Advances in Adaptive Data Analysis
Vol. 2, No. 4 (2010) 415–428

DOI: 10.1142/S1793536910000550

World Scientific
www.worldscientific.com

# MODEL VALIDATION BASED ON ENSEMBLE EMPIRICAL MODE DECOMPOSITION

YU-MEI CHANG

*Department of Statistics, Tunghai University, Taichung, Taiwan 40704*

ZHAOHUA WU

*Department of Meteorology and Center, for Ocean-Atmospheric Prediction Studies, Florida State University, Tallahassee, FL 32306, USA*

JULIUS CHANG

*Department of Atmospheric Sciences, National Central University, Zhongli, Taiwan 32001*

NORDEN E. HUANG

*Research Center for Adaptive Data Analysis, National Central University, Zhongli, Taiwan 32001*

We proposed a new model validation method through ensemble empirical mode decomposition (EEMD) and scale separate correlation. EEMD is used to analyze the nonlinear and nonstationary ozone concentration data and the data simulated from the Taiwan Air Quality Model (TAQM). Our approach consists of shifting an ensemble of white noise-added signal and treats the mean as the final true intrinsic mode functions (IMFs). It provides detailed comparisons of observed and simulated data in various temporal scales. The ozone concentration of Wan-Li station in Taiwan is used to illustrate the power of this new approach. Our results show that, at an urban station, the ozone concentration fluctuation has various cycles that include semi-diurnal, diurnal, and weekly time scales. These results serve to demonstrate the anthropogenic origin of the local pollutant and long-range transport effects were all important. The validation tests indicate that the model used here performs well to simulate phenomena of all temporal scales.

*Keywords*: Ensemble empirical mode decomposition; model validation; ozone concentration; significant test.

## 1. Introduction

Traditionally, validation of numerical models for natural or engineering systems has all based on holistic comparisons. Specifically, the methodology is based on the comparisons of the trends, the cycles, the means, and the standard variations of the observed and simulated data. If they are similar or the correlation

between them is high, we will deem the model as good, and use it for estimations or predications. However, it may happen that the correlation is high since the simulated data followed the same general trend of the observed data very well, but they tend to underestimated or overestimated all the time. It may also happen that the overall correlation is low but there are some intrinsic mode components agreeing well. More importantly, model validation should provide diagnosis of the model physics or the driving physical processes. In a complicate system, there might be many different forces and processes acting together. Different processes should produce reactions of different time scales. Any given model could have considered some processes well and neglected others. Therefore, if the validation method could sort out the cause and effects of the different reactions, it would provide a valuable guidance for improving the model. It is with this idea in mind that we propose the present approach: to compare the model results with the observations not holistically, but their respective intrinsic mode components. We will use the air quality mode as an example.

The Taiwan Air Quality Model (TAQM) was developed by Chang *et al.* (1987), which described atmospheric transport, transformation, and deposition of airborne chemical species via a set of species conservation equations that include dynamics, thermodynamics, and chemistry, to understand and analyze the ozone transport process. This model is one of the first to include a complete suit of physical and chemical processes; it provides a framework to examine the relative importance and sensitivity of numerous physical and chemical processes responsible for the formation and deposition of tropospheric acidity. Therefore, we can use this model to sort out the details of each active processes and determine which one is dominating at any given time and location. Since the atmospheres are complex system consisting of a large number of factors interacting with one another, it is difficult to identify which specific factor is or is not modeled properly. We decided to use the recently developed empirical mode decomposition (EMD) method [Huang *et al.*, 1998, 1999; Wu and Huang, 2009] to overcome this difficulty.

EMD is an adaptive decomposition method. It is designed for analyzing data from nonlinear and nonstationary processes. The decomposition is based on the time scale separation in physical space. Specifically, the step is as follows: for any data, all the local extrema were identified first. Then, the local maxima (minima) were connected by a cubic natural spline line to form the upper (lower) envelope. A mean of the upper and lower envelopes is then computed, which is used as the temporal axis. The deviation from this axis is the component of the shortest time scale. The procedure can be repeated many times till the extracted component satisfies the definition of intrinsic mode function (IMF) defined as any function has the same number of extrema and zero-crossing, and that the mean of the upper and lower envelopes is zero. The above procedure could be repeated to the residue repeatedly till the final residue becomes a monotonic function. The EMD method as originally proposed suffered a difficulty: the mode mixing, defined as any IMF having large disparity in time scales. That shortcoming was basically resolved by

the Ensemble EMD (EEMD) [Wu and Huang, 2009], which is a modification of the basic EMD with added white noise to the data at the beginning. Through the ensemble of many trials with different realization of white noises, the final ensemble would not suffer the mode mixing, for the white noise contains all the scales. By adding the white noise, the scale separation is guaranteed by the added noise. The effects of the added noise on the resulting decomposition would be canceled out through the ensemble procedure.

The EEMD method and ensure the resulting IMFs being orthogonal and satisfying the proper definition of IMF. Here, we use the modified EEMD to analyze the nonlinear and nonstationary ozone concentration data and the data simulated from the TAQM. The ozone concentration of 2003 of Wan-Li station in Taiwan is used to illustrate the power of this new method.

## 2. Data

The data used here consisted of two types: the observed and the simulated. Asian continental air pollutant outflow is increasing with the rapid industrialization in China, which drastically changed the emission quality and quantity everywhere locally and even globally. Such change had directly impacted the air quality in downwind neighboring areas. To examine the degree of impact of the changing situation in China, we decided to use the model to study the springtime northern Taiwan air quality in terms of ozone concentration. As Taiwan is facing the direction of the predominant East Asia outflow, the pollutants in this flow becomes a deciding factor on ozone concentration in Taiwan. In order to understand this process, Chiang *et al.* (2009) conducted a coupled study using ozone observations in northern Taiwan and regional-scale chemical transport TAQM model simulations. Modeling scenario is simulated for one-month period from April 13 to May 13, 2003, which covered 75% of northern Taiwan springtime high ozone episodes. Through analyzing model simulation result of ozone distribution over this period, we can obtain the characteristic of observed ozone concentration in northern Taiwan with various scenarios in three different cases: the first is the base case includes all emission in East Asia (denoted as Sim.Total); second is the case includes the emission in East Asia (denoted as Sim.DelT) only, and the last is the case with the Taiwan emission without the East Asia emissions (denoted as Sim.Local) in modeling system. The observational data consist of hourly observations of the standard meteorological parameters and the ozone concentration.

Figure 1 shows the ozone simulation results comparing with the observations. The differences are obvious: as the Sim.Local does not contain the East Asia emission, it models the local fluctuation of ozone well, but lacks the large structures. The Sim.DelT, on the other hand, omitted Taiwan local inputs, models the large structure well but totally missed the daily variation due to the local urban activities. Not surprisingly, the Sim.Total contains all the inputs and portraits all the

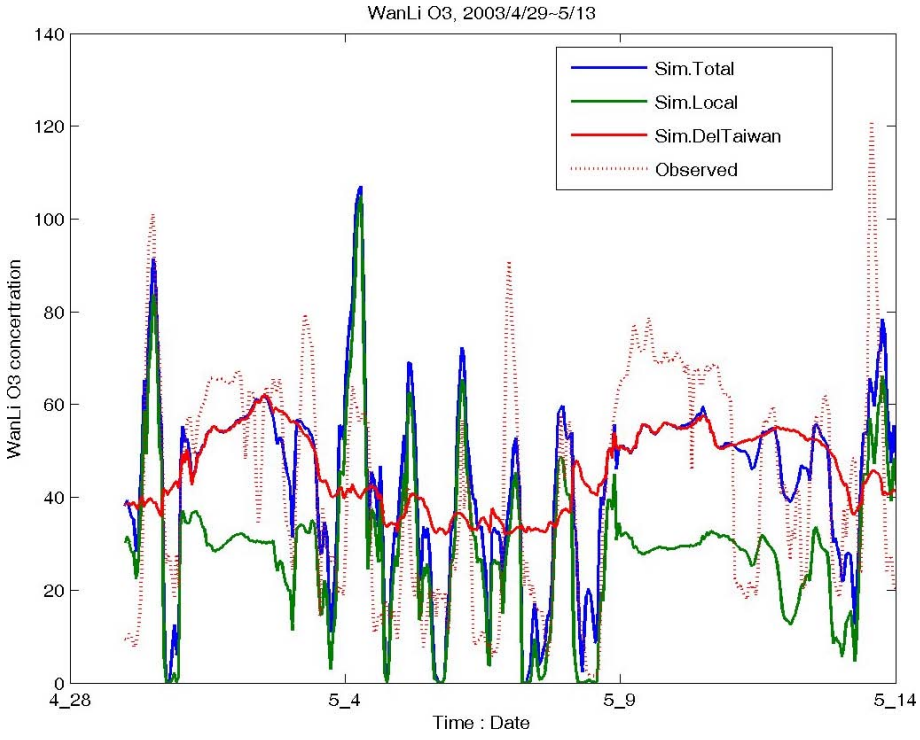



Fig. 1. The observed and simulated ozone concentration data of Wan-Li station over the period from April 29, 2003 to May 13, 2003 in Taiwan.

large and local scale phenomena. All the simulation results have certainly captured some specific pattern shown in the observation during the simulation period, but the different simulated results missed various scale phenomena at different locations. Part of the mismatch can be attributed to the emission uncertainty as shown in some previous researches [Chiang *et al.*, 2009]: the simulation results of ozone always underestimate the maximum values and overestimate the minimum values in East Asia. In general, the simulated results and observations are consistent. Based on the traditional approach of model validation, the simulation here would be designated as a success. However, this conclusion would miss the crucial diagnosis of which scale had been modeled successfully and which one still needs improvements. To achieve this goal, we have to do more detailed comparisons.

Now, we let us explore the difference in details. Hopefully, through analyzing the differences of these numerical experiments with observations, we can quantify the causes of the disagreements whether they are due to deficiencies in the observation program or the numerical model. Here, we used the observed and simulated ozone concentration data of Wan-Li station over the period from April 29, 2003 to May 13, 2003 in Taiwan.

## 3. Methods

### 3.1. *Ensemble empirical mode decomposition*

In this section, we modify the EEMD (Wu and Huang, 2009) method and employ it to analyze the correlation among the observed and simulated ozone concentration data. First, we briefly review the EEMD method.

In general, all data are amalgamations of signal and noise, i.e.,

$$x(t) = s(t) + n(t), \tag{1}$$

where $x(t)$ is the recorded data and $s(t)$ and $n(t)$ are the true signal and noise, respectively. To improve the accuracy of measurements, the ensemble mean is a powerful approach, where data are collected by separate observations, each of which contains different noises. Though the ensemble average is a powerful idea and useful tool in the laboratory, it is impossible to implement under most real observational program, for the real conditions are not repeatable. To generalize this ensemble idea, noise is introduced in EEMD to the single data set, $x(t)$, as if separate observations were indeed being made as an analog to a physical experiment that could be repeated many times. The added white noise is treated as the possible random noise that would be encountered in the measurement process. Under such conditions, the $i$th "artificial" observation will be:

$$x_i(t) = x(t) + w_i(t). \tag{2}$$

with these properties of the EMD (Huang *et al.*, 1998) in mind, the procedure of the EEMD is described as follows:

(1) Add a white noise series, $w_i(t)$, to the targeted data;
(2) Decompose the data with added white noise into IMFs using the regular EMD;
(3) Repeat Steps 1 and 2 to a predetermined numbers of times, but with different white noise series generated each time; and
(4) Obtain the (ensemble) means of corresponding IMFs of the decompositions as the final result.

The effects of the decomposition using the EEMD are these: the added white noise series cancel each other; forced by the added white noise, the IMFs in each trial have to stay within the natural dyadic filter windows. Consequently, the final ensemble mean IMFs will also stay in the dyadic filter band and, therefore, significantly reducing the chance of mode mixing and preserving the dyadic property. However, the IMFs decomposed from EEMD may not satisfy the proper definition of IMF, because the sum of IMF may not be IMF itself. As the summation operation had been used repeatedly in the ensemble procedure, the final result cannot be guaranteed to be IMFs themselves. Consequently, the overall orthogonal index amongst the components could be violated also. To remedy these shortcomings, we have instituted a rectifying step, which consisted of an additional decomposition of the

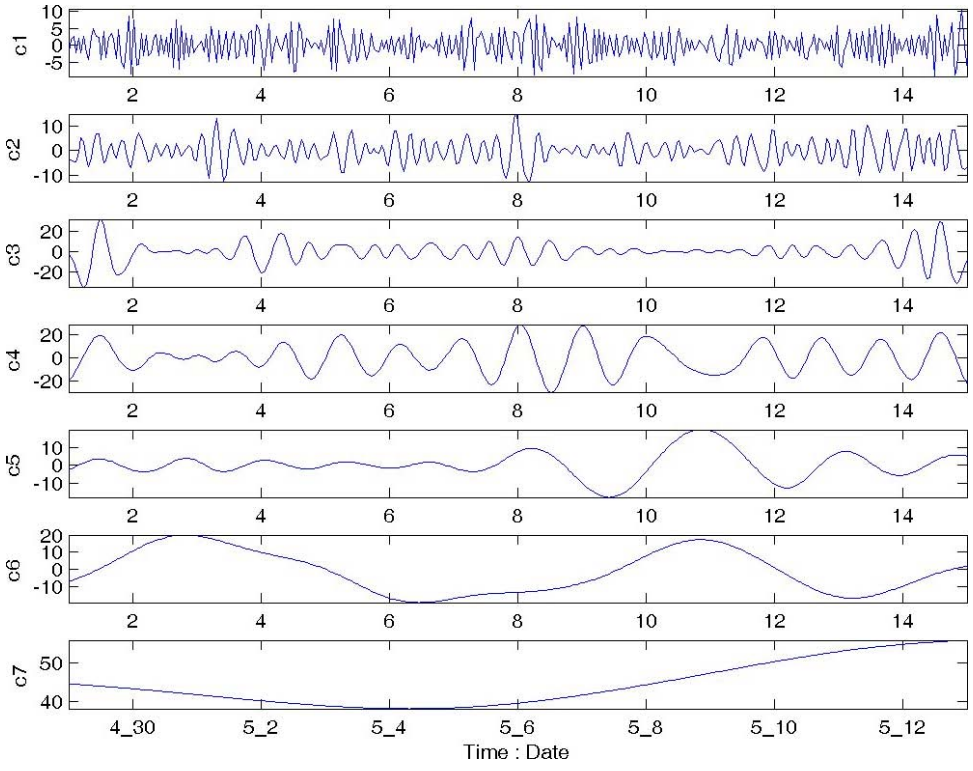



Fig. 2. IMFs of the observed data obtained by using EEMD (4, 0.6, 50) and EMD again.

ensemble IMFs using the original EMD to remove the irregularity from the already narrow-banded result to conform with the definition of IMF.

Let EEMD (S, $\sigma$, N) defined as the numbers of zero-crossing and extrema are the same for consecutive S siftings, an ensemble member of N is used, and the added white noise in each ensemble member has a standard deviation of $\sigma$ in the EEMD. After using the modified EEMD (S, $\sigma$, N), we have seven IMFs for observed and simulated data shown in Figs. 2–5 respectively. The corresponding time scales (day) of IMF3-6 for each data are given in Table 1. We can see that, from the results, at the selected urban station, the ozone concentration fluctuation has distinct cycles that include semi-diurnal, diurnal, weekly, and seasonal time scales. Moreover, the corresponding Hilbert spectra, given in Fig. 6, reveal that the diurnal cycles of Figs. 6(a)–6(c) are more significant than that of 6(d) for the Sim.DelT 03 result, but the long-term time scale is the most eminent in the Sim.DelT 03 result. The prominent time cycles of semi-diurnal, diurnal, and weekly certainly reflect the anthropogenic origin of the pollutant, for they agreed with the traffic and working patterns of the urban area selected. We next conducted a statistical test to measure the statistical significance of the results obtained.

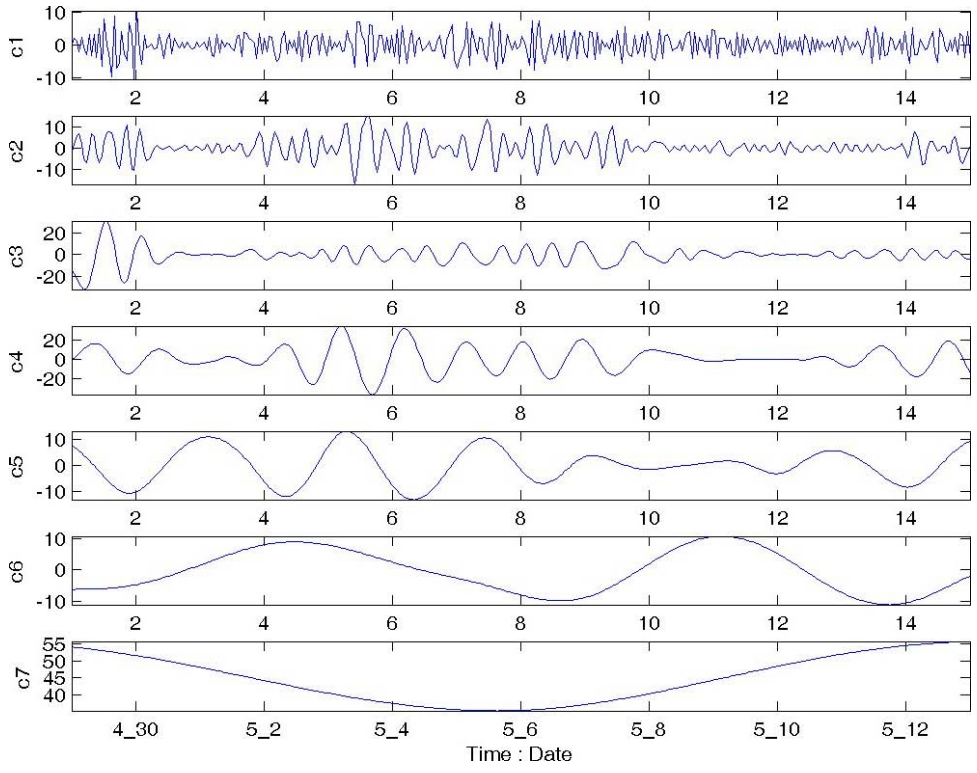



Fig. 3. IMFs of the simulated data from all emission in East Asia obtained by using EEMD (6, 0.5, 50) and EMD again.

### 3.2. *Statistical significant test*

As the data are highly variable, it is important to examine their statistical properties in details. The first step is to determine which IMFs from a dataset contain information and which IMFs may be only components of pure noise. We use the statistical significant test proposed by Wu and Huang (2004, 2005). They derived the relationship of average energy density $E$ and average period $T$ as

$$\ln E + \ln T = 0,$$

and they show that the distribution of $\ln E$ is approximately a Gaussian with a standard deviation of

$$\nu^2 = \frac{2}{nE} = \frac{2T}{n},$$

where $n$ is the length of data to be composed. Therefore, the spread lines were be approximately defined as

$$\ln E = -\ln T \pm z_{\alpha/2}\sqrt{\frac{2}{n}}e^{\ln T/2},$$

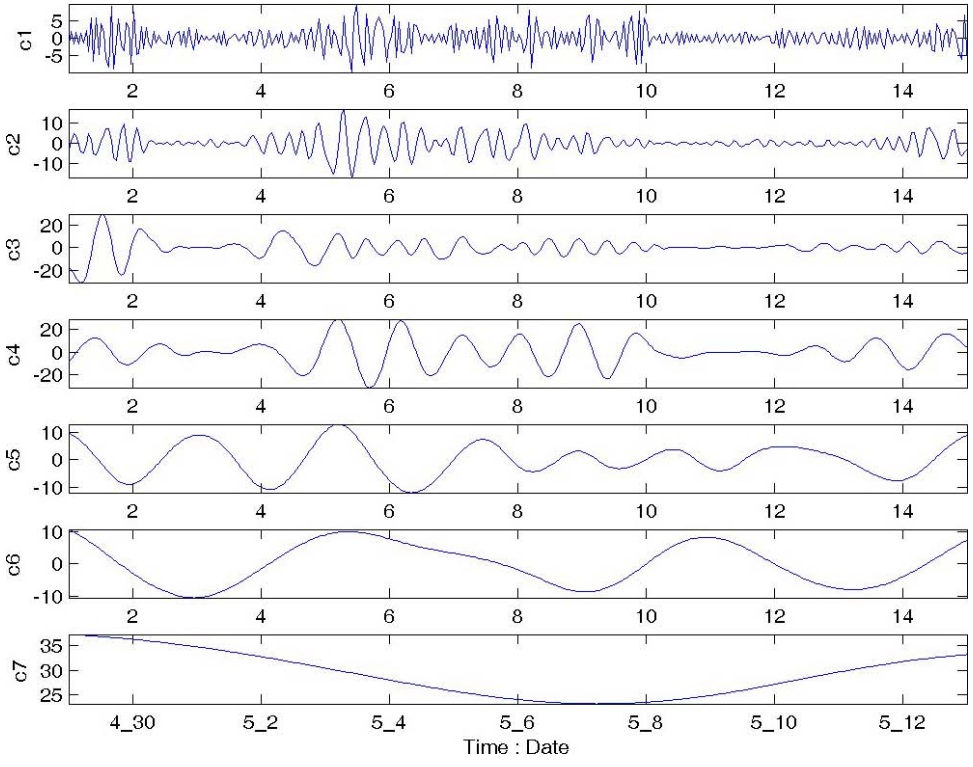



Fig. 4. IMFs of the simulated data from Taiwan only obtained by using EEMD (4, 0.5, 50) and EMD again.

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution. The method is based on the rejection of a null hypothesis, which states that all IMFs are the components of a pure white noise. If the null hypothesis is rejected, we conclude that the corresponding IMFs contain signal information.

The results of the testing of the ozone concentration data are presented in Figs. 7(a)–7(d). Notice that, both the statistical significant tests show that the first IMF can be treated as white noise. Moreover, we can see that the diurnal and weekly cycles in Fig. 7(a) are very obvious, the diurnal cycle is the most significant in Figs. 7(b) and 7(c), while the 10-days cycle is most significant in Fig. 7(d). Once the statistical significance is established, we will further examine the correlations between the observation and the simulated results to quantify the validity of the model used.

### 3.3. *Correlation coefficient*

For the validity test, we will employ the Pearson correlation coefficient to analyze the correlation among each IMF of observed and simulated data. The results reveal

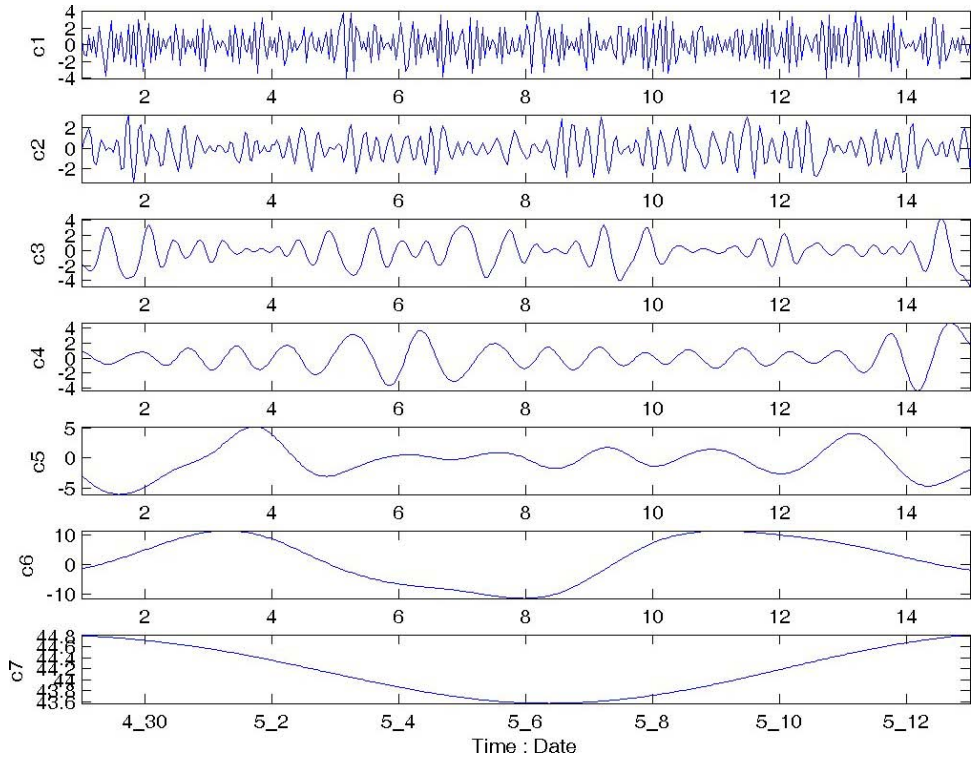



Fig. 5. IMFs of the simulated data from the emission in East Asia only obtained by using EEMD (5, 0.4, 50) and EMD again.

Table 1. The time scales (day) of IMF3-6 for each data.

| Data | IMF | | | |
|---|---|---|---|---|
| | 3 | 4 | 5 | 6 |
| Observed | 0.5 | 1 | 2 | 7.5 |
| Sim.Total | 0.5 | 1 | 2 | 7 |
| Sim.Local | 0.5 | 1 | 1 | 5.5 |
| Sim.DelT | 0.5 | 1 | 0.5 | 10 |

that the observed and simulated ozone concentrations are highly correlated (higher than 0.50) in the total signal for all the simulations. Yet the missing of local events in the Sim.DelT and large-scale events in the Sim.Local deserve additional examination and quantification. We decide to examine the correlations through their corresponding IMFs, which might show the details of the differences in each time scale. The correlation coefficients are shown in Fig. 8. As expected, IMF3, IMF4, and IMF6 of the observed and Sim.Total data and the IMF3 and IMF4 of the observed and Sim.Local data have very high correlations, but the correlation of the IMF6 of

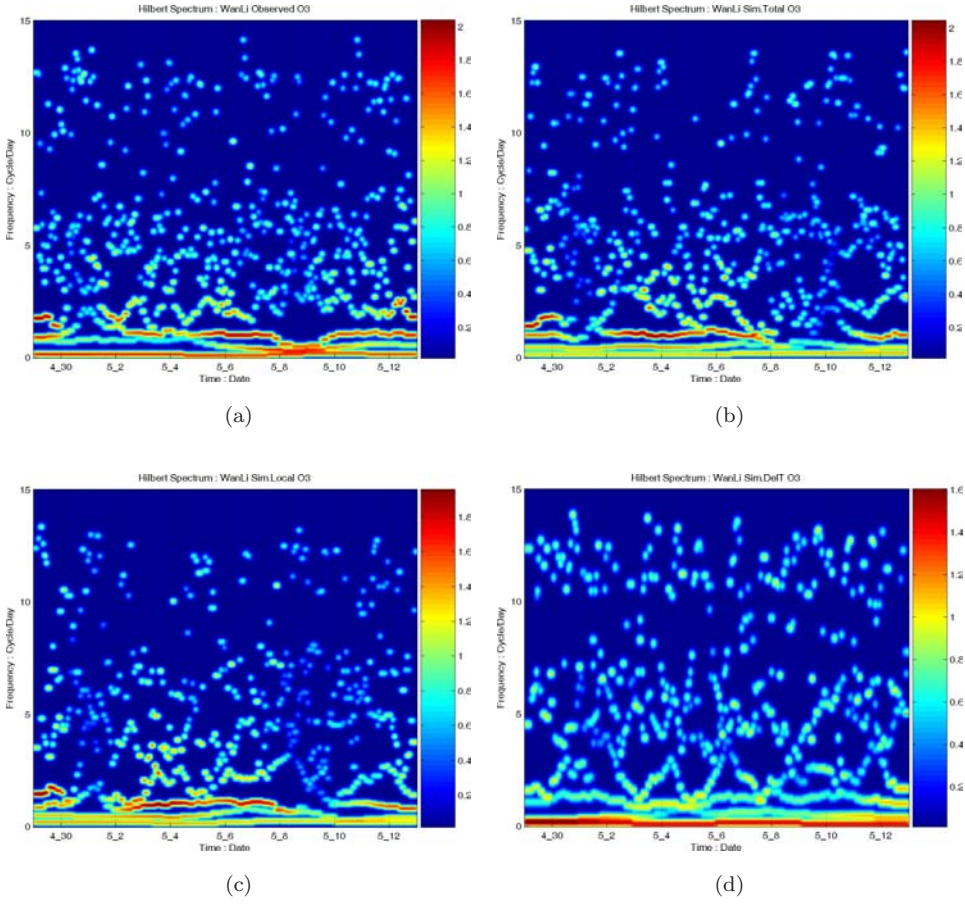
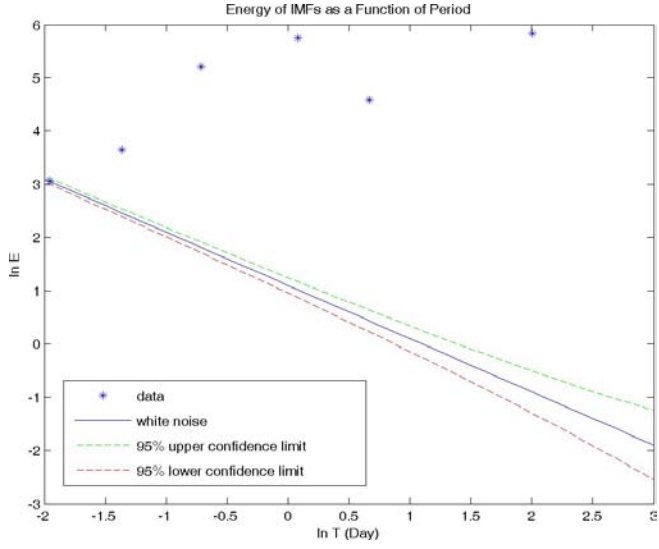
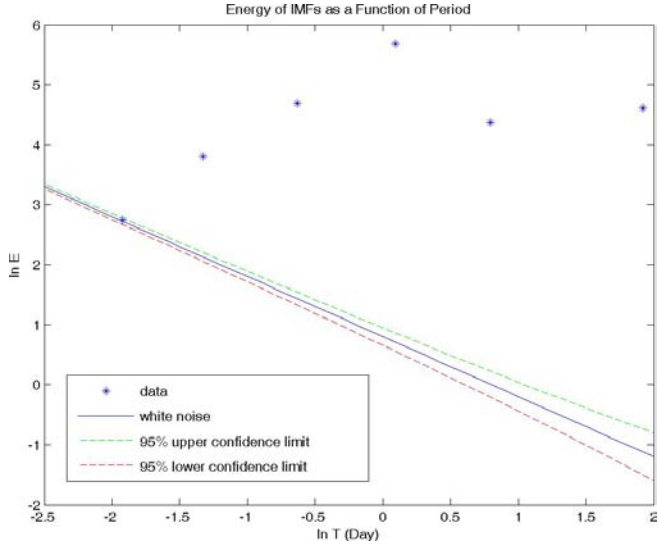



Fig. 6. Hilbert spectra of (a) the observed and simulated data, (b) Sim.Total 03, (c) Sim.Local 03, and (d) Sim.DelT 03.

the observed and Sim.Local data is only −0.04. Moreover, the IMF6 of the observed and Sim.DelT data has the highest correlation coefficient. It is worthy to point out that the correlation coefficients for IMF1, IMF2, and IMF5 are universally low for all the models. The explanation for IMF1 and IMF2 can be attribute to short time noise, for the temporal scale is shorter than semi-diurnal, and there just is any event of those scales. The lack of correlation for IMF5, representing a period of two days, can also be explained as lacking real physical phenomena at this scale. Indeed the magnitude of IMF5 is low in all the cases; the energy resided in this IMF component could result from the leakage from the dyadic filter bank formed by the EEMD. These results serve to demonstrate the anthropogenic origin of the pollutant and long-range transport effects are correctly model by the models. Of course, the model results are critically dependent on the quality and quantity of the input data. Collectively, the results also demonstrate the power of the present



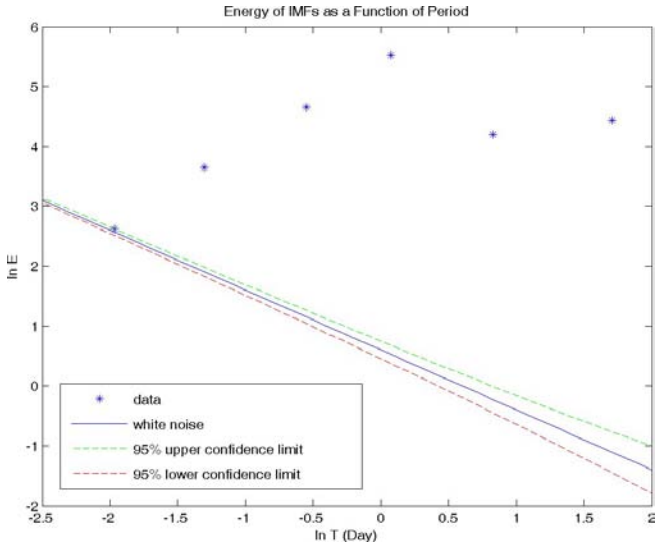

(a)



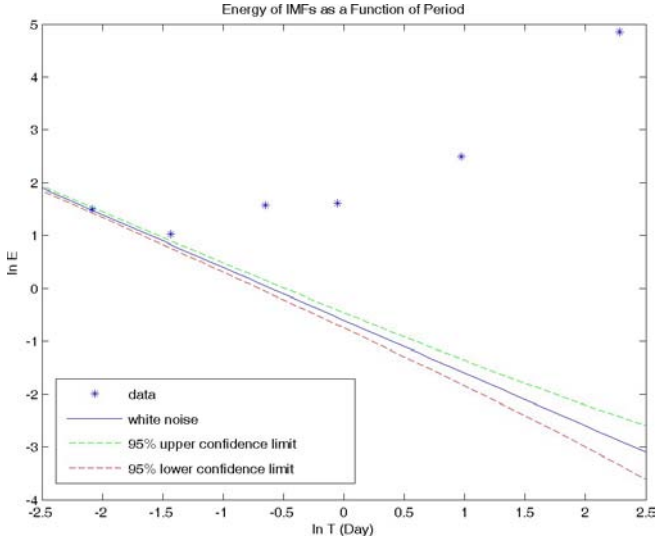

(b)

Fig. 7. Significant test of the IMFs of the ozone concentration data, (a) Observed data, (b) Sim.Total data, (c) Sim.Local data, and (d) Sim.DelT data.

approach of examining the model results in the constituting components through EEMD. As the IMF components are nonstationary and nonlinearly distorted, the correct components could not be extracted by other form of filters based on linear and stationary assumptions.




(c)




(d)

Fig. 7. (*Continued*)

## 4. Conclusions

EEMD provides detailed comparisons of observed and simulated data in various temporal scales. Using EEMD method, we established a detailed validation of the model not holistically but at various scales. We believe that this new detailed

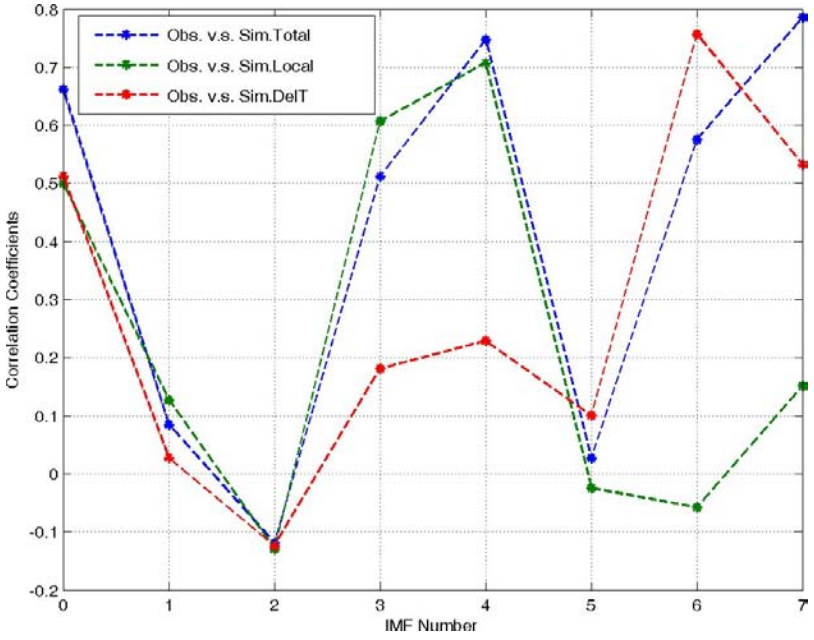



Fig. 8. The correlation coefficients of the observed and simulated ozone concentration and their corresponding IMFs. IMF0 here means the total original signal.

and mechanism-based comparison is better and reveals more detailed performance characteristics of the model. Furthermore, these detailed analyses could provide references, base for suggesting new parameters, improvements in observations, and fine-tunings for the models. Based on our study, the performance of the model is indeed good over all the scales. The missed two-day cycle could be the result of the leakage from the EMD dyadic filter, which give a small but nonzero component of relatively insignificant energy content. The results suggest that the ozone pollution source for short cycle is predominantly local, the same as the finding of Chiang *et al.* (2009).

It should also be pointed out that the newly developed time-dependent intrinsic correlation method by Chen *et al.* (2010) cold also be used here to examine the correlation in even more details. Such an approach is recommended if the goal is to understand the temporal fluctuation of the phenomena. As the present goal is model validation, the well-separated scale of the IMF component had provided a pretty uniform record for us to carry out the over all correlation; therefore, the more detailed time-dependent approach was not implemented. In the future study, the new approach should be explored.

## Acknowledgments

N. E. Huang was supported by a grant from Federal Highway Administration, DTFH61-08-00028, and the grants NSC 98-2627-B-008-004 (Biology) and NSC

98-2611-M-008-004 (Geophysical) from the National Science Council, and finally a grant from NCU 965941 that has made the conclusion of this study possible. He is also supported by a KT Lee endowed Chair at NCU.

## References

Chang, J. S., Brost, R. A., Isaksen, I. S. A., Madronich, S., Middleton, P., Stockwell, W. R. and Walcek, C. J. (1987). A three-dimensional eulerian acid deposition model: physical concepts and formulation. *J. Geophys. Res.*, **92**(D12): 14,681–14,700, doi:10.1029/JD092iD12p14681.

Chiang, C.-K., Fan, J.-F., Li, J. and Chang, J. S. (2009). Impact of Asian continental outflow on the springtime ozone mixing ratio in northern Taiwan. *J. Geophys. Res.*, **114**: D24304, doi:10.1029/2008JD011322.

Chen, X.-Y., Wu, Z. and Huang, N. E. (2010). The time-dependent intrinsic correlation based on the empirical mode decomposition, *Adv. Adaptive Data Anal.*, **2**: 233–265.

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 method and the Hilbert spectrum for non-stationary time series analysis. *Proc. Roy. Soc. London*, **454A**: 903–995.

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.

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

Wu, Z. and Huang, N. E. (2005). Statistical significant test for intrinsic mode functions, *Hilbert-Huang Transform and Its Applications*, Eds. N. E. Huang and S. S. P. Shen, World Scientific, pp. 107–127.

Wu, Z. and Huang, N. E. (2009). Ensemble empirical mode decomposition: A noise assisted data analysis method. *Adv. Adaptive Data Anal.*, **1**: 1–46.