Advances in Adaptive Data Analysis
Vol. 3, Nos. 1 & 2 (2011) 95–113

DOI: 10.1142/S1793536911000660

World Scientific
www.worldscientific.com

SOME CONSIDERATIONS ON PHYSICAL ANALYSIS OF DATA

ZHAOHUA WU*

Department of Earth, Ocean, and Atmospheric Science
and Center for Ocean-Atmospheric Prediction Studies
Florida State University, Tallahassee, FL 32306-4520, USA
zwu@fsu.edu

NORDEN E. HUANG

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

XIANYAO CHEN

First Institute of Oceanography
State Oceanic Administration, Qingdao
Shandong 266061, China

In this paper, we present some general considerations about data analysis from the perspective of a physical scientist and advocate the physical, instead of mathematical, analysis of data. These considerations have been accompanying our development of novel adaptive, local analysis methods, especially the empirical mode decomposition and its major variation, the ensemble empirical mode decomposition, and its preliminary mathematical explanations. A particular emphasis will be on the advantages and disadvantages of mathematical and physical constraints associated with various analysis methods. We argue that, using data analysis in a given temporal domain of observation as an example, the mathematical constraints imposed on data may lead to difficulties in understanding the physics behind the data. With such difficulties in mind, we promote adaptive, local analysis method, which satisfies fundamental physical principle of consequent evolution of a system being not able to change the past evolution of the system. We also argue, using the ensemble empirical mode decomposition as an example, that noise can be helpful to extract physically meaningful signals hidden in noisy data.

Keywords: Physical analysis of data; global domain analysis; adaptivity; locality; noise-assisted data analysis; empirical mode decomposition; ensemble empirical mode decomposition.

*Corresponding author.

1. Introduction

> "*Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.*"

When Henry Poincaré (1905) made this statement, he was calling for all the means to convert facts (data) into the edifice of science. In one of these means, the data obtained from observation, experimentation, and measurements, are inspected, denoised, transformed, and decomposed, and organized into testable hypotheses, theories, and laws. This process of determining the nature and relationship from the raw data is called data analysis, which leads to the mathematical modeling of the data and prediction of a system's evolution. The correctness of the mathematical models and predictions are further tested by the results of the analysis of new data.

Just as a heap of stones can be built into different houses based on different blueprints, different methods of data analysis applied to raw data can lead to extracting different useful information about the system described by the raw data. Often, the data analysis reflects the expertise and interests of a researcher and exaggerates the points of interest in the data. For example, from a statistician's perspective (Lindley, 2000), "*it is only the manipulation of uncertainty that interests us. We are not concerned with the matter that is uncertain. Thus we do not study the mechanism of rain*; *only whether it will rain.*" However, from the perspective of atmospheric scientists, the physical mechanism of rain is of more interests. For them, understanding the physical laws governing rain is essential to constructing mathematical models and making predictions of rain. From this stand point, how to analyze data becomes crucial.

In this paper, we will present some general considerations about data analysis from the perspective of a physical scientist and promote the physical, instead of mathematical, analysis of data. These considerations originated from studies of many well-established and widely used methods, such as statistical analysis (e.g., Lindsey, 2004), Fourier spectral analysis (e.g., Bloomfield, 2000), and wavelet analysis (e.g., Daubechies, 1992), and from our development of novel adaptive, local analysis methods, especially the empirical mode decomposition (Huang *et al.*, 1998, 1999, 2003; Flandrin *et al.*, 2004; Wu and Huang, 2004; Wu *et al.*, 2007; Huang and Wu, 2008; Huang *et al.*, 2009; Wu *et al.*, 2009), its major variation, the ensemble empirical mode decomposition (Wu and Huang 2009; Wu *et al.* 2009), and its mathematical explanations (Hou *et al.*, 2009; Daubechies *et al.*, 2009). The particular emphasis will be on the mathematical and physical constraints associated with various analysis methods. We argue, using data analysis in a given temporal domain of observation as an example, that the mathematical constraints hidden in a data may lead to difficulties in understanding the physics behind the data to be analyzed. With such difficulties in mind, we promote adaptive, local analysis method, which satisfies fundamental physical principle that consequent evolution of a system should not be able to change the past evolution of the system.

The paper is arranged as the following: Sec. 2 will discusses the limitations of global domain analysis, which is the foundation for many currently widely used methods. Section 3 discusses the advantages of adaptive basis of a method. We further argue in Sec. 4 that, to extract physically explainable components of signals, adaptivity is not enough, and locality should be a basic characteristic of a method so as to effectively reflect the nonlinear, nonstationary features of the signal hidden in data. Section 5 discusses the harmony between noise and signal and noise-assisted data analysis. A summary is presented in the last section.

2. Global Domain Data Analysis, Mathematical Rigor, and Physical Intuition

The importance of mathematics in natural sciences was well discussed in Eugene Wigner's seminal essay (Wigner, 1960). In his essay, Wigner argued for the role mathematics played in the formation of theories of physics, emphasizing the "unreasonable effectiveness" of mathematics in accurately expressing and generalizing important discoveries of physics although the rationales behind this unreasonable effectiveness are hard to explain and remain a mystery, which led to the searching for the answers to the question of whether God is a mathematician (Livio, 2009). There are opposite views on the role of mathematics in sciences, for example, Rohrlich (1996), using similar examples described in Wigner's essay, argued for the importance of physical intuitions that are disconnected from mathematics in the creation of theories of physics. The views on the effectiveness of mathematics (opposite to Wigner's) of scientists working in different fields of science have even led to the creation of a catchphrase "unreasonable ineffectiveness of mathematics" (e.g., Borovik, 2009; Velupillai, 2005).

In the field of data analysis, we have seen similar perspectives in the developments and usages of analysis methods. To many, a method that follows well-established mathematical rules is justified for its application to any real-world problem for the rigor of mathematics provides the rigor in the analysis of data. To others, they feel more the rigidness rather than rigor of the mathematical rules behind an analysis method and use methods that follow more of their physical intuitions to solve practical data analysis problem. For sure, these two perspectives can hardly be united to perfection into one. The question is, thereby, which perspective should be preferred more when searching the physical essence hidden in data is the main purpose of data analysis.

To answer this question, we need to understand the drawbacks behind each perspective. Currently, many data analysis methods widely used in practical data analysis are global domain analysis methods which have well established mathematical rules. One of the most widely used mathematical properties in data analysis is the orthogonality, which, for two state vectors (series), u_i and $v_i, i = 1, N$, is defined as

$$\langle u, v\rangle = \sum_{i=1}^{N} u_i v_i = 0 \tag{1a}$$

Similarly, for two (piecewise) continuous functions $f(t)$ and $g(t)$, they are said to be orthogonal if the inner product of them within a given interval $[a, b]$ satisfies

$$\langle f, g\rangle = \int_a^b f(t)g(t)dt = 0 \tag{1b}$$

To illustrate the side effect of such a mathematically well established property in data analysis, we use the Fourier transform as an example. The Fourier transform is named after Joseph Fourier's pioneering work of using a series of trigonometric functions to expand a function when he studied the problem of heat transport in solid bodies. Fourier's work was first reported to Paris Institute (the later French Academy) in 1807 and got little notice. It took about 15 years for him to formally publish one of the greatest contributions to sciences (mathematical and nonmathematical) and engineering in human history (Grattan-Guinness, 1972; Herivel, 1975). It took another few decades for scientists to realize the great potential usage of Fourier series in mathematical understanding and modeling of a physical system. A milestone work of using Fourier transform to understand a physical system was Von Helmholtz's theoretical analyses of the operation of the ear in 1863. The Von Helmholtz model was based on a Fourier spectral analysis, in which the spectral representation in frequency domain was introduced. Since then, the applications of Fourier spectral analysis as a data analysis tool took off and became extremely popular after Cooley and Tukey's work (Cooley and Tukey, 1965) on fast Fourier transform.

In Fourier spectral analysis, the orthogonality is the key constraint. A data series of a given length N are expanded into the sum of trigonometric functions of given frequencies (if the data is a time series) or of given wavelengths (if the data is a spatial series). All the frequencies/wavelengths are determined by the data length N, i.e., frequencies/wavelengths of $2n\pi/N$, where $n = 1, 2, \ldots, \mathrm{mod}(N, 2)$. When data length N changes, all frequencies/wavelengths change as well. This expansion is unique and complete. The selected frequencies/wavelengths provide that the product of the two trigonometric functions are orthogonal to each other so that the Parseval's theorem, the sum (or integral) of all the square of individual data values is equal to the sum of the square of amplitude coefficients of trigonometric functions, is automatically satisfied. While this constraint brings many conveniences in mathematical generalization of the decomposition of data, it also brings serious side effect in understanding a data series if the series is from the observation of a natural physical system.

This side effect can be understood through the Fourier analysis of data from a simple physical system. Suppose that we are working with a water wave tank in a laboratory. We have two wave generators at the one end of the tank that excite surface water waves of different frequencies and a probe near the wave generator to measure the height of the water surface. Assuming that the excited waves are sinusoidal functions of time when they reach the probe, we have the height of water

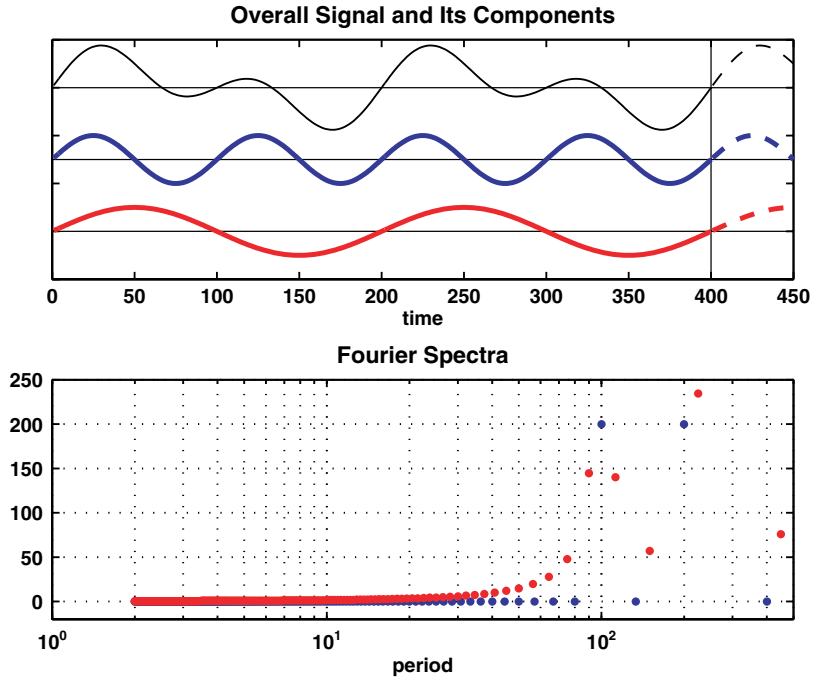

Fig. 1. The signals and their Fourier spectra. In the upper panel, the bold black line is the sum of two sinusoidal curves (blue and red, including both the solid and dashed parts). In the lower panel, the blue dots are the Fourier spectrum of the solid part of the bold black line in the upper panel over the time interval from 0 to 400; the red dots are the same as the blue dots but the spectrum is calculated for the bold black line over the interval from 0 to 450.

surface $h(t)$ given by

$$h(t) = a_1 \sin\left(\frac{2\pi t}{T_1}\right) + a_2 \sin\left(\frac{2\pi t}{T_2}\right) \tag{2}$$

where a_1 and a_2 are constant amplitudes and T_1 and T_2 are periods of water waves excited by two different wave generators. For simplicity, the measurement is made at integer time intervals of one unit and T_1 and T_2 are set to be 100 and 200 units, respectively; and a_1 and a_2 are set to be 1. The measured surface height is plotted in the upper panel of Fig. 1.

When Fourier spectral analysis is applied to the measured surface height for the time span 1–400, we obtain two perfect nonzero spectral values, as displayed by the two blue dots in the lower panel of Fig. 1. In this case, the two components derived from the Fourier spectral analysis are orthogonal and unrelated to each other over this interval, which truthfully reflects the physical properties of the system of the sum of two waves. However, if the Fourier spectral analysis is performed on the same measurement for the time span 1–450, we obtain tens of obviously nonzero spectral values and none of them corresponds to the actual waves excited by the

two wave generators. Since the orthogonality of each pair of the Fourier components to some degree implies that each wave of a particular frequency is not correlated with any other wave, the result from the analysis of the recorded surface height of an arbitrarily selected length could lead to the misinterpretation that the surface height is caused by tens rather than two wave generators. It should be also noted here that the above system, albeit simple, is typical in many application fields, and similar examples could be easily found.

The physically untruthful interpretations based on the analysis of the above system cast shadows on the usefulness of Fourier spectral analysis in the analysis of nonstationary data. The problem, to a large degree, is caused by the analysis that is over the global domain of the observed data combined with orthognality constraint built in the Fourier spectral analysis, which implicitly assumes that the observed data will be forever repetitive and therefore, from a mathematical perspective, the physical systems described by the observed data over different selected observational records are different although they are indeed the same. This implicit assumption caused the components in one domain being not the components in the other domains even over the part of the overlapped domain. It should be noted here that for the analysis of data from a physically fixed spatial domain, such as the analysis of the vibration of a string with two fixed ends, the selection of the global spatial domain leads to the Fourier components in spatial domain also reflecting truthfully the dominant resonant spatial oscillation patterns. However, for the analysis of a time series which records the evolution of a physical system, the luxury having two fixed ends in temporal domain does not exist. Therefore, we lack legitimate physical justifications for the preference of one temporal domain over others.

In general, the analysis methods that can only be applied over global domain often cannot avoid the sensitivity of results to the selection of the domain. In this sense, global domain analysis with well established mathematical foundation may not be effective in extracting essential physical information of a physical system that evolves. Therefore, the alternative perspective that emphasizes physical intuition may be a more natural choice of analysis.

It should be noted that a perfect mix of rigorous mathematics and physical intuitions often leads to a revolution in data analysis. An illustrative example is the many times of the rediscovery of wavelet analysis. Undoubtedly, Harr constructed a complete and orthogonal set now called Haar wavelet in 1909. However, it was the intuitive rediscovery of wavelet and its applications to geophysical data analysis by Jean Morlet, a French geophysicist working in an oil company who was tired of using windowed Fourier transform to analyze data that led to the later extensive research of wavelet in mathematics and popular applications of wavelet in almost all the fields of science. However, the perfect mix often follows a route from building an analysis method based on physical intuition to mathematical understanding of the method and then using currently available mathematics or newly developed mathematical tools to generalize the method for more great use.

3. Adaptiveness

The ultimate goal of scientific data analysis is to extract useful information on the evolution of a system and to predict the future evolution of the system. For many natural systems, such as the Earth's climate system, their evolutions involve numerous nonlinear interactions and are under nonstationary external forcings. In addition to that, the data that reflect the evolution of a natural system are often collected in a noisy environment and hence containing noise. When such data are analyzed, numerous difficulties need to be overcome and the problems of nonlinearity, nonstationarity, and noise associated with data need to be properly handled. In these situations, the adequacy of a data analysis method becomes a crucial issue; understanding the assumptions behind a method and its advantage and disadvantages also becomes important.

Many previous data analysis methods, such as Fourier transform-based methods, discrete wavelet analysis, etc., involve decomposition of data using a complete and orthogonal set. When such methods are used to analyze data, often, there are hidden assumptions of data to be analyzed. For example, in Fourier spectral analysis, data is decomposed into components of trigonometric functions of different amplitude and frequency. Since the trigonometric functions used in the analysis are all periodic functions of the length of the data, when the method is applied, an implicit assumption that the future evolution will repeat the evolution is made. This assumption is usually not consistent with the data collected from the observation of a natural system. In addition to that, when a selection of a complete and orthogonal set to decompose data is made, the preference of one complete and orthogonal set to others is not based on physical reasons for the suitability of a set is not a known *a priori*. In this sense, the selection of a complete and orthogonal set is quite subjective. Often, different selections lead to different physical interpretation of the data.

An alternative approach of data analysis is to let data determine the basis that are used to decompose themselves, i.e., using adaptive basis. However, to define adaptive basis of data, some *a priori* determined constraints need to be applied. For example, in empirical orthogonal function (EOF) analysis (also called principal component analysis, PCA), the covariance matrix of temporal–spatial data is used to obtain the orthogonal basis. The first principal component obtained in this way can be considered as the result of maximizing the variance of the data explained, which is consistent with the results obtained using a variational method that maximize variance (Monahan *et al.*, 2009). While EOF analysis avoids using *a priori* basis, it invokes a new mathematical (but not physical) quantity, the covariance matrix, to determine a complete and orthogonal basis, in which physical considerations are not integrated into the determination of the *a posteriori* basis. Since covariance matrix is derived from data over the global temporal/spatial domain of the data, often, the results from EOF analysis are quite sensitive to the selections of spatial or/and temporal domain. A variation of the EOF analysis, singular

spectrum analysis (SSA, see Ghil *et al.*, 2002) for time series analysis, which uses phase space reconstruction based on the time series to obtain equivalent (phase) space–time domain data set, has the similar sensitivity problem to the length of the time series and the *a priori* determined dimension of the reconstructed phase space. With the above concerns, a natural question that follows is whether it is possible to define adaptive basis that do not invoke seemingly unreasonable mathematical constraints for scientific data analysis.

Our answer to the above question is positive in light of the recent development of the Hilbert–Huang transform (HHT) (Huang and Wu, 2008). In HHT, the key component is the empirical mode decomposition (EMD), which decomposes the data in terms of "natural" waves, called intrinsic mode functions (IMFs). An IMF is an amplitude-frequency modulated wave of which the timescale of amplitude modulation is much longer than the timescale of the carrier wave at any local time. Indeed, the principles behind EMD are very simple and intuitive, as illustrated by Fig. 2. Suppose, there are two signals of different frequencies and amplitudes at any time come from different sources, the blue lines in the top two panels. The overall signal recorded by a signal receiver will be the black line in the bottom panel. If all

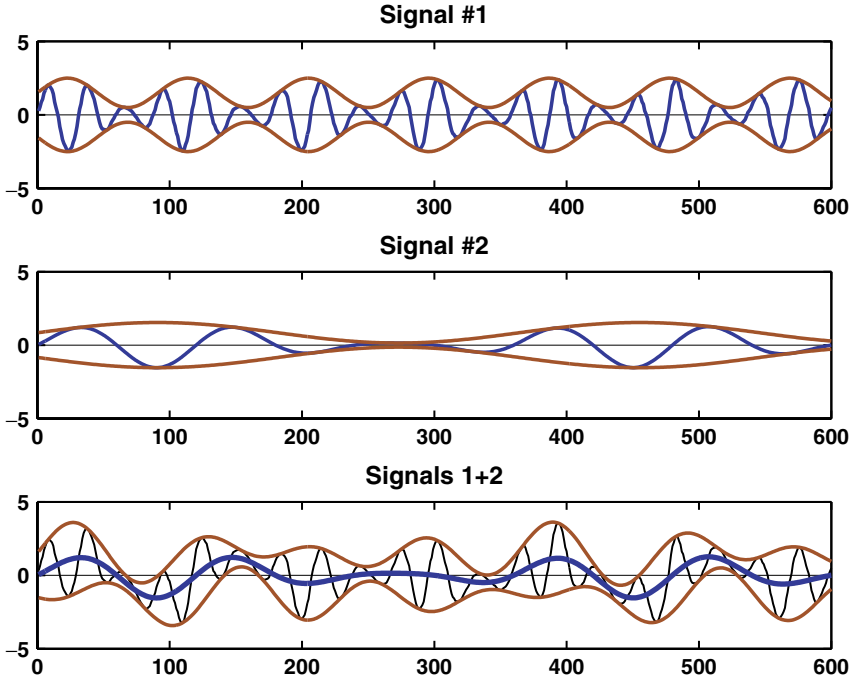

Fig. 2. The sum and the separation of amplitude–frequency modulated "natural" waves. In the top and middle panels, natural waves of different frequencies (blue lines) and their envelopes (positive and negative amplitudes, brown lines) are plotted. In the bottom panel, the sum (black line) of the two natural waves, the envelopes (brown lines) of the sum, and the mean (blue line) of the envelopes are plotted.

the maxima and minima are connected with smooth brown lines separately, it is found that the mean of the brown lines at any temporal location is the blue line in the lower panel, which is almost identical to the blue line in the middle panel. Subtracting the mean of the envelopes from the overall signal will naturally isolate the riding wave of high-frequency from the low-frequency wave on which it is riding. In this way, waves of high-frequency and of low-frequency are naturally separated in terms of natural waves. For a more complicated signal $x(t)$, such a separation process can be repeated level by level to extract riding waves from high frequencies to low frequencies, i.e., $x(t)$ can be expressed as

$$x(t) = \mathrm{Re}\left[\sum_{j=1}^{n} a_j(t) e^{i\int \omega_j(t)dt}\right] + r_n \tag{3}$$

where Re[•] represents the real part of "•," $a_j(t)$ and $\omega_j(t)$ are instantaneous amplitude and frequency of jth IMF, respectively. In Eq. (3), the residue, r_n, is not expressed in terms of an simple oscillatory form on purpose, for it is either a monotonic function, or a function with only one extremum, not containing enough information to confirm whether it is an oscillatory component whose frequency is physical meaningful.

An interesting feature of EMD is the way it determines the amplitude–frequency modulated riding waves. From pure mathematical perspective, a continuous function $x(t)$ can be expressed as the product of an arbitrarily given constant amplitude (frequency) wave (regardless whether frequency changes or not) and a corresponding changing frequency (amplitude) function if there is no constraint being applied to the amplitude and frequency at any local time. The requirement of timescale of amplitude modulation being much longer than the timescale of the carrier wave at any local time, which is intuitively based on our observation of natural waves, provides a useful constraint for determining (uniquely) the envelopes of riding waves. This requirement also provides the justification for using only the extrema information to determine the amplitudes of IMFs for EMD: since the distance between two neighboring maxima (minima) provides a wavelength (timescale) of the riding wave, the envelopes defined based on the connection of neighboring maxima (minima) with a low-order polynomial (spline) automatically satisfies the timescale requirement for the riding wave. It is also fair to say that the extrema are the data points of a wave that carries most efficiently the wave information: connecting neighboring extrema even with straight lines without using any other data of a wave could provide qualitative information of a wave, such as approximate changes of wave frequency and amplitude. As a contrast, zero-crossing points could only provide some information on the change of the frequency of waves, but not amplitude. For complicated data that include many waves of different frequencies, the zero-crossing information could not be determined, however, the extrema of the riding wave tend to emerge with respect to the sum of waves of significant lower frequencies. In this sense, using extrema information to determine the riding wave

is a natural selection, and is expected to carry more physical information of the data. It should be noted here that when the frequency of the riding wave is not significantly higher than its reference low frequency waves, EMD faces difficulties in separating waves, even for a simple case of a composite signal of two tones. In such a case, the amplitude ratio of two-tone signal becomes crucial for clean separation. A systematic study of such a problem can be found in Rilling and Flandrin (2008).

In summary, it is arguable that EMD does not invoke any significant mathematical constraint but is based on physical intuitions. Wave is the basic ingredient of our interpretation of the physical world and the wave packet behaves like a particle. In this sense, the EMD is a method that decomposes the complicated signal in terms of the basic ingredients of the physical world. The inclusion of amplitude and frequency modulation in a component provides capability of reflecting the complication of the physical world caused by nonlinear interactions and nonstationary external forcings.

4. Locality

In the previous two sections, we have illustrated the problems of various analysis methods that use *a priori* basis or adaptive basis derived from satisfying mathematical constraints. We argued that the intuitively developed EMD, which uses adaptive basis of "natural" waves and invoke few mathematical constraints, provides an alternative paradigm of data decomposition and can alleviate some problems caused by imposing mathematical constraints. Indeed, there are numerous studies to show that EMD does have many advantages over many other widely used data analysis methods (e.g., Huang and Attoh-Okine, 2005; Huang and Shen, 2005) when they are used to extract physically meaningful signals. However, in previous sections, we have not discussed how EMD handles the nonlinear and nonstationary problems. In this section, we argue that the temporal/spatial locality should be a requirement of any data analysis method if the goal of data analysis is to understand the physics of a system behind the data. We also argue that EMD can handle effectively nonlinear, nonstationary problem due to its temporal/spatial locality.

As we have known, the subsequent evolution of a physical system cannot alter the reality that has already happened. Suppose that the evolution of a physical system involves nonlinear and nonstationary change, the physics behind that evolution in a given period may be unique to that period. For example, the suspended lamp in the cathedral of Pisa did not always swing, when Galileo made observation of it. If Galileo had calculated the averaged period of a swing by dividing the total observation time with the total number of swings, he probably would have made a wrong discovery, for there must be some occasions the lamp did not swing when Gelileo made the observation. For him, the discovery was based on the periods when the lamp indeed swung. Since the swing of the lamp was most likely caused by the combination of intermittent winds and the gravity while the static lamp is under the balanced forcing between the gravity and suspension, the swing case and static

case actually had different physics. From this example, it is indicative that the derived physical interpretation based on the analysis of the data should be unique for a particular period of observation if it reflects the true physics behind a system evolution, unless the physics repeats again and again or remains the same. From this argument, one can continue to infer that the analysis of the data extended to the prior or/and later periods should not change that physical interpretation of the system evolution in that period, unless a physical process cannot be determined due to the shortness of the data of that period. In this sense, an analysis method must satisfy the "temporal locality" requirement based on the physical intuition and it is expected that such a method would provide a better chance to reveal true physics.

The popularity and wide applications of the wavelet analysis to scientific and engineering fields since 1970s is indeed rooted in its locality, although the mathematical development of the wavelet analysis can be traced back to the earlier 20th century (Jaffard *et al.*, 2001). The importance of the locality in data analysis and processing is well reflected by a US National Academy report written by Dana Mackenzie (2002). In her introduction of wavelets to general public, Machenzie used the title "*Wavelets*: *Seeing the forest and the trees*." While the emphasis of that article was to illustrate the greatness of the wavelet analysis in data analysis and data processing (e.g., image compression), the article title indeed pointed out the most significant feature of the wavelet analysis: seeing the trees by extracting spatially/temporally local features of data, although the details of trees may be distorted due to the use of particular type of wavelet. This locality is also the major feature of EMD to handle the nonlinear, nonstationary problems of data.

To clarify how various methods treat nonlinearity and nonstationarity hidden in data, here we discuss three types of methods: Fourier analysis, wavelet decomposition, and EMD-based methods. To illustrate, we use a daily atmospheric CO_2 concentration data observed by National Oceanic and Atmospheric Administration (NOAA) at Mauna Loa (MLO), Hawaii. The data to be analyzed is an arbitrarily selected 5-year section of that observation (starting from some time around mid-May), which is plotted at the top of Figs. 3 and 4. As pointed out by previous studies (e.g., Buermann *et al.*, 2007), the data are dominated by strong annual cycle related to the "breathing" of the northern hemisphere biosphere (i.e., the seasonal asynchrony between photosynthetic drawdown and respiratory release of CO_2 by terrestrial ecosystems) and a trend mostly related to the increasing fossil fuel burning. The seasonal cycle of atmospheric CO_2 concentration at the MLO, with a maximum at the beginning of the growing season (May) and a minimum at the end of the growing season (September/October), has a relatively shorter increasing period and a relatively longer decreasing period over each year. Therefore, this is asymmetric with respect to the ridge or trough of the CO_2 concentration.

In Figs. 3 and 4, the decomposition results using discrete wavelet (in this case Meyer wavelet) and ensemble empirical mode decomposition (EEMD), a major improvement of the original EMD which we discuss in more details later, are plotted,

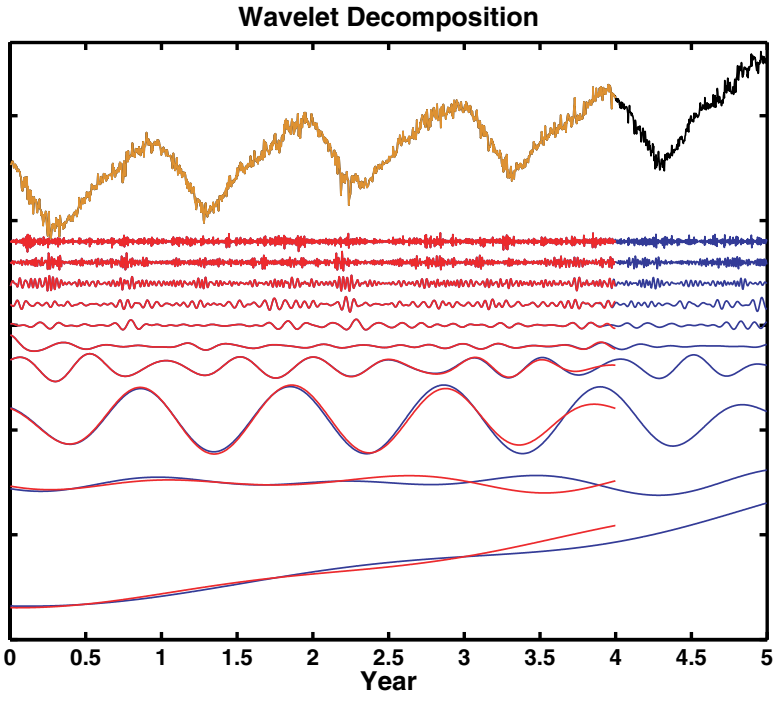

Fig. 3. Five years of Mauna Loa CO_2 (black line) and its discrete wavelet decomposition (blue lines). The brown and red lines are the same as the black and blue lines, respectively, but resulted from applying the same decomposition of the first 4 years of the same data.

respectively. To test the temporal locality of each method, we applied each method to the CO_2 concentration of the whole five years and of the first four years. The normalized Fourier spectra (i.e., the sum of the Fourier spectrum of a time series over the whole discrete frequency domain of the time series is one), as shown in Fig. 5, have noticeable difference at almost all the frequencies. However, for discrete wavelet analysis and EEMD, the difference below the timescale of semiannual for the overlapped period is negligibly small, indicating superb temporal locality at shorter period components. However, when the period becomes longer, the end effect comes into the calculation, and some decomposition errors emerge. In wavelet analysis, the noticeable errors are generally defined using cone of influence. In EMD/EEMD, with the end-effect treatment discussed in Wu and Huang (2009), noticeable errors appear usually within a half-period of a component at the data end.

The temporal locality of analysis makes the stationarity assumption irrelevant, for the analysis results are not affected by the data far away. Therefore, the temporal locality of an analysis method automatically bypasses the stationarity assumption, which is applied over the global domain.

Concerning how nonlinearity of a physical system is dealt within a data analysis method, we again use the above analysis of CO_2 concentration. From a physical

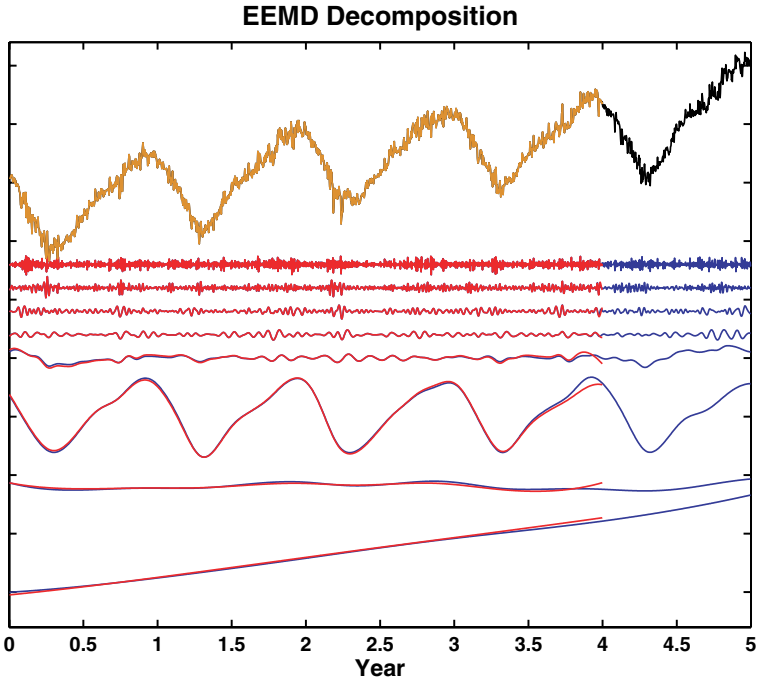

Fig. 4. The same as Fig. 3, but using ensemble empirical mode decomposition instead of discrete wavelet decomposition.

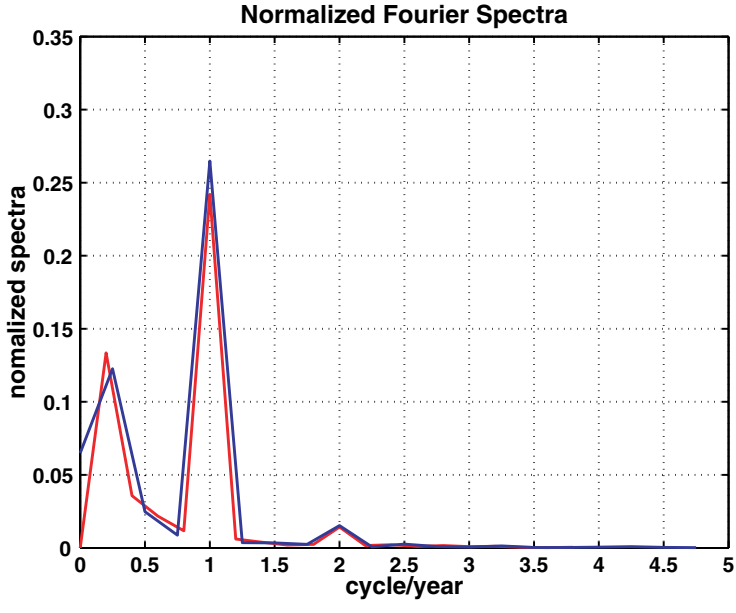

Fig. 5. The normalized Fourier spectra of the whole five years of Mauna Loa CO_2 (blue line) and of the first four years (red line), respectively.

perspective, the photosynthetic drawdown and the respiratory release of CO_2 by terrestrial ecosystems represent the two phases of the annual cycle of CO_2. It contains no semiannual cycle. When the data is decomposed using Fourier transform, the asymmetry of the MLO CO_2 concentration is qualitatively represented by a sinusoidal wave of 1-year period, which is symmetric with respect to its maxima, and its harmonics, especially seminannual cycle (figure not shown). To represent the nonlinear asymmetry of the CO_2 concentration, summation of the sinusoidal annual cycle and its harmonics are necessary, which implies that an individual component of annual cycle is not capable of reflecting the nonlinear evolution of the physically meaningful annual cycle of CO_2 concentration. Similar problem exists for discrete wavelet decomposition, which is displayed in Fig. 3 when a symmetric wavelet (in this case Meyer wavelet) is used.

When an EMD/EEMD is applied, the harmonic problem is no longer valid. In EMD, the nonlinearity is expressed by the modulations of amplitude and frequency (scale), as illustrated by simple nonlinear oscillators (Huang *et al.*, 1998, 1999). Since EMD approximates the envelope of the riding wave using only extrema information, the waveform between two neighboring maxima (minima) is well preserved, therefore, the harmonics become unnecessary. Figure 4 illustrates such a result.

5. Noise and Signal

The word "noise" can possibly be traced back to the Latin word "nausea", meaning "seasickness, feeling of sickness." In scientific community, noise refers to disturbance, especially a random disturbance that obscures or reduces the clarity of a signal. With the deficiencies of any observation system and noisy environment, it is always true that any data based on the observation of a natural system is an amalgamation of signal and noise

$$x(t) = s(t) + n(t), \tag{4}$$

where $x(t)$ is data, and $s(t)$ and $n(t)$ are signal and noise, respectively.

The detection of signal of a noisy data set is fundamental to information extraction and decision making. However, there are no clear definitions of "noise" and "signal" of a noisy data, for noise to someone may be signal to others, such as in the "cocktail party problem" (McDermott, 2009). From a probabilistic perspective, the "signal" of data is the invariant part in the implicit modeling of the data from a non-repeatable observation that can be exactly reproduced while the "noise" is everything else except that invariant part and is one particular realization of some stochastic process (Gardner, 1986). Therefore, complete separation of signal and noise in data requires known characteristics of either signal or noise. When the noise in the data has distinct characteristics from those of the signal, effective filters may be designed based on the characteristics of the signal and those of the noise to separate with sufficient accuracy the signal of data from the noise. However, such cases are relatively rare since knowledge of signal in a noisy data set is

limited prior to the data being analyzed. The problem could be more complicated if the signal is nonlinear, nonstationary and noise is intermittent. In these cases, the nonlinearity of the system itself can cause the system to behave like noise at some stage of the evolution. Overall, the noise in data can be quite detrimental while extracting useful information from data.

However, in some cases, noise is welcome in data analysis. When the observation is made of a non-repeatable experiment or of the evolution of a natural system, what we know about the data before analysis is that it contains noise. Even if noise is not filterable, we still need to draw inferences on the hidden physics in data from the analysis of data. In such a case, a cautious step is taken to check whether the inferences are sensitive to noise. Often, we determine the sensitivity of the inferences to the noise by analyzing the data with added noise so as to estimate the robustness of our inferences. Such an approach is indeed widely used in statistical analysis. In most of such situations, the noise added is of small amplitude. For those methods of *a priori* selected basis, such as the Fourier transform and wavelet analysis, since the determination of the components are based on the convolution, the sensitivity to small amplitude noise is usually small. Such insensitivity to the added small noise provides some confidence on the robustness of the inferences when the Fourier transform or wavelet analysis is used to analyze data. However, as is discussed in Secs. 2 and 3, the Fourier transform and wavelet analysis lack adaptivity and the robust inferences are often not related to the essence of the physics processes hidden in data. Therefore, the usefulness of noise in many analysis using traditional methods is small.

As we discussed in previous sections, EMD uses the extrema information to separate the riding "natural" wave and its reference, and hence, the change of extrema locations and values could lead to significantly different results, which makes the extracted "natural" wave sometimes appears bizarre and very sensitive to any added noise, for noise could alter the local extrema both in location and value. One of the consequences of this nature of EMD is that the analysis of two virtually almost identical data series (suppose the records collected by two almost identical observations of the same phenomena) using EMD may lead to totally different physical interpretations. This lack of the "physical uniqueness" essentially paved way for many complaints about the ineffectiveness of EMD. One of the solutions to the above problem was proposed in Huang *et al.* (2003): to use intermittency test to alleviate the problem of sensitivity to intermittent noise and to the problem of one component of data containing waves of drastically different temporal scales (the "mode mixing" problem). However, this solution introduced external intervention of the analysis by the analyzer.

Recently, we explored a new usage of noise to remedy the problem of high sensitivity to noise and lack of "physical uniqueness." The new solution is called noise-assisted data analysis (Wu and Huang, 2005, 2009; Huang and Wu, 2008) based on the understandings of characteristics of noise using EMD (Flandrin *et al.*, 2004; Wu and Huang, 2004, 2005). In Flandrin *et al.* (2004) and Wu and

Huang (2004, 2005, 2009), it is shown that the EMD is effectively a dyadic filter bank: the Fourier spectra of the natural wave components (called IMFs) are all identical and cover the same area on semi logarithm period scale, with the peak frequencies of neighboring components being doubled (or halved). This property of EMD is essentially the base for the development of the new noise-assisted data analysis method called the EEMD. In EEMD, the decomposition consists of shifting an ensemble of white noise-added signal and treating the mean of the corresponding components from the EMD decomposition of data with added different realizations of noise as the final result. The effect of the added white noise in EEMD is to provide a dyadic filtering reference frame in the time-frequency space; therefore, the added noise collates the portion of the signal of comparable scale in one IMF, significantly reducing the chance of mode mixing and leading to the stability of decomposition. As the EMD is a time–space analysis method, the white noise is averaged out with sufficient number of trials. In this sense, the added noise plays a role of mimicking multiple observations of a phenomenon recorded by a single observation and serves as a "catalyst" in the decomposition that leads to stable and more physically interpretable results.

An important practical issue in EEMD is how large the amplitude of the added noise should be. For most time series that we are going to study, the modern computational power is not a barrier for EEMD even if the ensemble number becomes very large. However, when the amplitude of the added noise is too large, the extrema distribution of data themselves may lost its identity, leading to the decomposition being dictated by the dyadic filter bank of the added noise and lose the adaptive nature of EMD by not taking the information of the extrema distribution of the signal in the original data in the process of obtaining components. On the other hand, the amplitude of the added noise cannot be too small so that the distribution of extrema is not affected by the noise originally observed. Therefore, the amplitude of the added noise in EEMD should be moderate. However, this is a qualitative but not a quantitative argument. The difficulty of a quantitative determination of the amplitude of the added noise remains in EEMD when it is applied to a particular time series. There is no universally applicable amplitude of added noise. Fortunately, from our experience, a range of the amplitude of the added noise that leads to stable results exists. Since the results of the EEMD analysis can be always supplemented by the knowledge of a physical system we already have, the determination of the amplitude of the added noise can be assisted by whether the decomposition leads to a self-consistent theory that already exists or is newly proposed.

The concept of the noise-assisted data analysis seems to fit into the Taoism philosophy well. The appeared opposite of "Ying" and "Yang" finds harmony in EEMD. Of course, this is not a surprise, for the parallels of the modern physics and the Eastern philosophy has been noticed (Capra, 1975). EEMD, as an analysis method for isolating physically interpretable signals, provides a new case of the harmony between signal and noise: to extract signal using noise. Of course, the rigorous mathematical understanding of such harmony remains to be developed.

6. Concluding Remarks

In previous sections, we have presented some considerations on the physical analysis of data. These considerations are quite preliminary. Historical developments of data physics have pointed to the harmony between mathematical generalizations of physical systems that lead to great understanding beyond physical intuition. Since the mathematical understanding of EMD/EEMD is still at its infancy, it remains to be seen to what degree the physical constraints and mathematical constraints will come to a harmony in data analysis. We wish future studies in both mathematical and physical analysis of data will eliminate the gaps between them and make a seamless merge.

Acknowledgments

The authors are indebted to Prof. P. Flandrin of the École Normale Supérieure de Lyon and an anonymous reviewer for their numerous suggestions. Z. Wu and N. E. Huang were sponsored by the NSF of USA grant ATM-0653136 and ATM-0917743, Federal Highway Administration of USA grant DTFH61-08-00028, and grants from NSC of Republic of China, NSC95-2119-M-008-031-MY3, NSC97-2627-B-008-007, and from a grant from NCU, NCU 965941. X. Chen was supported by the National Basic Research Program of China 2007CB816002, National Science Foundation of China 40776018, National Key Technology RandD Program 2006BAB18B02, and Chinese Polar Science Strategy Foundation 20070208.

References

Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction* (2nd edition), Wiley-Interscience, New York, 288 p.

Borovik, A. (2009). *Mathematics Under the Microscope*, American Mathematical Society, 317pp.

Buermann, W., Lintner, B. J., Koven, C. D., Angert, A., Pinzon, J., Tucker, C. J. and Fung, I. Y. (2007). The changing carbon cycle at Mauna Loa Observatory, *Proc. Natl. Acad. Sci. USA*, **104**: 4249–4254.

Capra, F. (2000). *The Tao of Physics: An Exploration of the Parallels Between Modern Physics and Eastern Mysticism* (4th edition), Shambhala Publications, Inc., 366 p.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series, *Math. Comput.*, **19**: 297–301.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, Cambridge University Press, 377 p.

Daubechies, I., Lu, J. and Wu, H.-T. (2010). Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool. *Applied and Computational Harmonic Analysis*, **30**: 243–261.

Flandrin, P., Rilling, G. and Gonçalvès, P. (2004). Empirical mode decomposition as a filter bank, *IEEE Signal Processing Lett.*, **11**: 112–114.

Gardner, W. A. (1986). *Statistical Spectral Analysis: A Nonprobabilistic Theory*, Prentice-Hall, Englewood Cliffs, NJ, 480 p.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F. and Yiou, P. (2002). Advanced

spectral methods for climatic time series, *Rev. Geophys.*, **40**(1): 3.1–3.41, doi: 10.1029/2000RG000092.

Grattan-Guinness, I. (1972). *Joseph Fourier, 1768–1830: A Survey of his Life and Work*, The MIT Press, Combridge, MA, 516 p.

Herivel, J., (1975). *Joseph Fourier, the Man and the Physicist*, Clarendon Press, Oxford, 350 p.

Hou, T. Y., Yan, M. P. and Wu, Z. (2009). A variant of the EMD method for multi-scale data, *Adv. Adap. Data Anal.*, **1**: 483–516.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, E. H., Zheng, Q., Tung, C. C. and Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis, *Proc. Roy. Soc. Lond.*, **454A**: 903–993.

Huang, N. E., Shen, Z. and Long, S. R. (1999). A new view of nonlinear water waves — the Hilbert Spectrum, *Ann. Rev. Fluid Mech.* ,**31**: 417–457.

Huang, N. E., Wu, M. L., Long, S. R., Shen, S. S., Qu, W. D., Gloersen, P. and Fan K. L. (2003). A confidence limit for the empirical mode decomposition and Hilbert spectral analysis, *Proc. Roy. Soc. Lond.*, **459A**: 2317–2345.

Huang, N. E. and Attoh-Okine, N. O. (ed.) (2005). *Hilbert–Huang Transforms in Engineering*, CRC Taylor and Francis Group, Boca Raton, FL, 313 p.

Huang, N. E. and Shen, S. S. P. (ed.) (2005). *Hilbert–Huang Transform and Its Applications*, World Scientific, Singapore, 311 p.

Huang, N. E. and Wu. Z. (2008). A review on Hilbert–Huang transform: The method and its applications on geophysical studies, *Rev. Geophys.*, **46**, RG2006, doi:10.1029/2007RG000228.

Huang, N. E., Wu, Z., Long, S. R., Arnold, K. C., Chen, X. and Blank, K. (2009). On instantaneous frequency, *Adv. Adap. Data Analy.* **1**: 177–229.

Jaffard, S., Meyer, Y. and Ryan, R. D. (2001). *Wavelets: Tools for Science and Technology*, Society for Industrial and Applied Mathematics, 256 p.

Lindley, D. (2000). The philosophy of statistics, *J. Roy. Statist. Soc. Ser. D* (The Statistician) **49**: 293–337.

Lindsey, J. K. (2004). *Statistical Analysis of Stochastic Processes in Time*, Cambridge University Press, New York, 338 p.

Livio, M. (2009). *Is God a Mathematician?*, Simon & Schuster, New York, 320 p.

Mackenzie, D. (2002). Wavelets: Seeing the forest and the trees, *Beyond Discovery*, National Academy of Sciences. http://www.beyonddiscovery.org/content/view.txt.asp?a=1952.

McDermott, J. H. (2009). The cocktail party problem, *Curr. Biol.*, **19**: R1024–R1027.

Monahan, A. H., Fyfe, J. C., Ambaum, M. H. P., Stephenson, D. and North, G. R. (2009). Empirical orthogonal functions: The medium is the message, *J. Climate*, **22**: 6501–6514.

Rilling, G. and Flandrin, P. (2008). One or two frequencies? The empirical mode decomposition answers, *IEEE Trans. Signal Proc.*, **56**(1): 85–95.

Rohrlich, F. (1996) The unreasonable effectiveness of physical intuition: Success while ignoring objections, *Found. Phys.*, **26**: 1617–1626.

Poincaré, H. (1905). *Science and Hypothesis*, The Science Press, New York, 271 p.

Velupillai, V. (2005). The unreasonable ineffectiveness of mathematics in economics, *Camb. J. Econ.*, **29**: 849–872.

Wigner, E. (1960). The unreasonable effectiveness of mathematics in the natural sciences, *Comm. Pure App. Math.*, **13**: 1–14.

Wu, Z. and Huang, N. E. (2004). A study of the characteristics of white noise using the empirical mode decomposition method, *Proc. R. Soc. Lond. A*, **460**: 1597–1611.

Wu, Z., Huang, N. E., Long, S. R. and Peng, C.-K. (2007). On the trend, detrending, and variability of nonlinear and nonstationary time series, *Proc. Natl. Acad. Sci. USA*, **104**: 14889–14894, doi: 10.1073/pnas.0701020104.

Wu, Z. and Huang, N. E. (2009). Ensemble empirical mode decomposition: A noise-assisted data analysis method, *Adv. Adap. Data Analy.*, **1**: 1–41.

Wu, Z., Huang, N. E. and Chen, X. (2009). Multi-dimensional empirical mode decomposition based on ensemble empirical mode decomposition, *Adv. Adap. Data Analy.*, **1**: 339–372.