Using Lomb–Scargle Analysis to Derive Empirical Orthogonal Functions from Gappy Meteorological Data[✉]

CHRISTOPHER DUPUIS

Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

COURTNEY SCHUMACHER

Department of Atmospheric Sciences, Texas A&M University, College Station, Texas

(Manuscript received 3 September 2017, in final form 5 July 2018)

ABSTRACT

The Lomb–Scargle discrete Fourier transform (LSDFT) is a well-known technique for analyzing time series. In this study, a solution for empirical orthogonal functions (EOFs) based on irregularly sampled data is derived from the LSDFT. It is demonstrated that this particular algorithm has no hard limit on its accuracy and yields results comparable to those of complex Hilbert EOF analysis. Two LSDFT algorithms are compared in terms of their performance in evaluating EOFs for precipitation observations from the Tropical Rainfall Measuring Mission satellite. Both are shown to be able to capture the pattern of the diurnal cycle of rainfall over the complex topography and diverse land cover of South America, and both also show other consistent features in the 0–12-day frequency band.

1. Introduction

Empirical orthogonal functions (EOFs) are an important part of spatial analysis of oscillations and have been an especially popular technique in the atmospheric sciences since the 1950s (North et al. 1982). However, EOFs have more frequently been applied to model datasets since there is currently no theoretically exact way to connect EOF analysis to nonequispaced data. “Nonequispaced” (alternatively “gappy,” “irregularly spaced,” or “unevenly spaced”) in this context refers to a time series that is sampled at irregular time intervals, a situation which typically arises for two reasons: 1) the sampling technique does not generally result in evenly spaced data or 2) the sampling technique does not always lead to valid results, even when sampling is otherwise equispaced. Examples of this kind of data might include satellite data, radar data, paleoclimate proxy data (like ice cores or speleothems), or a regular dataset that has been quality controlled. In

these cases, the reason for uneven spacing is a property extrinsic to the analysis technique in question.

For field quantities, nonequispaced still applies independently at each spatial point; that is, the gappiness of an individual time series in a field does not have to be related to the characteristics of gappiness at other points in the field. Whether these fields are gappy with respect to time or space can be more a matter of perspective if the temporal gappiness is not tightly coupled across the field. It may be appropriate to refer to this kind of data as “porous” rather than gappy to better evoke the general distribution of gaps.

The fundamental difficulty with connecting unevenly spaced time series to the EOF problem lies in calculating the covariance matrix. Approximate methods exist; however, they rely on various theoretical frameworks that are most appropriate only for specific cases, usually equispaced data. The simplest of these methods is to use pairwise-complete observations. There are also numerous interpolation methods, which are designed to fit unevenly spaced data into an even spacing. Both of these methods inherently result in data loss.

An additional pathway remains: to evaluate the covariance matrix elements in spectral space. Spectral analysis allows for the use of all available data and can also manage large gaps in sampling times. The simple discrete Fourier transform (DFT) as well as the fast

[✉] Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JAMC-D-17-0250.s1>.

Corresponding author: Christopher Dupuis, christopher.dupuis@noaa.gov

Fourier transform (FFT) family offer ideal techniques for equispaced datasets. However, Fourier analysis for gappy data requires specialized techniques.

a. Time-series methods

Lomb (1976) was the first to propose an exact solution to the problem of applying a Fourier transform to unevenly spaced data. Scargle (1982) offered a similar analysis, derived from a least squares approach. The techniques developed by Lomb and Scargle were recognized to be essentially identical, and were from then on known as the Lomb–Scargle periodogram (LSP).

Expanding the scope of nonequispaced time-series analysis, Scargle (1989) introduced a complex formulation of the LSP. This version was also derived through alternative methods in Mathias et al. (2004). Since this complex formulation denotes a true discrete Fourier transform, it shall hereinafter be referred to as the Lomb–Scargle discrete Fourier transform (LSDFT) to distinguish it from the LSP, as well as other discrete Fourier transforms. The LSDFT preserves phase information and therefore becomes necessary when investigating propagating patterns.

To obtain more precise results, Schulz and Stattegger (1997) apply Welch's overlapped segment averaging (WOSA) method to Lomb–Scargle analysis. Their approach does not address all the issues of periodogram biases, but it does offer a much-improved signal-to-noise ratio. Conveniently, Welch's method is applied to the individual transformed time-series segments, rather than time series itself, and is therefore abstracted from the data enough so that it applies regardless of which transform technique is used to derive the transformed segments. This offers the possibility of using the LSP as the core Fourier analysis, as used in "SPECTRUM" (Schulz and Stattegger 1997), or the LSDFT for complex-valued results, or even some other form of spectral analysis (Zechmeister and Kürster 2009).

The classic LSDFT kernel is computationally expensive relative to FFT and FFT-based methods, making it worthwhile to translate the LSDFT algorithms into "CUDA," the Nvidia Corporation's platform for programs designed to run on graphical processing units (GPUs; Nickolls et al. 2008). The code and files used to implement the LSDFT are available in the online supplemental material. Faster algorithms exist, notably the algorithm pioneered by Press and Rybicki (1989) and a number of similar nonequispaced FFT methods (Dutt and Rokhlin 1993; Steidl 1998; Leroy 2012), but GPU acceleration of the classic LSDFT (similar to Townsend 2010) proved satisfactory for this research.

b. The empirical orthogonal function problem

A variety of EOF techniques exist, and many of the major variations have previously been outlined (e.g.,

Kim and Wu 1999; Hannachi et al. 2007). Hasselmann (1988) states that EOFs only fully explain time-series oscillations when the complex cross-spectral covariance matrix is used, and is either evaluated at every frequency band or at each frequency in the spectral window, as in North (1984). Note that the analyses of noise-forced linear models in North (1984) are basically applicable, but with two caveats: first, that the lag-autocovariance function in Eq. (11) from North (1984) would have to be calculated in spectral space; and second, that discrete spectral techniques include an inherent error relative to the true continuous spectrum, and potentially more so for Lomb–Scargle techniques than for others. The EOF technique used herein is relatively simple, using the complex-valued covariance of convolved spectra for the covariance matrix elements (thereby fully describing the oscillations within the bandwidth), and solving as in a classic EOF analysis. Conceptually, this differs from the frequency-dependent EOFs of North (1984) in that the frequency range is accounted for during integration, rather than solving the EOF problem for each frequency in question. This has the advantage of being more computationally parsimonious, as well as potentially capturing dispersive oscillations, but carries the disadvantage that information about how EOF variance is partitioned in spectral space is lost.

The Lomb–Scargle EOF (LSEOF) algorithm is closely related, but not identical, to the complex Hilbert EOF (HEOF) method. Since the HEOF method relies on a Hilbert transform of the raw data (Horel 1984; Kim and Wu 1999), followed by a classic EOF matrix multiplication, it arrives at the EOF covariance matrix in a different way from the LSEOF. Nevertheless, the Hilbert transform is a multiplier operator of the Fourier transform; that is,

$$F[H(u)] = -i\Theta(\omega)F(u), \quad (1)$$

where $H(u)$ is the Hilbert transform of the function u , $F(u)$ represents the Fourier transform of u , ω is the frequency, and Θ represents the Heaviside step function as classically defined, or more simply, the sign of its operand. The complex EOF results from adding real-valued data to its Hilbert transform and treating the result of the data matrix of interest (Horel 1984; Kim and Wu 1999):

$$u' = u + H(u), \quad (2)$$

so

$$F(u') = F(u) + F[H(u)] = [1 - i\Theta(\omega)]F(u), \quad (3)$$

and HEOF covariance matrix element σ'_{xy} is therefore

$$\sigma'_{xy}(t_{\text{lag}}) = \int_{-\infty}^{\infty} \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega\tau_{\text{lag}}} d\omega \quad (4)$$

$$= \int_{-\infty}^{\infty} [1 - i\Theta(\omega)] \tilde{X}(\omega) [1 + i\Theta(\omega)] \tilde{Y}^*(\omega) e^{i\omega\tau_{\text{lag}}} d\omega \quad (5)$$

$$= \int_{-\infty}^{\infty} [1 + \Theta^2(\omega)] \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega\tau_{\text{lag}}} d\omega, \quad (6)$$

where X and Y represent two time series and the tilde represents Fourier transforms of the respective time series. The asterisk indicates the complex conjugate of $\tilde{Y}$. Since $\Theta^2(\omega) = 1$ as long as $\omega \neq 0$,

$$\begin{aligned} \sigma'_{xy}(t_{\text{lag}}) &\approx \int_{-\infty}^{\infty} \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega\tau_{\text{lag}}} d\omega \\ &+ \int_{0^+}^{\infty} \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega\tau_{\text{lag}}} d\omega; \end{aligned} \quad (7)$$

therefore, if ω is always greater than 0,

$$\sigma'_{xy}(t_{\text{lag}}) \approx 2 \int_{0^+}^{\infty} \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega\tau_{\text{lag}}} d\omega, \quad (8)$$

which implies that $\sigma'_{xy}(t_{\text{lag}}) \approx 2\sigma_{xy}(t_{\text{lag}})$. Since the covariance matrices are identical aside from a scalar multiple, the eigenvalues and eigenvectors of complex EOF analysis are identical to the results from LSEOF analysis.

The basic EOF technique is described in [section 2a](#), followed by a description in [section 2b](#) of the Lomb–Scargle periodogram and related techniques, with some important modifications. A specific case study applying these techniques to rainfall data over South America as well as their results is described in [section 3](#). A brief discussion of mathematical and computational aspects is given in [section 4](#). A summary follows in [section 5](#).

2. Methods

a. Empirical orthogonal functions

For a vector of time series $\mathbf{S}_{s \times t}$, where s is the number of spatial points and t is the number of temporal points arranged as

$$\mathbf{S}_{s \times t} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1t} \\ s_{21} & s_{22} & \cdots & s_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ s_{s1} & s_{s2} & \cdots & s_{st} \end{pmatrix}, \quad (9)$$

wherein the rows can be treated as individual time series, the covariance matrix is defined as

$$\mathbf{C} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1j} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{i1} & \sigma_{i2} & \cdots & \sigma_{ij} \end{pmatrix} = \mathbf{S}\mathbf{S}^T. \quad (10)$$

The EOF eigenvalue problem is thus given as $\Lambda\mathbf{C} = \lambda\mathbf{C}$, where Λ_{nj} denotes the n th EOF's j th spatial element and λ_n^2 denotes the n th EOF's explained variance. Solving an evenly sampled EOF problem is fairly straightforward, but to solve the EOF problem directly with unevenly sampled data is very difficult, if not impossible. The LSDFT offers some respite, because it is possible to relate the covariance of each element to its respective transform through the cross-correlation theorem:

$$\begin{aligned} \sigma_{xy}(t_{\text{lag}}) &= \int_0^{\infty} X(t) Y^*(t - t_{\text{lag}}) dt \\ &= \int_0^{\infty} \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega t_{\text{lag}}} d\omega. \end{aligned} \quad (11)$$

The covariance matrix [as in Eq. (10)] can be estimated by calculating each covariance element individually. Typically, each element of the covariance matrix is estimated by the first integral in Eq. (11), but the second integral proves that it is also possible to estimate the covariance matrix elements by taking the convolution of the spectra of both time series. With the covariance matrix known, it is possible to solve the EOF problem.

Note that, by using Eq. (11), it becomes unnecessary to calculate the cross-spectral Lomb–Scargle transforms; only univariate spectra are needed. In addition, although EOFs are typically calculated with $t_{\text{lag}} = 0$, it is possible to calculate EOFs for nonzero lags ([Monahan et al. 1999](#)). For a transformed cross-spectrum of time series X and Y , denoted as $\tilde{X}\tilde{Y}$, the relationship is

$$\begin{aligned} \sigma_{xy}(t_{\text{lag}}) &= \int_0^{\infty} \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega t_{\text{lag}}} d\omega \\ &= \int_0^{\infty} |\tilde{X}\tilde{Y}(\omega)|^2 e^{i\omega t_{\text{lag}}} d\omega. \end{aligned} \quad (12)$$

Using the first integral (henceforth the univariate integral) rather than the second (henceforth the bivariate integral) will reduce the number of LSDFT calculations from $O(N^2)$ to $O(N)$ plus a relatively small overhead for multiplication of the transformed time series.

Examination of the statistical significance of each EOF is made possible by North's rule of thumb ([North et al. 1982](#)). The variance confidence interval $\Delta\lambda^2$ is defined as

$$\Delta\lambda^2 = \lambda^2(2/N)^{1/2}, \quad (13)$$

where N is the number of independent samples. Although North's rule of thumb implicitly assumes a constant number of samples across the data field, a mean sample number can offer an estimate. An EOF with a variance confidence interval that does not overlap with any other EOF's confidence interval is generally considered to be linearly independent of all other EOFs; those that do overlap generally interact with each other. North's rule can then be used as a measure of validation for physically relevant EOFs.

b. The Lomb–Scargle discrete Fourier transform

The periodogram of a time series is classically defined as $P(\omega) = N^{-1}|\tilde{S}(\omega)|^2$, where $P(\omega)$ denotes the spectral power density and N_0 is the number of samples in the original time series. While useful for its simplicity, the modulus function in the periodogram ensures that half the spectral information from a complex-valued DFT will be lost. While the imaginary part is never strictly necessary, it is useful for reducing the number of necessary computations further along in the process of creating a covariance matrix.

The Lomb–Scargle periodogram as classically derived can be found in [Scargle \(1982\)](#), as well as numerous other papers on the topic. For evenly spaced data, t is simply the time between samples, but to preserve the “invariance of time translation,” as Scargle phrases it, a constant τ must be defined as

$$\tau = \frac{1}{2\omega} \arctan \left[\frac{\sum_j \sin(2\omega t_j)}{\sum_j \cos(2\omega t_j)} \right]. \quad (14)$$

One can argue by comparison with the classic periodogram and DFT that the logical formulation for the complex-valued LSDFT is the following:

$$\tilde{S}(\omega) = \frac{1}{\sqrt{2\sigma^2}} \sum_j \left\{ \frac{(S_j - \bar{S}) \cos[\omega(t_j - \tau)]}{\sqrt{\sum_j \cos^2[\omega(t_j - \tau)]}} + \frac{i(S_j - \bar{S}) \sin[\omega(t_j - \tau)]}{\sqrt{\sum_j \sin^2[\omega(t_j - \tau)]}} \right\}. \quad (15)$$

This formulation is essentially the same as that of [Scargle \(1989\)](#) and [Mathias et al. \(2004\)](#).

Superior implementations of the LSDFT are more complicated. As discussed in [section 1a](#), it is possible

to dramatically improve the signal-to-noise ratio using the WOSA technique ([Schulz and Stattegger 1997](#)). Note that, for this technique, σ^2 is omitted from the periodogram or DFT function itself. The cross-spectrum from WOSA ($\hat{G}_{xy}$) is calculated by

$$W_{xy}(\omega_j) = \frac{2}{n_{50} \sqrt{N_{\text{seg}}^{(x)} N_{\text{seg}}^{(y)} \Delta f_{xy}}} \times \sum_{n=1}^{n_{50}} H[\tilde{X}_n(\omega_j) I_n] H[\tilde{Y}_n^*(\omega_j) I_n] \quad \text{and} \quad (16)$$

$$\hat{G}_{xy}(\omega_j) = \frac{\sigma_x \sigma_y}{\Delta f \sum_j W_{xy}(\omega_j)} W_{xy}(\omega_j), \quad (17)$$

where I denotes the time interval,

$$I = \left[\frac{t_{\max}(n-1)}{n_{50} + 1} \leq t \leq \frac{t_{\max}(n+1)}{n_{50} + 1} \right], \quad (18)$$

and where H is a windowing function. In this study, the Hamming window is used, which is defined as

$$H(S_j) = \frac{(S_j - \bar{S}) \left[\alpha - \beta \cos \left(2\pi \frac{t_j - t_{\min}}{t_{\max} - t_{\min}} \right) \right]}{\frac{1}{N_{\text{seg}}} \sum_j \left[\alpha - \beta \cos \left(2\pi \frac{t_j - t_{\min}}{t_{\max} - t_{\min}} \right) \right]^2}, \quad (19)$$

where $\alpha = 0.53836$ and $\beta = 1 - \alpha$. It is important to note that this windowing function's bounds are defined by minimum and maximum time values. While in the evenly spaced case it is inconsequential, conflation of windowing the increment difference (as is common) and windowing the time difference will result in an incorrect window function, distorted in the time domain depending on the temporal differences between individual points. Therefore, while it is somewhat nonstandard, the above definition was chosen. The same observation should be considered for other windowing functions as well.

A caveat of using a complex form of the LSDFT is that [Schulz and Stattegger \(1997\)](#) offer no method to deal with complex data. Since phase data are defined on a circular domain, segment averaging will require the computation of circular means for phase. Phase is typically defined as

$$\phi(\tilde{S}) = \text{atan2}[\text{Im}(S_n), \text{Re}(S_n)], \quad (20)$$

where atan2 denotes the special function returning the arctangent with the sign ambiguity resolved. The phase of the LSDFT is therefore simply

$$\phi(\tilde{S}) = \text{atan2} \left\{ \frac{\sum_j (S_j - \bar{S}) \sin[\omega(t_j - \tau)]}{\sqrt{\sum_j \sin^2[\omega(t_j - \tau)]}}, \frac{\sum_j (S_j - \bar{S}) \cos[\omega(t_j - \tau)]}{\sqrt{\sum_j \cos^2[\omega(t_j - \tau)]}} \right\}. \quad (21)$$

For the WOSA algorithm, [Schulz and Stattegger \(1997\)](#) offer the formula

$$\phi(\hat{G}_{xy}) = \text{atan2}[\text{Im}(\hat{G}_{xy}), \text{Re}(\hat{G}_{xy})] \quad (22)$$

for calculating the phase, but the phase defined this way only describes the phase difference between the two time series in a bivariate analysis and not the actual phase per se.

For reasons that have been explained in [section 2a](#), the EOF problem is computed much more quickly when it is expressed in terms of univariate time series as opposed to bivariate time series. Thus, we must seek an alternative way to calculate the phase such that it gives meaningful results in the univariate case as well. Because the phase is a periodic quantity, the circular mean must be used instead of the arithmetic mean. The circular mean is defined as

$$\bar{\phi} = \text{atan2}[\overline{\sin(\phi_n)}, \overline{\cos(\phi_n)}], \quad (23)$$

which results in the following formula for WOSA phase:

$$\bar{\phi}[\hat{G}_{xy}(\omega_j)] = \text{atan2}[\overline{\sin\{\phi[g_n(\omega_j)]\}}, \overline{\cos\{\phi[g_n(\omega_j)]\}}], \quad (24)$$

where

$$g_n(\omega_j) = H[\tilde{X}_n(\omega_j)I_n]H[\tilde{Y}_n^*(\omega_j)I_n]. \quad (25)$$

In the next section, the spectra and EOFs from the LSDFT and WOSA techniques are demonstrated with a long-term precipitation dataset from the Tropical Rainfall Measuring Mission (TRMM) satellite, specifically the “3B42” product spanning from 1998 to 2016, which is gappy in time and space by nature of the satellite’s orbit.

3. Diurnal rain patterns over South America

To demonstrate the functionality of the time-series algorithms, the diurnal cycle of precipitation across South America is analyzed using observations from the TRMM satellite. The diverse topography and land cover of South America represent a rich set of forcings of diurnal rain patterns ([Garreaud and Wallace 1997](#)).

FIG. 1. The mean gap length at each point.

The diurnal cycle over land is one of the most prominent and predictable oscillations in atmospheric science, but many questions remain as to the most relevant physical mechanisms driving the diurnal cycle of precipitation ([Derbyshire et al. 2004](#); [Khairoutdinov and Randall 2006](#); [Wu et al. 2009](#)), and large-scale models have difficulty in accurately simulating the diurnal rain pattern over land ([Betts and Jakob 2002](#); [Dai 2006](#); [Dirmeyer et al. 2012](#)). Accurate representation of the diurnal cycle in global models has implications for mean climate biases and capturing variations in clouds and radiation in a changing climate ([Del Genio 2012](#)).

We use rain amounts from the 3-hourly, $0.25^\circ \times 0.25^\circ$ resolution TRMM 3B42 passive microwave “high quality precipitation” product from 1 January 1998 to 31 December 2016 ([Huffman et al. 2007](#)). The actual satellite observation times are used for the time series, which vary by up to 90 min from the 3-hourly granule time. This dataset also contains many missing observations per 3-hourly granule because of the satellite’s orbit, which makes for a gappy time series at each spatial point. The mean gap length can be seen in [Fig. 1](#) and largely corresponds to number of samples (which ranges between 27 000 and 48 000), median gap length, and standard deviation of gap length. An example histogram of these gaps for one time series is presented in [Fig. 2](#), showing spikes at roughly 3-hourly intervals and a roughly Poisson-distributed curve elsewhere.

Similar EOF analyses of TRMM 3B42 data can be found in [Kikuchi and Wang \(2008\)](#) and [Pritchard and Somerville \(2009\)](#). [Kikuchi and Wang \(2008\)](#) use TRMM precipitation data and interpolated local solar times for use with real-valued EOF analysis on the entire dataset, as well as on winter and summer subsets.

FIG. 2. Example histogram of time-series gap lengths.

This analysis generates pairs of EOFs representing propagating diurnal and semidiurnal cycles. Pritchard and Somerville (2009) use least squares analysis to fit sinusoidal curves to diurnal and diurnal harmonic composites of TRMM data, which are compared with a similar analysis of model data. This analysis shows large differences between observed and modeled diurnal cycle phases.

Although the TRMM mission was discontinued in April 2015, the continued addition of data from other sources does not have any noticeable negative impacts on the analysis. Notably, the other two 3B42 rainfall analyses, the “infrared precipitation” and the aggregate “precipitation” products, performed significantly worse at both the spectral analysis and EOF stages for reasons that remain unclear. The methods described in section 2 are used to analyze each time series in sequentially ordered $10^\circ \times 10^\circ$ boxes of TRMM 3B42 data over South America. An example WOSA spectrum is presented in Fig. 3 at a location near the coast with strong diurnal and semidiurnal cycles. These strong signals yield prominent peaks at the proper frequencies.

The results for the classic LSDFT and WOSA are presented as maps in Fig. 4. This analysis shows coherent spatial patterns at the frequency corresponding to the

diurnal cycle in terms of both relative amplitude (Fig. 4, left), highlighting areas with strong diurnal cycles, and phase (Fig. 4, right), showing areas with propagating diurnal storm systems as rainbow patterns, such as over Argentina and coastal Brazil. The WOSA method relative to the LSDFT results has stronger relative amplitudes in most places, and also appears to produce more granular spatial patterns. This is consistent with the fact that WOSA spectra show larger peak-to-floor ratios than LSDFT spectra (see Fig. 5), implying that WOSA is much more successful at evaluating spectra. Notably, WOSA is only about twice as time-consuming to compute as LSDFT, so WOSA appears to offer the superior analysis between the two.

The spatial patterns of the diurnal cycle over South America resulting from the LSDFT and WOSA analyses in Fig. 4 are consistent with previously published satellite climatologies and specific meteorological mechanisms (Yang and Slingo 2001; Hirose et al. 2008; Biasutti et al. 2012; Venugopal et al. 2016). Notably, while a sea-breeze mechanism is evident over land near the coasts during the day, nocturnal offshore propagation is also present in many areas. For example, the strong amplitude off the west coast of Colombia is consistent with the results of Mapes et al. (2003), and the phase data derived herein present a new way

FIG. 3. WOSA spectrum at 3°S , 45°W , showing a prominent diurnal cycle and at least two possible harmonics.

FIG. 4. The (top) classic LSDFT analysis with a complex formulation and (bottom) WOSA analysis of TRMM 3B42 3-hourly rain rates over South America. Amplitudes are relative to the 5th-percentile value in the domain, which is set to 0 dB.

to view its propagation. The Amazon and its tributaries also show strong river breeze signals (Negri et al. 2002).

The EOF results for classic LSDFT and WOSA time-series analyses are presented in Figs. 6 and 7, respectively. Because of memory constraints in evaluating the EOF problem, the domain was reduced to an area covering most of the Amazon basin. In addition, integrating over the whole spectrum of interest can reveal more chaotic

patterns that occupy a wide frequency band. However, unlike in principal oscillation pattern (POP) analysis (Hasselmann 1988), the resulting EOFs do not yield information about the frequency bandwidths they are associated with. In this case though, it can safely be assumed that the diurnal cycle will be the strongest oscillation, and indeed, the first EOFs in each analysis (top rows, Figs. 6 and 7) appear essentially identical in both amplitude and

FIG. 5. Map of the ratio of WOSA to LSDFT diurnal peak-to-floor ratios (dB). WOSA shows a large advantage over LSDFT in regions known to have strong diurnal cycles.

phase, and account for the vast majority of the variance contained in the 0–12-day bandwidth. The diurnal pattern from both the raw spectral data and the EOF analyses show close agreement with previous research [notably Burleyson et al. (2016)—their Fig. 6 matches our phase results, and their Fig. 9 matches our amplitude results].

The next EOFs in Figs. 6 and 7 reveal distinct patterns within the 0–12-day bandwidth that cannot be found in an EOF analysis of just the diurnal LSDFT and WOSA results. The second EOF from both analyses seems to be a harmonic of the diurnal cycle and is similar to the EOF pattern at the half-day frequency (Fig. 8). This pattern exhibits a more homogeneous pattern in the interior Amazon, as well as a propagating pattern that is more tightly bound to the coast than the inland-propagating portion of the diurnal cycle. Kikuchi and Wang (2008) show similar results in the Amazon basin, finding diurnal and semidiurnal propagating modes, with the semidiurnal mode more tightly bound to the coast. A pattern that appears to be consistent with strong cold surges from farther south (e.g., Lupo et al. 2001) corresponds to the third EOFs (Figs. 6 and 7, bottom rows). The phase differences between the different analyses herein suggest that this mode is dispersive, which is known to be generally true of cold surges. Higher-index EOFs appear to have multipolar forms that bear little relation to known real-world mechanisms.

Because the dataset contains a high number of observations, the error bars derived from North's rule of thumb are difficult to resolve relative to the spread of eigenvalues. Therefore, the confidence intervals estimated by North's

rule (in units of percent variance, i.e., squared eigenvalues normalized such that their sum is equal to 1) for the three main analyses are presented in tabular form in Tables 1 and 2. As explained before, EOF variances that lie outside of any other EOF's error bars (some multiple of North's rule) are considered to be linearly independent from all other EOFs, and therefore statistically significant (North et al. 1982). These results then show that interaction between the major EOFs should be minimal. WOSA EOF analysis appears to have the more rapid decay of variance with respect to EOF number, indicating that in the present research, WOSA offers the more minimally descriptive EOF analysis. Note that, as per Eq. (13), reducing the sample count necessarily enlarges the confidence interval.

4. Discussion

The techniques presented can be hampered by at least two factors. First, the LSDFT assumes a sinusoidal waveform, which means that a nonsinusoidal signal will not be captured if an oscillation only occurs in an unobserved period of time. This could be a problem for datasets with long gaps, although it does not seem to pose a significant issue in this study.

Second, a situation where the observation times are correlated to the principal component time series for an oscillation will necessarily reduce the strength (and possibly alter the character) of the resulting EOF. This is exactly the definition of cyclostationarity (Kim et al. 1996). Therefore, such a dataset should be treated as containing a cyclostationarity constraint as an inherent feature.

Choosing to use the univariate integral in LSEOF analysis may complicate and muddle the phase data unnecessarily compared to the bivariate integral; however, since the bivariate integral is very computationally expensive, a spatial comparison of TRMM data processed through both methods remains unfeasible at the present. Further, the present research only makes use of the most basic EOF analysis, so the effects of using LSDFT analysis together with more advanced EOF/PCA analyses remain unknown.

More advanced techniques are possible on the time-series-analysis side as well. Pritchard and Somerville (2009) in particular analyze the wave shape in terms of diurnal peak broadness, and this consideration could potentially be implemented by using the generalized LSP (Zechmeister and Kürster 2009).

The inherently discrete nature of observational data implies that spectral algorithms and their derivatives will never be integrable across the entire range of frequencies. Therefore, it is important to note that the

FIG. 6. Relative (left) amplitude and (right) phase of the (top) first, (middle) second, and (bottom) third EOFs of TRMM 3B42 3-hourly rain rates over the Amazon basin from classic LSDFT analysis for the 0–12-day bandwidth. Amplitudes are relative to the 5th-percentile value in the domain, which is set to 0 dB.

EOFs derived from integrating LSDFTs inherit the frequency window used in calculating the LSDFT and cannot account for all possible variance in a time-series set. However, this is only a minor concern as long as the

EOFs of interest correspond to frequencies within the LSDFT frequency window. EOFs corresponding to frequencies outside of the frequency window will not be present, unless by aliasing. This can be advantageous

FIG. 7. As in Fig. 6, but for WOSA.

if specific oscillations would otherwise obscure more subtle phenomena.

Computing the covariance matrix, either by direct calculation or by some other method, can be very memory intensive. To reduce the memory overhead, it is

possible to compute covariance elements in block form; that is, by only importing the time-series data relevant to the area of the covariance matrix the program is iterating through. This requires opening either one or more data blocks at a time, rather than requiring the

TABLE 1. Variances and error estimates for LSDFT EOF analysis.

EOF no.	LSDFT variance	LSDFT North's rule
1	61.4%	$4.24 \times 10^{-1}\%$
2	1.26%	$8.73 \times 10^{-3}\%$
3	0.585%	$4.04 \times 10^{-3}\%$
4	0.433%	$2.99 \times 10^{-3}\%$
5	0.319%	$2.20 \times 10^{-3}\%$
6	0.231%	$1.60 \times 10^{-3}\%$

TABLE 2. Variances and error estimates for WOSA EOF analysis.

EOF no.	WOSA variance	WOSA North's rule
1	99.5%	$6.88 \times 10^{-1}\%$
2	0.231%	$1.60 \times 10^{-3}\%$
3	$5.97 \times 10^{-3}\%$	$4.13 \times 10^{-5}\%$
4	$3.06 \times 10^{-3}\%$	$2.12 \times 10^{-5}\%$
5	$2.04 \times 10^{-3}\%$	$1.41 \times 10^{-5}\%$
6	$1.68 \times 10^{-3}\%$	$1.16 \times 10^{-5}\%$

outer product of the entire dataset to be allocated at once. The trade-off is that this block calculation of the covariance matrix can be time-intensive, so care should be taken to make the blocks large enough that this technique does not result in a high number of block iterations.

Although the above method provides a path to large covariance matrices, the whole covariance matrix may be so large as to be impossible to load at once. The classic eigenvalue solvers usually require the entire covariance matrix to be loaded at once, so large EOF analyses through spectral integration will primarily be limited by the scalability of the eigenvalue solver.

5. Conclusions

A novel method for calculating EOFs from gappy data is presented. This method uses a complex-valued extension of the Lomb–Scargle periodogram, referred to as the Lomb–Scargle discrete Fourier transform, to calculate the covariance matrix. This method can be improved with a complex-valued extension of WOSA. These LSDFT EOFs can then be used to find the amplitude, phase, and variance of the primary modes of oscillation in a dataset. However, a current weakness is

that LSDFT EOF analysis does not explicitly yield the frequency range of each EOF, as in principal oscillation pattern analysis.

A key advantage of using LSDFT EOF analysis as compared with other gappy EOF techniques such as data-interpolating EOF analysis and pairwise-complete truncations is that all degrees of freedom contained in the data are explicitly preserved up to the integration over bandwidths. Therefore, there is no hard limit to the accuracy of LSDFT EOF analysis as it is formulated in this text. However, practical limits may make other options with hard limits on accuracy more appealing, like using nonequispaced FFTs instead of the LSDFT for greater frequency resolution relative to computational cost.

Two LSDFT techniques (classic and WOSA) are applied to rainfall observations from the TRMM satellite over South America. Both spectral analyses successfully identify real-world oscillations over a bandwidth of 0–12 days, most prominently the diurnal cycle, but also a mode consistent with cold surges. It is also proven that the LSDFT EOF method yields results that are nearly identical to complex Hilbert EOF analysis, but with the added benefit that LSDFT EOF analysis can accommodate gappy data.

FIG. 8. Relative amplitude and phase of the half-day WOSA EOF. Amplitudes are relative to the 5th-percentile value in the domain, which is set to 0 dB.

Acknowledgments. This work was partially supported by the DOE ASR (Grant DE-SC0016245) and NASA PMM (Grant NNX16AE34G) programs. We also thank the anonymous reviewers for their insightful commentary.

REFERENCES

- Betts, A. K., and C. Jakob, 2002: Evaluation of the diurnal cycle of precipitation, surface thermodynamics, and surface fluxes in the ECMWF model using LBA data. *J. Geophys. Res.*, **107**, 8045, <https://doi.org/10.1029/2001JD000427>.
- Biasutti, M., S. E. Yuter, C. D. Burleyson, and A. H. Sobel, 2012: Very high resolution rainfall patterns measured by TRMM Precipitation Radar: Seasonal and diurnal cycles. *Climate Dyn.*, **39**, 239–258, <https://doi.org/10.1007/s00382-011-1146-6>.
- Burleyson, C. D., Z. Feng, S. M. Hagos, J. Fast, L. A. Machado, and S. T. Martin, 2016: Spatial variability of the background diurnal cycle of deep convection around the GoAmazon2014/5 field campaign sites. *J. Appl. Meteor. Climatol.*, **55**, 1579–1598, <https://doi.org/10.1175/JAMC-D-15-0229.1>.
- Dai, A., 2006: Precipitation characteristics in eighteen coupled climate models. *J. Climate*, **19**, 4605–4630, <https://doi.org/10.1175/JCLI3884.1>.
- Del Genio, A. D., 2012: Representing the sensitivity of convective cloud systems to tropospheric humidity in general circulation models. *Surv. Geophys.*, **33**, 637–656, <https://doi.org/10.1007/s10712-011-9148-9>.
- Derbyshire, S., I. Beau, P. Bechtold, J.-Y. Grandpeix, J.-M. Piriou, J.-L. Redelsperger, and P. Soares, 2004: Sensitivity of moist convection to environmental humidity. *Quart. J. Roy. Meteor. Soc.*, **130**, 3055–3079, <https://doi.org/10.1256/qj.03.130>.
- Dirmeyer, P. A., and Coauthors, 2012: Simulating the diurnal cycle of rainfall in global climate models: Resolution versus parameterization. *Climate Dyn.*, **39**, 399–418, <https://doi.org/10.1007/s00382-011-1127-9>.
- Dutt, A., and V. Rokhlin, 1993: Fast Fourier transforms for nonequispaced data. *SIAM J. Sci. Comput.*, **14**, 1368–1393, <https://doi.org/10.1137/0914081>.
- Garreaud, R., and J. M. Wallace, 1997: The diurnal march of convective cloudiness over the Americas. *Mon. Wea. Rev.*, **125**, 3157–3171, [https://doi.org/10.1175/1520-0493\(1997\)125<3157:TDMOCC>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<3157:TDMOCC>2.0.CO;2).
- Hannachi, A., I. Jolliffe, and D. Stephenson, 2007: Empirical orthogonal functions and related techniques in atmospheric science: A review. *Int. J. Climatol.*, **27**, 1119–1152, <https://doi.org/10.1002/joc.1499>.
- Hasselmann, K., 1988: Pips and pops: The reduction of complex dynamical systems using principal interaction and oscillation patterns. *J. Geophys. Res.*, **93**, 11 015–11 021, <https://doi.org/10.1029/JD093iD09p11015>.
- Hirose, M., R. Oki, S. Shimizu, M. Kachi, and T. Higashiuwatoko, 2008: Finescale diurnal rainfall statistics refined from eight years of TRMM PR data. *J. Appl. Meteor. Climatol.*, **47**, 544–561, <https://doi.org/10.1175/2007JAMC1559.1>.
- Horel, J. D., 1984: Complex principal component analysis: Theory and examples. *J. Climate Appl. Meteor.*, **23**, 1660–1673, [https://doi.org/10.1175/1520-0450\(1984\)023<1660:CPCATA>2.0.CO;2](https://doi.org/10.1175/1520-0450(1984)023<1660:CPCATA>2.0.CO;2).
- Huffman, G. J., and Coauthors, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeorol.*, **8**, 38–55, <https://doi.org/10.1175/JHM560.1>.
- Khairoutdinov, M., and D. Randall, 2006: High-resolution simulation of shallow-to-deep convection transition over land. *J. Atmos. Sci.*, **63**, 3421–3436, <https://doi.org/10.1175/JAS3810.1>.
- Kikuchi, K., and B. Wang, 2008: Diurnal precipitation regimes in the global tropics. *J. Climate*, **21**, 2680–2696, <https://doi.org/10.1175/2007JCLI2051.1>.
- Kim, K.-Y., and Q. Wu, 1999: A comparison study of EOF techniques: Analysis of nonstationary data with periodic statistics. *J. Climate*, **12**, 185–199, <https://doi.org/10.1175/1520-0442-12.1.185>.
- , G. R. North, and J. Huang, 1996: EOFs of one-dimensional cyclostationary time series: Computations, examples, and stochastic modeling. *J. Atmos. Sci.*, **53**, 1007–1017, [https://doi.org/10.1175/1520-0469\(1996\)053<1007:EOODCT>2.0.CO;2](https://doi.org/10.1175/1520-0469(1996)053<1007:EOODCT>2.0.CO;2).
- Leroy, B., 2012: Fast calculation of the Lomb-Scargle periodogram using nonequispaced fast Fourier transforms. *Astron. Astrophys.*, **545**, A50, <https://doi.org/10.1051/0004-6361/201219076>.
- Lomb, N. R., 1976: Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci.*, **39**, 447–462, <https://doi.org/10.1007/BF00648343>.
- Lupo, A. R., J. J. Nocera, L. F. Bosart, E. G. Hoffman, and D. J. Knight, 2001: South American cold surges: Types, composites, and case studies. *Mon. Wea. Rev.*, **129**, 1021–1041, [https://doi.org/10.1175/1520-0493\(2001\)129<1021:SACSTC>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<1021:SACSTC>2.0.CO;2).
- Mapes, B. E., T. T. Warner, M. Xu, and A. J. Negri, 2003: Diurnal patterns of rainfall in northwestern South America. Part I: Observations and context. *Mon. Wea. Rev.*, **131**, 799–812, [https://doi.org/10.1175/1520-0493\(2003\)131<0799:DPORIN>2.0.CO;2](https://doi.org/10.1175/1520-0493(2003)131<0799:DPORIN>2.0.CO;2).
- Mathias, A., F. Grond, R. Guardans, D. Seese, M. Canela, H. H. Diebner, and G. Baiocchi, 2004: Algorithms for spectral analysis of irregularly sampled time series. *J. Stat. Software*, **11**, 1–30, <https://doi.org/10.18637/jss.v011.i02>.
- Monahan, A. H., F. T. Tangang, and W. W. Hsieh, 1999: A potential problem with extended EOF analysis of standing wave fields. *Atmos.–Ocean*, **37**, 241–254, <https://doi.org/10.1080/07055900.1999.9649628>.
- Negri, A., L. Xu, and R. Adler, 2002: A TRMM-calibrated infrared rainfall algorithm applied over Brazil. *J. Geophys. Res.*, **107**, 8048, <https://doi.org/10.1029/2000JD000265>.
- Nickolls, J., I. Buck, M. Garland, and K. Skadron, 2008: Scalable parallel programming with CUDA. *Queue*, **6**, 40–53, <https://doi.org/10.1145/1365490.1365500>.
- North, G. R., 1984: Empirical orthogonal functions and normal modes. *J. Atmos. Sci.*, **41**, 879–887, [https://doi.org/10.1175/1520-0469\(1984\)041<0879:EOFANM>2.0.CO;2](https://doi.org/10.1175/1520-0469(1984)041<0879:EOFANM>2.0.CO;2).
- , T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110**, 699–706, [https://doi.org/10.1175/1520-0493\(1982\)110<0699:SEITEO>2.0.CO;2](https://doi.org/10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2).
- Press, W. H., and G. B. Rybicki, 1989: Fast algorithm for spectral analysis of unevenly sampled data. *Astrophys. J.*, **338**, 277–280, <https://doi.org/10.1086/167197>.
- Pritchard, M. S., and R. C. J. Somerville, 2009: Assessing the diurnal cycle of precipitation in a multi-scale climate model. *J. Adv. Model. Earth Syst.*, **1**, 12, <https://doi.org/10.3894/JAMES.2009.1.12>.
- 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, <https://doi.org/10.1086/160554>.
- , 1989: Studies in astronomical time series analysis. III—Fourier transforms, autocorrelation functions, and cross-correlation

- functions of unevenly spaced data. *Astrophys. J.*, **343**, 874–887, <https://doi.org/10.1086/167757>.
- Schulz, M., and K. Stattegger, 1997: SPECTRUM: Spectral analysis of unevenly spaced paleoclimatic time series. *Comput. Geosci.*, **23**, 929–945, [https://doi.org/10.1016/S0098-3004\(97\)00087-3](https://doi.org/10.1016/S0098-3004(97)00087-3).
- Steidl, G., 1998: A note on fast Fourier transforms for nonequispaced grids. *Adv. Comput. Math.*, **9**, 337–352, <https://doi.org/10.1023/A:1018901926283>.
- Townsend, R. H. D., 2010: Fast calculation of the Lomb–Scargle periodogram using graphics processing units. *Astrophys. J., Suppl. Ser.*, **191**, 247–253, <https://doi.org/10.1088/0067-0049/191/2/247>.
- Venugopal, V., K. Virts, J. Sukhatme, J. Wallace, and B. Chattopadhyay, 2016: A comparison of the fine-scale structure of the diurnal cycle of tropical rain and lightning. *Atmos. Res.*, **169**, 515–522, <https://doi.org/10.1016/j.atmosres.2015.09.004>.
- Wu, P., J.-I. Hamada, M. D. Yamanaka, J. Matsumoto, and M. Hara, 2009: The impact of orographically-induced gravity waves on the diurnal cycle of rainfall over southeast Kalimantan Island. *Atmos. Ocean. Sci. Lett.*, **2**, 35–39, <https://doi.org/10.1080/16742834.2009.11446773>.
- Yang, G.-Y., and J. Slingo, 2001: The diurnal cycle in the tropics. *Mon. Wea. Rev.*, **129**, 784–801, [https://doi.org/10.1175/1520-0493\(2001\)129<0784:TDCITT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0784:TDCITT>2.0.CO;2).
- Zechmeister, M., and M. Kürster, 2009: The generalised Lomb–Scargle periodogram: A new formalism for the floating-mean and Keplerian periodograms. *Astron. Astrophys.*, **496**, 577–584, <https://doi.org/10.1051/0004-6361/200811296>.