Advances in Adaptive Data Analysis
Vol. 1, No. 3 (2009) 447–460

REDUCTIONS OF NOISE AND UNCERTAINTY IN ANNUAL GLOBAL SURFACE TEMPERATURE ANOMALY DATA

NORDEN E. HUANG

Research Center for Adaptive Data Analysis
National Central University
Chungli 32001, Taiwan, ROC

ZHAOHUA WU

Department of Meteorology
Florida State University
Tallahassee, FL 32306-4520, USA

JORGE E. PINZÓN

Science Systems and Applications, Inc.
Laboratory for Terrestrial Physics, NASA Goddard Space Flight Center
Greenbelt, Maryland 20771, USA

CLAIRE L. PARKINSON

Laboratory for Hydrospheric Processes, Goddard Space Flight Center
Greenbelt, Maryland 20771, USA

STEVEN R. LONG

Laboratory for Hydrospheric Processes, Wallops Flight Facility
Wallops Island, Virginia 23337, USA

KARIN BLANK

Microelectronic and Signal Processing Branch, Goddard Space Flight Center
Greenbelt, Maryland 20771, USA

PER GLOERSEN

Laboratory for Hydrospheric Processes, Goddard Space Flight Center
Greenbelt, Maryland 20771, USA

XIANYAO CHEN

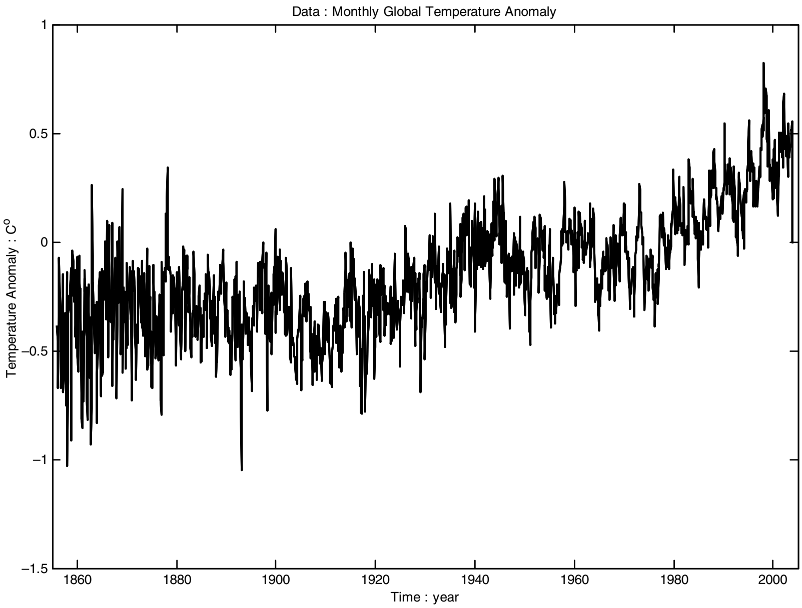
Laboratory of Marine Science and Numerical Modeling
First Institute of Oceanography
State Oceanic Administration
Qingdao, Shandong 266061, China

Global climate variability is currently a topic of high scientific and public interest, with potential ramifications for the Earth's ecologic systems and policies governing world economy. Across the broad spectrum of global climate variability, the least well understood time scale is that of decade-to-century.[1] The bases for investigating past changes across that period band are the records of annual mean Global Surface Temperature Anomaly (GSTA) time series, produced variously in many painstaking efforts.[2–5] However, due to incipient instrument noise, the uneven distribution of sensors spatially and temporally, data gaps, land urbanization, and bias corrections to sea surface temperature, noise and uncertainty continue to exist in all data sets.[1,2,6–8] Using the Empirical Mode Decomposition method as a filter, we can reduce this noise and uncertainty and produce a cleaner annual mean GSTA dataset. The noise in the climate dataset is thus reduced by one-third and the difference between the new and the commonly used, but unfiltered time series, ranges up to 0.1506°C, with a standard deviation up to 0.01974°C, and an overall mean difference of only 0.0001°C. Considering that the total increase of the global mean temperature over the last 150 years to be only around 0.6°C, we believe this difference of 0.1506°C is significant.

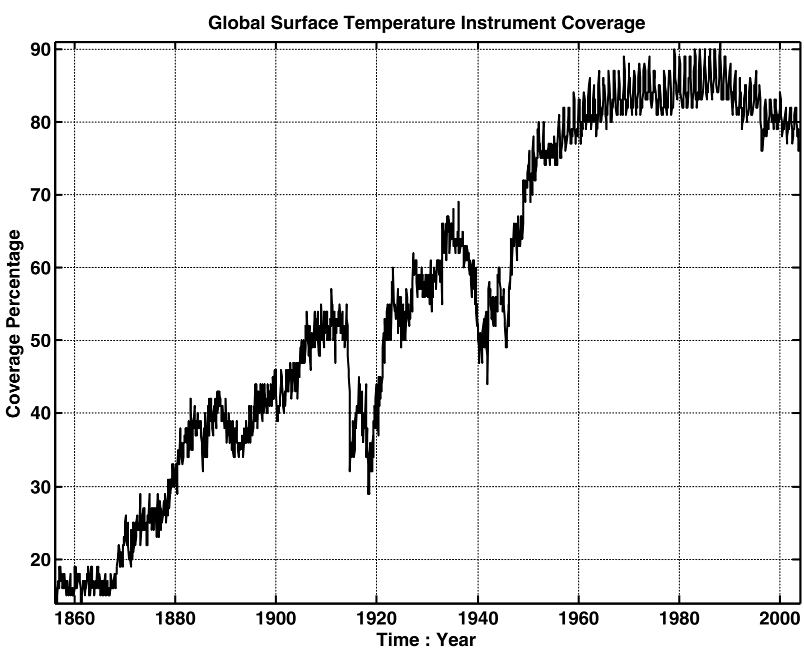
Keywords: Global temperature change; down sampling; HHT filtering.

1. Introduction

Currently used annual Global Surface Temperature Anomaly (GSTA) datasets are all based on the instrument records of monthly global surface temperature over the past 150 years. Traditionally, to highlight the variability, a climatologic mean seasonal cycle (defined as the mean from 1961 to 1990) is removed to produce the monthly GSTA data. An annual mean GSTA calculated by simple yearly averaging has been used in most climate variability studies.[1,6] In order to make the time series as clean and free of noise and error as possible, several efforts have further assessed and calibrated the time series, resulting in a considerable reduction in the uncertainties.[7,8] Nevertheless, the record may still not be as clean as is technically possible, so we take a new approach to processing the dataset. The GSTA time series that we consider is based on the well known monthly time series prepared by Jones *et al.*[2] (posted by the Climate Research Unit (CRU), University of East Anglia, UK, at the website maintained jointly by the CRU and the UK Meteorological Office Hadley Centre), extensively referred in the Intergovernmental Panel on Climate Change (IPCC), Fourth Assessment Report.[6] In this study, we use the data covering 1856 to the end of 2004 for a total of 148 years. This CRU-GSTA time series, together with the percentage coverage of the globe, is shown in Figs. 1(a) and 1(b). Clearly, the percentage of instrument coverage of the global surface is quite uneven in time as shown in Fig. 1(b). The data are under sampled during many periods, particularly in the early stage of the CRU-GSTA (hereafter CRU) time series. Additionally, the two periods during the World Wars also caused the coverage to decrease. Other than the temporal nonuniformity, the spatial coverage is also uneven, but it is not shown in this presentation. With such uneven coverage, we should expect problems in computing a fair mean value of the data to indicate the temporal change of the global temperature. Furthermore, the effects of the



(a)



(b)

Fig. 1. Monthly global surface temperature anomalies, GSTA (a) and the percent data coverage of the global area (b). The anomalies are calculated using the 30-year period 1961–1990 as the baseline. (Data from the University of East Anglia Climate Research Unit website at http://www.cru.uea.uk/cru/data/temperature/).

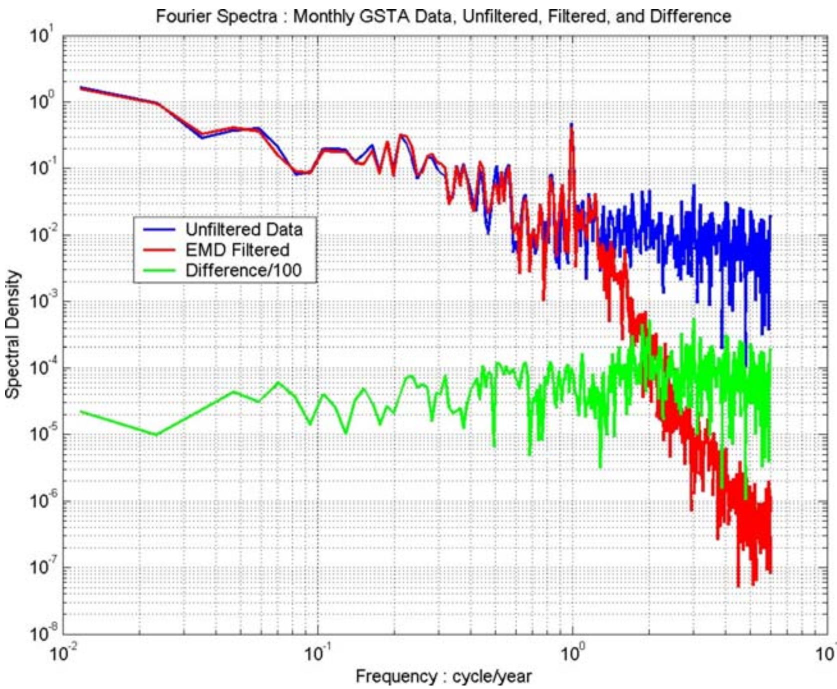


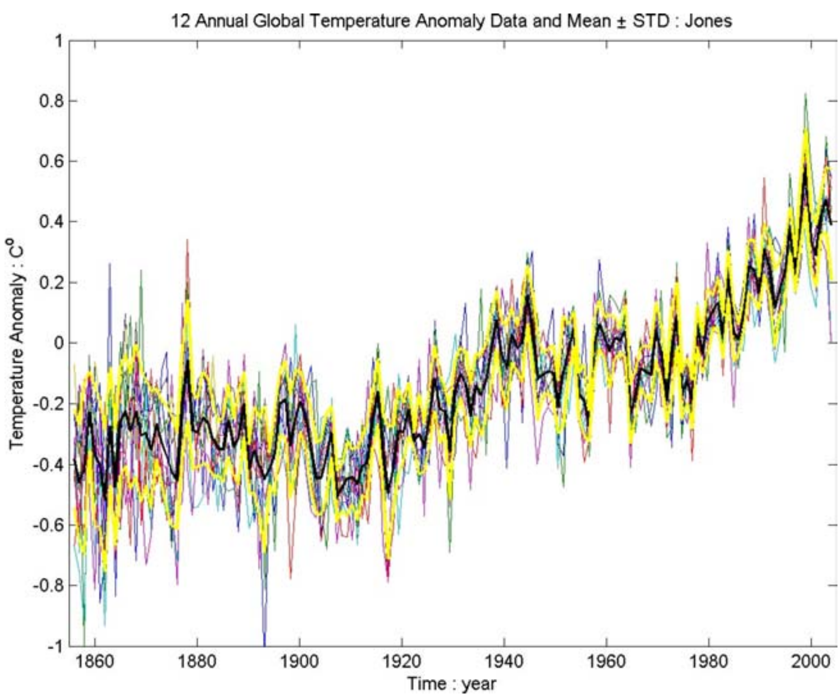
Fig. 2. The Fourier Spectra of the original monthly GSTA (blue), the EMD-filtered data (red), and their difference (green). The EMD-filtered data and the difference are plotted in Fig. 5. The substantive differences between the original and the EMD-filtered spectra are all in the high frequency range. The spectrum of the difference is essentially white, indicating that no useful information has been filtered out.

unknown source of noise in the mean could also cause problems, if the data are analyzed without special attention to reduce these effects to the minimum.

Let us examine the potential problems of the CRU time series quantitatively. First, the Fourier Spectrum (FS) of the CRU series, presented in Fig. 2, indicates that CRU still contains a visible annual cycle and some discernible semi-annual cycle harmonics signal even though a climatologic annual cycle has already been removed. The FS also indicates that the dataset provides little useful information at frequencies higher than the annual cycle, for that part of the spectrum is essentially flat. Next we examine the annual mean, defined as the simple arithmetic mean over the period of a year; it is essentially the mean of the 12 sets of time series each representing a particular month of the year: for example, the temperature of all Januarys and Februarys, and so on. As the annual cycle had been removed, these time series of specific month should be similar and without seasonal variations. Indeed, from the CRU down sampled data, 12 mean and standard deviation values, each representing the mean of a specific month over the 148 years can be computed. These 12-to-1 down sampling time series of the specific monthly data are shown in Fig. 3(a). Indeed the seasonal variation is not noticeable here.

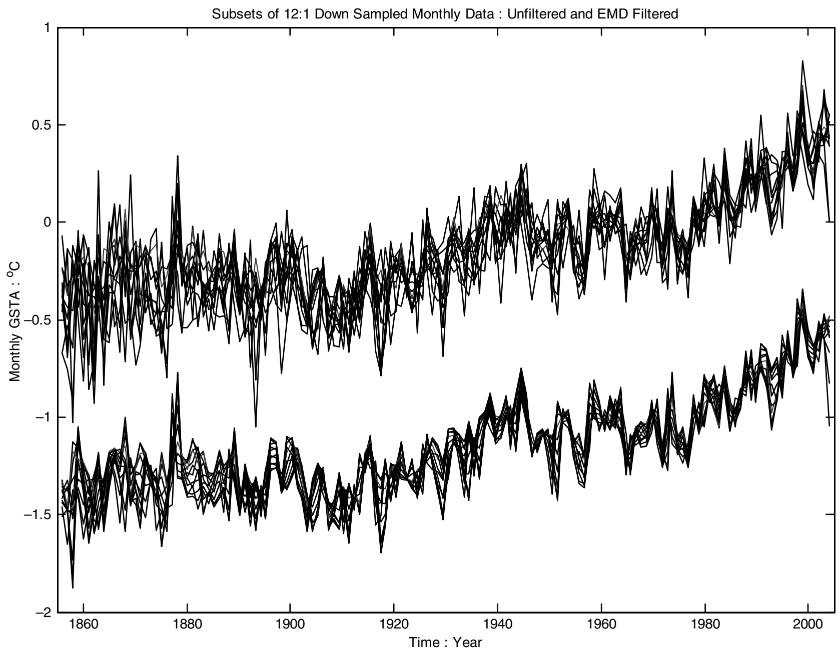
It should be noted that, in down sampling of data, higher frequency noise should always be removed first.[9,10] As the data could contain noise, uneven distribution temporally and spatially, and the data also suffer incomplete removal of annual cycle; therefore, each of down sampled series would suffer from aliasing effects. There is no guarantee that a simple arithmetic mean of a cluster of monthly data could cancel out the alias and provides a clean annual mean. From the variance of the 12 monthly series shown in Fig. 3(c), we can see that the aliasing effects are pretty serious for part of the data especially at the beginning of the dataset.

As we are deriving the mean from a 12:1 down sampled data, variation with a period shorter than 12 month should be removed before the down sampling to avoid aliasing, i.e. to avoid this down sampling induced alias, we first have to filter

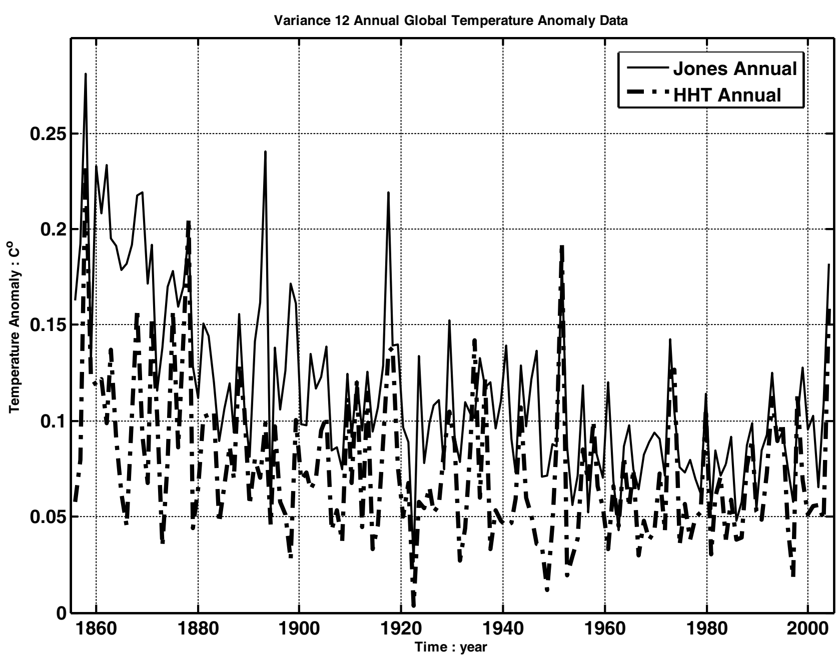


(a)

Fig. 3. (a) The 12:1 down sampled GSTA data plotted for each of the 12 months of the year (given in the colored thin lines). The cluster represents the 12 subsets from the original monthly data; the arithmetic mean (given in the broad black line) is the annual mean temperature used in most climate studies. The standard deviation values are given in the solid yellow line. (b) The filtered and unfiltered data of the 12:1 down sampled series for specific month. Note the reduction of the data spreading of the filtered data than the original data. (c) The Variance computed for the cluster of 12 down sampled monthly dataset (thin solid line) and the HHT smoothed set (bold dash-dot line). Note the grossly uneven distributed variance as a function of time for the original data. The variance for the filtered data has near a factor 2 reduction in the variance at the beginning of the data series and the variance values are almost uniformly distributed through out the time span.



(b)



(c)

Fig. 3. (*Continued*)

the data. The traditional method to remove noise in a dataset is through Fourier analysis-based filtering in frequency space.[9,10] Although the Fourier-based filtering might be useful for many other applications, the method has shortcomings and is not applicable here. Fundamentally, as climate variations involve nonstationary and nonlinear processes, the underlying assumptions of stationarity and linearity of Fourier Analysis are violated. The possible consequences of such violations are several. First, the Fourier filter will remove higher harmonics associated with nonlinear waveforms of the fundamentals, thereby distorting the low frequency fundamental waveforms. Second, temporally local noises in the data have wide spectra in frequency space; therefore, the low pass filter will never remove all of their influence. As an alternative, we employ the newly developed adaptive time domain Hilbert–Huang Transform (HHT) method as a filter. HHT has been described in detail in previous publications.[11–15]

Now we decided to use HHT as filter to remove any wave component with period shorter than 12 months. The reason is simple: Fourier filter is based on linear and stationary assumption. For any nonlinear and nonstationary data, we would inviolably cut-off the harmonics of longer period components, such as the semi-annual cycle components. Using HHT, however, we can devise a time space filtering.[16] For example, a low pass filtered results of a signal having n-intrinsic mode function (IMF) components can be simply expressed as

$$X_{lk}(t) = \sum_{k}^{n} c_j + r_n; \tag{1}$$

a high pass results can be expressed as

$$X_{hk}(t) = \sum_{1}^{k} c_j; \tag{2}$$

and a band pass result can be expressed as

$$X_{bk}(t) = \sum_{b}^{k} c_j. \tag{3}$$

The advantage of this time space filtering is that the results preserve the full nonlinearity and nonstationarity in the physical space.

To implement this filtering procedure, we first perform the Empirical Mode Decomposition with the intermittence test.[12,13] The resulting IMF components are given in Fig. 4. Here we have completely exhausted the shorter period IMF components: in the first four IMF for periods shorter than 4 months, and the subsequent 5th to 13th IMF components for periods shorter than 8 months. As the annual cycle could also vary, we decide to give ample space between the cut-off period and the annual cycle. IMF components from 14th to 21st are for period equal or longer than a year. The last IMF represents the over all trend of the global temperature change. Its properties had been discussed separately.[17]

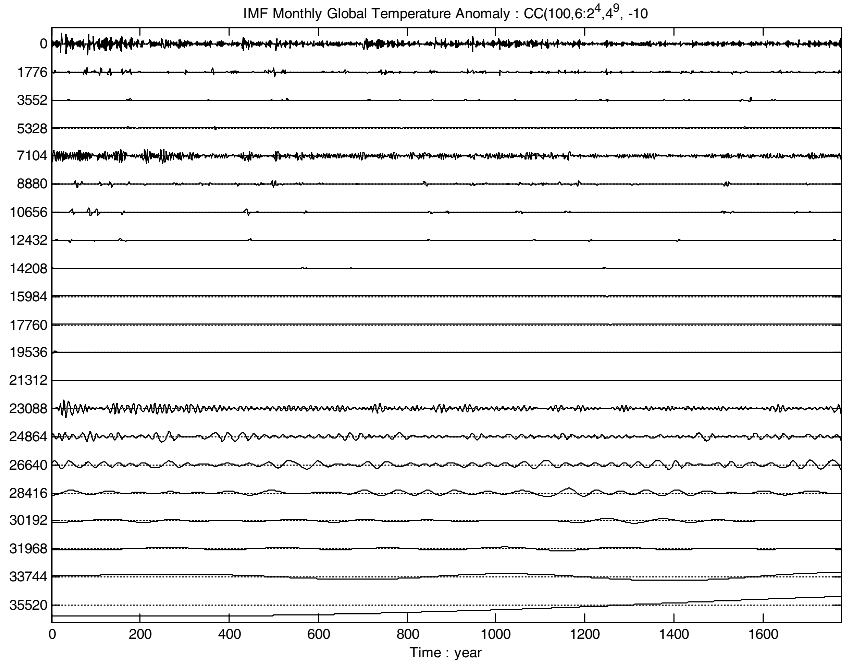


Fig. 4. The Intrinsic Mode Function (IMF) components of the monthly CRU data using the intermittency test to separate the IMF into 4 and 8 months components (the first 4 IMF and the subsequent 5th to 13th IMF components, respectively).

To examine in detail the effect of HHT filtering, we plot the Fourier spectra of the EMD-filtered data and of removed noise as well also in Fig. 2. The spectral magnitudes of the original and filtered data are almost identical over the range from the low frequency end all the way up to the annual cycle. All substantial differences are in the higher frequency range: the spectrum of the filtered data shows a drastic decrease in energy immediately beyond the annual cycle, while the spectral density suffers negligible degradation in the low frequency range. Meanwhile, the spectrum of the removed noise is nearly white, an indication that no useful information has been removed. Although we have only removed oscillations with periods shorter than 8 months in the time domain, there is still energy in the frequency range lower than the annual cycle. The residue low frequency energy is due to the local property of the noise temporal but not in frequency space, for local in the temporal space is uniform in the frequency space. Therefore, even the short period noise removed could yield a near white spectrum. Consequently, the noise could not be removed by a low pass Fourier filter.

As our goal is to produce a cleaner annual GSTA dataset, we have removed any variations with periods less than 8 months to avoid down sampling alias. Also since we want to derive the annual mean, any period shorter than the annual cycle is not of interest. Then, Fourier analysis indicates that there is no information higher than the annual cycle (Fig. 2). An independent study indicated that the IMFs having mean periods shorter than a year of the original monthly CRU-GSTA are not statistically significant.[16,17,18,19] Finally, recognizing that the time of yearly temperature maxima/minima can fluctuate by weeks, if not months, we conservatively select an 8-month cut-off rather than a 9–12-month cut-off, to avoid excess filtering. The difference between the filtered and unfiltered monthly data is shown in Fig. 5, with the difference plotted with an off-set of 1.5°C. The result shows that the difference of individual monthly data series is substantial, with a peak-to-peak range of 1°C. The magnitude of the removed noise is larger at the beginning of the dataset and also slightly, but significantly, larger over the two World Wars than in their neighboring periods, in line with the paucity of observations in the 1800s, and with the relative irregularity of observations embracing the periods of the two World Wars.

From the HHT-filtered database, the variance of the 12 down sampled monthly subsets of data can be computed. The results are shown in Fig. 3(c). As each member of the cluster of the 148-point HHT-filtered time series has a lower alias error than the original cluster, we can expect that the mean derived would also be less influenced by the alias error. Indeed this effect could be seen from the variance values. The variance for the filtered data has near a factor of two reduction in the variance values at the beginning of the data series. Furthermore, the variance values

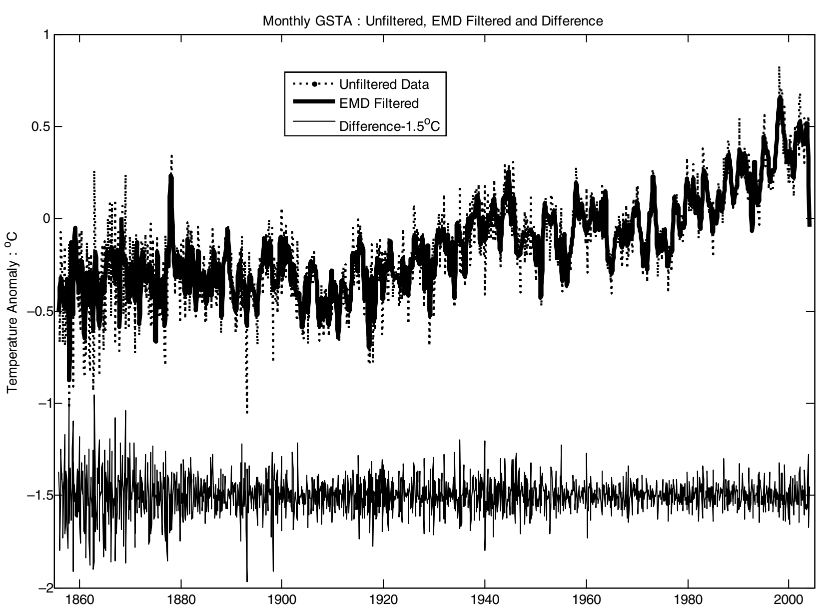


Fig. 5. The original monthly GSTA (dotted), the HHT-filtered values (solid bold), and their differences (solid thin) off-set by 1.5°C. These monthly data are the base for the annual GSTA computation. Notice that the difference is quite large especially at the beginning of the data coverage period, which reflects the paucity of observations in the early period. Other periods for the large differences are during the World Wars just before 1920 and just after 1940.

are almost uniformly distributed through out the time span, while the distribution of the original data shows a clear uneven distribution reflecting the paucity of data at the early period. As the variance is an indicator of confidence limits, these results indicate that there is a reduction of the uncertainty of the climate variability by about a third.

The difference between the mean of the filtered and original data is also shown in Fig. 6. The difference of the two spans a range up to 0.1506°C with a variance of the difference of 0.01974°C. Interestingly, the overall mean of the difference is only 0.0001°C, indicating that the new method has not introduced any bias through the data processing procedures, and that the differences are all due to the uneven data distribution and random noises.

As a check of the effect of the filtering, we examine the seasonal variations of the data next. Theoretically, once the annual cycles are removed, there should not be any seasonal variations left. Unfortunately, the traditional method of removing the annual cycle depends on a "climatologic mean" derived from 30 years average

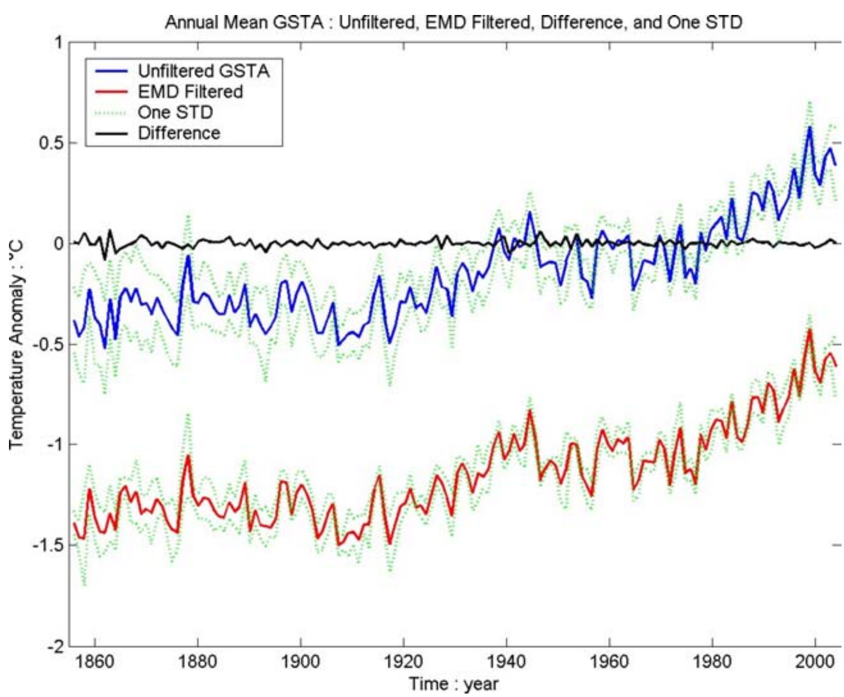


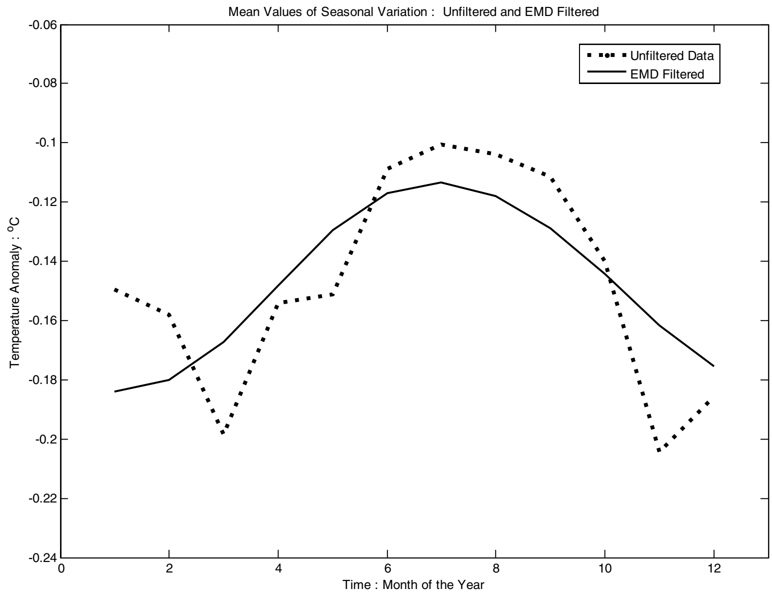
Fig. 6. The unfiltered (blue) and the EMD-filtered (red) annual mean GSTA are shown with one standard deviation uncertainty bounds (green dash) and their difference (black). The filtered data have been shifted by −1°C. Notice that the standard deviation values are uneven, with the largest values at the beginning of the data period due to smaller percent data coverage. Notice also that the standard deviation values are much smaller for the EMD-filtered data. As the confidence is measured by the standard deviation values, this filtering represents an overall improvement of the uncertainty in the mean and the variability by a third.

covering 1961–1990. The shortcomings of this method of annual cycle removal had been studied thoroughly elsewhere.[20] The correct way for removing the annual cycle should relay on an adaptively defined annual variation rather than a fixed annual mean, for the climate changes are nonstationary and the annual cycle would also modulate from year to year. A clear annual peak in the FS as shown in Fig. 2 is a clear indication of incomplete removal of the annual cycle. We will use the residual annual cycle to check the effect of the filter here. By computing the mean and variance of each of the twelve 12-to-1 down sampled specific monthly time series subsets, we should be able to see the seasonal, or month to month changes. The results of this computation are shown in Figs. 7(a) and 7(b).

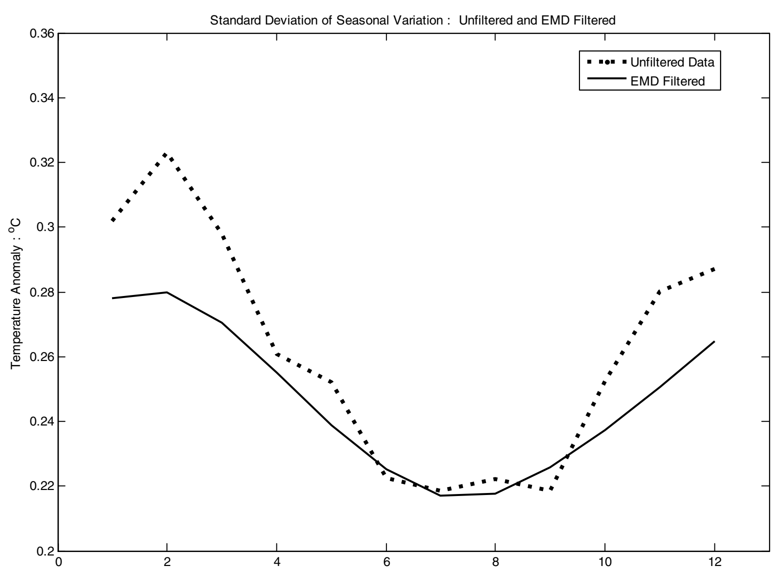
The results from the original are presented as the blue lines in Fig. 7(a) for mean values and Fig. 7(b) for the variance. Means and variance both show a general seasonal variation, with high mean values and low variance in the boreal summer and low mean values and high variance in the boreal winter months. The existences of these seasonal variations indicate two problems as discussed above: First is the imperfect removal of the seasonal cycle. As the climate variations involve nonlinear and nonstationary processes that result in varying seasonal cycles in both strength and phase, it is impossible to remove the seasonal cycle using any local temporal mean seasonal cycle. Second alias in the data is quite severe, so much so that the residual seasonal trend shown here is only correct in general, and quite irregular in the details.

Next, we also computed the mean and the variance of the corresponding seasonal variations from the HHT-filtered data, which is also shown in Fig. 7. Now, both the seasonal variation and the SD values are smooth, regular, and physically reasonable. As the long-term average seasonal cycle reflects the effect of the earth rotation axial inclination relative to the plane of the ecliptic, it should be regular and smooth after averaging over some 150 years. This test offers another direct evidence of the improvement offered by the HHT-filtered time series.

We have used the HHT filter on the extremely important long-term (1856–2004) global monthly surface temperature anomaly data to produce a clean annual dataset by removing alias errors associated with down sampling before computing the mean. We attribute the irregularity of the seasonal variation primarily to the alias error caused by the random noise and also from sparse coverage in the original data, especially near the beginning of the dataset in the mid-1800s. By using a temporally based filter, we have removed a substantial portion of these errors and obtained a cleaner global annual temperature anomaly dataset that reduces the uncertainty of the mean as well as the variability by about a third. Understanding the variations in the Earth's climate, as registered in the time series of planetary surface temperatures is fundamental to properly address present and future issues related to the potential impacts of climate variations on natural and human systems. Here we suggest that this HHT filtered, cleaner dataset should be considered as the basis for climate investigations. We further suggest that the annual cycle should also be removed using the adaptive method as discussed elsewhere.[20] Unfortunately, we



(a)



(b)

Fig. 7. (a) The average seasonal cycle of global temperatures over the period 1856–2004 based on two methods: First, averaging the 148 monthly anomaly values for each of the 12 months (dotted line); and second, the HHT-filtered data (solid line). All of the anomalies are negative because of the choice, by Jones *et al.*,[2] of 1961–1990 as the baseline period. (b) The standard deviations of the monthly GSTA data. The variation of the original data (dotted line) is somewhat irregular, with the highest standard deviation in February and the lowest in September. The variation of the HHT-filtered data (solid line) shows a regular and smooth variation.

do not have the original dataset before the annual cycle had been removed. It would be a worthy project for the climatologists in the future.

Acknowledgments

Norden E. Huang was supported by a grant from Federal Highway Administration, DTFH61-08-00028, and grants from NSC, NSC95-2119-M-008-031-MY3, NSC97-2627-B-008-007, and finally a grant from NCU 965941 that have made the conclusion of this study possible. He is supported by a TSMC endowed chair at NCU.

This effort was sponsored in part by the National Aeronautics and Space Administration Ocean and Ice Processes Programs.

References

1. National Research Council, *Decade-to-Century-Scale Climate Variability and Change* (National Academy Press, Washington, DC, 1998).
2. P. D. Jones *et al.*, Adjusting for sampling density in grid-box land and ocean surface temperature time series, *J. Geophys. Res.* **106** (2001) 3371–3380.
3. T. C. Peterson and R. S. Voss, An overview of the global historical climatology network temperature data base, *Bull. Am. Met. Soc.* **78** (1997) 2837–2849.
4. J. Hansen *et al.*, GISS analysis of surface temperature change, *J. Geophys. Res.* **104** (1999) 30997–31022.
5. K. Y. Vinnikov *et al.*, Empirical data on contemporary global climate changes (temperature and precipitation), *J. Climate* **3** (1990) 662–677.
6. S. Solomon, D. Qin, M. Manning, M. Marquis, K. Averyt, M. M. B. Tignor, H. L. Miller, Jr. and Z. Chen (eds.), *IPCC, Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007).
7. C. K. Folland *et al.*, Global temperature change and its uncertainties since 1861, *Geophys. Res. Lett.* **28** (2001) 2621–2624.
8. T. R. Karl and K. E. Trenberth, Modern global climate change, *Science* **302** (2003) 1719–1723.
9. J. S. Bendat and A. G. Piersol, *Random Data, Analysis and Measurement Procedures* (Wiley Interscience, New York, 1986).
10. H. Von Storch and F. W. Zwiers, *Statistical Analysis in Climate Research* (Cambridge University Press, Cambridge, UK, 1999).
11. N. E. Huang *et al.*, The empirical mode decomposition method and the Hilbert spectrum for non-stationary time series analysis, *Proc. Roy. Soc. London* **A454** (1998) 903–995.
12. N. E. Huang *et al.*, A new view of nonlinear water waves — the Hilbert spectrum, *Ann. Rev. Fluid Mech.* **31** (1999) 417–457.
13. N. E. Huang *et al.*, A confidence limit for the empirical mode decomposition and the Hilbert spectral analysis, *Proc. Roy. Soc. London* **A459** (2003) 2317–2345.
14. Z. Wu and N. E. Huang, A study of the characteristics of white noise using the empirical mode decomposition method, *Proc. Roy. Soc. London* **A460** (2004) 1597–1611.
15. D. A. T. Cummings, R. A. Irizarry, N. E. Huang, T. P. Endy, A. Nisalak, K. Ungchusak and D. S. Burke, Traveling waves in the occurrence of dengue haemorrhagic fever in Thailand, *Nature* **427** (2004) 344–347.

16. N. E. Huang, M. Brenner, C.-G. Pak and L. Salvino, An application of Hilbert–Huang transform in the stability study of airfoil flutter. *AIAA J.* **44** (2006) 772–786.
17. Z. Wu, N. E. Huang, S. R. Long and C. K. Peng, On the trend detrend and variability of nonlinear and nonstationary time series, *Proc. Natl. Acad. Sci.* **104** (2007) 14889–14894.
18. P. Flandrin, G. Rilling and P. Gonçalvès, Empirical mode decomposition as a filter bank, *IEEE Signal Process. Lett.* **11** (2004) 112–114.
19. Z. Wu and N. E. Huang, Statistical significance test of intrinsic mode functions, *EMD and Its Applications*, eds. N. E. Huang and S. S. P. Shen (World Scientific, Singapore, 2005), pp. 107–127.
20. Z. Wu, E. K. Schneider, B. P. Kirtman, E. S. Sarachik, N. E. Huang and C. J. Tucker, The modulated annual cycle: An alternative reference frame for climate anomalies, *Clim. Dyn.* **31** (2008) 823–841.