# Rainfall Forecasting Based on Ensemble Empirical Mode Decomposition and Neural Networks

Juan Beltrán-Castro[1], Juliana Valencia-Aguirre[1], Mauricio Orozco-Alzate[1], Germán Castellanos-Domínguez[1], and Carlos M. Travieso-González[2]

[1] Universidad Nacional de Colombia - Sede Manizales, Colombia - Signal Processing and Recognition Group - Km. 7, Vía al Magdalena, Campus La Nubia - Manizales, Colombia
{jdbeltranc,jvalenciaag,morozcoa,cgcastellanosd}@unal.edu.co
[2] Universidad de Las Palmas de Gran Canaria, Technological Centre for Innovation in Communications (CeTIC), Campus Universitario de Tafira, s/n, Las Palmas de Gran Canaria, Spain
ctravieso@dsc.ulpgc.es

**Abstract.** In this paper a methodology for rainfall forecasting is presented, using the principle of decomposition and ensemble. In the proposed framework, the employed decomposition technique is the Ensemble Empirical Mode Decomposition (EEMD), which divides the original data into a set of simple components. Each component is modeled with a Feed Forward Neural Network (FNN) as a forecasting tool. Finally, the individual forecasting results for all components are combined to obtain the prediction result of the input signal. Experiments were performed on a real-observed rainfall data, and the attained results were compared against a single FNN model for the raw data, showing an improvement on the system performance.

**Keywords:** Forecasting, Neural Networks, Ensemble Empirical Mode Decomposition, Rainfall.

## 1 Introduction

Rainfall forecasting is an interesting field of research with important applications, such as management of water resources, and the evaluation of drought and flooding events. In particular, the analysis of this variable is relevant in tropical countries like Colombia where the agriculture is an essential part of the economy. Moreover, unanticipated flash floods are very destructive and threaten human lives and properties. Then, among all weather happenings, rainfall plays an imperative part in human life [1]. However, obtaining accurate forecast results is a challenging task, since rainfall is difficult to understand and model, due to the complexity of the atmospheric processes that generate it [3].

In the last decades, it has been shown that Artificial Neural Networks (ANNs) are suitable to predict weather variables with accurate results. This technique is

I. Rojas, G. Joya, and J. Cabestany (Eds.): IWANN 2013, Part I, LNCS 7902, pp. 471–480, 2013.

able to capture the complex nonlinear relation between input and output without the physics being explicitly provided. Another advantage of ANNs is that they avoid the use of differential equations [6, 11]. Regarding practical applications, ANNs have also been used in rainfall forecasting [1, 9, 13].

In spite of the generalization ability of ANNs and due to the nonlinear and non-stationary nature of the rainfall time series, it is necessary the search for analysis alternatives that improve the accuracy of predictions. For instance, it has been used the decomposition and ensemble principle introduced by Huang et al. in [7], which aims to simplify the forecasting task by dividing it into forecasting subtasks [2,12]. The goal of the ensemble is to formulate a consensus forecasting on the input data. In general, this technique is suitable for time series analysis, presenting even better results than those obtained by other techniques such as Wavelet and Fourier decomposition [8, 10].

This paper proposes a methodology for rainfall forecasting, adopting the above-mentioned decomposition and ensemble principle. In the proposed framework, the employed decomposition technique is the Ensemble Empirical Mode Decomposition (EEMD), and as a forecasting tool, the Feed Forward Neural Network (FNN). The experiments are performed on a real-observed rainfall signal. The obtained results are compared against a single FNN model for the raw data.

The remaining part of the paper is organized as follows. The proposed methodology is detailed in Sect. 2, where a brief description of Empirical Mode Decomposition (EMD), EEMD and FFNs are also presented. Section 3 presents the experiments and the obtained results. Finally, the paper is concluded in Sect. 4.

## 2 Methodology

### 2.1 Empirical Mode Decomposition (EMD)

The EMD method was firstly introduced in [7]. The essence of the method is to empirically identify the intrinsic oscillatory modes by their characteristic time scales in the data in order to decompose them accordingly. This guided the authors to the definition of a class of functions, based on their local properties, designated as intrinsic mode function (IMF) for which the instantaneous frequency can be defined everywhere. According to [7], an IMF is a function that satisfies two conditions:

1. In the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one.
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.

Using the definition, any data series $x(t)(t = 1, 2, \ldots, n)$, can be decomposed according to the following sifting procedure.

1. Identify all the local extrema of $x(t)$.
2. Connect all local extrema by a cubic spline line to generate its upper and lower envelopes $x_{up}(t)$ and $x_{low}(t)$, respectively.
3. Compute the mean $m(t)$ as $m(t) = (x_{up}(t) + x_{low}(t))/2$.
4. Extract $m(t)$ from the data series and define $c(t) = x(t) - m(t)$.
5. Check the properties of $c(t)$: (i) if $c(t)$ meets the above two requirements, an IMF is derived and then replace $x(t)$ with the residual $r(t) = x(t) - c(t)$; (ii) if $c(t)$ is not an IMF, replace $x(t)$ with $c(t)$.
6. Repeat Steps 1 to 5 until some stopping criterion is satisfied.

Finally, after the above procedure is complete, the original data series $x(t)$ can be expressed by

$$x(t) = \sum_{j=1}^{n} c_j(t) + r_n(t), \tag{1}$$

where $n$ is the number of IMFs, $r_n(t)$ is the residue, which is the main trend of $x(t)$, and $c_j(t)(j = 1, 2, \ldots, n)$ are the IMFs. All IMFs are nearly orthogonal to each other, and all have nearly zero means. Thus, the data series can be decomposed into $n$ IMFs and one residue. The IMF components contained in each frequency band are different and change with variation of the data series $x(t)$, while $r_n(t)$ represents the central tendency of the data series $x(t)$.

### 2.2 Ensemble Empirical Mode Decomposition (EEMD)

One of the major drawbacks of the original EMD is the mode mixing. It is defined as any IMF consisting of oscillations of dramatically disparate scales, often caused by intermittency of the driving mechanisms. When mode mixing occurs, an IMF can cease to have physical meaning by itself, suggesting falsely that there may be different physical processes represented in a mode. Ensemble empirical mode decomposition (EEMD) represents a major improvement of the EMD method, eliminating largely the mode mixing problem and preserving physical uniqueness of decomposition [15]. One of the basic principles is that adding white noise to the data series will provide a relatively uniform reference scale distribution to facilitate EMD, making EEMD a truly noise-assisted data analysis method. The EEMD is developed as follows:

1. Add a white noise series to the targeted data.
2. Decompose the data with added white noise into IMFs usign EMD.
3. Repeat step 1 and step 2 again and again, but with different white noise series each time.
4. Obtain the (ensemble) means of corresponding IMFs of the decompositions as the final result.

One of the effects of the decomposition using the EEMD are that the added white noise series cancel each other in the final mean of the corresponding IMFs. This effect should decrease following the well-established statistical rule:

$$\varepsilon_n = \frac{\varepsilon}{\sqrt{N}}, \tag{2}$$

where $N$ is the number of ensemble members, $\varepsilon$ is the amplitude of the added noise, and $\varepsilon_n$ is the final standard deviation of the error, which is defined as the difference between the input signal and the corresponding IMFs. In most cases, as indicated in [15], it is suggested to add noise of an amplitude that is about 20% of the standard deviation of the data. However, when the data is dominated by high-frequency signals, the noise amplitude may be smaller; similarly, when the data is dominated by low-frequency signals, the noise amplitude may be increased.

### 2.3 Feed Forward Neural Network (FNN)

The main reason for selecting a FNN as a predictor is that it is often viewed as a "*universal approximator*" [5]. In [5] and [14] it was found that a three-layer FNN with an identity transfer function in the output unit and logistic functions in the middle-layer units can approximate any continuous function arbitrarily well, given a sufficient amount of middle-layer units. That is, neural networks have the ability to provide a flexible mapping between inputs and outputs [16].

### 2.4 Proposed EEMD-FNN Scheme

The central idea of the proposed methodology is to address the task of predicting a complex signal, first decomposing it into simpler parts that can be modeled more accurately, and then aggregating these results into the final prediction of the original signal. The method follows these main steps:

1. Decompose the rainfall signal (time series) using EEMD to obtain a set of IMF components and a residue.
2. Use the FNN model as a forecasting tool for each extracted IMF and the residue component, and make the individual one step ahead prediction for each one.
3. Combine (aggregation) the individual forecasting results of all IMF components to obtain the final prediction result of the input signal.

The parameter determination of the neural networks are based on the methodology proposed in [4], in which the Partial Autocorrelation Function (PACF) is used to determine the input variables of each FNN model, looking for those lags at which the PACF is outside of the confidence interval, and the number of hidden nodes is equal to $2p + 1$, where $p$ is the number of inputs.

## 3 Experimental Analysis

### 3.1 Data Description and Performance Measures

The rainfall data were registered in a meteorological station at Manizales city, Colombia. The observations were taken with a time step of 5 minutes, and then

a daily time series is derived summing the whole register per day. The time series covers the period from 01-Jan-2005 to 31-Dec-2008, for a total of 1461 days, see Fig. 1. The data are randomly divided into three parts for training, validation and test, respectively assigning 70%, 15%, 15%. These divisions are used to train each FNN model with its respective IMF, referred hereafter as component, in a more general way.

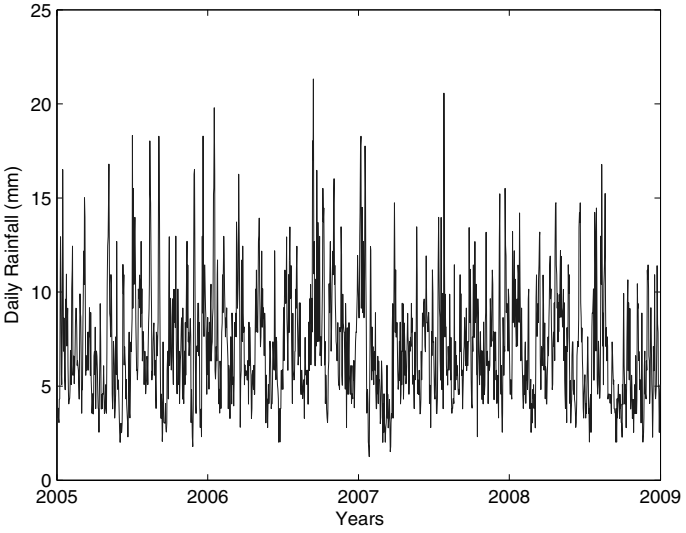



**Fig. 1.** Daily rainfall computed for a meteorological station at Manizales city

Three criteria are used for evaluating the forecasting performance: the Mean Absolute Percentage Error (MAPE), the Mean Square Error (MSE) and the Mean Absolute Error (MAE), calculated as

$$MAPE = \frac{1}{M} \sum_{t=1}^{M} \left| \frac{x_t - y_t}{x_t} \right| \cdot 100 \tag{3a}$$

$$MSE = \frac{1}{M} \sum_{t=1}^{M} (x_t - y_t)^2 \tag{3b}$$

$$MAE = \frac{1}{M} \sum_{t=1}^{M} |x_t - y_t| \tag{3c}$$

where $M$ is the number of elements in the set, $x_t$ is the actual value and $y_t$ the predicted one. In these performance measures, the smaller, the better.

### 3.2 Experimental Results

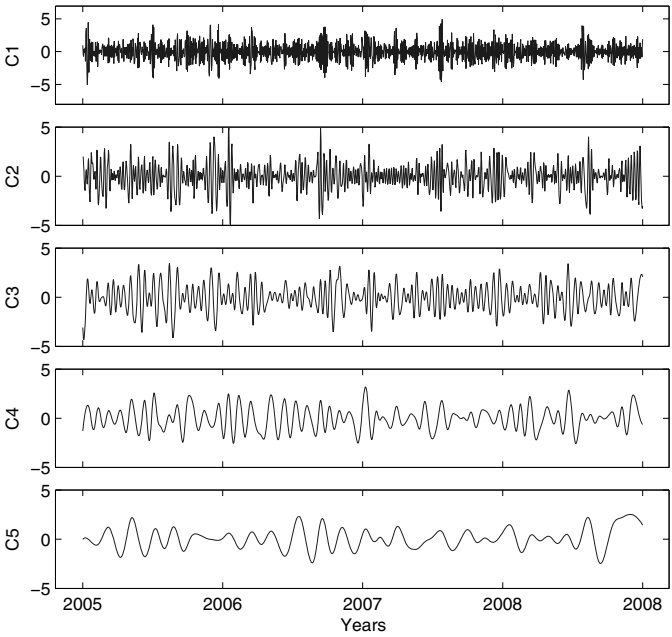
The decomposition using EEMD is presented in Fig. 2, for a total of 9 components and the residue. The parameters used were an amplitude of added noise equals to 0.2 and an ensemble of 100 trials. Regarding to this technique, the EEMD code package proposed by [15] was used, which is available at `http://rcada.ncu.edu.tw/` in MATLAB language. Minor changes were made to the code relative to enhance computation time, but in essence the original procedure was preserved. For the FNNs, the Neural Network Toolbox-Version 7.0.1 (R2011a) for MATLAB was used for the modeling task.

The entire process of training all models is repeated 10 times with a different division of the data. The box plot for each FNN model for the training set is shown in Fig. 3; on each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points and outliers are plotted individually. The horizontal axis represents each predicted component. As shown, the high frequency components exhibit worse performance than the lower ones, probably because of their complexity. It can also be noted that, in general the lowest components have a higher dispersion than the highest ones, and that the outliers represent cases in which the tests were significantly worse than the others.

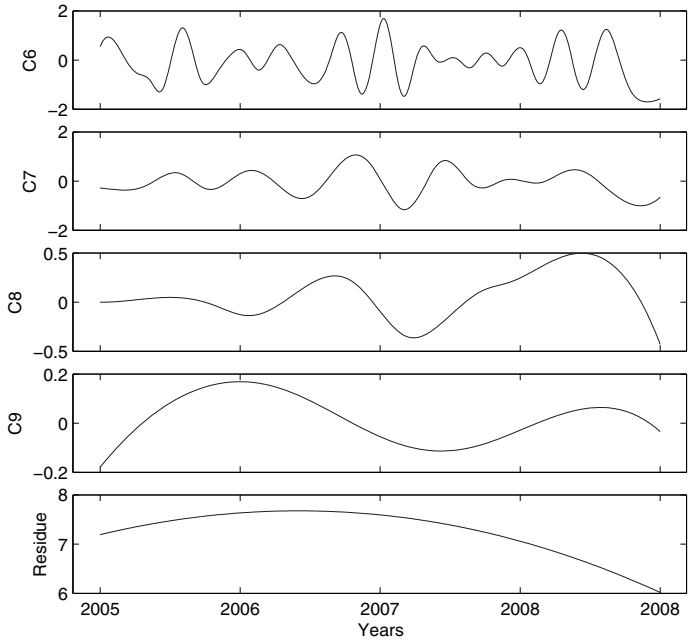
The final step in the proposed EEMD-FNN is the combination of the individual forecasting results of all components to obtain the final prediction result of the input signal. This can be done summing it all, according to (1). For all the 10 experiments, the mean for each performance measure and its Standard Error (SE) are computed and thus a reliable estimate of the performance is obtained. Table 1 summarizes these estimates for the proposed methodology, which is compared against a single FNN model for the original signal, as a basic benchmark. This FNN model was estimated in the same way for all components, i.e. using the PAFC to determine the inputs and setting the hidden nodes to $2p+1$, where $p$ is the number of inputs. The proposed EEMD-FNN scheme outperforms the single FFN model for original signal, showing a considerable difference. The MAPE criterion is almost halved even when the SEs are quite similar, the MSE and MAE criteria and their SEs are reduced too. This proves that decomposing the original signal enhances the forecasting results, specifically, in the case of the studied rainfall data.

**Table 1.** Mean performance measures and their standard errors

| | EEMD-FNN | | Single FNN | |
|---|---|---|---|---|
| | Mean | SE | Mean | SE |
| MAPE(%) | **14.0908** | 0.4365 | 28.1798 | 0.4378 |
| MSE | **1.5126** | 0.0637 | 5.4430 | 0.2131 |
| MAE | **0.9177** | 0.0191 | 1.7906 | 0.0314 |




(a) Components 1 to 5 resulting from EEMD decomposition.




(b) Components 6 to 9 resulting from EEMD decomposition and the residue.

**Fig. 2.** Components resulting from the EEMD decomposition

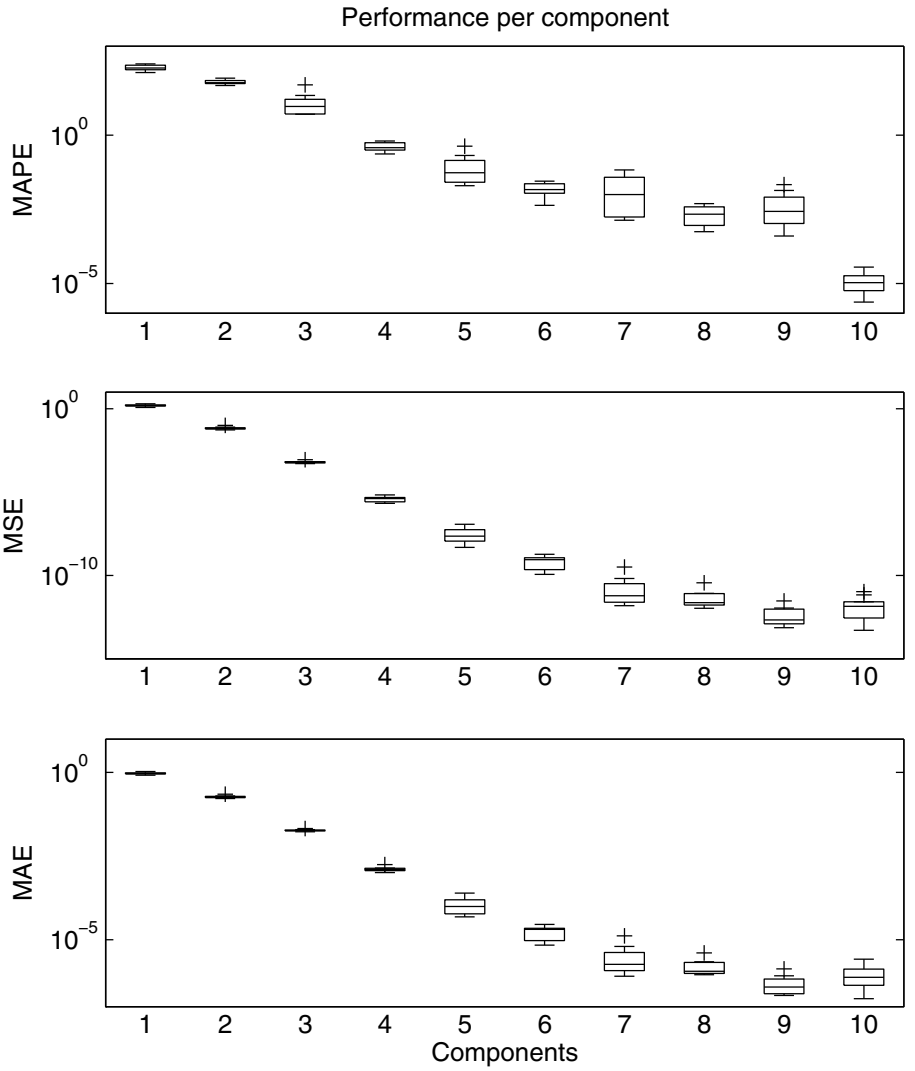



**Fig. 3.** Box plot of performance measures for the test data set per each FNN model or component predicted, the last one (10) is the residue. The vertical axis are in *log* scale.

## 4 Conclusions

In this paper a rainfall forecasting scheme has been proposed, which integrated EEMD and FNN. The proposed methodology decomposes the rainfall signal into more stationary and regular components (IMFs) using the EEMD technique. Moreover, the single model applied to each component is simple. According to the obtained results, the EEMD-FNN scheme improves the forecasting results and offers a simple approach for the stable prediction of non-stationary data.

The results were compared with a single FNN model for the original signal, using MAPE, MSE and MAE as their criteria. As a future work, it would be interesting to explore the possibility of employing different aggregation methods as well as performing an additional and more significant test that exposes more reliable results, maybe considering other data sets. In addition, it would be useful to perform a comparison with traditional forecasting techniques and testing the scheme with different long term predictions.

**Acknowledgments.** This study is supported by the "Programa Jóvenes Investigadores e Innovadores 2011, convenio especial de cooperación No. 0043 de 2012 suscrito entre la Fiduciaria Bogotá S.A. como vocera del patrimonio autónomo denominado Fondo Nacional de Financiamiento para la Ciencia, la Tecnología y la Innovación Francisco José de Caldas y la Universidad Nacional de Colombia", the research program "Fortalecimiento de capacidades conjuntas para el procesamiento y análisis de información ambiental" (code Hermes-12677) and "Grupo de Control y Procesamiento Digital de Señales Código 20501007205" funded by Universidad Nacional de Colombia. Instituto de Estudios Ambientales (IDEA) from Universidad Nacional de Colombia - Sede Manizales is also acknowledged for kindly supplied the data used in the experiments.

## References

1. Abhishek, K., Kumar, A., Ranjan, R., Kumar, S.: A rainfall prediction model using artificial neural network. In: 2012 IEEE Control and System Graduate Research Colloquium (ICSGRC), pp. 82–87 (July 2012)
2. Chen, C.F., Lai, M.C., Yeh, C.C.: Forecasting tourism demand based on empirical mode decomposition and neural network. Knowledge-Based Systems 26, 281–287 (2012)
3. French, M.N., Krajewski, W.F., Cuykendall, R.R.: Rainfall forecasting in space and time using a neural network. Journal of Hydrology 137(1), 1–31 (1992)
4. Guo, Z., Zhao, W., Lu, H., Wang, J.: Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model. Renewable Energy 37(1), 241–249 (2012)
5. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2(5), 359–366 (1989)
6. Hsu, K.L., Gupta, H.V., Sorooshian, S.: Artificial neural network modeling of the rainfall-runoff process. Water Resources Research 31(10), 2517–2530 (1995)
7. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society A Mathematical Physical and Engineering Sciences 454(1971), 903–995 (1998)
8. Huang, N., Shen, Z., Long, S.: A new view of nonlinear water waves: The Hilbert spectrum. Annual Review of Fluid Mechanics 31(1), 417–457 (1999)
9. Hung, N.Q., Babel, M.S., Weesakul, S., Tripathi, N.K.: An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrology and Earth System Sciences 13(8), 1413–1425 (2008)

10. Li, X.: Temporal structure of neuronal population oscillations with empirical model decomposition. Physics Letters A 356(3), 237–241 (2006)
11. Luk, K., Ball, J., Sharma, A.: A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting. Journal of Hydrology 227(1), 56–65 (2000)
12. Tang, L., Yu, L., Wang, S., Li, J., Wang, S.: A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting. Applied Energy 93, 432–443 (2012)
13. Vovoras, D., Tsokos, C.P.: Statistical analysis and modeling of precipitation data. Nonlinear Analysis: Theory, Methods & Applications 71(12), e1169–e1177 (2009)
14. White, H.: Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. Neural Networks 3(5), 535–549 (1990)
15. Wu, Z., Huang, N.E.: Ensemble empirical mode decomposition: A noise-assisted data analysis method. Advances in Adaptive Data Analysis 1(1), 1–41 (2009)
16. Yu, L., Wang, S., Lai, K.K.: Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics 30(5), 2623–2635 (2008)