J3.1 A NEW SIGNIFICANCE TEST FOR EMPIRICAL ORTHOGONAL FUNCTIONS

M. S. Santhanam[1], B. Aditya[2], G. Anil Kumar[3]
IBM-Research, India Research Laboratory, Block-1, IIT-Campus, New Delhi 110 016, India.
[2]Dept. of Computer Science and Engineering., IIT, Mumbai, India
[3]Dept. of Electrical Engineering, IIT-Madras, Chennai, India.

1. Introduction

The aim of Empirical Orthogonal Function (EOF) analysis [1] is two fold; (1) to determine the principal modes of variability of the studied data and (2) to obtain a low-dimensional representation for the data. In analysis of this type, the eigenvalues fall exponentially as a function of mode number and most of the variability is captured only by a few modes (as compared to the dimensionality of the space considered). Thus, by projecting the data only along these dominant modes, we can reduce the dimensionality of the considered space. In atmospheric applications, once the modes are found, it is important to distinguish the modes that represent physically relevant patterns from those that represent only noisy variations. This paper aims to provide a statistical basis for the choice of only a certain set of modes over others using techniques borrowed from random matrix theory that has its origins in quantum physics of complex systems [2]. We study four cases of atmospheric correlations in the EOF framework and compare our results with the standard technique for EOF selection, namely, rule-N [1].

2. Correlations and EOF

Consider any atmospheric anomoly $z'(x,t)$ that varies with both space(x) and time(t). If there are p space points and n time points, $\mathbf{Z}$ is the matrix of anomaly of order $p \times n$. The data is also rescaled to zero mean and its variance $< z'(x)^2 >$ is unity. Then, the spatial correlation matrix is given by,

$$\mathbf{S} = \frac{1}{n}\mathbf{Z}\mathbf{Z}^T \tag{1}$$

The eigenfunctions of the Hermitian matrix $\mathbf{S}$ form an orthonormal basis to represent $\mathbf{Z}$ and are hence called empirical orthogonal functions. The eigenvalue, λ_m is a measure of the percentage variability represented by mth EOF.

3. The Selection Rule

The study of random matrices arose from the need to understand the spectra of high-dimensional quantum system with complex interactions [2]. Recently it has been shown that the atmospheric correlation matrices can be modelled as random matrices from an appropriate ensemble [3]. The starting point is the result in random vector theory [2]; assuming that norm is the only invariant under an orthogonal transformation it can be shown that the eigenvector components $r_i, (i = 1, 2....N)$ of a random matrix of order N are Gaussian distributed,

$$P(r) = \frac{1}{\sqrt{2\pi}} \exp(-r^2/2) \tag{2}$$

Thus, we expect the components of a purely noisy EOFs to be Gaussian distributed. The EOFs which represent physically significant modes will deviate from this distribution. Hence by comparing the cumulative distribution function(CDF) of an EOF and the Gaussian CDF, one can distinguish between significant EOFs and those that could be regarded as purely noise.

We perform the Kolmogrov-Smirnov test [4] to compare the CDFs of the EOF with the Gaussian CDF. All those EOFs with KS-Probability less than 0.05 are retained as being significant while the rest are eliminated. It is important to note that this procedure is not a dominant variance rule like rule-N and picks out EOFs based on their distributions. It takes more information into account as compared to Rule N as every component of the eigenfunction is considered instead of only the eigenvalues. Thus, it can be thought of as a detailed probe of the signal carried by the EOF.

4. Data sets

We apply the test to the following cases ;

1. pseudo wind stress over the tropical pacific region $(20^oS - 20^oN, 130^oE - 70^oW)$
2. Sea level pressure(SLP) over the North Atlantic region $(20^oN - 90^oN, 120^oW - 30^oE)$
3. Geopotential height at pressure levels ranging from 1000mb to 100mb over the North Atlantic region $(20^oN - 90^oN,\ 120^oW - 30^oE)$

[1]Corresponding author; email : msanthan@ in.ibm.com

All these data sets, spanning 52 years from 1948 to 2000, were obtained as monthly and daily means available from the NOAA-CIRES/NCEP reanalysis archives. EOF analysis for data sets 1 and 2 was performed on monthly means [3] and for data 3 on weekly averaged data. The North Atlantic Oscillation (NAO) is a large scale seesaw in atmospheric mass between the subtropical high and the polar low and has been extensively studied in the literature [5]. For capturing this oscillation, geopotential height data ranging from 1000mb to 100mb was used and composite 3-dimensional (winter time) EOFs as shown in fig 1 were obtained. This 2nd dominant EOF, plotted as an isosurface, captures the opposing behaviour in the sub-tropical and polar regions. The second principal component was also calculated (not shown here) and it closely follows the documented NAO index. EOF analysis of SLP and wind-stress were done and will be published elsewhere. See ref [3] for some of the results.

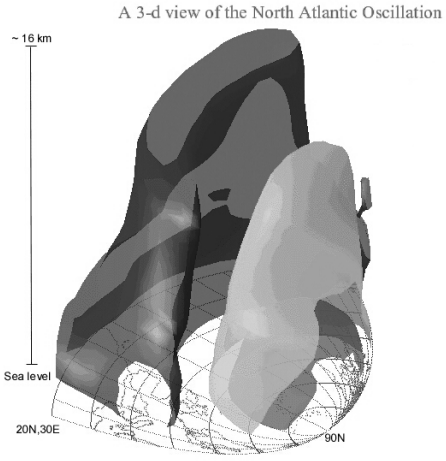

FIG 1: Second dominant EOF from the analysis of geopotential height in 3 dimensions.

5. Results

The cumulative distribution of EOF components for the geopotential height correlation matrix is shown in fig 2. The solid curve is the CDF of the standard Gaussian distribution given in eq (2). Note that the dot-dashed and dashed lines in fig(2) correspond to first two dominant EOFs that deviate significantly from the Gaussian. The long-dashes curve corresponds to 100th EOF and follows the Gaussian closely. The deviating EOFs correspond to physically significant patterns like the one shown in fig(1). From this, it can be seen that the random EOFs follow the Gaussian CDF and the principal EOFs show significant deviation from Gaussian.

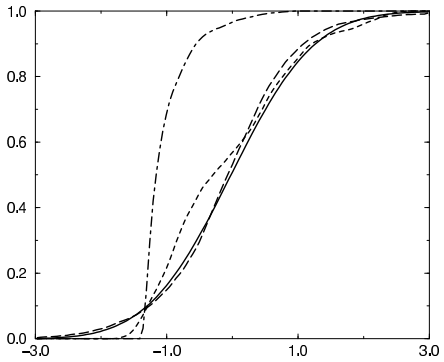

FIG 2 :The cumulative distribution of EOF components for geopotential height EOF analysis.

The results were compared with Rule-N [1]. In general, Rule N provides a conservative estimate of the number of significant EOFs while our analysis provides a more realistic estimate (see Table below). In most cases, the EOFs that pass the test correspond closely with patterns that have physical significance.

Data Set	**G.Hgt**	**SLP**	**W.Stress**
Rule N	30	16	15
RMT based rule	6	6	3

Thus, based on the Gaussian nature of distribution of EOFs, significant EOFs and distinguished from the random ones.

Acknowledgements: The NCEP reanalysis data provided by the NOAA-CIRES Climate Diagnostics Center, from their website at http://www.cdc.noaa.gov is thankfully acknowledged.

References :

[1] Preisendorfer and Mobley, *Principal Component Analysis in Meteorology and Oceanography*, (Elsevier, 1988).

[2] T. Guhr et. al., Phys. Rep., **299** 189 (1999).

[3] M. S. Santhanam and Prabir K. Patra, Phys. Rev. E, **64** (016102), 2001.

[4] Daniel J. Wilks, *Statistical Methods in Atmospheric Sciences* (Academic, 1995).

[5] P.J. Lamb and R.A. Peppler, Bull. Am. Met. Soc. **68** 1218 (1987).