

## ANALYSIS OF THE PRECIPITATION CLIMATE SIGNAL USING EMPIRICAL MODE DECOMPOSITION (EMD) OVER THE CASPIAN CATCHMENT AREA

F. Sabzehee<sup>1,\*</sup>, V. Nafisi<sup>2</sup>, S. Iran Pour<sup>3</sup>, B. Dutt Vishwakarma<sup>4</sup>

<sup>1</sup> Geomatics Engineering, Faculty of Civil Engineering and Transportation, University of Isfahan, Isfahan, Iran - sabzfarideh@gmail.com

<sup>2</sup> Geomatics Engineering, Faculty of Civil Engineering and Transportation, University of Isfahan, Isfahan, Iran - vnaresi@gmail.com

<sup>3</sup> Institute of Geodesy, University of Stuttgart, Stuttgart, Germany - iranpour.sivash@gmail.com

<sup>4</sup> School of Geographical Sciences, University of Bristol, University Road, Bristol BS8 1SS, UK - bd.vishwakarma@bristol.ac.uk

### Commission VI, WG VI/4

**KEY WORDS:** Empirical Mode Decomposition, Precipitation, Intrinsic mode functions, Trend, Hilbert Transform, Frequency.

### ABSTRACT:

In this paper, we employ Empirical Mode Decomposition (EMD) together with Hilbert Transform to analyze precipitation time series over the Caspian Sea catchment. Several studies have shown that EMD can extract nonlinear and non-stationary signals better than Fast Fourier Transform (FFT) and Wavelet Transform. EMD decomposes the time series into a finite number of Intrinsic Mode Functions (IMFs) in the time-frequency domain, while FFT helps us operate either in the time or the frequency domain, which fuels limitations such as the inability of nonstationary signal processing and the lack of time transparency. Although Wavelet Transform is shown to be better than FFT, it fails to detect the instantaneous frequencies and needs to have prior information about characteristics of the data. On the other hand, EMD has shown that it is almost able to determine the signal characteristics with no previous assumptions to estimate the instantaneous frequencies of the signal. In this work, EMD is applied to identify the main frequencies of precipitation time series. Thereafter, a statistical procedure is used to identify the prominent IMF of the original signal.

We use the correlation coefficient, Minkowski distance and variance test to extract the relevant and prominent IMFs. The results show that IMF 1-3 are the relevant components and are related to annual and biennial variations of precipitation time series over the Caspian catchment during 2003-2016, respectively.

### INTRODUCTION

Decomposing time series to identify dominant drivers of the observation is a routine analysis in climate studies. Decomposition can be achieved either in the time domain, the frequency domain, or time-frequency domain. Each one has its own advantages and disadvantages. The time domain methods are still considered because of time transparency and simplicity, but these methods alone cannot provide all the signal information. Fourier Transform is used to extract the frequency content of a time signal. Fast Fourier Transform (FFT) is one of the most important analyses methods of frequency domain processing. In spite of the remarkable performance of this approach, it can be noted that there are some disadvantages such as inability of nonstationary signal processing and the lack of time transparency.

For nonstationary signals, the frequencies change over time, so we need a varying window length size to extract the frequencies. The changing length window is indeed employed to achieve a better separation of time-frequency components of the signal. One solution to this problem is to use wavelet transform. Wavelet Transform analysis is a tool used to detect the time-frequency content of the signal. However, despite its good performance for nonstationary signals, the method is unable to distinguish the

instantaneous frequencies. In fact, wavelet transform needs to predefine some assumptions about data which limits the applicability of this technique. That is especially the case where the underlying behavior of the data is completely unknown. Empirical Mode Decomposition (EMD), on the other hand, is not completely based on prior assumptions. This technique attempts to split time series data into the numbers of specific frequencies, while the results are estimated in the time domain. Through this method, it will be possible to interpret the data better.

EMD is a method which was introduced by Huang et. al., 1996 for analyzing nonlinear and nonstationary time series. The method is used to analyze multivariate signals that can be separated into a number of modulated signals of amplitude and frequency, called Intrinsic Mode Functions (IMFs), with frequency bandwidth reduction [9]. An IMF is similar to a harmonic function with the difference that the IMF does not have a constant amplitude and frequency like a harmonic function (i.e. IMF has different frequencies with different amplitudes).

The first IMF component contains the highest frequency information and the last IMF contains lowest frequency due to the subtraction of local mean from the signal during various iterations [13]. The first IMF illustrates highest frequency oscillations in the original signal [1].

This process continues until the lowest frequency component remains as it is called residual. For analyzing non-stationary and non-linear data, EMD is an efficient method that can decompose a time-series into physically meaningful modes.

## DATA

The geographical focus for this study is the catchment of the Caspian Sea which is located in the northern provinces of Iran. The Caspian Sea catchment has an area of about 170,000 km<sup>2</sup> which is about 10% of the total area of the country [8]. The Caspian Sea is surrounded by five countries. Figure 1 shows location map of the Caspian Sea catchment. The river discharge, precipitation and evaporation are the processes that affect the water level in the Caspian Sea. Therefore, it is important to study the climate variations over the region through the time series analysis of the involved parameters. This can help us to identify the driver of changing Caspian Sea level and thus predict and manage concerned regional water resources in a sustainable way. We use datasets obtained from the land surface model (Global Land Data Assimilation System) GLDAS, which uses forcing data (e.g., precipitation, near-surface air temperature, specific humidity, wind speed, surface pressure and etc.), to provide estimates of land surface states (e.g., soil moisture, surface runoff, and subsurface runoff), and flux quantities (e.g., evaporation and sensible heat flux) [15]. GLDAS products are provided by four land surface models (CLM, Mosaic, Noah and VIC). In this study, GLDAS precipitation data is used to analyze the components of original signals with a temporal resolution of 1 month and a spatial resolution of 0.5° × 0.5°.



Figure 1. Location of Caspian (Khazar) Catchment.

## PROPOSED METHOD

Using a sifting process, the empirical mode decomposition (EMD) allows us to separate the observed signal ( $x(t)$ ) into a set of intrinsic mode functions (IMFs) and a residual function, given as follows:

$$x(t) = \sum_{i=1}^n C_i + \text{Residual}, \quad (1)$$

wherein  $C$  indicates IMF [9]. Each IMF demonstrates a mono-component signal. An Intrinsic Mode Function (IMF) is a function that satisfies two conditions: (1) In the whole dataset, the number of extrema and the number of zero crossings must either equal or differ at most by one; and (2) At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero [9]. Often, this method yields a higher number of IMFs than the true number of mono-component contained in the original signal.

At the end of the sifting process, the original signal is often decomposed to some IMFs. Not all of these IMFs are significant. That means, some IMFs may not have physical meaning. A lot of techniques are proposed to identify and select significant IMFs. In this paper, we employ two criteria called Minkowski Distance (dmink) and the correlation coefficient to determine genuine IMFs from the set of all extracted IMFs.

The Minkowski Distance (dmink) calculates the Euclidian distance between vectors  $X$  and  $Y$  and is defined as follows:

$$d_k = \sqrt[k]{\sum_{i=1}^n |x_i - y_i|^k}. \quad (2)$$

Where  $x_i$  and  $y_i$  are the  $i$ -th respective samples of the observed signal and the extracted IMF [5][7]. The dmink computes for the set of all extracted IMFs. It should be noted that the redundant IMFs may increase the redundant information and therefore reduce the Signal to Noise Ratio (SNR).

On the other hand, when an IMF is inappropriate, the  $d_{\text{mink}}$  is a maximum value due to differences in shape and frequency content rather than the original signal [5][7].

The second criterion is the correlation coefficient that as below equation [5][12]:

$$\rho = \frac{\sum_{i=1}^N x_i C_i}{\sqrt{\sum_{i=1}^N x_i^2 \sum_{i=1}^N C_i^2}}, \quad (3)$$

where  $N$ ,  $x_i$  and  $C_i$  are the total number of data points, the  $i^{\text{th}}$  data point of the original signal and the  $i^{\text{th}}$  data point of the IMF, respectively. High correlation displays the similarity between the appropriate IMFs and the original signal and low correlation expresses that weak relationship exists between the inappropriate IMFs and the original signal. For the purposes of this paper, we need to determine the appropriate IMF of the considered signal. So, if the  $i$ -th IMF and the original signal are similar, it can be concluded that the correlation coefficient  $\rho$  will be very high while dmink will be very low.

In this section, we briefly mention two important mathematical properties of EMD: a) the completeness of EMD can be evaluated so that the original signal are easily reconstructed by sum of all IMFs and residuals. The magnitude of the difference between the original signal and the reconstructed signal is approximately between  $10^{-15}$  and  $10^{-16}$ . b) The orthogonality of IMFs helps us to obtain the independent extracted components that are not

overlapping with each other and therefore it avoids the mode mixing problem. The orthogonality of the EMD method is still not guaranteed theoretically, however the extracted IMFs are locally orthogonal [9]. In addition to the mathematical model, EMD is an empirically based data-analysis method which means any physical meaning of the resulting IMFs can not be reliable [2].

An invaluable advantage of the EMD technique is that each IMF component after performing a Hilbert transform can be transformed to time-frequency space and the amplitude and instantaneous frequency can be extracted. The instantaneous frequency (IF) of a signal is the variation of the phase shift of the sinusoidal signal. The easiest way to compute the instantaneous frequency is by using the Hilbert transform which can be determined as follows:

$$Z_i(t) = IMF_i(t) + jH[IMF_i(t)] = a_i(t)e^{j\theta_i(t)}, \quad (4)$$

where  $H[\cdot]$  is the Hilbert Transform operator,  $a_i(t)$  and  $\theta_i(t)$  are instantaneous amplitude and phase of the  $i$ -th IMF, respectively. The Hilbert Transform provides a phase shift of  $\pm\pi/2$  to all frequency components, while leaving the magnitudes unchanged [6]. The corresponding instantaneous frequency of the  $i$ -th IMF can be extracted as follows [6]:

$$\omega_i(t) = \frac{d\theta_i(t)}{dt}, \quad (5)$$

Hilbert spectrum describes the joint distribution of the amplitude and frequency content of the signal as a function of time and frequency [10][11]. The resulting time-frequency-energy representation of the data is called the Hilbert-Huang spectrum and provides information similar to that identified in a continuous wavelet transform [14].

The mode mixing is an annoying problem in EMD method that disturbs the physical uniqueness of decomposition. When phenomena of mode mixing occur, an IMF can discontinue to have physical meaning by itself, suggesting incorrectly that there may be different physical processes behind? a mode. The Ensemble Empirical Mode Decomposition (EEMD) method is developed to improve the performance of the EMD method and reduce mode-mixing [16].

## RESULT

The precipitation sequence is divided into five IMFs and one residual. Figure 2 shows the fluctuation characteristics from high frequency to low frequency at different time scales.

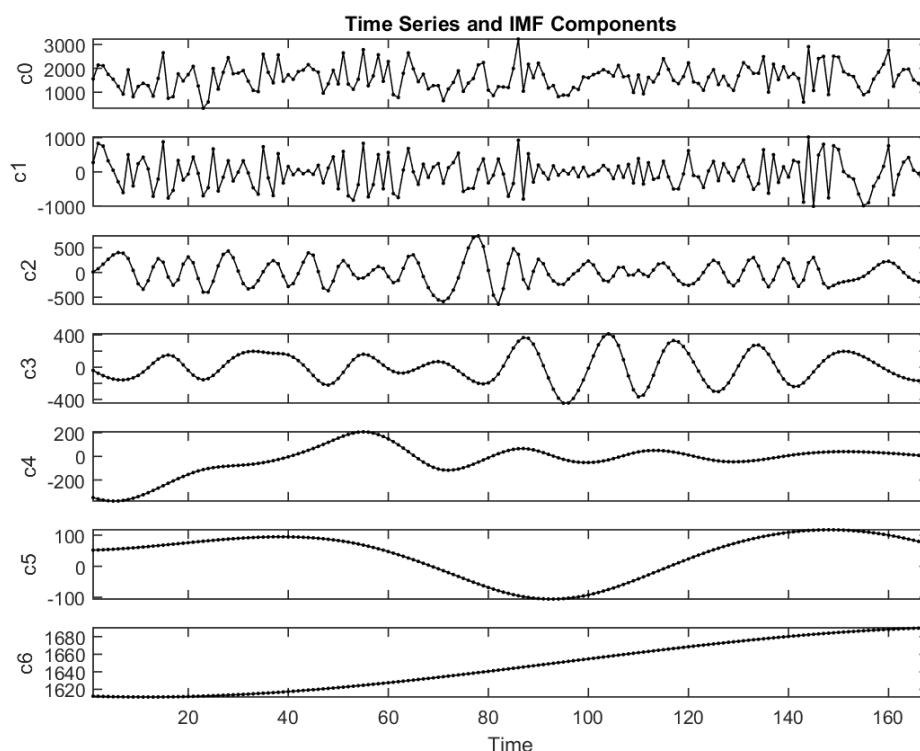


Figure 2. EMD of precipitation data over Caspian catchment, ( $C_0$ ) 14-year monthly precipitation, ( $C_1$ - $C_6$ ) decomposed five IMF components and the final residue( $C_6$ ).

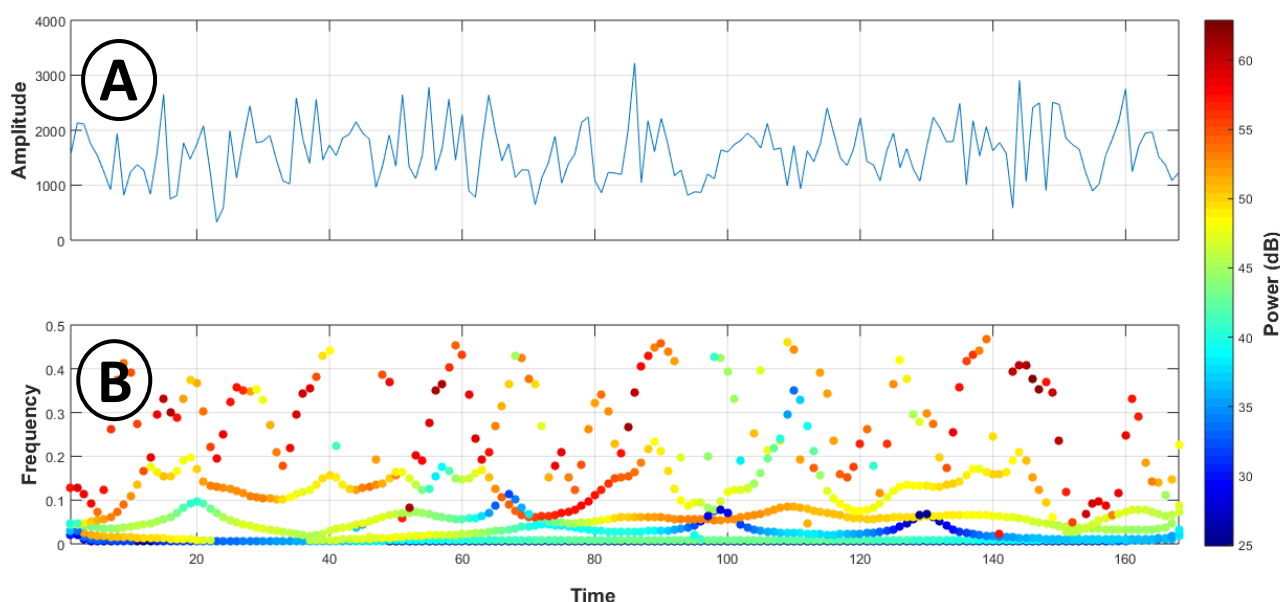


Figure 3. Hilbert transform of the signal decomposed using EMD; A) The synthetic time series by summing IMFs, and B) signal power plotted in time-frequency.

The residual indicates the long-term trend in the average of the precipitation signal over the Caspian catchment. Figure 3 presents the synthetic precipitation time series by summing IMFs and its residual (Figure 3. A), and signal power plotted in time-frequency (Figure3.B). It appears as though the signal consists of broadband energy and several frequencies. This can be seen in Figure 3, the instantaneous frequencies are widely distributed and the number of frequencies are continuous phenomenon along the time axis and some of them happen in the moment. It is important to select stable and low frequencies components of the original signal. The variance of each IMF is shown in Figure 4: variances of IMF 1-3 are higher than other IMFs. The highest variance among the IMFs represents the greatest impact on the shape of the original time series. Each IMF has a specific meaning. The results of the coefficient and dmink are summarized in Table 1 and demonstrated by Figure 5. The values of the coefficient of IMF1, IMF2 and IMF3 are higher than others and may be considered as relevant IMFs. Figure 5 shows that for dmink, the values of the first three IMFs have smaller distance and for  $\rho$ , the first three IMFs have higher correlation than other IMFs. A method introduced to identify relevant and irrelevant IMFs that can lead to improve overall EMD process [12].

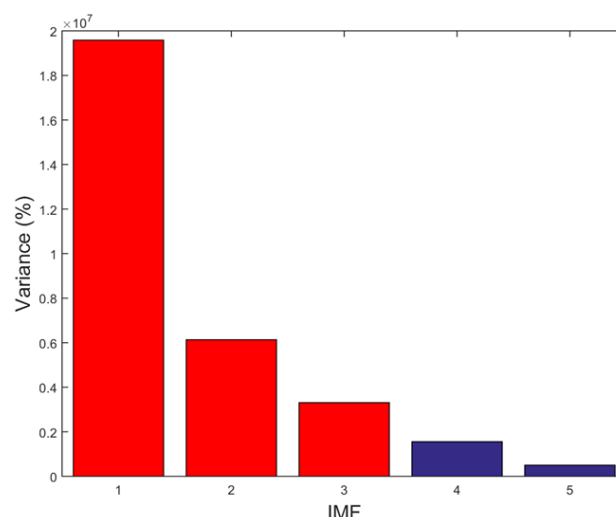


Figure 4. Variance of IMFs of precipitation using EMD algorithm over Caspian catchment. IMF1, IMF2 and IMF3 (red bars) are significant IMFs among others IMFs.

Real Signal ( $C_0$ )		
	$\rho$	$d_{mink}$
$C_1$	0.7814	67976.66
$C_2$	0.3990	101016.70
$C_3$	0.2932	105098.38
$C_4$	0.1595	108360.66
$C_5$	0.1216	109073.64

Table 1. The values of the correlation coefficient and the Minkowski distance for the precipitation signal.

Since the generated IMFs with a finite number are considered to be nearly orthogonal to each other and sum of the IMF components plus the its residue lead to reconstructing and an orthogonal representation of the original signal. Each IMF will have a relatively strong correlation with the original signal and the irrelevant components will have weak correlation with the original signal.

Following the proposal of [3] a threshold value is derived (equation 6) as  $\tau$  to extract relevant IMFs:

$$\tau = \frac{\max(\mu_i)}{10 \times (\mu_i) - 3}, \quad (6)$$

where  $\mu_i$  is the correlation coefficient. If  $\mu_i \geq \tau$ , the  $i$ -th IMF is suitable and keep it, else remove it and add it to the residual (see Figure 7). Each IMF component has physical meaningful interpretation and illustrates the oscillation of inherently characteristics at different time scales in the original series. The last component (residual) demonstrates the trend of the original data over time. The actual physical meaning contained in each IMF component can be determined by the significance test, and different confidence levels indicate the strength of the physical meaning [4].

The IMFs extracted from the EMD technique are not all significant. It is important for users to be aware of the significant components in order to reduce data redundancy. For identifying significant IMFs, there is a method based on the results of the determination of higher energy IMFs. The energy density of the  $i$ -th IMF component can be calculated as follows [4]:

$$E_i = \frac{1}{N} \sum_{t=1}^N |c_i(t)|^2, \quad (7)$$

Where  $N$  is the length of the IMFs and  $c_i(t)$  is the  $i$ -th IMF component. Figure 6 shows the energy density of the IMF 1-3 and residual (IMF 6) have higher values than others and can be considered as the significant components. The vertical scale is logarithmic (equal scale between powers of 10) and the horizontal scale is linear (IMFs components). It is concluded that the significant components are IMF 1–3 and IMF 6 (residual). In Figure 8 shows that IMF1, IMF2 and IMF3 are respectively related to semi-annual, annual and biennial variations and the residual is more or less associated with the trend. It should be noted that trend signal for precipitation is ascending (a positive slope) over the Caspian catchment during 2003 until 2016.

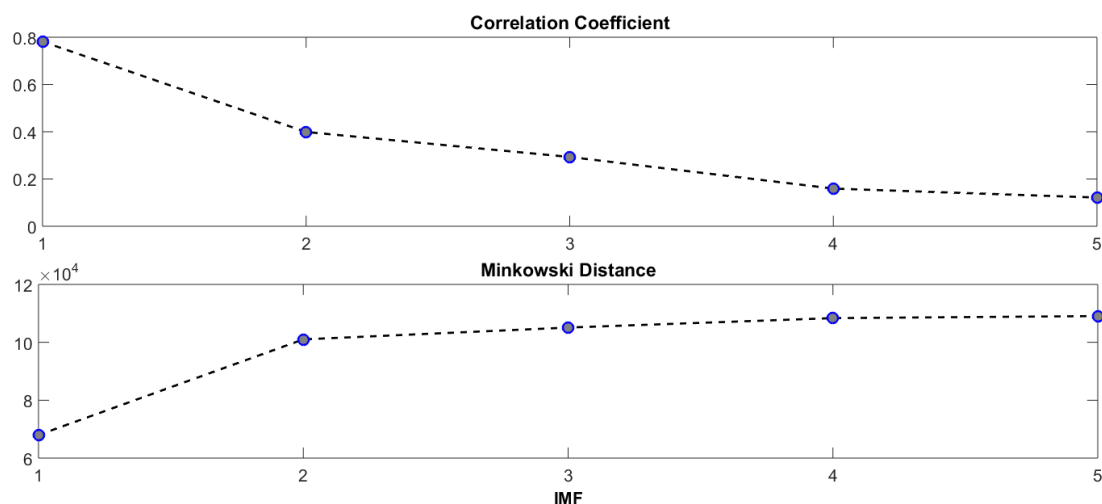


Figure 5. The values of correlation coefficient, Minkowski distance for the five IMFs.

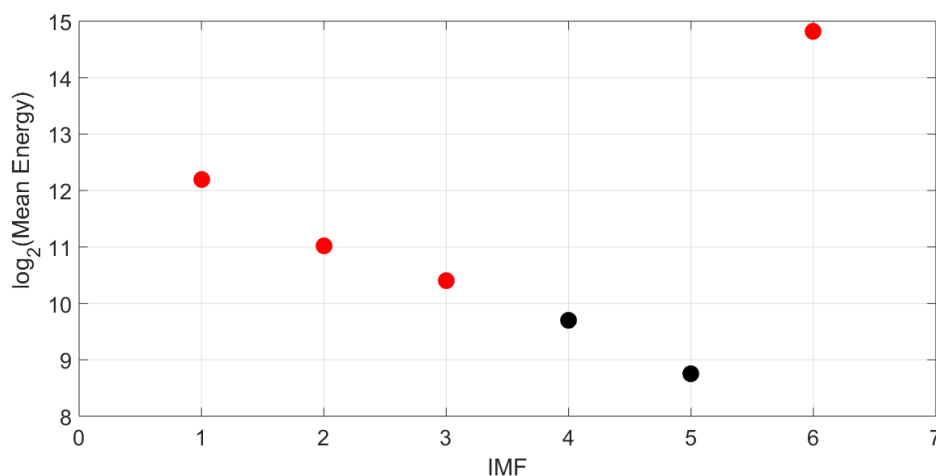


Figure 6. The energy density of IMFs component of average monthly precipitation time series during 2003-2016 in Caspian catchment.

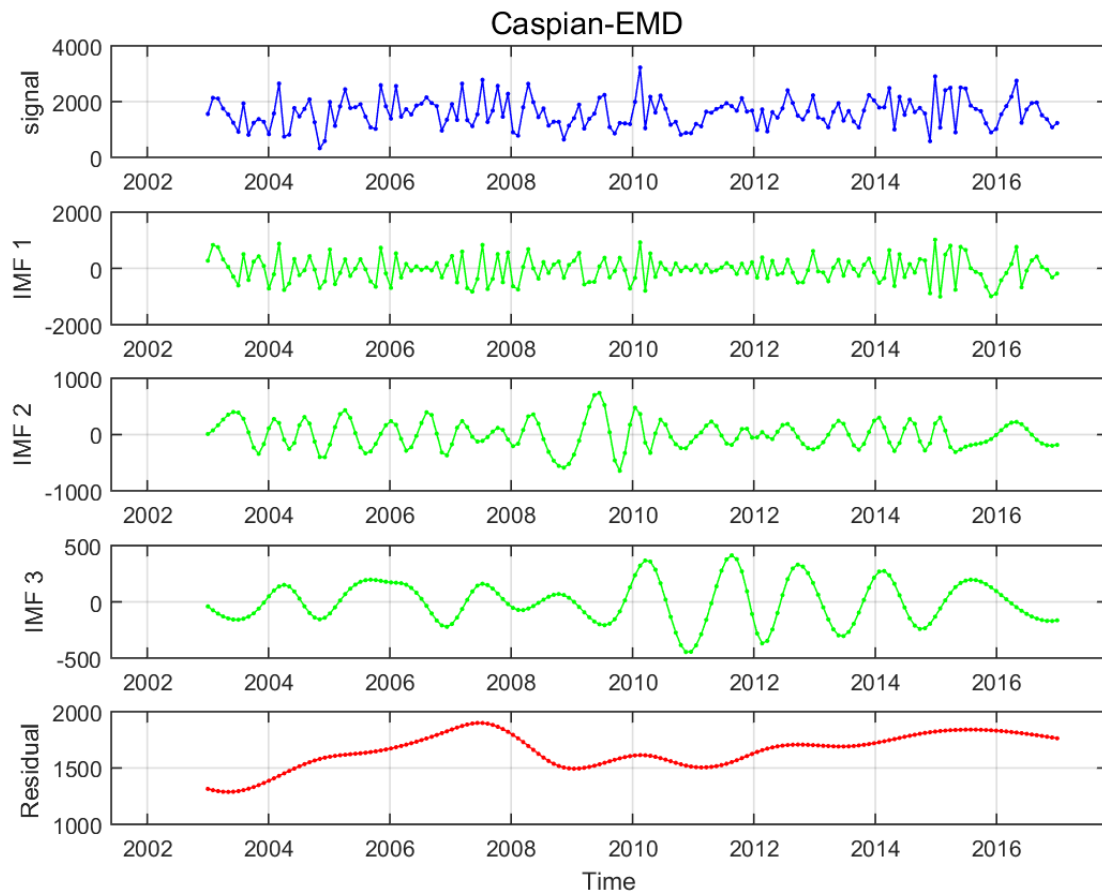


Figure 7. Top row is the original signal, middle rows are the significant IMFs and last row is summing other IMFs and residual.

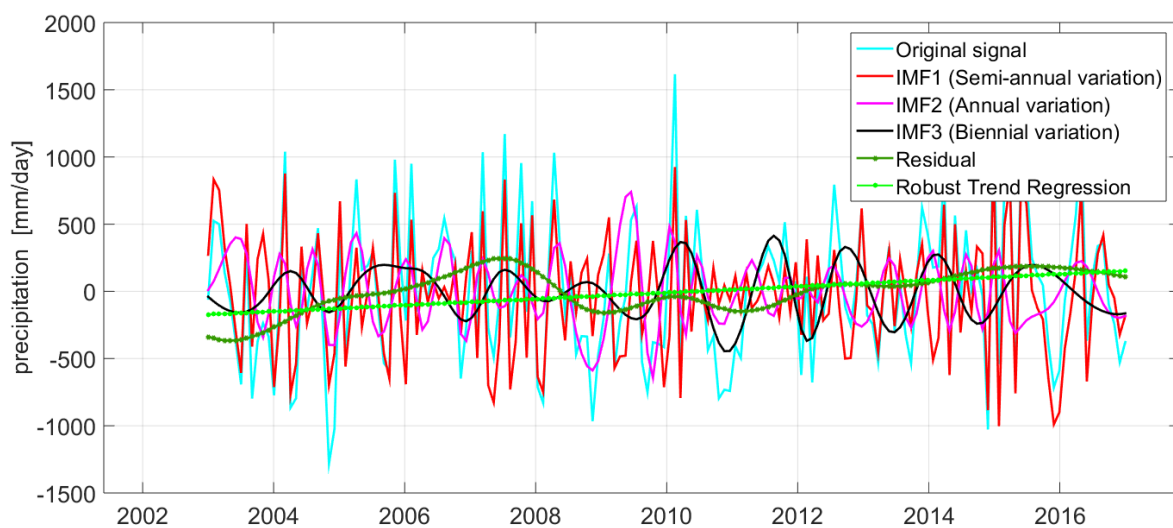


Figure 8. Semi-Annual, Annual and Biennial variations of precipitation and their comparisons with the original precipitation anomaly over Caspian catchment from 2003 until 2016.

## CONCLUSION

A separation process of EMD is based on the intrinsic and the physical properties of the data. The technique does not require

any pre-definition of parameters and data processing is entirely automated without an interpreter. In this paper, we used the correlation coefficient, Minkowski distance and variance test to extract the relevant and prominent IMFs. The results show that

IMF 1-3 are the valuable and significant components and are related to semi-annual, annual and biennial variations of precipitation time series over the Caspian catchment during 2003-2016, respectively. The so-called varying trend of the precipitation time series has a slight ascending slope.

## REFERENCES

- [1] Oztuerk, A., Seker, S., 2010. On the Frequency Resolution of Improved Empirical Mode Decomposition Method, vol. 5.
- [2] Al-Subari, K. S. A., 2017. A Study of Biomedical Time Series Using Empirical Mode Decomposition: Extracting event-related modes from EEG signals recorded during visual processing of contour stimuli. PhD, Universität Regensburg.
- [3] Ayenu-Prah, A. and Attah-Okine, N., 2010. A criterion for selecting relevant intrinsic mode functions in empirical mode decomposition, *Advances in Adaptive Data Analysis, Theory and Applications*, vol. 2, no. 1, pp. 1–24.
- [4] Bai, L., Xu, J., Chen, Zh., Li, W., Liu, Z., Zhao, B., Wang, Z., 2015. The regional features of temperature variation trends over Xinjiang in China by the ensemble empirical mode decomposition method. *INTERNATIONAL JOURNAL OF CLIMATOLOGY Int. J. Climatol.* 35: 3229–3237.
- [5] Boutana, D., Benidir, M., Barkat, B., 2010. On the selection of Intrinsic Mode Function in EMD method: Application on heart sound signal. In: *3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, Rome, Italy, pp. 1-5.
- [6] Cooke, M., 1993. *Modeling Auditory processing and organization*, Cambridge University Press, Cambridge.
- [7] Moctezuma, L. A., Molinas, M., 2018. EEG-based subjects identification based on biometrics of imagined speech using EMD, *Proc. Int. Conf. Brain Inform*, pp. 458-467.
- [8] <https://water.fanack.com/iran/water-resources-in-iran/>.
- [9] Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. R. Soc. London, Ser. A*, 454, 903–993.
- [10] Khademul Islam Molla, M.d., Rani Ghosh, P., and Hirose, K., 2011. Bivariate EMD-Based Data Adaptive Approach to the Analysis of Climate Variability, *Discrete Dynamics in Nature and Society*, vol. 2011, Article ID 935034, 21 pages.
- [11] Khademul Islam Molla, M.d., Hirose, K., Minematsu, N., 2004. Audio source separation from the mixture using empirical mode decomposition with independent subspace analysis, *Proc. ICSLP*, pp. 2449-2452.
- [12] Peng, Z.K., W. Tse, P. and Chua, F.L., 2005. An improved Hilbert–Huang transform and its application in vibration signal analysis, *Journal of Sound and Vibration* 286, 187–205.
- [13] Saini, M.K., Beniwal, R.K., 2018. Recognition of Multiple PQ Issues using Modified EMD and Neural Network Classifier. *IJEEE*, 14 (2) :188-203.
- [14] Torrence, C. and Compo, G. P., 1998. A practical guide to wavelet analysis, *Bull. Am. Meteorol. Soc.*, 79, 61–78.
- [15] Wang, W., Cui, W., Wang, X., & Chen, X., 2016. Evaluation of GLDAS-1 and GLDAS-2 forcing data and Noah model simulations over China at the monthly scale. *Journal of Hydrometeorology*, 17(11), 2815–2833.
- [16] Wu, Z.H., Huang, N.E., 2009. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.* 2009, 1, 1-41.