2B.10 A NEW METHOD FOR TIME SERIES FILTERING NEAR ENDPOINTS

Anthony Arguez *
NOAA National Climatic Data Center, Asheville, North Carolina

Peng Yu and James J. O'Brien
Center for Ocean-Atmospheric Prediction Studies, Tallahassee, Florida

1. INTRODUCTION

Time series filtering can be done in the spectral domain without loss of endpoints. However, filtering is commonly performed in the time domain using convolutions, resulting in lost points near the series termini. Multiple incarnations of a least squares minimization approach are developed that retain the endpoint intervals that are normally discarded due to filtering with convolutions in the time domain. The techniques minimize the errors between the pre-determined frequency response function (FRF) of interior points with FRF's that are to be determined for each position in the endpoint zone. The least squares techniques are differentiated by their constraints: (1) unconstrained, (2) equal-mean constraint, and (3) an equal-variance constraint. The equal-mean constraint forces the new weights to sum up to the same value as the pre-determined weights. The equal-variance constraint forces the new weights to be such that, after convolution with the input values, the expected variance is identical to the expected variance of the interior points.

2. TYPICAL FILTERING METHOD AND THE NEW LEAST SQUARES APPROACH TO RETAINING ENDPOINTS

Consider an input time series, *x(t)*, that is to be filtered to produce *y(t)*, the output time series. Filtering in frequency space is accomplished by applying a Fourier Transform to the time series (*X*), multiplying by the frequency response function (FRF; *H*), and back- transforming into time space. The FRF is defined, using filter weights denoted as $h(\tau)$, as follows:

$$H(f) = \sum_{\tau=a}^{\tau=b} h(\tau) e^{i2\pi f\tau} \qquad (1)$$

Typically, however, time series filtering is computed in time space. This requires a convolution between the input time series and the filter weights:

$$y(t) = \sum_{\tau=a}^{\tau=b} h(\tau) x(t+\tau) \qquad (2)$$

The filter weights are determined a priori by the user based on the spectral characteristics one wants to transfer to the output. The parameters *a* and *b* are integers that are usually chosen such that *a* is the negative of *b*. The parameter τ represents a time lag. Filtering in the time domain results in lost points in the left and right endpoint intervals; this is a consequence of the convolution. Specifically, the convolution cannot be defined at the first *b* points and the last *b* points of the time series. In these regions, at least one of the lags is associated with an unavailable point in the time series (points beyond the terminal values). These points where the full convolution cannot be computed are customarily dropped from consideration. For example, if a monthly ENSO time series is defined using a five-month running average, and SST data are available through December 2006, the last filtered value that can be computed with full convolutions is Oct. 2006 (in this case b=2); Dec. 2006 value cannot be computed until Feb. 2007 value is available.

This dilemma leads to the following question: can variable-length filter weights be determined in the endpoint intervals such that the signal extraction replicates that observed in the interior points? For this project, a penalty function is constructed to minimize the squared error between the FRF in the interior and a new FRF in each point of the endpoint intervals. Following (1) the FRF of interior points are denoted as follows:

* *Corresponding author address:* Anthony Arguez, National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28801; email: Anthony.Arguez@noaa.gov.

$$G(f) = \sum_{k=-b}^{b} \alpha(k) e^{i2\pi fk} \quad (3)$$

where $\vec{\alpha}$ represents the new filter weights to be determined. The value of *b* is incrementally decreased to force a symmetric filter at each point (i.e. the number of filter weights for a particular point in the time series will always equal twice the distance to the terminal end plus one). The result is a set of filter weights for each point in the endpoint interval. Without constraints, this minimization technique truncates the interior filter weights (see Bloomfield 2000).

A practical constraint is to force the new weights to sum to 1 (see Arguez et al. 2005), thereby preserving the expectation of the mean of the output series. The cost function for a given point *L* in the endpoint interval is a function of $\vec{\alpha}$ and λ (the Lagrange multiplier that imposes the constraint):

$$J(\vec{\alpha},\lambda) = \sum_f [H(f) - G(f)]^2 + \lambda[\sum_{\tau=-b}^{b} h(\tau) - \sum_{k=-L}^{L} \alpha(k)] \quad (4)$$

An alternative, albeit more complicating, constraint is to force the total variance to be preserved. Variance is related to the spectrum via the following relation:

$$\sigma_y^2 = \int S_{yy}(f) df \quad (5)$$

where S_{yy} is the raw power spectrum of *y(t)*. The following equality will be imposed to insure that total variance in the interior will match the total variance of each point in the endpoint interval:

$$\sum_f H(f)H^*(f)X(f)X^*(f) = \sum_f G(f)G^*(f)X(f)X^*(f) \quad (6)$$

The asterisks indicate complex conjugates. Note that the constraint is dependent on the input time series, unlike all other methods employed in the present study. Modifying (4) to impose the new constraint, and incorporating the definition in (6), results in the new cost function:

$$J(\vec{\alpha},\lambda) = \left(\sum_f \; [H(f) - \sum_{k=-L}^{L} \alpha_k e^{i2\pi fk}]^2 \right) + \lambda\left(T - \sum_f \left(S_{xx}(f) \left[\sum_{j=-L}^{L} \sum_{m=-L}^{L} \alpha_j \alpha_m e^{i2\pi f(j-m)} \right] \right) \right) \quad (7)$$

where *T* is the variance of interior points and is proportional to the left hand side of (6). The equations are solved using the Newton method available in the IDL programming language. We term this method the equal-variance method.

3. PRELIMINARY RESULTS

The 3 least squares methods are each tested under three separate filtering scenarios involving time series of well-known climate modes: the Arctic Oscillation (AO), the Madden Julian Oscillation (MJO), and the El Niño Southern Oscillation (ENSO). For each case, the least squares methods are compared to each other as well as to the spectral filtering method – the standard of comparison. This is accomplished by using several thousand simulated time series of each climate mode, which are computed by imposing the FRF of each climate series onto white noise time series and back-transforming into the time domain. These artificially-created time series are intentionally truncated to utilize the above methods to compute the estimations near the endpoints (termed the estimation zone), but the discarded points are used to compute the true filtered endpoints (using full convolutions). RMS errors and the variance of estimates for each endpoint position are calculated to measure the viability of all endpoint techniques.

The results indicate that all 4 methods (including the spectral method) possess skill at determining suitable endpoints estimates. However, both the unconstrained and equal-mean schemes exhibit bias toward zero near the terminal ends due to problems with appropriating variance. The equal-variance method does not show evidence of this attribute and was never the worst performer. The equal-variance method showed great promise in the ENSO project involving a 5-month running mean filter, with regard to both RMSE (Fig. 1) and appropriating variance (Fig. 2).

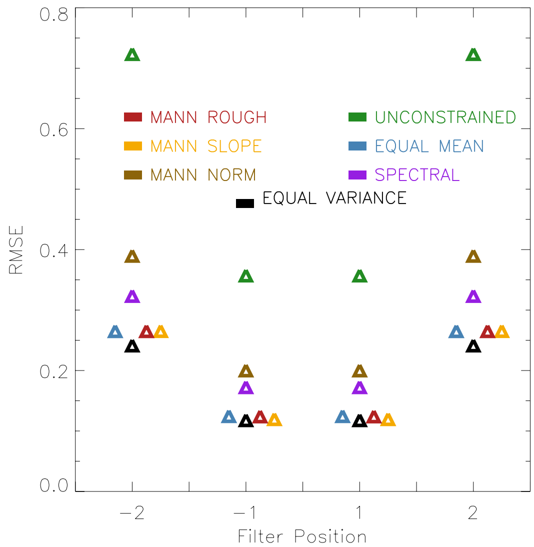

Figure 1: ENSO RMS Errors. Root mean square errors as a function of filter position for the unconstrained, equal-mean, equal-variance, norm, slope, roughness, and spectral methods for the ENSO project (top). A liberal skill threshold based on random filter output is an RMSE value of about 0.4.

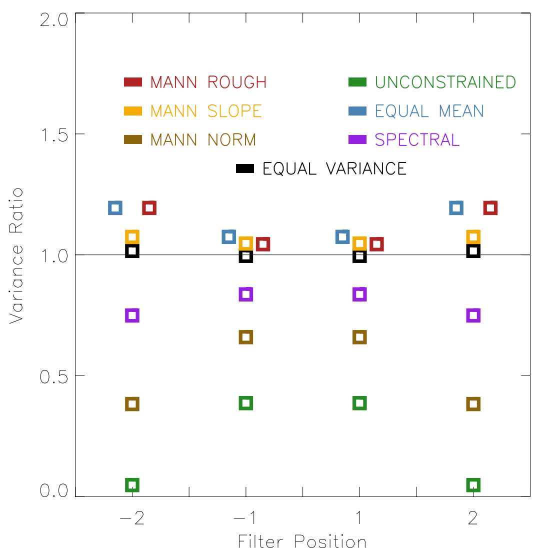

Figure 2: ENSO Endpoint Variances. Variance of estimated endpoint values as a function of endpoint location.

The equal-variance method performed at least on par with the other methods for almost all time series positions in all three filtering scenarios. Results were also compared to the 3 boundary constraints (norm, slope, and roughness) utilized in Mann (2004) for smoothing non-stationary time series. Although these three constraints occasionally exhibited the least RMS errors, the outputs were plagued by misappropriated variances.

4. REFERENCES

Arguez, A., M. A. Bourassa, and J. J. O'Brien, 2005: Extraction of the MJO signal from QuikSCAT. *Journal of Atmospheric and Oceanic Technology*, **22**, 1885-1894.

Bloomfield, P., 2000: *Fourier Analysis of Time Series – An Introduction*. John Wiley and Sons, Inc. 261 pp.

Mann, M. E., 2004: On smoothing potentially non-stationary climate time series. *Geophys. Res.Lett.*,**31**,L07214,doi:10.1029/2004GL019569.