Point–process statistical analysis for the ECMWF Ensemble Prediction System

Fabrizio Nerozzi
ARPA Emilia–Romagna, Servizio IdroMeteoClima, fnerozzi@arpa.emr.it

Abstract: The possibility of applying mathematical tools of point–process statistics to the ECMWF Ensemble Prediction System (EPS) is exploited in this work, in order to provide a different way to reduce ensemble information. The first two empirical orthogonal functions enable to represent 5–day ensemble forecasts as point processes in a plane. These planar representations are hence compared to a sample of Gaussian random point patterns, obtained by a Montecarlo method. The estimations of the nearest–neighbour distribution function and of the reduced second order momentum function for point processes relative to the ensemble predictions are in good agreement with the corresponding estimations of Gaussian random point processes.

Keywords: Ensemble predictions, Principal Component Analysis, Point–process statistics

1 Introduction

Ensemble predictions appear to be the only feasible method to predict the evolution of the atmospheric probability distribution function beyond the range in which error growth can be prescribed by linearized dynamics (Molteni *et al.*, 1996).

However, the large amount of information contained in the ECMWF Ensemble Prediction System (EPS) can be hardly managed in the whole, and two different strategies, clustering and tubing, are adopted for reducing the 51 EPS members to few alternative scenarios. For clustering, “similar” EPS forecasts are collected in clusters, whose probabilities of occurrence are provided by the cluster sizes (Molteni *et al.*, 2001). As concerns the tubing technique, this consists in an averaging of all ensemble members close to the ensemble mean, while the excluded members are grouped together in a number of tubes. Each tube is represented by its most extreme member belonging to it.

One of the principal shortcomings of the clustering technique is the empirical distribution of the ensemble members. Although tubing allows a better visualization of the most different scenarios in the ensemble than clustering, tubes do not provide probabilities of occurrence. In order to provide a different way to condense information, which tries to overcome these shortcomings of clustering and tubing techniques, it is here exploited the possibility to represent ensemble forecasts as a finite set of random points distributed in a plane. In particular, it is tested the hypothesis according to which these ensemble point processes can be treated statistically equivalent to Gaussian random point processes.

2 Materials and Methods

2.1 The Principal Component Analysis

The Principal Component Analysis (PCA) enables to transform a data set, characterized by a large number of variables in a new one, where the number of variables is highly reduced. The new variables are calculated as the eigenvectors of the covariance matrix and they are orthogonal among them (Preisendorfer, 1988).

For each day of the whole meteorological winter season 2006–2007, starting from the 1st December 2006 to the 28th February 2007, it has been computed the covariance matrix of the 51 ECMWF EPS 5–day forecasts at 12 UTC of the 500 hPa geopotential height. The geopotential height is a meteorological field, here defined over a regular grid at 1 degree of resolution and covering the European area (33N–74N; 27W–45E). Then, for each one of the 51 ensemble members the PCA technique has been applied, and the first two normalized principal components have been considered. The explained variance by these first two principal components ranges from 46% to 69% of the total variance.

Eventually, in order to represent the ensemble forecasts as a a random point–process (hereafter called EPS point–process), for each one of the 90 winter days the 51 ensemble members are represented as single points lying over the plane formed by the first two PCA eigenvectors, whose coordinates are provided by the first two principal components of the ECMWF EPS members.

2.2 The point–process statistics

The Gaussian random point–process is here taken as reference model. In particular, for each winter day 199 bidimensional Gaussian random point–processes with 51 points, zero mean and variance equal to 1, have been simulated by a Montecarlo method. Hence, for the corresponding EPS point–process and for these 199 simulations the nearest–neighbour distribution function $D(r)$ has been defined computing the distance from the analysis point (the "observed" 500 hPa geopotential height reduced to the first two principal components), chosen as the arbitrary event, to its nearest event belonging to each one of the 200 random patterns. Analogously, the derivative L of the reduced second order momentum, K function, is computed counting for each point pattern the number of events within a distance r from the analysis point.

The nearest–neighbour distribution function $D(r)$ describes the probability that distance from a randomly chosen event to its nearest event is less than or equal to $r > 0$. This function can be heuristically estimated from the observed pattern:

$$\hat{D}(r) = \frac{\sum_{i=1}^{n} I(r_{i,A} \leq r, d_i > r)}{\sum_{i=1}^{n} I(d_i > r)} \tag{1}$$

where d_i denotes the distance of the event from the nearest boundary of the closed set A and $r_{i,A}$ is the distance from the nearest event in A (Cressie, 1991).

The K function uses information in the pattern over a wide range of scales than the nearest–neighbour distribution function. Its definition is related to the number of extra events within distance r from an arbitrary event. Estimating of K from an observed pattern in a bounded $A \subset \Re^2$ is complicated by edge effects. Here the Ripley's edge–corrected estimator is considered (Cressie, 1991):

$$\hat{K}(r) = \frac{1}{n\lambda} \sum_{i=1}^{n} \sum_{j=1'}^{n} w(\mathbf{s}_i, \mathbf{s}_j)^{-1} I(\|\mathbf{s}_i - \mathbf{s}_j\| \leq r) \tag{2}$$

The estimator $\hat{K}$ is approximately unbiased provided that the n events are approximately independent. Estimates of the derivative of the K function, $\hat{L}(r)$, are computed by the formula (Stoyan *et al.*, 1987):

$$\hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}} \tag{3}$$

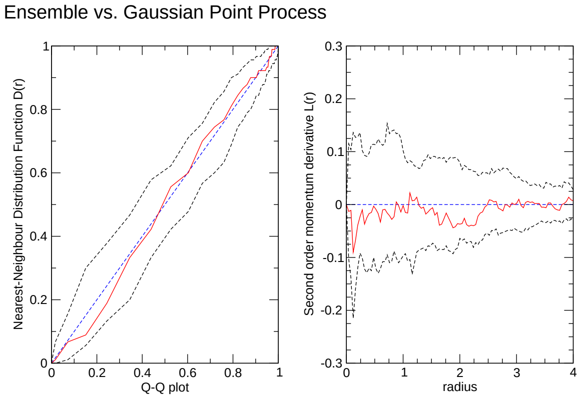

Figure 1: EPS against Gaussian Point–Processes: nearest–neighbour distribution function Q–Q plot (left panel), and the derivative of the K function (right panel).

3 Results and Concluding remarks

The possibility of applying mathematical tools of point–process statistics to the ECMWF Ensemble Prediction System has been exploited in this work, in order to look for a different way to condense the large amount of ensemble information. It has been proved the first two empirical orthogonal functions enable to represent about, or more, 50% of the ensemble spread. Therefore the ensemble forecasts have been represented as point processes in the plane of the first two PCA eigenvectors and compared to Gaussian random point processes, obtained by a Montecarlo method and considered as reference models.

In figure 1, it is reported on the left side the Q-Q plot, where the quantiles of the median of the 199 nearest–neighbour functions relative to 90 Gaussian random point processes are in the abscise axis. In the ordinate axis there are the quantiles of the nearest–neighbour function relative to the 90 ensemble point processes (continuous red line), the median (dashed blue line), the minimum and maximum (dashed black lines) of the 199 nearest–neighbour functions. On the right side, it is instead reported the derivative of the K function relative to the 90 ensemble point processes minus the median of the 199 simulations (continuous red lines). Analogously, the confidence interval is represented by the minimum and maximum of the 199 simulations, again subtracted by the median (dashed black lines).

The good agreement between ensemble and Gaussian random point processes, in terms of the nearest–neighbour distribution function and of the reduced second order momentum function estimations, coming out of the present work, could render plausible to consider the probability distribution function of ensemble members as asymptotically normal.

References

Cressie N. (1991) *Statistics for spatial data*, J. Wiley & Sons, New York.

Molteni F, Buizza R., Palmer T. N., Petroliagis T. (1996) The ECMWF Ensemble Prediction System: Methodology and validation. *Q. J. R. Meteorol. Soc. 122, 73–119.*

Molteni F., Buizza R. (1999) Validation of ECMWF Ensemble Prediction System using empirical orthogonal function. *Mon. Wea. Rev., 127, 2346–2358.*

Molteni F., Buizza R., Marsigli C., Montani A., Nerozzi F. and Paccagnella T. (2001) A strategy for high–resolution ensemble prediction. Part I: Definition of Representative Members and Global Model Experiments. *Q. J. R. Meteorol. Soc., 127, 2069–2094.*

Preisendorfer R. W. (1988) *Principal component analysis in meteorology and oceanograpy*, Curtis D. Mobley, New York.

Stoyan D., Kendall W. S., Mecke J. (1987) *Stochastic geometry and its applications*, J. Wiley & Sons, New York.