Geophysical Journal International

Geophys. J. Int. (2019) **218,** 2165–2178
Advance Access publication 2019 June 14
GJI General Geophysical Methods

doi: 10.1093/gji/ggz280

A multitaper spectral estimator for time-series with missing data

Alan D. Chave

Department of Applied Ocean Physics and Engineering, Woods Hole Oceanographic Institution, Woods Hole, MA 02543*, USA. E-mail: achave@whoi.edu*

Accepted 2019 June 13. Received 2019 May 1; in original form 2019 May 23

SUMMARY

A multitaper estimator is proposed that accommodates time-series containing gaps without using any form of interpolation. In contrast with prior missing-data multitaper estimators that force standard Slepian sequences to be zero at gaps, the proposed missing-data Slepian sequences are defined only where data are present. The missing-data Slepian sequences are frequency independent, as are the eigenvalues that define the energy concentration within the resolution bandwidth, when the process bandwidth is $[-1/2,\ \ 1/2)$ for unit sampling and the sampling scheme comprises integer multiples of unity. As a consequence, one need only compute the ensuing missing-data Slepian sequences for a given sampling scheme once, and then the spectrum at an arbitrary set of frequencies can be computed using them. It is also shown that the resulting missing-data multitaper estimator can incorporate all of the optimality features (i.e. adaptive-weighting, *F*-test and reshaping) of the standard multitaper estimator, and can be applied to bivariate or multivariate situations in similar ways. Performance of the missing-data multitaper estimator is illustrated using length of day, seafloor pressure and Nile River low stand time-series.

Key words: Fourier analysis; Numerical approximations and analysis; Statistical methods; Time-series analysis.

1 INTRODUCTION

The analysis of time-series where the data are sampled at constant intervals in time (or space) is well understood, and constitutes a core capability in many fields of science and engineering. A major issue in time-series analysis is devising a spectral estimator that operates on a finite sample such that the estimate is not dominated by bias, is statistically consistent, has a measurable variance and is relatively immune to small departures from the underlying assumptions. This problem becomes especially acute when the time-series is short (i.e. when the required resolution is of order the inverse of the time-series length), is a mixture of stochastic and deterministic components or when the spectral dynamic range is large. Under these circumstances, the multitaper estimator of Thomson (1982) is the gold standard; see also Percival & Walden (1993, §7–9). The advantages of the multitaper method include (1) it is a small sample theory with sample size explicit, (2) its bias is quantifiable, (3) the resolution bandwidth is well-defined, (4) the variance efficiency is high, (5) it is data adaptive, so yields a low bias result even where the spectrum is weak, (6) deterministic or spectral line components can be accommodated in a straightforward manner and (7) for Gaussian data, it is approximately maximum likelihood (Stoica & Sundin 1999). In this context, bias primarily means spectral leakage from frequencies where the spectrum is large to those where it is small, the resolution bandwidth defines the ability to resolve closely spaced spectral features, and high variance efficiency refers to the ability to make use of most of the data. A maximum likelihood estimator has desirable statistical optimality properties. For a more detailed review of these statistical concepts, see Chave (2017, §5.2 and 5.4).

The analysis of time-series with missing data (i.e. gaps) is less well understood. Most of the literature concerns time-series with one or more periodic components contained in noise. For example, the Lomb–Scargle periodogram (Lomb 1976; Scargle 1982) was introduced to analyse astronomical data containing gaps for periodic components. However, it has been known since the time of Schuster (1898) that the periodogram is badly biased; see Thomson & Haley (2014, Fig. 1a) for a spectacular example. Other parametric and resampling methods have been introduced to accommodate missing data, as recently reviewed by Babu & Stoica (2010). None of these provides the performance required for general applications.

However, power spectra of stochastic processes, or mixtures of stochastic and deterministic processes, are more commonly of interest in the earth and ocean sciences. Fodor & Stark (2000) modified the standard multitaper estimator by forcing the ordinary Slepian sequences (OSSs) for a complete time-series to be zero where there are data gaps. Smith-Boughner & Constable (2012; hereafter SC12) investigated this approach more deeply, and will be used for comparisons in this paper, where it will be shown that such an approach yields suboptimal data tapers, in the sense that the spectral window main lobe shape is not substantially square and the sidelobes are significantly elevated relative to those of the estimator proposed here.

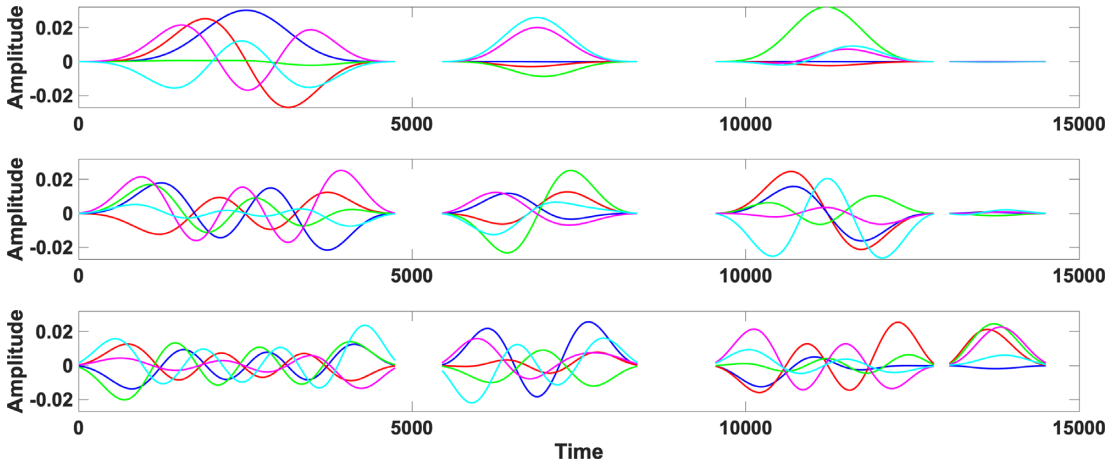

Figure 1. The first 15 missing-data Slepian sequences for a time-bandwidth of 12 and 12 400 data derived from a regular sampling scheme 14 500 long. From top to bottom, the panels show the tapers for indices of 0–4, 5–9 and 10–14. The curve colour sequence is blue, red, green, magenta and cyan, so that for the top panel blue = 0, red = 1, green = 2, magenta = 3 and cyan = 4.

As an alternative, the extension of the multitaper method to irregular sampling introduced by Bronez (1985, 1988) is re-examined. A significant problem with the Bronez approach is that the generalized Slepian sequences that it yields are analogous to the point-by-point product of an OSS and the complex exponentials in the Fourier transform. This makes the generalized Slepian sequences frequency-dependent, and poses numerical issues at high frequencies. It will be shown that a simple transformation of Bronez's result yields frequency-independent Slepian sequences, along with frequency-independent eigenvalues that define the energy concentration within the resolution bandwidth, whenever the process bandwidth is $[-1/2,\ 1/2)$ for unit sampling and the sampling scheme comprises integer multiples of unity. As a consequence, one need only compute the resulting missing-data Slepian sequences (MDSSs) for a given sampling scheme once, and then the spectrum at an arbitrary set of frequencies can be computed using them. It will also be shown that the resulting missing-data multitaper estimator can incorporate all of the optimality features (i.e. adaptive-weighting, F-test and reshaping) of the standard multitaper estimator, and can be applied to bivariate or multivariate situations in similar ways.

The following section outlines the standard multitaper estimator and the Bronez extension to irregular sampling, ending by showing that a simple transformation results in missing-data Slepian sequences that are frequency independent. Section 3 utilizes length of day data with missing values as analysed by SC12, illustrating the superior performance of the present approach. Section 4 describes the effect of varying lengths of data gaps using seafloor pressure data that have a high dynamic range, showing that performance with a single long gap is typically superior to many gaps for the same number of missing data. Section 5 analyses a ∼1300 yr long time-series of the Nile River low stands that contains several gaps. Section 6 is a discussion of the results, and Section 7 contains conclusions.

2 SPECTRUM ESTIMATION

Let a time sequence having a sample interval of one be given by $x_t,\ \ t = 0, \ldots N-1$. Assuming that $\{x_t\}$ is derived from a harmonizable random process (i.e. a process that can be represented as the superposition of random, infinitesimal harmonic oscillators) implies a spectral representation (Cramér 1940)

$$x_t = \int_{-1/2}^{1/2} e^{i2\pi\,\xi t}\,\mathrm{d}X(\xi)\,, \tag{1}$$

where $X(f)$ is an unobservable increments process whose statistical moments are of interest. If the process is weakly stationary, the increments are orthogonal, so that for distinct frequencies f_1 and f_2

$$\mathcal{E}[\mathrm{d}X(f_1)\,\mathrm{d}X^*(f_2)] = S(f_1)\,\delta(f_1 - f_2)\,\mathrm{d}f_1\mathrm{d}f_2, \tag{2}$$

where $\mathcal{E}$ denotes the expected value, $S(f)$ is the true or population power spectral density and $\delta(x)$ is the Dirac function.

The discrete Fourier transform of the time sequence is given by

$$y(f) = \sum_{t=0}^{N-1} x_t \mathrm{e}^{-i2\pi ft} \tag{3}$$

whose inverse is

$$x_t = \int_{-1/2}^{1/2} \mathrm{e}^{i2\pi ft}\,y(f)\,\mathrm{d}f. \tag{4}$$

Frequency f is a continuous rather than discrete variable because the Fourier transform is an entire function of frequency.

Convert (1) and (3) to a time-centred form by replacing t with $t-(N-1)/2$, and combine them to yield the fundamental equation of spectral analysis

$$y(f) = \int_{-1/2}^{1/2} \frac{\sin N\pi\,(f-\nu)}{\sin \pi\,(f-\nu)}\,\mathrm{d}X(\nu)\,. \tag{5}$$

The essence of multitaper spectral analysis is treatment of (5) as a Fredholm integral equation of the first kind, with $\mathrm{d}X(f)$ as the unknown function whose moments are to be estimated, and $y(f)$ as the data. It is well known that first kind integral equations do not have unique solutions, and thus the 'best' solution in some specified sense is sought. The criterion used in the multitaper estimator is minimization of the bias (i.e. spectral leakage) outside an interior domain $[-W,\ W)$, where W is a free parameter that specifies the resolution bandwidth $2W$ of a multitaper estimate. W is typically chosen to be a few times the Rayleigh resolution $1/N$. In other

Table 1. MDSS and OSS eigenvalues for LoD data.

Index	15 per cent MDSS eigenvalue	30 per cent MDSS eigenvalue	OSS eigenvalue
0	0.999999999999692	0.999999999983198	1.000000000000000
1	0.999999999975732	0.999999998473752	1.000000000000000
2	0.999999999437882	0.999999900175601	1.000000000000000
3	0.999999998932286	0.999999454849053	1.000000000000000
4	0.999999991499431	0.999997362604173	1.000000000000000
5	0.999999936867818	0.999986527805390	1.000000000000000
6	0.999999854130043	0.999956480699381	1.000000000000000
7	0.999999316879772	0.999905584824261	1.000000000000000
8	0.999993727855227	0.999365644651816	1.000000000000000
9	0.999989586069465	0.998939918841159	1.000000000000000
10	0.999964956826422	0.997782028780468	1.000000000000000
11	0.999901751191980	0.996739055257954	1.000000000000000
12	0.999778843706515	0.986104098451560	1.000000000000000
13	0.999219282505405	0.983280719474481	0.999999999999990
14	0.998449788311167	0.950649184136669	0.999999999999822

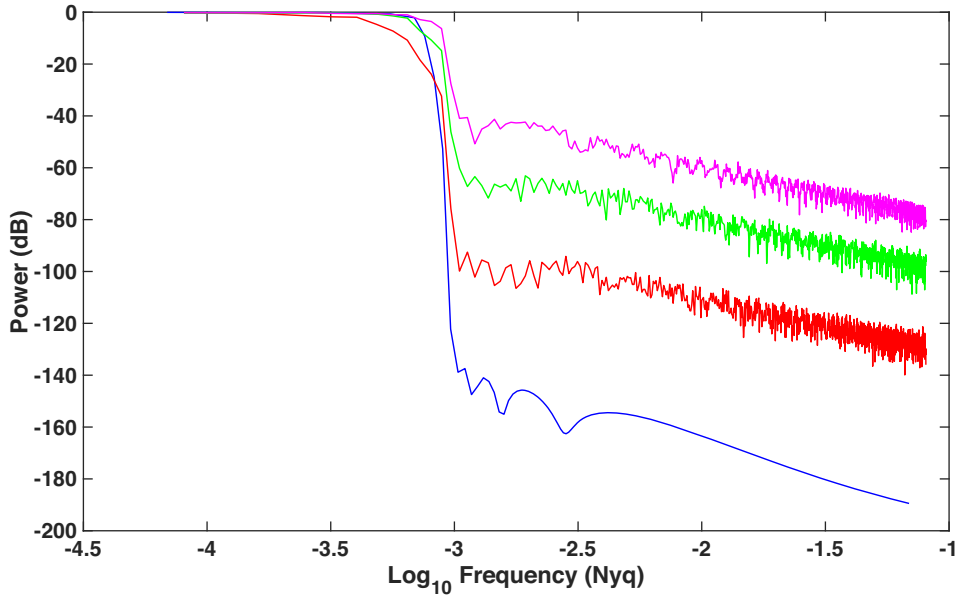

Figure 2. The spectral window, or power spectrum of the Slepian sequences, obtained from the average of the absolute square of the Fourier transforms of 15 ordinary Slepian sequences (blue) and 5 (red), 10 (green) and 15 (magenta) missing-data Slepian sequences as shown in Fig. 1. The half bandwidth is 0.00097 d^{-1} for all cases.

words, the multitaper estimator yields the integrated average of the power over a user-specified bandwidth.

The kernel function in (5) is the Dirichlet kernel whose eigenfunctions are the Slepian functions $U_k(N, W; f)$ that are orthonormal on the process domain $[-1/2,\ 1/2)$ and orthogonal on the interior domain $[-W,\ W)$ (Slepian 1978). The Dirichlet kernel eigenvalues $\{\lambda_k(N, W)\}$ give the fractional energy concentration in the interior domain.

The first $\lfloor 2NW \rfloor$ eigenvalues are near 1, and then fall off exponentially to zero. The Fourier transforms of the Slepian functions are the Slepian sequences $\mathrm{v}_t^k(N, W)$ that serve as data tapers in multitaper estimates, and are the solutions to

$$\sum_{s=0}^{N-1} \frac{\sin 2\pi W(t-s)}{\pi(t-s)} \mathrm{v}_s^k(N, W) = \lambda_k(N, W)\, \mathrm{v}_t^k(N, W) \tag{6}$$

for $t = 0, \ldots N-1$. As is apparent from (6), both the eigenvalues and eigenvectors are independent of frequency. The matrix form of (6) has a Toeplitz structure that makes its numerical solution fast and accurate, although Slepian (1978) presents an equivalent tridiagonal form that has even better numerical properties, and is typically used computationally.

The derivation of the multitaper power spectrum estimator using a Slepian function basis is described in Thomson (1982) and Percival & Walden (1993, §7). The form used in this work for the high resolution spectral estimator averaged over the interior domain is

$$\bar{S}(f) = \frac{\sum_{k=0}^{K-1} d_k^2(f)\, \hat{S}_k(f)}{\sum_{k=0}^{K-1} d_k^2(f)}, \tag{7}$$

where $K \leq \lfloor 2NW \rfloor$. The eigenspectra $\{\hat{S}_k(f)\}$ are direct estimates using a Slepian sequence as the data taper, and are obtained by taking the absolute square of

$$\hat{a}_k(f) = \sum_{n=0}^{N-1} \mathrm{v}_n^k(N, W)\, x_n \mathrm{e}^{-i2\pi f\left(n-\frac{N-1}{2}\right)}. \tag{8}$$

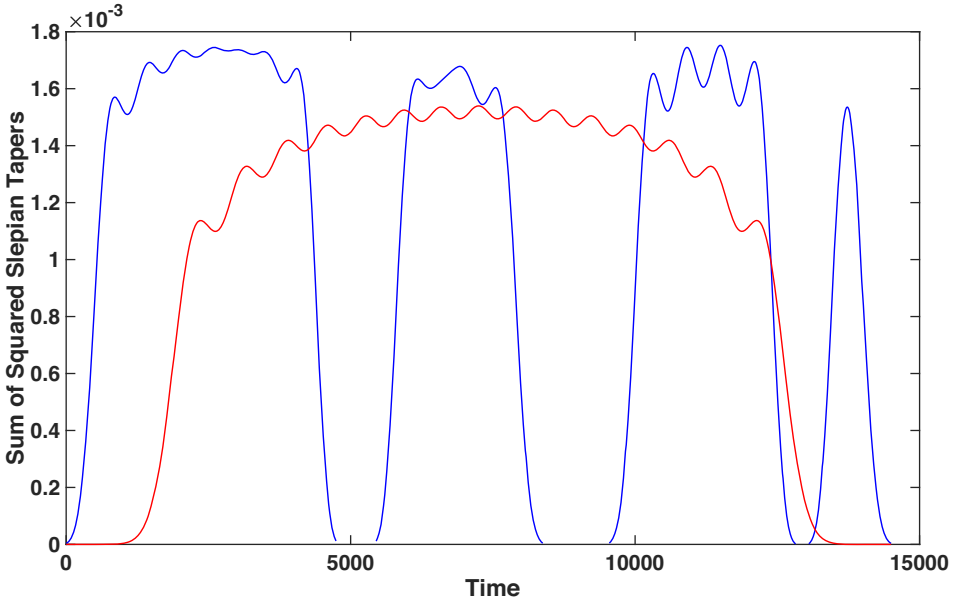

Figure 3. The energy concentration given by the sum of the squares of the 15 missing data Slepian sequences (blue) from Fig. 1 and 15 ordinary Slepian sequences (red) for a half-bandwidth of 0.00097 d^{-1}.

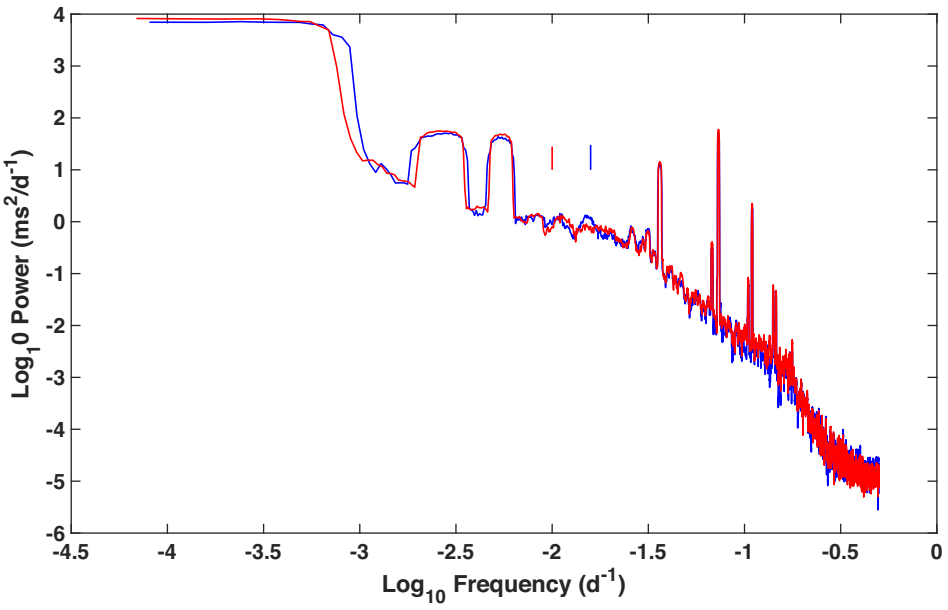

Figure 4. The base 10 logarithm of power spectral density for the length of day data in $ms^2\ d^{-1}$ plotted against the base 10 logarithm of frequency in d^{-1} using an adaptively weighted standard multitaper estimator (red) having about 30 degrees of freedom per frequency and an adaptively weighted 15 per cent missing-data multitaper estimator (blue) having about 26 degrees of freedom per frequency. The resolution bandwidth for both estimates is 0.0019 d^{-1}. The vertical red and blue lines at the centre of the figure are double-sided 95 per cent confidence limits on the spectra.

The data-adaptive weights are given by

$$d_k(f) = \frac{\lambda_k S(f)}{\lambda_k S(f) + \sigma^2 (1 - \lambda_k)}, \tag{9}$$

where σ^2 is the population variance and the second term in the denominator is an upper bound on the bias outside of the interior domain, or broad-band bias. The adaptive weights (9) are nearly unity at frequencies where the broad-band bias is small, and become small when the broad-band bias is dominant. When the broad-band bias is negligible, (7) is the arithmetic average of the eigenspectra. A spectrum estimate is obtained by substituting (9) into (7), replacing the population entities $S(f)$ and σ^2 with estimates, assuming that $d_k(f)$ is constant across the interior domain and iteratively solving the resulting non-linear equation. Once the weights are determined, the degrees of freedom is given by

$$\eta(f) = 2 \sum_{k=0}^{K-1} d_k^2(f). \tag{10}$$

In practice, better performance is achieved after prewhitening, typically using a short autoregressive filter.

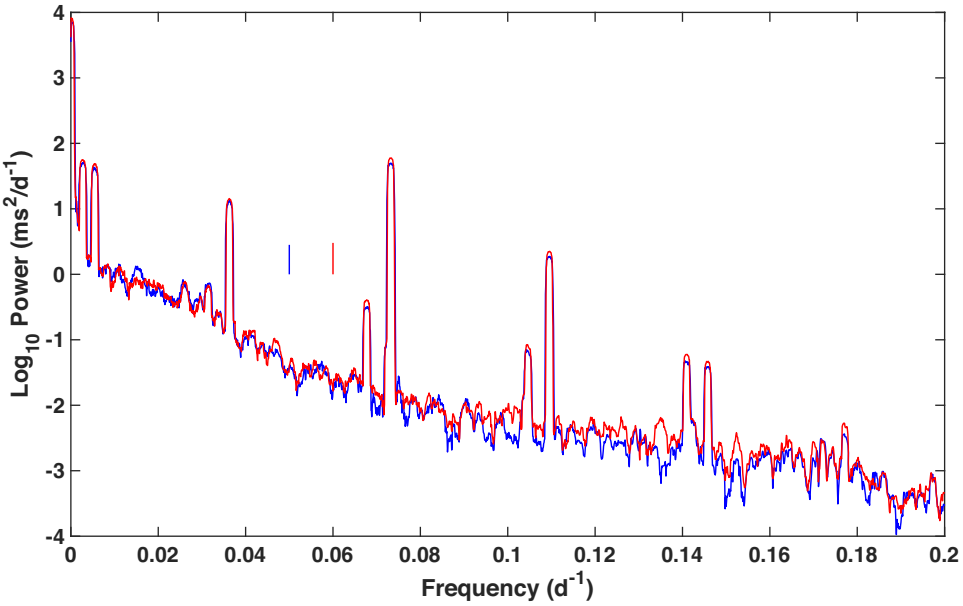

Figure 5. The base 10 logarithm of power spectral density for the length of day data in $ms^2\ d^{-1}$ plotted against frequency over the band 0.0–0.2 d^{-1} using an adaptively-weighted standard multitaper estimator (red) having about 30 degrees of freedom per frequency and an adaptively-weighted 15 per cent missing-data multitaper estimator (blue) having about 26 degrees of freedom per frequency. The resolution bandwidth for both estimates is 0.0019 d^{-1}. The vertical red and blue lines at the centre of the figure are double-sided 95 per cent confidence limits on the spectra.

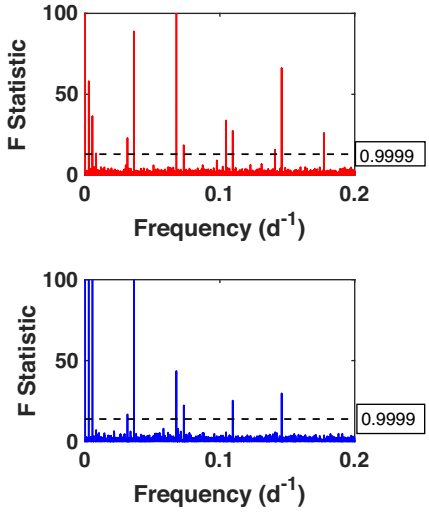

Figure 6. The multitaper *F*-test (14) for the complete length of day time-series (top panel) and for the 15 per cent missing data length of day time-series using the sampling scheme of Figs 1–3 (bottom panel). The *F* statistic has been truncated at 100 for clarity. The horizontal dashed lines show the 0.9999 significance level for the *F* statistic with 2,28 (top) and 2,24 (bottom) degrees of freedom.

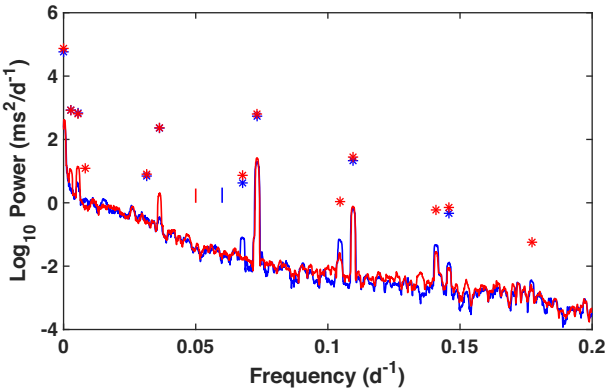

Figure 7. The reshaped power spectrum for the entire length of day record (red) and for the 15 per cent missing data example (blue). The power removed from the stochastic spectrum is shown by the red and blue asterisks assuming the corresponding bandwidth is the Rayleigh resolution. The vertical red and blue lines at the centre of the figure are doublesided 95 per cent confidence limits on the spectra assuming 30 (red) and 26 (blue) degrees of freedom per frequency.

The presentation to this point has focused on the second moment of the estimate of the orthogonal increment process $\mathrm{d}X(f)$. A significant advantage of the multitaper method is that it provides both a method to estimate the complex amplitude of deterministic or line components, and a statistical test for their presence. The Slepian functions define the shape that a harmonic component will have in a multitaper power estimate, and because of their properties, line components will appear smeared out over a band of width $2W$ that has a nearly rectangular shape. The expected value of (8) is

$$\mathcal{E}[\hat{a}_k(f)] = \sum_{i=1}^{L} \hat{\mu}_i(f)\, U_k(N, W; f - f_i), \tag{11}$$

where there are L line components and $\hat{\mu}_i(f)$ is a complex line amplitude. The least squares estimator for the line amplitude is

$$\hat{\mu}(f) = \frac{\sum_{k=0}^{K-1} U_k(N, W; 0)\, \hat{a}_k(f)}{\sum_{k=0}^{K-1} U_k^2(N, W; 0)}. \tag{12}$$

The power in the line at a given frequency is $|\hat{\mu}(f)|^2$ and has 2 degrees of freedom. The residual power spectral estimator with the

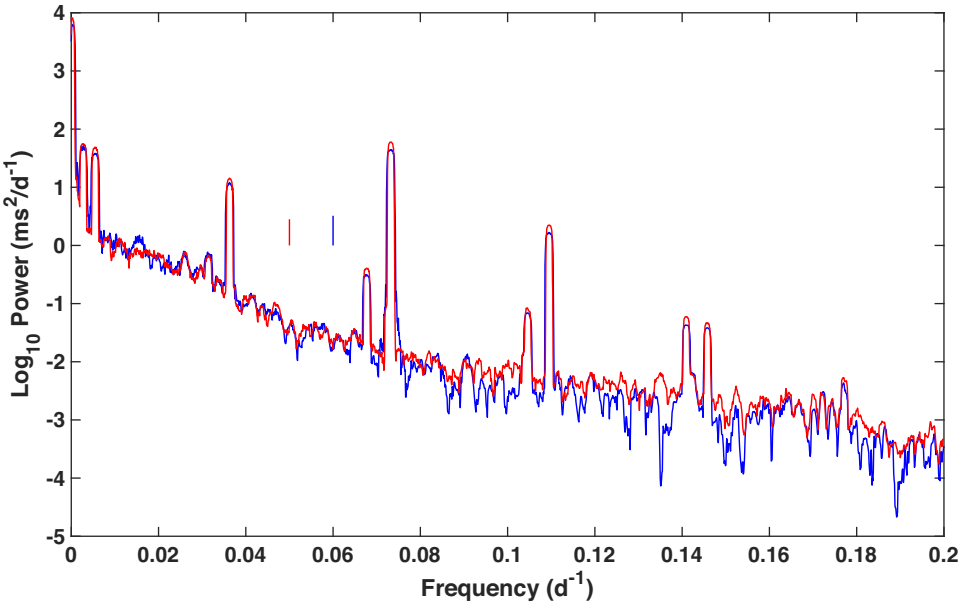

Figure 8. The base 10 logarithm of power spectral density for the length of day data in $ms^2\ d^{-1}$ plotted against frequency over the band 0.0–0.2 d^{-1} using an adaptively-weighted standard multitaper estimator (red) having about 30 degrees of freedom per frequency and an adaptively-weighted 30 per cent missing-data multitaper estimator (blue) having about 23 degrees of freedom per frequency. The resolution bandwidth for both estimates is 0.0019 d^{-1}. The vertical red and blue lines at the centre of the figure are double-sided 95 per cent confidence limits on the spectra assuming 30 (red) and 23 (blue) degrees of freedom per frequency.

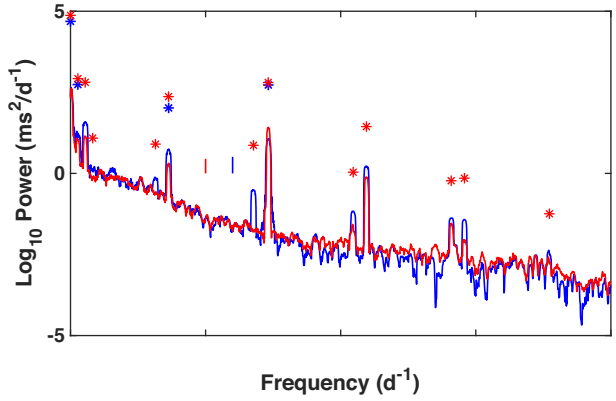

Figure 9. The base 10 logarithm of power spectral density for the length of day data in $ms^2\ d^{-1}$ plotted against frequency over the band 0.0–0.2 d^{-1} using an adaptively-weighted standard multitaper estimator (red) having about 30 degrees of freedom per frequency and an adaptively-weighted 30 per cent missing-data multitaper estimator (blue) having about 23 degrees of freedom per frequency. The resolution bandwidth for both estimates is 0.0019 d^{-1}. The vertical red and blue lines at the centre of the figure are double-sided 95 per cent confidence limits on the spectra assuming 30 (red) and 23 (blue) degrees of freedom per frequency.

line removed is

$$\bar{S}'(f) = \sum_{k=0}^{K-1} |\hat{a}_k(f) - \hat{\mu}(f)\, U_k(N, W; 0)|^2 \tag{13}$$

and has $\eta(f) - 2$ degrees of freedom. The process of removing a line component from the spectrum is called reshaping, and (13) is also called the reshaped spectrum. Thomson (1982, §XIII) defines an F statistic

$$\hat{F}(f) = \frac{[\eta(f) - 2]\,|\hat{\mu}(f)|^2 \sum_{k=0}^{K-1} U_k^2(N, W; 0)}{2\,\bar{S}'(f)} \tag{14}$$

that is asymptotically distributed as $F_{2,\,\eta(f)-2}$ to test for the presence of a line component, and can be assessed in the standard way. For example, one might choose to reshape the spectrum at the 0.999 probability level by assessing (14) against the critical values for a $F_{2,\,\eta(f)-2}$ distribution, computing (13) at those frequencies where they are exceeded, and then replacing the power in the line as $|\hat{\mu}(f)|^2$ under the assumption that it has the Rayleigh resolution bandwidth $1/N$.

Bronez (1985, 1988) extended multitaper analysis to irregularly spaced data in one or more dimensions. He suggested that the quantity of interest in spectral analysis is the integrated spectrum over a user-specified bandwidth; this is effectively the same as in (7) where the analysis band is the interior domain about any frequency of interest. Bronez proposed a quadratic spectral estimator

$$\bar{S}_W = \mathbf{x}^* \mathbf{Q}_W \mathbf{x}, \tag{15}$$

where $\mathbf{Q}_W$ is an $N \times N$ positive semidefinite Hermitian matrix that depends on the analysis half bandwidth W. It may be decomposed as

$$\mathbf{Q}_W = \mathbf{w}_W \mathbf{w}_W^*, \tag{16}$$

where $\mathbf{w}_W$ is $N \times K$ with rank K. The weights in (16) are chosen using a minimum bias criterion. Considering only the 1-D problem for simplicity, define the process half bandwidth β which is not necessarily $\frac{1}{2}$. For the parameters W and β, compute the eigenvalues λ_k and eigenvectors $\mathbf{w}^k$ at a given frequency for the generalized eigenvalue problem

$$\mathbf{A}w^k = \lambda_k \mathbf{B} w^k \tag{17}$$

where

$$A_{nm} = \int_{f-W}^{f+W} e^{i2\pi f(t_n - t_m)} df \tag{18}$$

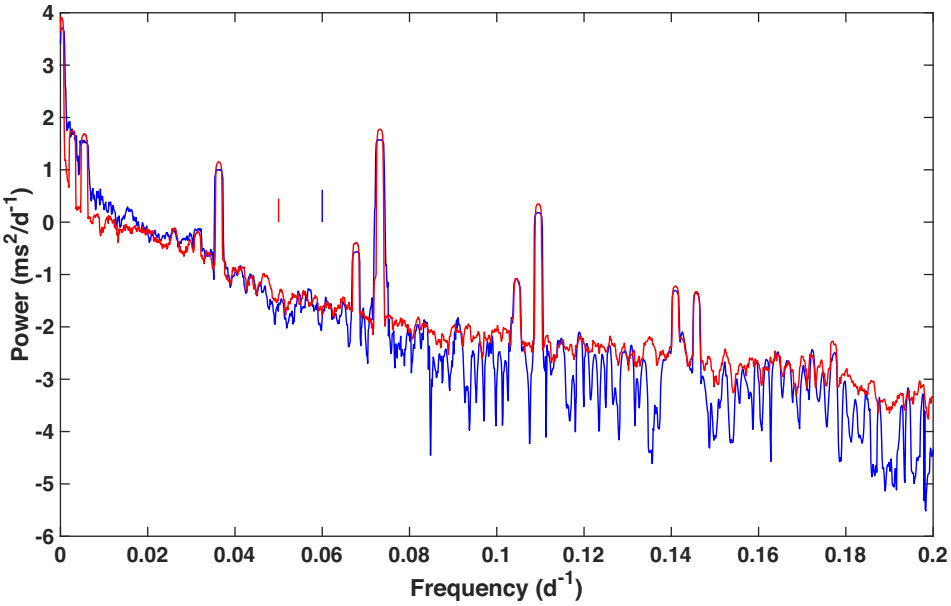

Figure 10. The base 10 logarithm of power spectral density for the length of day data in ms^2 d^{-1} plotted against frequency over the band 0.0–0.2 d^{-1} using an adaptively-weighted standard multitaper estimator (red) having about 30 degrees of freedom per frequency and an adaptively-weighted 45 per cent missing-data multitaper estimator (blue) having about 16 degrees of freedom. The resolution bandwidth for both estimates is 0.0019 d^{-1}. The vertical red and blue lines at the centre of the figure are double-sided 95 per cent confidence limits on the spectra assuming 30 (red) and 16 (blue) degrees of freedom per frequency.

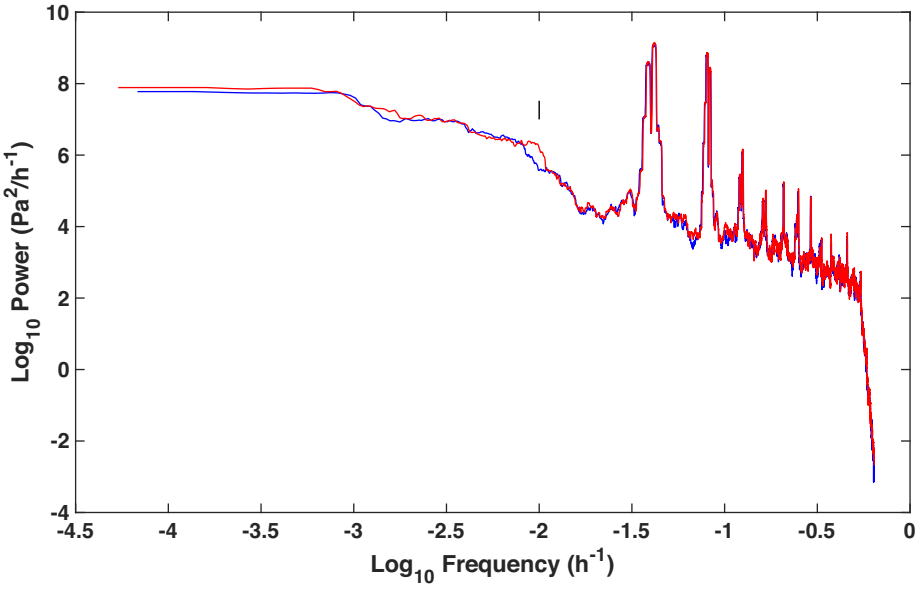

Figure 11. The base 10 logarithm of seafloor pressure power spectral density in Pa^2 hr^{-1} plotted against the base 10 logarithm of frequency in hr^{-1} using an adaptively weighted standard multitaper estimator (red) having a time-bandwidth of 13 and an adaptively weighted missing-data multitaper estimator (blue) having a time-bandwidth of 10 for 2580 missing values in the centre of the time-series. The resolution bandwidth for both estimates is 0.0014 hr^{-1}. The vertical black line at the centre of the figure is a double-sided 95 per cent confidence limit on the spectra assuming 24 degrees of freedom per frequency.

and

$$B_{nm} = \int_{-\beta}^{\beta} e^{i2\pi f(t_n - t_m)} df \tag{19}$$

and w^k is a column of $\mathbf{w}_W$. The time vector is not necessarily equally spaced. The integrated spectrum is then obtained as

$$\tilde{S}(W, \beta; f) = \frac{1}{K} \sum_{k=0}^{K-1} \left| \sum_{i=0}^{N-1} w_i^k x_i \right|^2 \tag{20}$$

Note the analogy to (7): in the absence of broadband bias from spectral leakage, the weights (9) become unity, and (7) is just the arithmetic average of the raw estimates using the *k*th Slepian sequence as a data taper. For uniformly sampled data, the eigenvectors from (17) used in (20) are the point-by-point product of an ordinary Slepian sequence with a complex exponential term in the Fourier transform at a given frequency. More generally, they are frequency-dependent eigenfunctions that have a similar character, and the eigenvalues may also be frequency-dependent. This creates numerical issues, particularly as the frequency approaches extremes within the process band.

Table 2. MDSS Eigenvalues for pressure data.

Index	1 Gap MDSS eigenvalue	5 Gap MDSS eigenvalue	12 Gap MDSS eigenvalue
0	0.999999999999946	0.999985910962565	0.986074561310057
1	0.999999999999850	0.999937210577215	0.979481224911075
2	0.999999999986706	0.999929729565371	0.979481224911060
3	0.999999999983828	0.999929589064199	0.979481224910006
4	0.999999999181086	0.999887583322106	0.979481224863396
5	0.999999998426710	0.999866809438120	0.979481223360637
6	0.999999971435184	0.999207423450190	0.979481186467767
7	0.999999907433297	0.996240475367617	0.979480477405569
8	0.999999146195923	0.991757457909191	0.979468852020767
9	0.999997248750588	0.991622177978471	0.979310358130167
10	0.999975945028074	0.990914514532513	0.977293857744475
11	0.999950940641856	0.990453321407156	0.959517073933613

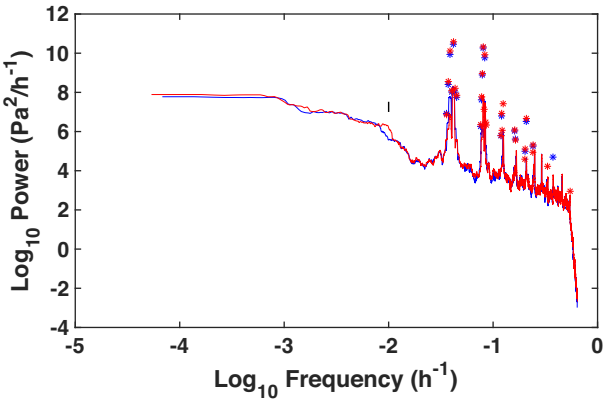

Figure 12. The 0.9999 level reshaped base 10 logarithm of seafloor pressure power spectral density in Pa^2/hr^{-1} plotted against the base 10 logarithm of frequency in hr^{-1} using an adaptively weighted standard multitaper estimator (red) having a time-bandwidth of 13 and an adaptively weighted missing-data multitaper estimator (blue) having a time-bandwidth of 10 for 2580 missing values in the centre of the time-series. The resolution bandwidth for both estimates is $0.0014\,hr^{-1}$. The vertical black line at the centre of the figure is a double-sided 95 per cent confidence limit on the spectra assuming 24 degrees of freedom per frequency. The extracted line energy is represented by the red and blue asterisks under the assumption that their bandwidths are the Rayleigh resolution of the estimates.

This problem can be minimized by writing $w^k = v^k \exp[-i2\pi f(t_n - t_m)]$, yielding the new eigenvalue problem

$$\mathbf{A}' v^k = \lambda_k \mathbf{B}' v^k \tag{21}$$

where

$$A'_{nm} = \frac{\sin 2\pi W (t_n - t_m)}{\pi (t_n - t_m)} \tag{22}$$

and

$$B'_{nm} = e^{-i2\pi f(t_n - t_m)} \frac{\sin 2\pi\beta (t_n - t_m)}{\pi (t_n - t_m)} \tag{23}$$

General solutions to (21) are obtained by computing the Cholesky factors $\mathbf{B}' = \mathbf{R}\mathbf{R}^T$, and transforming (21) into a standard eigenvalue problem

$$\mathbf{R}^{-1}\mathbf{A}'\left(\mathbf{R}^T\right)^{-1}\mathbf{R}^T v^k = \lambda_k \mathbf{R}^T v^k \tag{24}$$

where the $\mathbf{R}^T v^k$ are orthonormal. It is straightforward to show that (23) reduces to the identity matrix when $\beta = 1/2$ and $t_n - t_m = k$, where k is any integer, and so the Cholesky factors in (24) are also identity matrices and hence become irrelevant. Consequently, the eigenvalue problem (24) becomes frequency independent for any regularly sampled time sequence that contains arbitrary gaps, so that the transformed tapers v^k need only be computed once for a given sampling scheme. However, the Toeplitz structure of the regularly sampled version of (22) is lost in the presence of time-series gaps.

The power spectrum analogous to (20) is

$$\tilde{S}(f) = \frac{1}{K}\sum_{k=0}^{K-1}\left|\sum_{i=0}^{N-1} v_i^k x_i e^{-i2\pi f t_i}\right|^2 \tag{25}$$

and is similar to (7) when the broadband bias is negligible. The frequency-independent tapers v^k are missing-data Slepian sequences (MDSSs) for the time variable t_i and time sequencex_i that contain gaps, and reduce to the OSSs that are the solutions to (6) when gaps are absent. It follows directly that the same principles used to obtain the gap-free adaptively weighted spectrum (7), the F-test (14), estimates for line components (12) and the reshaped power spectrum estimator (13) pertain equally to time sequences containing gaps.

SC12 modified (6) to account for time-series gaps by adding a multiplicative indicator function on both sides of the eigenvalue problem that takes on a value of 0 when data are absent and 1 when they are present; these will be called indicator function Slepian sequences (IFSSs) in the sequel. In other words, they assumed that a missing-data Slepian sequence is an OSS that is forced to be zero when data are absent. However, the MDSSs from (21) are only defined where data are present in the time sequence, hence are distinct from IFSSs, as is further demonstrated below.

3 LENGTH OF DAY DATA

In this section, the characteristics of the MDSSs and IFSSs are compared using the example presented in SC12 §3. The sampling scheme comprises 14 500 points with a sample interval of unity, but with gaps at 4745–5447, 8378–9545 and 12 823–13 051, or ∼15 per cent missing data. Fig. 1 shows the first 15 MDSSs using a time-bandwidth of 12 with 12 400 data, or a half bandwidth W of $0.00097\ d^{-1}$. This is the same as for the IFSS result in Fig. 1 of SC12. It is immediately apparent that the MDSSs and IFSSs bear little resemblance to each other. The IFSSs in SC12 are spread more widely across the sections where data are present, and all descend to zero at the data gap boundaries. By contrast, the MDSSs in Fig. 1 are more concentrated in the longest data sections for small order k, with the shortest data section showing up only for higher values of the order. For example, the two lowest order MDSSs resemble the two lowest order OSSs over only the first data section, with small

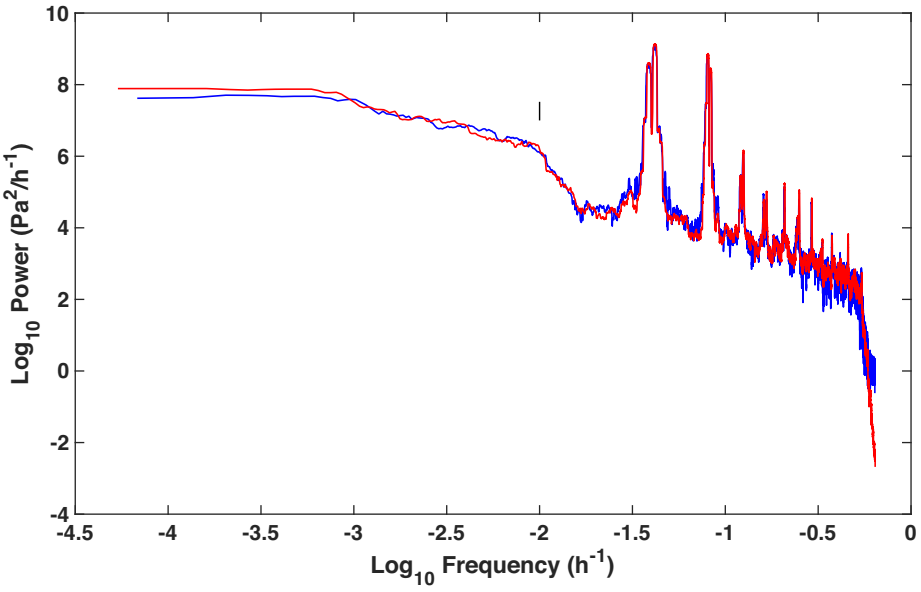

Figure 13. The base 10 logarithm of seafloor pressure power spectral density in $Pa^2\ hr^{-1}$ plotted against the base 10 logarithm of frequency in hr^{-1} using an adaptively weighted standard multitaper estimator (red) having a time-bandwidth of 13 and an adaptively weighted missing-data multitaper estimator (blue) having a time-bandwidth of 10 for 2580 missing values distributed into five intervals spaced evenly along the time-series. The resolution bandwidth for both estimates is 0.0014 hr^{-1}. The vertical black line at the centre of the figure is a double-sided 95 per cent confidence limit on the spectra.

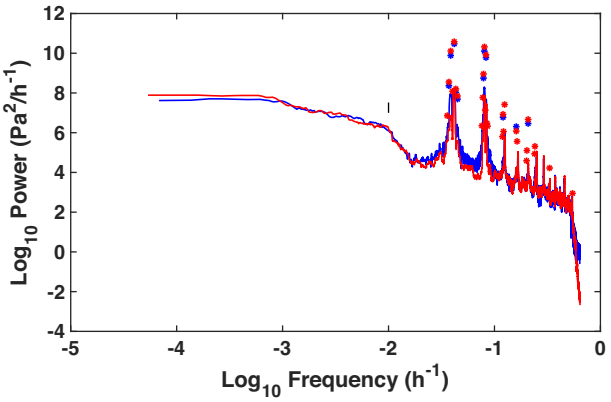

Figure 14. The base 10 logarithm of the reshaped seafloor pressure power spectral density in $Pa^2\ hr^{-1}$ plotted against the base 10 logarithm of frequency in hr^{-1} using an adaptively weighted standard multitaper estimator (red) having a time-bandwidth of 13 and an adaptively weighted missing-data multitaper estimator (blue) having a time-bandwidth of 10 for 2580 missing values distributed into five intervals spaced evenly along the time-series. The resolution bandwidth for both estimates is 0.0014 hr^{-1}. The vertical black line at the centre of the figure is a double-sided 95 per cent confidence limit on the spectra assuming 24 degrees of freedom per frequency. The extracted line energy is represented by the red and blue asterisks under the assumption that their bandwidths are the Rayleigh resolution of the estimates.

but nonzero amplitudes over the remaining three data sections. In addition, the MDSSs are not constrained to descend to zero at data gap boundaries, as is most apparent for order values above 10. Table 1 compares the MDSS eigenvalues with the OSS eigenvalues for a time-bandwidth of 14 with 14 500 data, which yields the same resolution bandwidth as for the MDSS example. There is obviously a penalty to be paid for gaps in a time-series, as reflected in the smaller MDSS eigenvalues. SC12 specify that the first ten IFSS eigenvalues are nearly one, which is also true for the MDSSs in Table 1.

The observed differences between the IFSSs and MDSSs provide little insight into their relative performance. Fig. 2 compares the average spectral windows for the first 15 OSS tapers with a time-bandwidth of 14 and a time-series length of 14 500 and the first 5, 10 and 15 MDSS tapers with a time-bandwidth of 12 and a time-series length of 12 400. The OSS spectral window has at least 140 dB of sidelobe protection. By contrast, the MDSSs provide at least 100, 70 and 40 dB of sidelobe protection for the average of 5, 10 and 15 tapers, which is substantially poorer. However, the MDSS spectral windows display the square main lobe and slow falloff with frequency characteristic of Slepian functions. Comparing to Fig. 3 in SC12 shows sidelobe protection of about 50, 50 and 30 dB for the average of 5, 10 and 15 IFSSs, and a main lobe that is not as square as for the MDSSs. Consequently, the MDSSs offer a substantial performance advantage compared to the IFSSs for comparable bandwidths, without accounting for adaptive-weighting that was not utilized in SC12.

Fig. 3 compares the energy concentration $\sum_{k=0}^{K-1}(v_n^k)^2$ for the OSSs and $\sum_{k=0}^{K-1}(\upsilon_n^k)^2$ for the MDSSs at the same resolution bandwidth with $K = 15$. This result is very similar to Fig. 2 in SC12. It is notable that the MDSSs pick up more energy at long data sections as compared to short sections, which is also reflected in the absence of individual MDSS amplitudes in the shortest section except for large order values in Fig. 1. The absence of sensitivity of the OSSs at the ends of the data sequence reflects the fact that only 15 out of 28 useful data tapers were used.

SC12 utilized the first 14 500 values of changes in the length of day obtained from the International Earth Orientation Reference System as an exemplar. These data extend from 01.01.1962 to 12.09.2001 (and not through 2009 as SC12 state), and data were removed corresponding to the sampling scheme of Figs 1–3. Fig. 4 compares the base 10 log frequency-log power utilizing adaptive-weighting for the complete time-series using OSSs and the missing data time-series using MDSSs with identical resolution bandwidths. Fig. 5 shows the 0.0–0.2 d^{-1} interval as linear frequency versus log

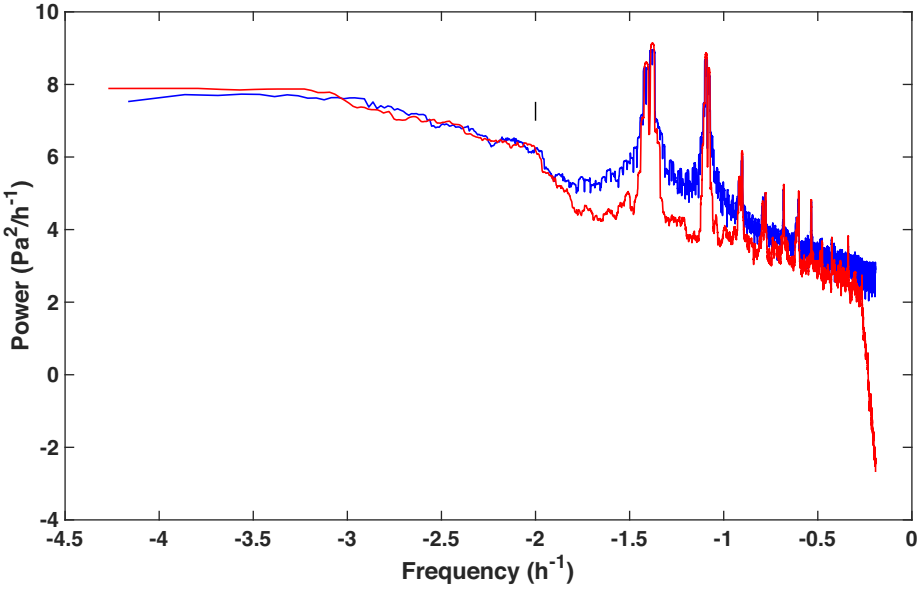

Figure 15. The base 10 logarithm of seafloor pressure power spectral density in Pa^2 hr^{-1} plotted against the base 10 logarithm of frequency in hr^{-1} using an adaptively weighted standard multitaper estimator (red) having a time-bandwidth of 13 and an adaptively weighted missing-data multitaper estimator (blue) having a time-bandwidth of 10 for 2580 missing values distributed into twelve intervals spaced evenly along the time-series. The resolution bandwidth for both estimates is 0.0014 hr^{-1}. The vertical black line at the centre of the figure is a double-sided 95 per cent confidence limit on the spectra.

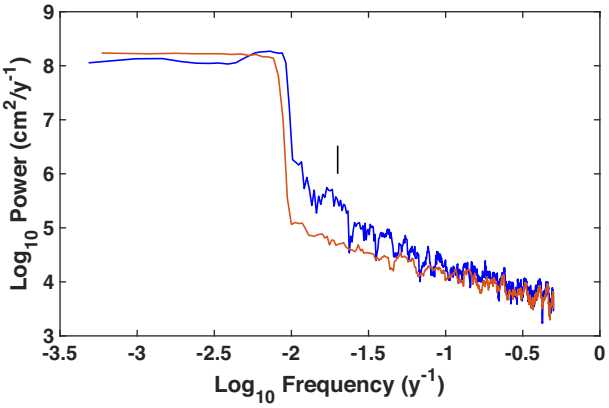

Figure 16. The base 10 logarithm of the power in cm^2 yr^{-1} against the base 10 logarithm of frequency in yr^{-1} for the entire Cairo nilometer data set using a missing-data multitaper estimator with a time bandwidth of 10 and a regular multitaper estimator using the data from 622 to 1470 CE with a time bandwidth of 8, in both cases utilizing 12 missing-data or ordinary Slepian sequences. The resolution bandwidth for both estimates is 0.0097 yr^{-1}. The vertical black line shows the double sided 95 per cent confidence interval on the estimate assuming 24 degrees of freedom per frequency.

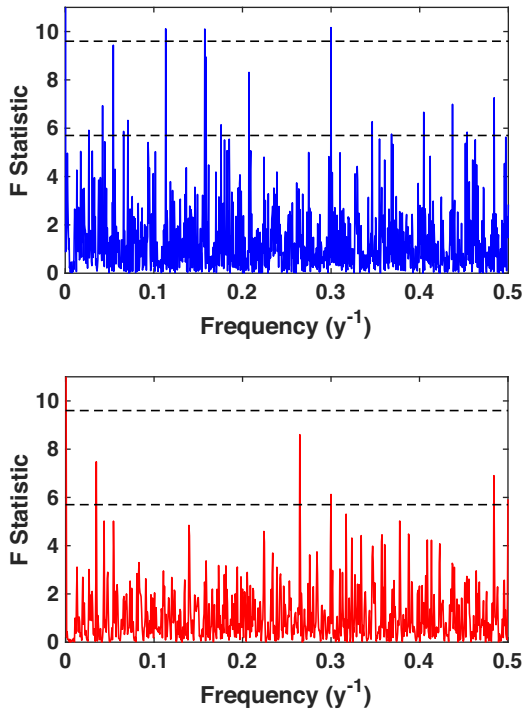

Figure 17. The missing-data multitaper F test for the complete nilometer data set (top) and the regular multitaper F test for the 622–1470 CE nilometer data (bottom) using the same parameters as in Fig. 16. The horizontal black dashed lines show the 0.99 and 0.999 probability thresholds for an $F_{2,22}$ distribution.

power for both estimates. The double-sided 95 per cent confidence intervals based on a scaled chi square distribution are shown at the centre of each figure. The missing-data spectrum is statistically indistinguishable from that obtained over the complete record in both figures. The peaks seen in the figures are due to the long period tides. Fig. 5 is similar to the inset in SC12 Fig. 9.

Fig. 6 shows the F statistic (14) over 0.0–0.2 d^{-1} for both the standard and missing-data multitaper methods, along with the 0.9999 probability threshold based on the F distribution. Both F-tests easily detect the major long period tides, but performance of the missing-data F-test is somewhat weaker than for the complete data set. For example, the Sta tidal species at ∼120 d (0.0083 d^{-1}) is barely significant in the complete data F-test at the 0.9999 level, and absent in the missing-data F-test. Since the multitaper F-test is sensitive to the signal-to-noise ratio within the interior domain about a frequency of interest, it is not surprising that gaps in the time-series will change it.

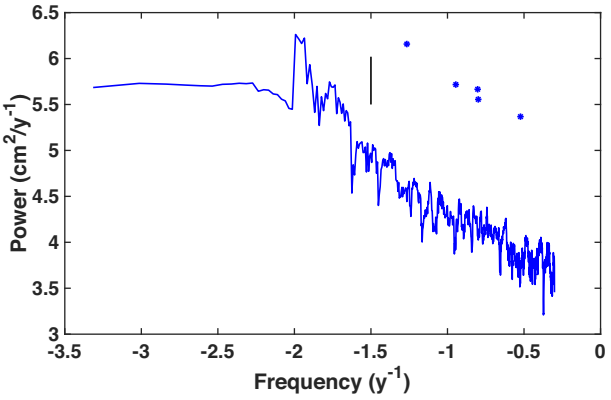

Figure 18. The missing-data multitaper spectrum for the nilometer data reshaped at the 0.998 level, along with the resulting line components inserted under the assumption that their bandwidths are the Rayleigh resolution. The vertical black bar shows a double-sided 95 per cent confidence interval on the spectrum assuming 24 degrees of freedom per frequency.

Fig. 7 contains the reshaped power spectrum estimate (13) after removing line components that are significant at the 0.9999 probability level, along with the power in the lines under the assumption that each has the Rayleigh resolution bandwidth. As for Fig. 5, the reshaped power spectra are statistically indistinguishable except at a few frequencies where the missing-data F test failed to detect a line component, such as ∼0.038, ∼0.0105 and ∼0.14 d^{-1}. The fortnightly tide at 0.073 d^{-1} is actually two lines spaced more closely than the Rayleigh resolution that the single frequency F-test (14) cannot resolve for either the complete or missing-data multitaper approach.

Fig. 8 compares the multitaper spectrum of the complete and a 30 per cent missing data example obtained by doubling the length of the gaps in the previous examples. Table 1 shows the eigenvalues for the 30 per cent missing-data MDSSs, which are smaller than for the 15 per cent missing data example. The spectra are statistically indistinguishable where the power is high, but there is the suggestion of downward bias in the missing-data spectrum above ∼0.1 d^{-1}.

Fig. 9 shows the reshaped power spectrum estimate (13) for the 30 per cent missing data example after removing line components that are significant at the 0.9999 probability level, along with the power in the line under the assumption that each has the Rayleigh resolution bandwidth. Comparing to Fig. 7, the missing-data F-test is significant at $p = 0.9999$ only below ∼0.08 d^{-1} such that no reshaping occurs above that point. This could be obviated by reducing the probability threshold.

Fig. 10 compares the multitaper spectrum of the complete and a 45 per cent missing data example obtained by tripling the length of the gaps in the initial case. The missing-data spectrum is statistically indistinguishable from the complete data one below ∼0.08 d^{-1}, but is increasingly downward biased relative to the complete data spectrum above that point.

4 SEAFLOOR PRESSURE DATA

Seafloor pressure data were collected as a component of the Hawaii Ocean Mixing Experiment (HOME) in 2001–2; see Pinkel *et al.* (2000) for details about HOME. The time-series in this study is from site PN2 located at 26°52.5′N, 161°56.7′W, north–northwest of the island of Kauai at 5235 m water depth. The data were sampled at 64/hr, and the record length is 389 d beginning in late April 2001. The data were decimated by a factor of 100, yielding a sample interval of 2812.5 s, or about 0.78 hr. These data have a high dynamic range due to the presence of the diurnal and semidiurnal tides, hence bias control during their spectral analysis is important.

A section of data 2580 long was removed from the middle of the pressure time-series comprising 21.6 per cent of the record. Fig. 11 compares the multitaper spectrum of the entire pressure time-series with a time-bandwidth of 13 against the missing-data multitaper spectrum with a time-bandwidth of 10, in both cases using 12 tapers and no prewhitening. The eigenvalues for the OSSs are all one, while Table 2 gives the smaller eigenvalues for the MDSSs in the second column, all of which are at least 0.9999. The two spectra in Fig. 11 are statistically identical, being dominated by the diurnal and semidiurnal tides, along with the overtides. Both of the estimates easily resolve the high frequency roll off caused by low pass filtering prior to decimation. Fig. 12 compares the multitaper spectra after reshaping at the 0.9999 probability level, with the line energy shown as asterisks replaced under the assumption that their bandwidth is the Rayleigh resolution of the estimate. The two spectra are very similar, although there is a tendency for the missing-data line amplitudes to be slightly smaller than those for the complete time-series.

The same number of missing data was changed into five gaps 516 values long evenly distributed along the time-series. Fig. 13 compares the standard multitaper estimate for the whole time-series with the missing-data estimate using the same parameters as for Fig. 11. The eigenvalues are given in the middle column of Table 2, and are substantially smaller than for a single gap. The spectra are statistically identical, although there is a suggestion for upward bias in the missing-data spectrum between the diurnal and semidiurnal bands and in the high frequency roll off. The degrees of freedom are reduced for the missing-data spectrum at high frequencies, as seen in the increased variability above ∼0.25 hr^{-1}. Fig. 14 compares the complete and missing-data spectrum after reshaping at the 0.9999 probability level, along with the line energy inserted with the Rayleigh resolution as its bandwidth. There is more evidence for bias between the diurnal and semidiurnal tidal bands as compared to Fig. 13, and underestimation of the line power by the missing-data estimator is enhanced compared to Fig. 12. In addition, the missing-data estimator does not detect line components at periods shorter than 4 hr due to the presence of gaps, but the standard multitaper estimator does in their absence.

The missing data were then distributed into 12 gaps 215 points long distributed evenly along the time-series; this corresponds to a week of data per 30-day month being absent. Fig. 15 compares the complete time-series multitaper estimate with the missing-data one using the same parameters as for Fig. 11. The eigenvalues for the MDSSs are given in the last column of Table 2, and are decidedly smaller than for the other examples, suggesting that bias problems may ensue even in the presence of adaptive weighting. The missing-data spectrum confirms this, as the spectrum is upward biased at all frequencies save those of the spectral peaks. The degrees of freedom are high (∼24) only at very low frequencies and over the diurnal and semidiurnal tide bands, dropping to ∼10 in their vicinity and to nearly zero above 0.1 hr^{-1}. Reshaping fails for this spectrum due to pervasive spectral leakage.

5 CAIRO NILOMETER DATA

A final example will utilize the arguably longest instrumental time-series in existence from the Cairo nilometer which measures the

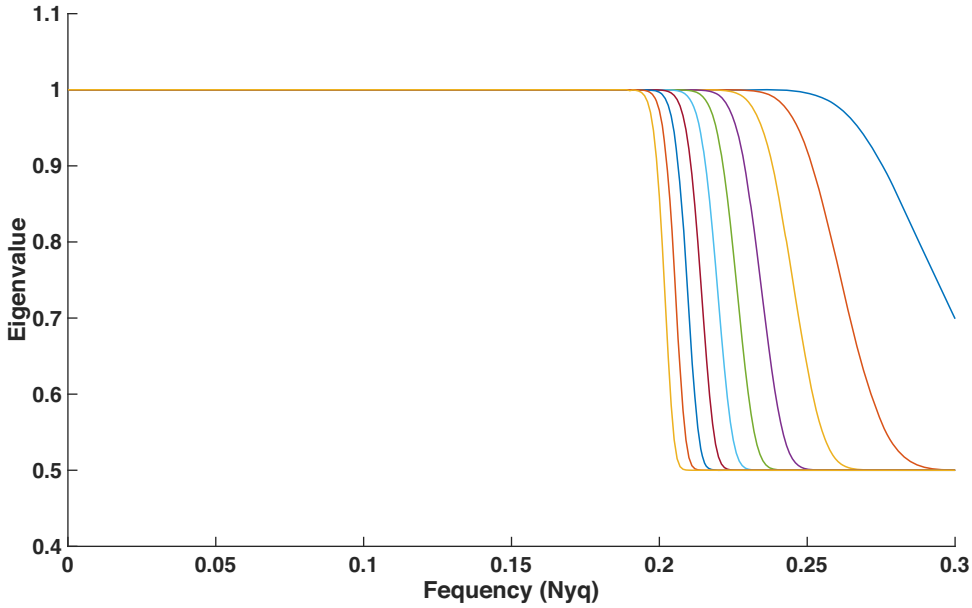

Figure 19. The first ten irregular sampling Slepian sequence eigenvalues for time sampling $t_i = i^{1.05}$, $i = 1, 1000$ and a resolution half bandwidth of 0.008. The eigenvalue index increases from right to left.

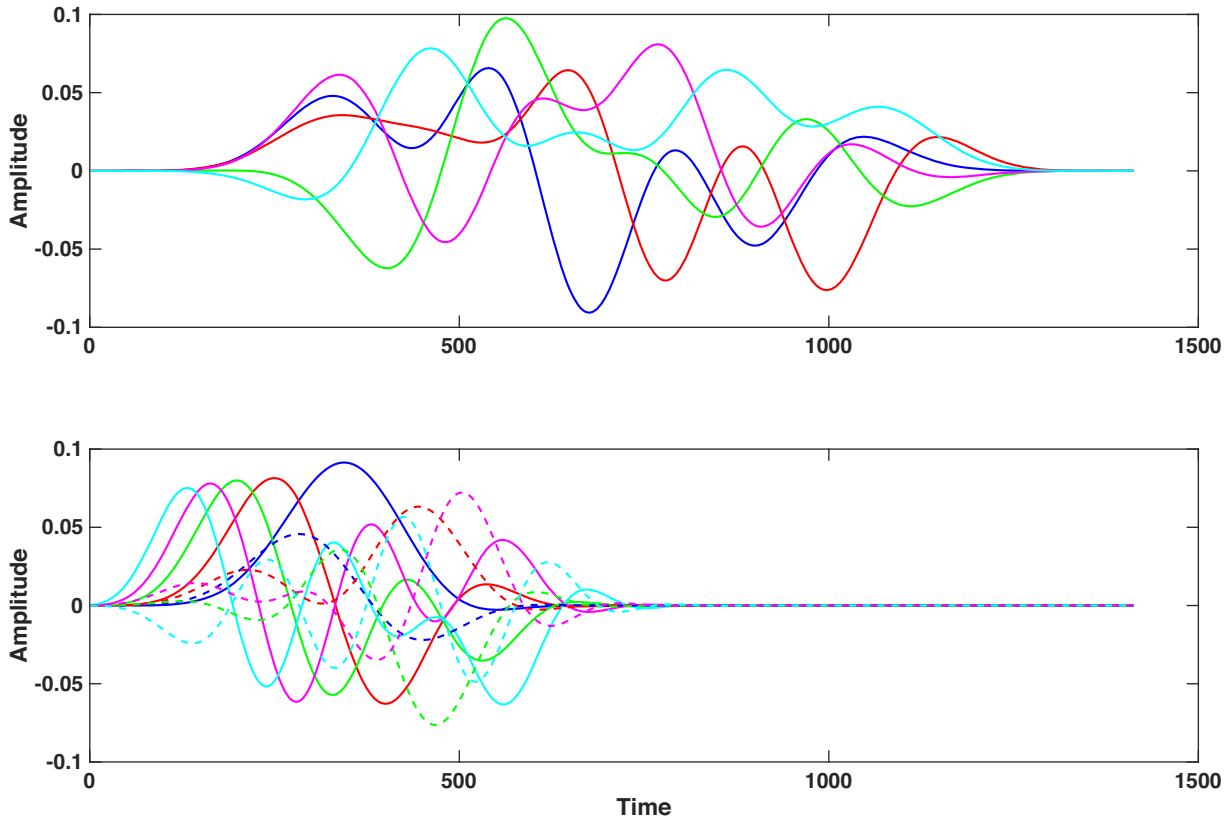

Figure 20. The first five irregular sampling Slepian sequences for time sampling $t_i = i^{1.05}$, $i = 1, 1000$ and a resolution half bandwidth of 0.008 at a frequency of 0 (top panel) and 0.2 (bottom panel). The colour sequence is blue, red, green, maroon and cyan correspond to the 0, 1, 2, 3 and 4 indexes. The solid and dotted lines in the lower panel are the real and imaginary parts, respectively.

height of the Nile River that is thoroughly described by Popper (1951) based on medieval Arabic source documents. Popper outlines many potential issues with these data and provides summary statistics, but it is frustrating that he did not gather his preferred values together into a single data set. Instead, the data given by Toussoun (1925) extending from 622 to 1921 CE will be utilized, although there are several long gaps beginning in 1470 CE. The data give the level of the Nile River near Cairo in late June of each year

that is taken as its annual low stand. After interpolating across a few 1–2 yr gaps and eliminating a few isolated values, the time-series comprises the years 622–1470, 1587–1623, 1703–1738, 1749–1771 and 1836–1921 CE. Descriptions and the principles of the nilometer are described in Borchardt (1906) and Popper (1951), and will not be given here. Several papers about hydrological and climatic inferences from these data are available, but will not be reviewed as the present purpose is illustrating the value of a missing-data estimator for their study.

Fig. 16 compares a regular multitaper spectrum of the data over 622–1470 CE with a time-bandwidth of 8 and a missing-data multitaper spectrum with a time-bandwidth of 10 for the entire data set. The resolution bandwidth of each spectrum is 0.0097 yr^{-1}. The missing-data spectrum is higher than the regular multitaper spectrum below 0.1 yr^{-1} or for periods longer than ∼10 yr. It also contains potentially interesting spectral features over the same frequency range. By contrast, the regular spectrum is quite featureless over the same band.

Fig. 17 compares multitaper F statistics for the regular and missing-data multitaper estimators, along with the 0.99 and 0.999 thresholds for an $F_{2,22}$ distribution. The F statistics are larger for the missing-data estimator, with 8.8, 6.3 and 3.3 yr exceeding the 0.999 threshold, and 18.4 yr reaching a probability of 0.99989 and hence nearly as significant. The latter is very close to the lunar nodal cycle of 18.6 yr. The regular multitaper F statistic exceeds the 0.99 level at only four frequencies, and never reaches the 0.999 level.

Fig. 18 shows the missing-data spectrum reshaped at the 0.998 probability level, along with the energy in the spectral lines under the assumption that their bandwidths are the Rayleigh resolution of the estimate. Each of the line components is about 1.5 decades higher than the residual spectrum. In addition, reshaping the DC value results in a substantial peak emerging at ∼90 yr period, which is very close to the Gleissberg cycle period of 88 yr (Peristykh & Damon 2003).

6 DISCUSSION

The three examples in Sections 3–5 demonstrate the utility of the missing-data multitaper estimator defined in Section 2. The length of day and seafloor pressure examples illustrate two different types of bias that can occur. The first is the pervasive downward bias seen in Figs 9 and 10 as the number of missing values in the length of day data rises to 30 and 45 per cent, respectively. This is not due to spectral leakage, which produces upward bias at low points in the spectrum, as seen in Fig. 15. The downward bias is due to an inability to obtain a meaningful spectral estimate from a limited data set that increases with the fraction of missing data.

Fig. 3 demonstrates the well-known phenomenon for standard multitaper estimators where the ends of a data set are included more fully in high order tapers than in low order ones. Truncating the number of tapers below $\lfloor 2NW \rfloor$ reduces the influence of the ends of a time sequence. This is less of an issue with the missing-data estimator, as seen in Fig. 3. The differences between the standard and missing-data spectra in Fig. 16 are due to a combination of the truncation effect on the standard estimator and the inclusion of data from after 1470 CE in the missing-data estimator. Since the differences are primarily at frequencies under 0.1 yr^{-1}, the addition of the 1836–1921 CE data in the missing data result has the most effect, and in fact the missing-data tapers 'favor' this data section over the shorter data segments between 1470 and 1836 CE.

This paper has investigated the missing data problem where (23) reduces to the identity matrix, and (21) becomes a standard eigenvalue problem with frequency-independent eigenvalues and eigenvectors. However, (21)–(23) are valid for a more general class of irregularly sampled data where gaps are not integer multiples of a fundamental sample interval, or when the process bandwidth is other than $[-1/2, 1/2)$. In these instances, four important distinctions are observed: (1) the eigenvalues and eigenvectors become frequency dependent, (2) the eigenvectors for nonzero frequency are complex, (3) the eigenvectors υ^kin (21) are not orthogonal, although their product with the Cholesky factor of $\mathbf{B}$ given by $\mathbf{R}^T\upsilon^k$ in (24) are and (4) the concept of a Nyquist frequency becomes ill defined, while aliasing becomes complicated. The first three of these will be illustrated, while further work is needed to understand the last point.

A somewhat contrived example where time is defined as $t_i = i^{1.05},\ i = 0, \ldots N-1$ serves to illustrate these phenomena. Fig. 19 shows the first ten eigenvalues as a function of frequency over [0, 0.3) for a resolution half-bandwidth of 0.008 under the assumption that the process bandwidth is $[-1/2,\ 1/2)$. The eigenvalues are nearly 1 at frequencies up to ∼0.18, and then fall off, asymptoting to a value of 0.5. The higher order eigenvalues fall off at lower frequencies than the lower order ones. The behaviour in Fig. 19 is quite different than for the MDSSs, where the eigenvalues are constant across frequency.

Fig. 20 shows the first five eigenvectors $\mathbf{R}^T\upsilon^k$ in (24), or irregular sampling Slepian sequences (ISSSs), at a frequency of 0 (top panel) and 0.2 (bottom panel) for the same parameters as used for Fig. 19. The zero frequency ISSSs extend across the time support, and undersample at the ends relative to the middle of the time interval, as expected for low order tapers. By contrast, the higher frequency ISSSs are concentrated at early time, and are complex. The concentration at early time is easy to understand, as higher frequencies are determined by the more closely sampled data at early time as compared to the more widely spaced data at later time.

The concepts introduced in this paper easily extend to bivariate or multivariate missing-data multitaper entities by analogy to the standard multitaper estimator. Further, since frequency is a continuous rather than discrete variable in (3), it is not necessary for the sampling scheme to be identical for all of the time-series under analysis. However, major differences in sampling may not yield meaningful results; computing the cross spectrum of disjoint time-series does not make much sense. The ISSSs are also well defined for two dimensional problems, yielding an optimal array processing solution in wavenumber space. The F test in wavenumber is a test for the presence of a plane wave, and since the multitaper F test has the Rayleigh resolution that is much better than that of the multitaper estimator, this will yield superior results as compared to beam forming, which has the resolution bandwidth. Further work is needed to understand this application.

7 CONCLUSIONS

This paper has presented a multitaper estimator that accommodates time-series containing gaps without using any form of interpolation. The missing-data multitaper estimator is an extension of the irregular sampling multitaper approach introduced by Bronez (1985, 1988). It has been shown that a simple transformation of Bronez's result yields frequency independent Slepian sequences, along with

frequency independent eigenvalues that define the energy concentration within the resolution bandwidth, whenever the process bandwidth is $[-1/2, 1/2)$ for unit sampling and the sampling scheme comprises integer multiples of one. As a consequence, one need only compute the resulting missing-data Slepian sequences for a given sampling scheme once, and then the spectrum at an arbitrary set of frequencies can be computed using them. It was also shown that the resulting missing-data multitaper estimator can incorporate all of the optimality features (i.e. adaptive-weighting, F test, reshaping) of the standard multitaper estimator, and can be applied to bivariate or multivariate situations in similar ways. The extension to higher dimensional (i.e. array) data is also straightforward.

Three time-series containing gaps were examined in the paper. The first is length of day data as utilized by Smith-Boughner & Constable (2012) in an alternative multitaper approach that forces standard Slepian sequences to be zero where data are missing. It was shown that the missing-data Slepian sequences behave quite differently from their result, and that the ensuing spectral window has a much squarer shape and substantially higher sidelobe protection. The proposed multitaper estimator accurately measures the spectrum with up to 30 per cent missing data, and then becomes downward biased.

The second example utilized seafloor pressure data that have a large dynamic range due to the signature of diurnal and semidiurnal tides. Using a fixed ∼21 per cent fraction of missing data, it was shown that the performance of the missing-data multitaper estimator is degraded as the number of gaps rises from one to five and then twelve due to increasing spectral leakage.

The final example used the arguably longest instrumental time-series in existence, the low stand of the Nile River covering 622–1921 CE, with substantial gaps after 1470 CE. For a fixed number of Slepian tapers, the missing-data estimator resolved several features that were missed by a standard multitaper estimator operating only on the 622–1470 CE interval.

ACKNOWLEDGEMENTS

The length of day utilized in Section 3 are available from http://hpiers.obspm.fr. The pressure data used in Section 4 are available from https://doi.org/10.1029/2018JC014586. A Matlab function MDmwps.m to compute missing-data power spectra is available from the Mathworks file exchange website. The author thanks Jeff Park and editor F.J. Simons for thorough reviews. This work was supported by an Internal Research and Development award at WHOI, and by the Walter A. and Hope Noyes Smith Chair for Excellence in Oceanography.

REFERENCES

Babu, P. & Stoica, P., 2010. Spectral analysis of nonuniformly sampled data – a review, *Dig. Sig. Proc.,* **20,** 359–378.

Borchardt, L., 1906. *Nilmesser und nilstandsmarken,* Verlag der Königl. Akad. der Wissenschaften, 66pp.

Bronez, T.P., 1985, Nonparametric spectral estimation of irregularly sampled multidimensional random processes, *PhD dissertation*, Arizona State U., 179pp.

Bronez, T.P., 1988. Spectral estimation of irregularly sampled multidimensional processes by generalized prolate spheroidal sequences, *IEEE Trans. Acoust., Speech Signal Process.,* **36,** 1862–1873.

Chave, A.D., 2017. *Computational Statistics in the Earth Sciences,* Cambridge U. Press.

Cramér, H., 1940. On the theory of stationary random processes, *Ann. Math.,* **41,** 215–230.

Fodor, I.K. & Stark, P.B., 2000. Multitaper spectral estimation for time series with gaps, *IEEE Trans. Signal Proc.,* **48,** 3472–3483.

Lomb, N.R., 1976. Least-squares frequency analysis of unequally spaced data, *Astrophys. Space Phys.,* **39,** 10–33.

Percival, D.B. & Walden, A.T., 1993. *Spectral Analysis for Physical Applications,* Cambridge U. Press.

Peristykh, A.N. & Damon, P.E., 2003. Persistence of the Gleissberg 88-year solar cycle over the last ∼12,000 years: evidence from cosmogenic isotopes, *J. geophys. Res.,* **108,** SSH 1–1-SSH 1–15, doi:10.1029/2002JA009390.

Pinkel, R. *et al.*, 2000. Ocean mixing studied near Hawaiian Ridge, *EOS, Trans Am. geophys. Un.,* **81,** 545–553.

Popper, W., 1951. The Cairo nilometer: studies in Ibn Taghrî Birdî's chronicles of Egypt, I, *Publications in Semitic Philology,* **12,** 1–269.

Scargle, J.D., 1982. Studies in astronomical time series analysis II. Statistical aspects of spectral analysis of unevenly spaced data, *Astrophys. J.,* **263,** 835–853.

Schuster, A., 1898. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena, *Terr. Magn.,* **3,** 13–41.

Slepian, D., 1978. Prolate spheroidal wave functions, Fourier analysis and uncertainty-V. The discrete case. *Bell. Sys. Tech. J.,* **57,** 1371–1430.

Smith-Boughner, L.T. & Constable, C.G., 2012. Spectral estimation for geophysical time series with inconvenient gaps, *Geophys. J. Int.,* **190,** 1404–1422.

Stoica, P. & Sundin, T., 1999. On nonparametric spectral estimation, *Circuits Syst. Signal Process,* **18,** 169–181.

Thomson, D.J., 1982. Spectrum estimation and harmonic analysis, *Proc. IEEE,* **70,** 1055–1096.

Thomson, D.J. & Haley, C.L., 2014. Spacing and shape of random peaks in non-parametric spectrum estimates, *Proc. R. Soc. Lon.,* **A470,** 20140101, doi:10.1098/rspa.2014.0101.

Toussoun, O., 1925. Mémoire sur l'histoire du Nil, *Mémoires àl'Institut d'Egypte,* **18,** 366–404.