A selection rule for parameter reduction using fuzzy membership roster method

Sreyasi Ghosh, Sarbari Ghosh and Pradip KR Sen

Abstract
In the present study, an attempt is made to propose a new operational technique for weather forecasting at Kolkata (22.53° N, 88.33° E), India, during the pre-monsoon season (March, April and May). The newly suggested technique is based on fuzzy membership roster method. It can handle inherent non-linearity in a physical phenomenon. It is interesting to note that for the prediction of weather of next 12 hours based on Radio/Rawin Sonde observation at 1200 UTC of a day, the fuzzy membership roster method is better than any previous technique. Although the previous methods are however almost equally suitable to predict the weather of the next 12 hours based on Radio/Rawin Sonde observation at 0000 UTC. Our main objective is to reduce the number of parameters without losing any primary information for predicting the future situation. It is interesting to note that the data has been reduced from 20 in number to 12 parameters in different situations to furnish more than 70% correct result. The degrees of compatibility are defined using a training data set for the period 1985-1996 and validated for the period 1997-1999.

Keywords: Convective development, forward selection rule, fuzzy membership roster method, instability

1. Introduction
Prediction of any atmospheric phenomenon is always of ultimate interest to the weather forecasters as well as researchers and others. Specially, in recent years, there has been growing interest in the prediction of pre-monsoon convective developments (CD), not only for the possible hazards caused by them, but also for their beneficial character like cooling due to rain during the hot summer days.

The convective developments occurring during March, April and May in Kolkata, India are usually termed as pre-monsoon thunderstorms. The principal component analysis (PCA) technique was applied by previous workers to identify the significant parameters for the occurrence of pre-monsoon thunderstorms (TS) in Kolkata. They have also shown how the linear discriminant analysis (LDA) technique alone as well as in conjunction with PCA can be successfully applied to the set of 20 parameters to predict the pre-monsoon thunderstorms for Kolkata (Ghosh et al., 1999, 2004; Chatterjee et al., 2009) [14, 15, 7]. Cluster analysis and LDA technique (Maryon and Storey, 1985) [24] were utilized to describe a multivariate statistical model for forecasting anomalies of surface pressure present over Europe and North Atlantic. In another study, multiple linear regression (Ward and Folland, 1991) [31] was compared with LDA for making hindcasts and real time forecasts of north-east Brazil wet season rainfall using sea surface temperature. Though a number of attempts (Showalter, 1953; Darkow, 1968) [30, 9] were made to establish empirical models for the prediction of atmospheric stability/instability, the work done on Kano (Oduro-Afriyie and Adefolalu, 1993) [29] is perhaps the first successful attempt for tropical region. Another attempt was made to predict the occurrence of CD at Dhaka (Bangladesh) in terms of stability indices (Chowdhury et al., 1996) [8]. In 1995 Murtha applied the fuzzy logic in operational meteorology. Yu and Tao (2000) [35] developed a fuzzy multi-objective function for rainfall-runoff model calibration in 2000. In 2002, Gomes and Casanovas reported a case study of solar irradiance which involved fuzzy logic and meteorological variables. In 2003, Mackay et al. used fuzzy logic in automated parameterization of land surface process models. Chang et al. (2005) [6] applied fuzzy theory in genetic algorithm to interpolate precipitation. Mitra et al. (2008) [26] used rule-based fuzzy inference system for weather forecasting.
Also Ma et al. (2009) [22] have applied the same technique for the verification of meso-scale NWP forecasts. Hubbert et al. (2009) [17] have developed a technique for real time identification and filtering using fuzzy logic. Dhanya and Kumar (2009) [9] have used a fuzzy rule based modeling approach for the prediction of monsoon rainfall in India. The power output forecasting of a photovoltaic system based on insolation forecasting at 24 hours ahead by using fuzzy theory was developed by Atsushi et al (2013). Recently Mohammad Iqbal et al predicted weather pattern using fuzzy rough clustering (2016).

Convective developments are generally favored by convective instability, abundant moisture at lower levels, strong wind shear, and a dynamical lifting mechanism that can release the instability (Kessler, 1982) [18]. Also, the vertical shear of the environmental winds has to match the value of the convective instability for proper development of a large convective cloud (Asnani, 1992) [3]. The presence of conditional instability is an essential criterion for supporting electrification and lightning (Williams and Reno, 1993) [33].

In addition to the parameters mentioned, two other parameters, viz. (θe–θes) and (P-PLCL) are also present, where θes and θe denote the saturated equivalent potential temperature and equivalent potential temperature respectively. P is a level pressure and PLCL is the pressure at the corresponding lifting condensation level.

The thermodynamic parameter (θe–θes) was introduced by Betts (1974) [4] as a measure of the unsaturation of the atmosphere. PLCL for the surface parcel was considered as the cloud base (Kuo, 1965) [21] and hence (P-PLCL) is taken as a forcing factor for the saturation of a parcel.

Fuzzy set theory, was originally proposed by Zadeh (1965a, b) [16, 37], aimed at imitating the model of human thought process. The basic premises of bi-valued true-false Boolean notion are redefined here.

In spite of strong resistance to fuzzy logic, many researchers started working in the field during 1965-1975. During the first decade, many mathematical structures were fuzzified by generalizing the underlying sets to be fuzzy, i.e. the sets with no sharp boundaries. The 90s was an era of new computational paradigms. The applications of fuzzy set theory include studies in many fields, e.g. meteorology, biology and others (Klir and Folger, 1998) [19].

We propose a new method based on fuzzy membership roster method. The technique suggested in the study helps one to select the most effective combination of significant parameters out of 20 to discriminate the two important situation convective development and fair weather during the pre-monsoon season of Kolkata. The methodology and the corresponding results are discussed in detail in the respective sections.

2. Data

The number of CD and fair weather (FW) days linked with the morning and afternoon Radio Sonde/Rawin Sonde (RS/RW) observations at Kolkata are presented in Table I. These data have been used to derive the required thermodynamic parameters, which have been used to construct the fuzzy rule. Any convective development occurring within the next 12 hours from the morning RS/RW observation taken at 0000 UTC (0530 Indian Standard Time) is considered as CD related to morning RS/RW, otherwise it is FW related to the same RS/RW. On many occasions the data, either at one or more of the significant levels i.e. 1000, 850, 700, 600 and 500 hPa were not available. Naturally those occasions could not be taken into consideration. The fuzzy rule (morning) for forecasting the convective development at Kolkata have been constructed utilizing all the available radiosonde data of 12 years (1985-1996) and for the validation of these techniques, the radiosonde data of 3 years (1997-1999) have been used.

<table>
<thead>
<tr>
<th>Nature of days</th>
<th>Time</th>
<th>No. of variables</th>
<th>No. of days involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>MORNING</td>
<td>20</td>
<td>123</td>
</tr>
<tr>
<td>FW</td>
<td>MORNING</td>
<td>20</td>
<td>280</td>
</tr>
</tbody>
</table>

In the literature of this paper, \( O_1 = (\theta_e - \theta_s) \) at 1000 hPa level; \( O_2 = (P - PLCL) \) at 1000 hPa level; \( O_3 = \partial \theta_e/\partial z \) at 1000-850 hPa layer; \( O_4 = \theta_e/\partial z \) at 1000-850 hPa layer; \( O_5 = \partial v/\partial z \) at 850 hPa layer; \( O_6 = \theta_e/\partial z \) at 850-500 hPa layer; \( O_7 = (\theta_e - \theta_s) \) at 850 hPa level; \( O_8 = (P - PLCL) \) at 850 hPa level; \( O_9 = \partial \theta_e/\partial z \) at 850-700 hPa layer; \( O_{10} = \partial v/\partial z \) at 850-700 hPa layer; \( O_{11} = (\theta_e - \theta_s) \) at 700 hPa level; \( O_{12} = (P - PLCL) \) at 700 hPa level; \( O_{13} = \partial \theta_e/\partial z \) at 700-600 hPa layer; \( O_{14} = \partial \theta_e/\partial z \) at 700-600 hPa layer; \( O_{15} = \partial v/\partial z \) at 700-600 hPa layer; \( O_{16} = (\theta_e - \theta_s) \) at 600 hPa layer; \( O_{17} = (P - PLCL) \) at 600 hPa level; \( O_{18} = \partial \theta_e/\partial z \) at 600-500 hPa layer; \( O_{19} = \partial v/\partial z \) at 600-500 hPa layer; \( O_{20} = \partial v/\partial z \) at 600-500 hPa layer. It is worth mentioning that the values of (θe–θs) and (P–PLCL) at the lower level of each layer have been treated as their respective values for that layer. Here, \( z \) stands for vertical height, \( \partial \theta_e/\partial z \) for conditional instability, \( \partial v/\partial z \) for convective instability and \( \partial v/\partial z \) for the vertical shear of horizontal wind.

3. Methodology

The present study considers separately the following two situations:

1. Prediction of convective development from the data of 0000 UTC (Morning CD or MCD)
2. Prediction of fair-weather from the data of 0000 UTC (Morning FW or MFW)

All the above mentioned predictions have been made for the next 12 hours from the time of observations.

Forward Selection Rule has been applied for structure specification.

Step 1: It involves starting with each parameter in the model. Selecting the best out of the parameters.

Step 2: Now the best parameter from the previous step is taken and the addition of each variable using the chosen model comparison criterion. Again the best is chosen.

Step 3: Adding the parameter (if any) that improves this model the most, and repeating this process until none improves the model.
As we see at each step the best parameter is chosen by forward selection rule, the next parameter is chosen keeping the first parameter fixed which is obtained from the previous stage. The process is continued till no better result is obtained in comparison to the previous one.  

The study involves the concept of multi dimensional degree of compatibility (Klir and Yuan, 2002). The main features of the techniques have been discussed, in short, in the subsequent section. 

**Basics of fuzzy-rule based technique suggested for the present analysis**

It is well known that a membership function is such that the values assigned to the elements of the universal set fall within a specified range and indicate the membership grade of these elements in the set in question. The set defined by such membership function is called a fuzzy set. 

Let S denote the universal set of the parameters. Then the membership functions μ_X and μ_Y by which the fuzzy sets X and Y are defined have the forms: 

μ_X: S⇒[0,1], μ_Y: S⇒[0,1] 

where [0,1] denotes the interval of real numbers from 0 to 1, inclusive (Klir & Folger, 2009). In the study the possible membership function for the fuzzy sets of parameters are chosen to be Gaussian Membership Function. The justification for this choice is given later. 

Here we consider the same two groups X and Y which are taken as two standard pattern classes. X and Y are termed as fuzzy sets since it is difficult to identify sharp boundaries between these two sets so far the parameters, viz., convective instability, conditional instability and vertical shear are concerned. Then the degrees of compatibility of a parameter, O_i (i=1 to 20) with the standard pattern classes, X and Y are computed as follows: 

\[
AY(O) = \prod_{i=1}^{20} AY(O_i) = \prod_{i=1}^{20} \exp \left[ -\frac{(O_i - m_{CD})^2}{\sigma^2_{CD}} \right] 
\]

\[
AX(O) = \prod_{i=1}^{20} AX(O_i) = \prod_{i=1}^{20} \exp \left[ -\frac{(O_i - m_{FW})^2}{\sigma^2_{FW}} \right] 
\]

where m and σ denote the center and width of the values of O_i respectively. Since, the numerical values of the selective parameters are not scattered, the respective means of the thermodynamic and dynamic parameters represent the two patterns in a reliable way.  

Finally, the degrees of compatibility of a day (i.e. a relevant pattern) defined by \( O = (O_1, O_2, O_3, \ldots, O_{20}) \) with the two standard pattern classes, Y and X are constructed as follows: 

\[
AY(O) = \prod_{i=1}^{20} AY(O_i) 
\]

\[
AX(O) = \prod_{i=1}^{20} AX(O_i) 
\]

In order to measure the degree of fuzziness of the pattern classes X and Y, the membership based validity measures, named as the partition coefficients (Yen & Langari, 2005) have been computed separately for the following four conditions under consideration. 

\[
V_{PCA} = \frac{1}{84} \sum [AX(U) + AX(U)] 
\]

\[
V_{PAC} = \frac{1}{44} \sum [AX(U) + AX(U)] 
\]

where Uj denotes an unknown pattern or day, which belongs to the dataset used for validation. 

The study includes the following two stages: 

Stage 1. The result obtained in stage 1 are not satisfactory to discriminate the situation so in this stage those combinations of the parameters are taken into consideration that produce almost 75% correct results in the two situations. The study is performed for the two situations. In this stage the combinations of common parameters are selected to identify an unknown day as CD or FW for morning respectively. (Table VI) It is worth mentioning that the second stage helps to reduce the number of parameters effectively. 

Now, an unknown pattern or a day, U is classified by the larger value of AY(U) or AX(U), i.e. if AY(U) > AX(U), then there is a possibility for U to be more of the pattern Y than of the pattern X for next 12 hours. So it may be
predicted that U is possibly a day with convective development for next 12 hours (Klir and Yuan, 2002) [20]. The programs for selecting the best parameters and finding the best possible combination using forward selection rule are developed by authors themselves.

3.2 Justification for choice of membership function
It is worth mentioning that there is no sound principle yet for guiding the choice of membership function or degree of compatibility. But, Gaussian membership function has been selected here because of the following reasons:
1. Since some of the parameters are found to follow Gaussian distribution and usually the physical parameters are assumed to be Gaussian or quasi Gaussian in nature, for each parameter, the Gaussian membership function has been chosen to construct the one dimensional or univariate degree of compatibility.
2. They are local although not strictly compact
3. The output is very smooth and not probabilistic
4. Gaussian membership functions are continuously differentiable as well as parameterizable.
5. Gaussian membership functions are factorizable. Hence, we may synthesize a multi dimensional or multivariate degree of compatibility as the product of one dimensional or univariate degree of compatibility. That is why the product forms have been used in the relations (6) and (7) to handle the nonlinearity (Yen and Langari, 2005) [34]. This product form of membership functions is used already by Dhanya and Kumar (2009) [9].

The nature of the membership function for an individual parameter is shown in the following graph as an example:

\[ \text{Graph plotting membership values v/s } \theta_e - \theta_c \]

3. Results and discussion
In the fuzzy membership roster method, the number of days involved are 280 for MNTS, 123 for MTS; those are used for constructing the means and standard deviations, using the relations (4) & (5). The 20 thermodynamic and dynamic parameters computed from the RS/RW data of 0000 UTC and 1200 UTC as already mentioned earlier are used here.

The degrees of compatibility computed using relations (6) and (7) are applied to classify a pre-monsoon day of unknown category (U) of the year 1997, 1998, 1999 and the results are presented in Table I.

Table I

<table>
<thead>
<tr>
<th>No. of Combinations</th>
<th>Combinations</th>
<th>MTS</th>
<th>No. of correct prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td></td>
<td>38</td>
</tr>
<tr>
<td>2</td>
<td>13,19</td>
<td></td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>13,19,10</td>
<td></td>
<td>44</td>
</tr>
<tr>
<td>4</td>
<td>13,19,10,20</td>
<td></td>
<td>44</td>
</tr>
<tr>
<td>5</td>
<td>13,19,10,20,8</td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>6</td>
<td>13,19,10,20,8,4</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>13,19,10,20,8,4,15</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>13,19,10,20,8,4,15,18</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>13,19,10,20,8,4,15,18,14</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>13,19,10,20,8,4,15,18,14,5</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>13,19,10,20,8,4,15,18,14,5,9</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>13,19,10,20,8,4,15,18,14,5,9,2,16</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13,19,10,20,8,4,15,18,14,5,9,2,16,1</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>13,19,10,20,8,4,15,18,14,5,9,2,16,1,3</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>13,19,10,20,8,4,15,18,14,5,9,2,16,1,3,11</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>13,19,10,20,8,4,15,18,14,5,9,2,16,1,3,11,19</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>
From the above table III it is noticed that a common set of parameters cannot be detected for discriminating the two situations in the morning with desired degree of accuracy. The result with all 20 parameters is not satisfactory in the 2 situations (Morning thunderstorm, Morning Non-thunderstorm). So we are confined to the combination of those parameters which produce almost 75% correct results to discriminate an unknown situation. Thus we the parameters O1 = (∇e-θe) at 1000 hPa level; O2 = (P-PLCL) at 1000 hPa level; O4 = ∂θe/∂z at 1000-850 hPa layer; O5 = ∂v/∂z at 1000 –850 hPa layer; O8=∂θes/∂z at 850-700 hPa layer; O10=∂v/∂z at 850-700 hPa layer; O13=∂θes/∂z at 700-600 hPa layer ; O14=∂θe/∂z at 700-600 hPa layer; O15=∂v/∂z at 700 –600 hPa layer; O18=∂θes/∂z at 600-500 hPa layer; O19=∂θe/∂z at 600-500 hPa layer; O20=∂v/∂z at 600 –500 hPa layer for the morning is selected. 

Now the program is repeated for just these many number of data.

The results of the categorical discrimination of an unknown day (U) belonging to the dataset used for validation, have been presented in Table IV using this technique, where the dataset consists of the pre-monsoon days of 1997, 1998 and 1999. 

The membership validity measures for MFW, MCD are computed with the help of relation (8) and are presented in Table V. 

Here, the study is performed for morning, since a previous study reveals that during the pre-monsoon season in Kolkata, the weather of the morning differs significantly from that of the afternoon (Ghosh et al., 1999) 14. The combinations of different parameter are selected, since it is well known that any atmospheric phenomenon is essentially complex and multivariate in nature.

One should also note that the fuzzy membership method has ample scope for further development in the field of pre-monsoon weather forecasting. Regarding the membership validity measures (Table III), which have been computed using (8) it can be stated that the classes X and Y, are not hard, as VPCMFW ≠1, VPCMC ≠1. Hence the fuzzy rule based technique as suggested here, has a possibility for improvement (Yen and Langari, 2005) [34].

Table 3: Summary of the results in Tables 1, 2

<table>
<thead>
<tr>
<th>Nature of Day</th>
<th>No. of Combination</th>
<th>Total Days</th>
<th>No. of Correct Prediction</th>
<th>% of Correct Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTS</td>
<td>20</td>
<td>44</td>
<td>21</td>
<td>47.73</td>
</tr>
<tr>
<td>MNTS</td>
<td>20</td>
<td>84</td>
<td>72</td>
<td>85.71</td>
</tr>
</tbody>
</table>

5. Conclusions

As discussed above the program has been developed by us and no software packages have been used. The observations made in the study indicate that the computationally simpler technique discussed in the literature may help one to reduce the number of parameters to analyse a situation almost successfully. But the limitations of our work lie in the fact that the entire technique is context bound. Also the data we have used here is secondary data and hence we have assumed it to be reliable. 

Future scope for our work lies in the fact that the work has been validated with atmospheric data but the technique is a general one hence can be used in other fields for structure specification.

~ 815 ~
6. Acknowledgements
The authors are grateful to the India Meteorological Department for supplying the necessary data for the present study. Also we would like to express our heartfelt gratitude to Prof. U.K. De, Atmospheric Science and Research Group, School of Environmental Studies Jadavpur University, S Chatterjee, Assistant Professor, Academy of Technology and M. Ghosh, Assistant Professor, RCC Institute of IT for helping us all through out.

7. References