c = 1500$, $B_N = 100$, and $r \in \{0, 0.4, 0.8\}$, the results based on 1000 realizations for each setting are summarized in Tables 1–3 for Scenarios 1–3 respectively. Additional simulation results for other choices of ρ, N_c and B_N are provided in the supplementary material. From the simulation results, we can observe the followings.

First, the PR10 method of Politis and Romano (2010) was developed for independent time series and is therefore expected to work when $r = 0$. When there exists dependence between the two time series to be compared, however, it is no longer guaranteed to work and it can be seen from Tables 1–3 that the PR10 method starts to exhibit a certain degree of size distortions when $r = 0.4$ and further deteriorates when $r = 0.8$. For example, if we consider the mean case with $r = 0.8$ in Table 1, then the empirical coverage probabilities of the PR10 method are distorted to 1.000, 1.000 and 1.000 at 90%, 95% and 99% nominal levels. Note that the PR10 method requires estimating the normalizing sequence of the test statistic, which relates to the long-run variance of a dependent process and we use the banding estimate as described in Zhang (2021). As a comparison, the S15 method of Shao (2015) uses a self-normalizer to pivotalize the two-sample test statistic that avoids direct estimation of the normalizing sequence. It can be seen from Table 1 that the S15 method performs reasonably well under Scenario 1 even when the two time series depend on each other with $r = 0.4$ or $r = 0.8$. However, as discussed in Section 2, the S15 method described in Shao (2015) is not directly applicable to address the situation when the two time series have different starting times unless they are independent. In particular, it can be seen from Table 2 that the S15 method

performs reasonably well under Scenario 2 when $r = 0$ but exhibits size distortions similar to the PR10 method when $r = 0.4$ and $r = 0.8$. For example, if we consider the mean case with $r = 0.8$ in Table 2, then the empirical coverage probabilities of the S15 method are distorted to 0.989, 0.995 and 1.000 at 90%, 95% and 99% nominal levels. In comparison, the empirical coverage probabilities of the proposed WSNS method for the same setting are 0.922, 0.964 and 0.993 which are reasonably close to their 90%, 95% and 99% nominal levels. The main reason is that, when the two time series exhibit joint dependence and are observed on staggered observation periods, the original self-normalizer as used in Shao (2015) may no longer be able to pivotalize the two-sample statistic; see also the discussion in Section 2. By incorporating an appropriate time-warping into the construction of a new time-warped self-normalizer as proposed in Section 3, our WSNS method can effectively handle time series exhibiting joint dependence with staggered observation periods. In addition, it provides a unified approach to handle the mean and other quantities such as the median and variance. It can be seen from Tables 1–3 that the proposed WSNS method seems to perform reasonably well as the empirical coverage probabilities are mostly close to their nominal levels for Scenarios 1–3 no matter if the two time series are independent ($r = 0$) or dependent ($r = 0.4$ and $r = 0.8$); see also

the additional simulation results in the supplementary material.

Table 1: Empirical coverage probabilities of the proposed WSNS method, the PR10 method of Politis and Romano (2010) and the S15 method described in Shao (2015) for Scenario 1 with standard normal innovation, $\rho = 0.3$, $N_c = 1500$, and $B_N = 100$.

		Mean			Median			Variance		
r	Method	90%	95%	99%	90%	95%	99%	90%	95%	99%
0	WSNS	0.893	0.937	0.982	0.895	0.950	0.992	0.874	0.925	0.979
	PR10	0.884	0.934	0.985	0.879	0.945	0.991	0.882	0.942	0.986
	S15	0.898	0.951	0.991	0.904	0.958	0.995	0.896	0.954	0.990
0.4	WSNS	0.889	0.938	0.983	0.894	0.944	0.981	0.871	0.927	0.972
	PR10	0.965	0.989	0.998	0.950	0.976	0.997	0.916	0.960	0.989
	S15	0.900	0.947	0.992	0.913	0.959	0.991	0.902	0.947	0.987
0.8	WSNS	0.887	0.926	0.983	0.868	0.929	0.984	0.887	0.928	0.974
	PR10	1.000	1.000	1.000	0.998	0.999	1.000	0.984	0.995	1.000
	S15	0.890	0.946	0.994	0.925	0.964	0.997	0.903	0.944	0.992

Table 2: Empirical coverage probabilities of the proposed WSNS method, the PR10 method of Politis and Romano (2010) and the S15 method described in Shao (2015) for Scenario 2 with standard normal innovation, $\rho = 0.3$, $N_c = 1500$, and $B_N = 100$.

		Mean			Median			Variance		
r	Method	90%	95%	99%	90%	95%	99%	90%	95%	99%
0	WSNS	0.900	0.953	0.988	0.902	0.954	0.987	0.881	0.932	0.983
	PR10	0.903	0.942	0.982	0.892	0.950	0.982	0.888	0.941	0.993
	S15	0.901	0.940	0.989	0.907	0.958	0.991	0.887	0.948	0.992
0.4	WSNS	0.909	0.954	0.991	0.908	0.961	0.994	0.876	0.937	0.985
	PR10	0.953	0.978	0.996	0.936	0.977	0.996	0.915	0.963	0.994
	S15	0.942	0.973	0.999	0.939	0.973	0.995	0.896	0.960	0.993
0.8	WSNS	0.922	0.964	0.993	0.892	0.957	0.994	0.890	0.937	0.981
	PR10	0.991	0.998	1.000	0.980	0.994	1.000	0.980	0.996	1.000
	S15	0.989	0.995	1.000	0.970	0.988	0.999	0.973	0.991	0.999

Table 3: Empirical coverage probabilities of the proposed WSNS method, the PR10 method of Politis and Romano (2010) and the S15 method described in Shao (2015) for Scenario 3 with standard normal innovation, $\rho = 0.3$, $N_c = 1500$, and $B_N = 100$.

		Mean			Median			Variance		
r	Method	90%	95%	99%	90%	95%	99%	90%	95%	99%
0	WSNS	0.894	0.945	0.992	0.903	0.961	0.991	0.866	0.933	0.986
	PR10	0.896	0.940	0.980	0.884	0.933	0.985	0.881	0.939	0.988
	S15	0.896	0.951	0.989	0.917	0.962	0.989	0.894	0.949	0.990
0.4	WSNS	0.877	0.938	0.988	0.896	0.948	0.990	0.866	0.931	0.982
	PR10	0.947	0.979	0.992	0.932	0.969	0.997	0.913	0.961	0.994
	S15	0.940	0.976	0.997	0.933	0.971	0.997	0.908	0.960	0.998
0.8	WSNS	0.895	0.952	0.990	0.877	0.936	0.989	0.869	0.927	0.973
	PR10	0.992	0.999	1.000	0.981	0.994	1.000	0.974	0.991	1.000
	S15	0.982	0.994	1.000	0.971	0.988	1.000	0.964	0.987	0.998

4.2 Application to a Precipitation Data

We shall here apply the proposed method to a precipitation data to compare the monthly precipitation series in the Northwest England and Wales

(NwEW) region and the Northern Ireland (NI) region of the United Kingdom (UK). The data is available from the Met Office Hadley Centre website at `https://www.metoffice.gov.uk`, and time series plots are provided in Figure 1. For the NwEW region, the precipitation series begins in January 1873, while for the NI region it begins in January 1931. Therefore, one of the time series has an earlier starting time than the other and the two time series are of different lengths. This makes it desirable to use the proposed method for their two-sample tests, and we use all the available data through December 2021 in the analysis. The sample size for the NwEW series is 1788 and the sample size for the NI series is 1092. We consider testing if the NwEW region and the NI region share the same mean, median and variance for their history precipitation. The subsampling bandwidth is selected using the minimum volatility method described in the supplementary material, and the selected bandwidths are $B_N = 43$, 79 and 132 for the mean, median and variance respectively. The results are summarized in Table 4, from which we can see that, at the 5% significance level, the NwEW region and the NI region seem to possess the same mean and median in precipitation with p-values 0.501 and 0.356 respectively. However, the p-value for the variance test is 0.000 indicating that the NwEW region and the NI region have statistically different variations in their precipitation. This

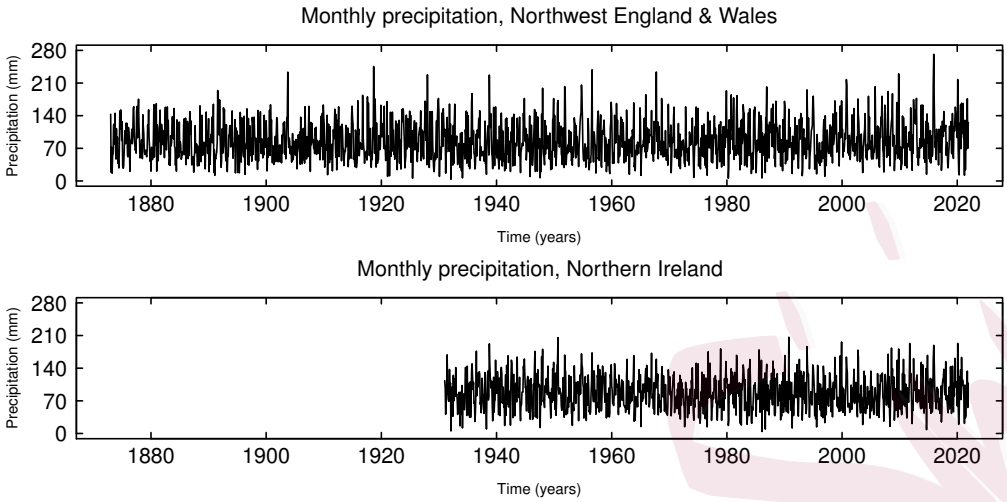

Figure 1: Time series plots for monthly precipitation records in the Northwest England and Wales (NwEW) region (top) and the Northern Ireland (NI) region (bottom) through December 2021.

could possibly be related to the mountainous terrain in the NwEW region that makes its rainfall vary a lot across its different subareas. In particular, according to the fact sheet published by the Met Office (2013), the moist westerly winds in general tend to produce orographic rainfalls that lead to more precipitation over the mountains. On the other hand, as commented in the climatological memorandum published by the Meteorological Office (1983), the protected areas behind the mountains can be much drier than even the driest zones of NI. Therefore, our analysis seems to provide the statistical support that complements the climatological intuition.

Table 4: Two-sample test results of the monthly precipitation in the Northwest England and Wales (NwEW) region and Northern Ireland (NI) region of the United Kingdom (UK).

Quantity	$T^{\theta}_{1,N}$	p-value
Mean	3.70	0.501
Median	9.12	0.356
Variance	253.59	0.000

5. Conclusion

Applications from various disciplines often desire the comparison of a given quantity from two time series in the form of a two-sample test. In practice, there are many circumstances that can cause the two time series to depend on each other and be recorded on different time periods. For example, climate science data recorded in different regions often exhibit joint dependence, and one region may get monitored earlier than the other. The joint dependence between the two time series together with their staggered observation periods make many existing methods not directly applicable, and it is desirable to propose a new method to handle such situations. In this paper, we consider incorporating a suitable time domain warping into

the construction of a new time-warped self-normalizer as proposed in Section 3, and the resulting method can effectively handle two-sample testing of time series that exhibit joint dependence with staggered observation periods. In addition, it provides a unified approach to handle the mean and other quantities such as the variance or quantiles which can be a convenient feature for practitioners. By applying the proposed method to a precipitation data, we found that the NwEW region and the NI region in UK share the same mean and median in their history precipitation but have statistically different variations. This provides the statistical support and complements the climatological intuition that the mountainous subareas in NwEW tend to receive much more rainfalls due to orographic precipitation than the protected subarea that is often much drier than the NI.

Supplementary Materials

Supplementary material contains technical proofs for our main results in Sections 2 and 3, additional simulation results, and the description of a minimum volatility bandwidth choice.

Acknowledgements

We are grateful to the editor, the associate editor and two anonymous referees for their helpful comments and suggestions. The research is supported by the NSF CAREER Award DMS-2131821.

References

Bai, S., Taqqu, M. S. and Zhang, T. (2016). A unified approach to self-normalized block sampling. *Stochastic Processes and their Applications*, **126**, 2465–2493.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, **6**, 311–329.

Berkes, I., Liu, W. and Wu, W. B. (2014). Komlós-Major-Tusnády approximation under dependence. *The Annals of Probability*, **42**, 794–817.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, **2**, 107–144.

Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*. Kendrick Press, Utah.

Cai, T. T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, 349–372.

Chen, L., Dou, W. W. and Qiao, Z. (2013). Ensemble subsampling for imbalanced multivariate two-sample tests. *Journal of the American Statistical Association*, **108**, 1308–1323.

REFERENCES

Chen, S. X., Li, J. and Zhong, P.-S. (2019). Two-sample and ANOVA tests for high dimensional means. *The Annals of Statistics*, **47**, 1443–1474.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, **38**, 808–835.

Cressie, N. and Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal*, **28**, 131–148.

Dette, H., Kokot, K. and Volgushev, S. (2020). Testing relevant hypotheses in functional time series via self-normalization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **82**, 629–660.

Dette, H. and Weissbach, R. (2009). A bootstrap test for the comparison of nonlinear time series. *Computational Statistics & Data Analysis*, **53**, 1339–1349.

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V. and Lahiri, S. N. (2015). A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association*, **110**, 837–849.

Hall, P. and Martin, M. (1988). On the bootstrap and two-sample problems. *Australian Journal of Statistics*, **30**, 179–192.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Rousseeuw, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hannan, E. J. (1979). The central limit theorem for time series regression. *Stochastic Processes and their Applications*, **9**, 281–289.

REFERENCES

Herrndorf, N. (1984). A functional central limit theorem for weakly dependent sequences of random variables. *The Annals of Probability*, 141–153.

Horváth, L., Kokoszka, P. and Reeder, R. (2013). Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 103–122.

Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics*, **2**, 360–378.

Keselman, H., Othman, A. R., Wilcox, R. R. and Fradette, K. (2004). The new and improved two-sample t test. *Psychological Science*, **15**, 47–51.

Met Office (2013). *Climate of the British Isles*. No. 4 in National Meteorological Library and Archive Fact Sheet. Met Office, Exeter, U.K.

Meteorological Office (1983). *The Climate of Northern Ireland*. No. 143 in Climatological Memorandum. Meteorological Office, Bracknell, U.K.

Phillips, P. C. and Durlauf, S. N. (1986). Multiple time series regression with integrated processes. *The Review of Economic Studies*, **53**, 473–495.

Politis, D. N. and Romano, J. P. (2010). K-sample subsampling in general spaces: The case of independent time series. *Journal of Multivariate Analysis*, **101**, 316–326.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, **42**, 43–47.

Shao, X. (2010). A self-normalized approach to confidence interval construction in time series.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), **72**, 343–366.

Shao, X. (2015). Self-normalization for time series: a review of recent developments. *Journal of the American Statistical Association*, **110**, 1797–1817.

Städler, N. and Mukherjee, S. (2017). Two-sample testing in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 225–246.

Volgushev, S. and Shao, X. (2014). A general approach to the joint asymptotic analysis of statistics from sub-samples. *Electronic Journal of Statistics*, **8**, 390–431.

Wu, W. B. (2007). Strong invariance principles for dependent random variables. *The Annals of Probability*, **35**, 2294–2320.

Xu, G., Lin, L., Wei, P. and Pan, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika*, **103**, 609–624.

Zhang, J.-T., Guo, J., Zhou, B. and Cheng, M.-Y. (2020). A simple two-sample test in high dimensions based on L^2-norm. *Journal of the American Statistical Association*, **115**, 1011–1027.

Zhang, T. (2021). High-quantile regression for tail-dependent time series. *Biometrika*, **108**, 113–126.

Zhang, T. and Lavitas, L. (2018). Unsupervised self-normalized change-point testing for time series. *Journal of the American Statistical Association*, **113**, 637–648.

Zhang, T., Lavitas, L. and Pan, Q. (2019). Asymptotic behavior of optimal weighting in generalized self-normalization for time series. *Journal of Time Series Analysis*, **40**, 831–

851.

Zhang, X. and Shao, X. (2015). Two sample inference for the second-order property of temporally dependent functional data. *Bernoulli*, **21**, 909–929.

Department of Mathematics and Statistics, Boston University, 665 Commonwealth Ave, Boston, MA 02215, U.S.A.

E-mail: weiliang@bu.edu

Department of Statistics, University of Georgia, 310 Herty Drive, Athens, GA 30602, U.S.A.

E-mail: tingzhang@uga.edu