MAUSAM, **67**, 3 (July 2016), 669-676

551.576.1 (540.27)

Clustering technique for interpretation of cloudburst over Uttarakhand

KAVITA PABREJA and RATTAN K. DATTA*

Maharaja Surajmal Institute, GGSIP University, New Delhi, India

** Mohyal Educational Research Institute of Technology, New Delhi, India*

(*Received 18 May 2015, Accepted 20 August 2015*)

e mail : kavita_pabreja@rediffmail.com

सार – पिछले कुछ वर्षों से कई कार्यों और वैज्ञानिक अनुप्रयोगों के लिए आंकड़ा संग्रहण का व्यापक रूप से प्रयोग किया गया है। आंकड़ा संग्रहण बहुत बड़े आंकड़ा बेसों में छिपे हुए तथ्यों को समझने की गहन जानकारी उपलब्ध कराता है। आंकड़ा संग्रहण कम्प्यूटर विज्ञान का अंत: अनुशासनिक उपयोग है जो कृत्रिम सूचना, मशीनी जानकारी, आंकड़े और डेटा-बेस प्रणालियों की अंतरधाराओं में पद्धतियों के उपयोग द्वारा वृहत आंकड़ा सेटों में पैटर्नों का पता लगता है। इस शोध पत्र में बादल फटने जैसी एक बहुत ही संकटपूर्ण मौसम परिघटना के लिए मौसम पूर्वानुमानों को बताने के लिए आंकड़ा संग्रहण तकनीक प्रयुक्त की गई है। प्रति वर्ष पहाड़ी और तटीय क्षेत्रों में बादल फटने से जान और माल की हानि होती है। इन घटनाओं का पूर्वानुमान लगाना और इनकी चेतावनी देना बहुत कठिन है। बादल फटने की घटना का पूर्वानुमान लगाने के लिए कोई भी तकनीक अपने छोटे पैमाने के कारण संतोषजनक नही पाई गई है। बादल फटने की संभावना का पता लगाने के लिए रेडार के एक अत्यंत सूक्ष्म संजाल की आवश्यकता है जो निषेधात्मक रूप से महंगे हो सकते है। मॉडल आउटपुट सांख्यिकीय (एम ओ एस) मॉडलों के उपयोग द्वारा अथवा नवीनतम उपग्रह चित्र आंकड़ा, शक्तिशाली रेड़ार (डॉप्लर वर्ग) यदि उपलब्ध हो तो उनकी व्याख्या के आधार पर कुछ घंटे पहले थोड़े से लीड समय में बादल फटने की चेतावनी दी जा सकती है। इस मौसम परिघटना के अन्य आयाम का मौसम पूर्वानुमान भूमंडलीय और क्षेत्रीय मॉडलों द्वारा पूर्वानुमानित आरंभिक आंकड़ों पर सामूहिक तकनीक का प्रयोग कर पता लगाया जा सकता है। उत्तराखंड में हाल ही में बादल फटने की घटना से हुई बहुत भारी हानि का आंकड़ा संग्रहण की K-मीन्स सामूहिक तकनीक का उपयोग कर विश्लेषण किया गया। इसमें पाया गया है कि सांख्यिकीय मौसम पूर्वानुमान मॉडल, पूर्वानुमान आंकडें के संग्रहण में बादल फटने की घटना के संकेत 3-4 दिन पहले ही देखे जा सकते है।

ABSTRACT. Data Mining has been used extensively in various business and scientific applications for last few years. Data mining has been found to be providing a deep insight into understanding the hidden facts in huge databases. Data mining is an interdisciplinary subfield of computer science that discovers patterns in large data sets by using methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. In this paper, data mining technique for Interpretation of Weather Forecasts for one of the most disastrous weather phenomenon *viz.* cloudburst has been applied. Every year, cloudburst over hilly areas and coastal regions causes loss of lives and property. The forecasting and warning of these events is very difficult. There is no satisfactory technique for anticipating the occurrence of cloudbursts because of their small scale. A very fine network of radars is required to be able to detect the likelihood of a cloudburst and this would be prohibitively expensive. The warning of cloudburst could only be provided at a small lead time say a few hours in advance based on the interpretation of latest satellite imagery data, powerful radar (Doppler category), if available, or by using Model Output Statistics (MOS) models. Another dimension to forecasting this weather event has been identified by applying clustering technique on primary data forecasted by global and regional models of weather forecasting. A recent case of Cloudburst over Uttarakhand that caused a huge loss has been analyzed using k-means clustering technique of data mining. It has been observed that with the mining of Numerical Weather Prediction model forecast data, the signals of formation of cloudburst can be found3-4 days in advance.

Key words – Data mining, Cloudburst, k-means clustering, Numerical weather prediction, Sub-grid scale weather systems.

1. Introduction

At present, 'Numerical Weather Prediction' (NWP) models which solve a close set of equations representing atmospheric flow, have been adopted by most of the meteorological services to issue day to day weather forecasts. These forecasts are issued for public in general, farmers, pilots, water works managers, health

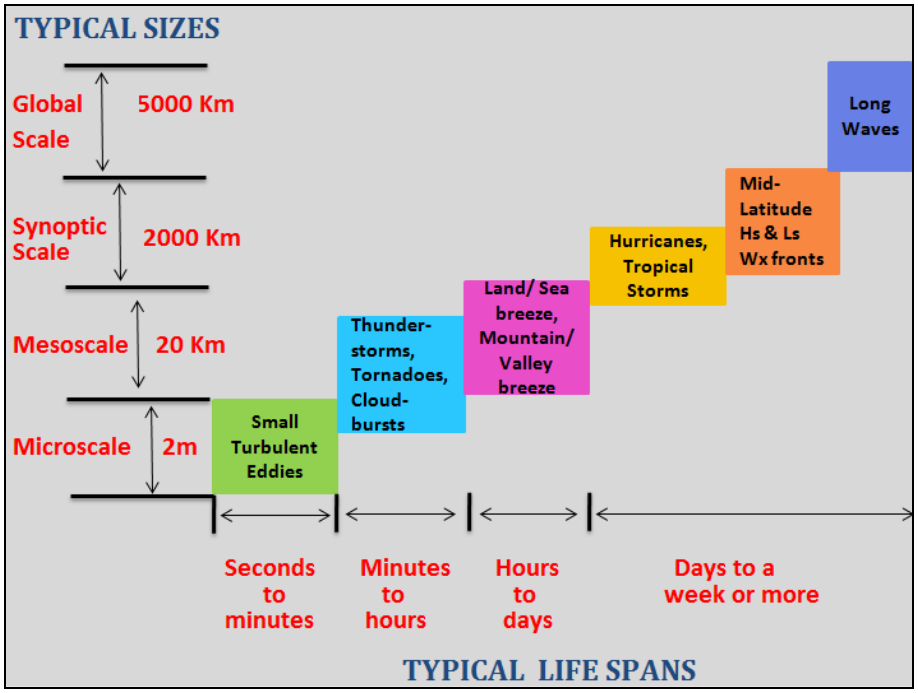


Fig. 1. Time and Space Scale of Atmospheric Motion (Short, 2005)

departments, planners, disaster management services etc. NWP models continue to improve in resolution (both horizontal and vertical) as well as sophistication in including various atmospheric processes. Despite this, there are various limitations, specifically:

(*i*) There are many important processes and scales of motion in the atmosphere especially Sub-Grid scale weather phenomenon that cannot be explicitly resolved with present models. A meteorological classification of weather systems at time and space scale as given by Short, 2005 is shown in Fig. 1.

Different weather systems exist at different time and space scales in the atmosphere, *viz*., Hurricanes, Tropical Cyclones exist at synoptic scale (large area of order of 100s to 1000s of km and time from days to weeks), land/sea breeze exist at meso-scale having time scales from hours to days and space scales of 10s to 100s of km, etc.

The NWP outputs enable prediction of weather events at a time scale of hours to days to a week or more and at a space scale of 100s and 1000s of km or more. This is because the grid size of NWP models is of order of 10s to 100s of km. For detection of presence of a weather system of wavelength λ, numerically the grid-size must be of order less than or equal to λ/4. The Sub-Grid Scale weather systems are of space scale 10s to 100s of meters and last for a few minutes only. So, indirect empirical methods are used to delineate the sub-grid scale weather systems as the forecast produced by NWP models can not directly provide an insight. Some of the significant Sub-Grid scale phenomena are tornadoes and cloudbursts. Various empirical techniques are used by the experienced forecasters to infer these events and one of them is Model Output Statistics as explained next.

(*ii*) The NWP output consists of mainly flow patterns namely wind, temperature, humidity and pressure fields at various temporal and spatial levels. The forecast of actual weather elements like rain/snow etc. are derived from the NWP output products through statistical relationship, known as Model Output Statistics (MOS). In India, both at National Centre of Medium Range Weather Forecast (NCMRWF) and Indian Meteorological Department (IMD), one or the other form of MOS has been used. General experience is that MOS products show improved skills over the raw model output as suggested by Neiley and Hanson, 2004. But MOS is not a theoretically stable process as it requires longer datasets on time scale to derive the MOS relationships for forecasting of weather elements. Longer and consistent datasets are not available because of frequent revisions of NWP model.

The other approach being adapted by meteorologist to detect storms and their intensity is using Doppler radar. It works by sending out radio waves from an antenna. These radio waves are reflected back to the antenna by objects in the air. Through this process Doppler radar can detect precipitation in the air. It even detects frequency differences based on whether an object is moving away from the antenna or towards it. After the antenna detects an object it sends the information back to a computer that brings up the different frequencies as different colors. The colors used represent speed and direction to the user. Using this method, meteorologists can tell a lot of things about the storms it detects, as explained online by Premium Weather service. Using this important information meteorologists are then equipped to send out warnings for specific areas. Doppler radar aids meteorologists in sending out cloudbursts warnings but the radar cannot forecast a Sub-Grid scale phenomenon few days in advance.

There is thus a strong need for searching supplementary tools for interpretation of weather patterns provided by NWP models into weather elements including cloudburst occurrence.

2. Cloudburst

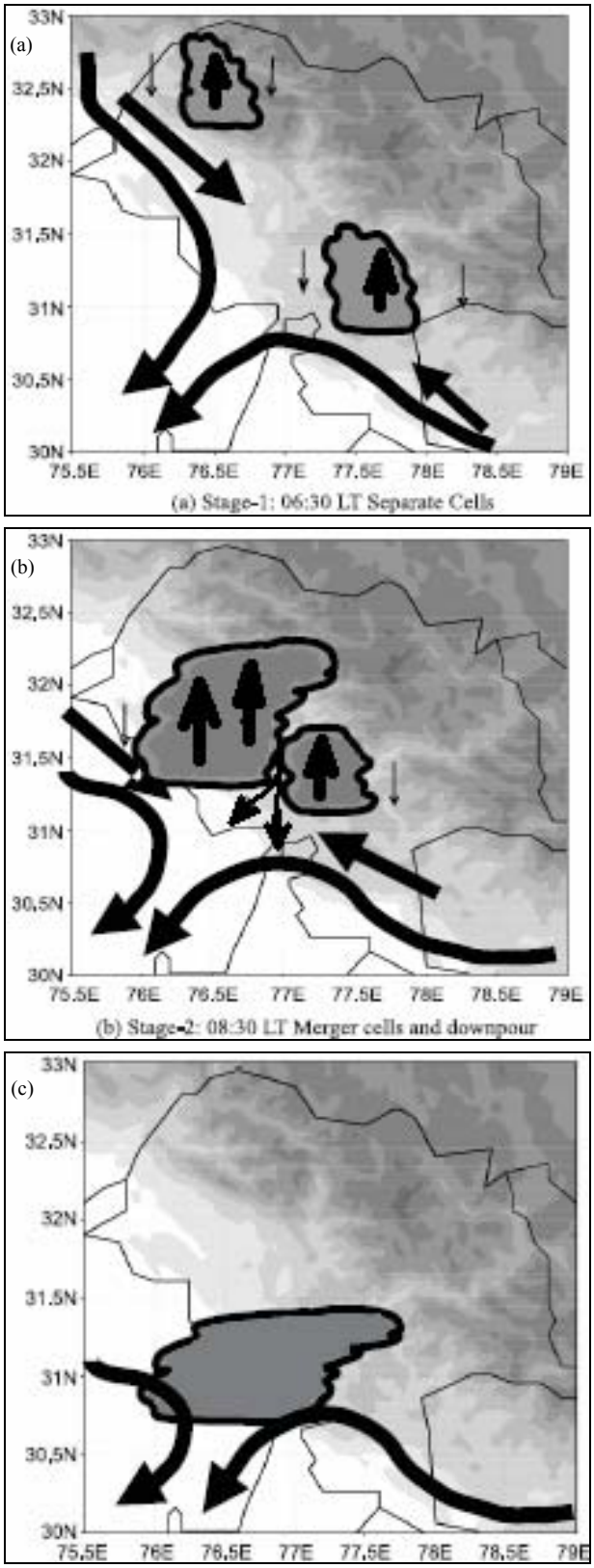
2.1. *About the Weather Event - Cloudburst*

Cloudbursts or downpours are sub-grid scale weather events and usually signify a sudden, heavy fall of rain over a short period of time as discussed by Heidorn, 2009. The rainfall rate in excess of 25 millimeters per hour (1 inch per hour) constitutes a downpour.

Most cloudbursts come from convective, cumulonimbus clouds that form thunderstorms and the air is generally rather warm in order to contain the amount of moisture required for a heavy downpour. Besides providing the proper conditions to spawn large quantities of liquid water drops, cumulonimbus clouds have regions of strong updrafts which hold raindrops aloft en masse and can produce the largest raindrops (those greater than 3.5 mm, in diameter).

These updrafts are filled with turbulent wind pockets that toss small raindrops around with large force. Within the turmoil of the randomly moving drops, there are more collisions among the drops and many of those close encounters result in their conglomeration into new drops larger in size.

Eventually all updrafts collapse, and when they do, the upheld raindrops descend unobstructed toward the surface, often forming a strong downdraft.



Figs. 2(a-c). A conceptual model of cloudburst Stage-1: Separate Cells, Stage-2: Merger of Cells and heavy downpour, Stage-3: Dissipation [Das *et al*., 2006]

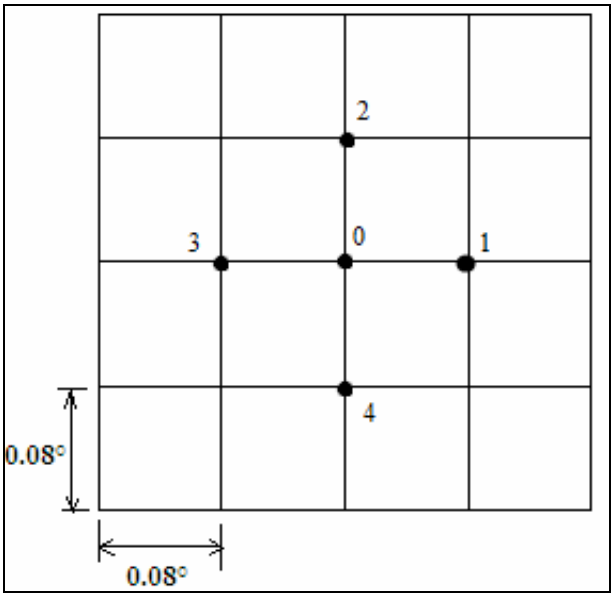


Fig. 3. A snapshot of gridded locations to calculate convergence

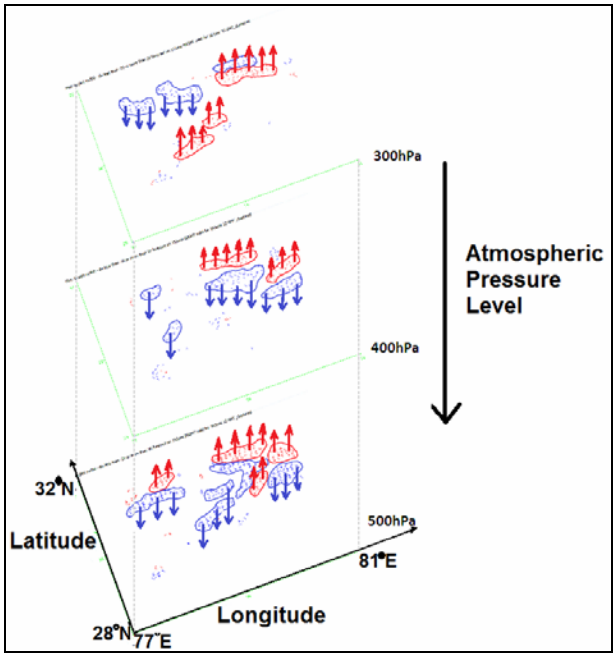


Fig. 4. 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 1200GMT 16 June 2013 (Red points and arrows correspond to convergence, blue points and arrows correspond to divergence)

2.2. *A conceptual model of the cloudburst*

A conceptual model of the cloudburst explained by Das *et al*., 2006, is based on the development of the vertical shear, vertical motion and the moisture distribution is shown in Figs. 2(a-c), which illustrates three stages in the lifecycle of the cloudburst. The drift of the cells towards each other and their vertical motion as explained in section IV can be related to the concept that follows.

In the first stage, the two convective cells are separate and drift towards each other as part of the mean flow [Fig. 2(a)]. Isolated heavy rain occurs during this stage. In the second stage [Fig. 2(b)], the two convective cells merge. Intensification follows due to strong wind shear and intense vertical motion. Heavy downpour and formation of the anvil also occurs. The storm moves rapidly southward due to strong steering flow. The third stage [Fig. 2(c)] is one of dissipation in which the two merged cells form one single large cell, which drifts westward and the cloudburst ceases.

The cloudburst in India occurs during monsoon season over the orographically dominant regions like Himalayan region, northeastern states and the Western Ghats.

3. Cloudbursts over Uttarakhand on 17th June, 2013

A series of cloudbursts wreaked havoc in 5 districts of Rudraprayag, Uttarkashi, Chamoli, Pithoragarh and Tehri, on 16-17 June, 2013. The cloudbursts led to flash floods that swept away mountainsides, villages, people, animals, houses, trucks, cars, roads. Ten days later, about 6,000 persons were rescued from Kedarnath and more than 800 bodies were recovered in and around Kedarnath, as given by a report by India today.

4. Case study using weather research and forecasting output products

4.1. *About weather research and forecasting model*

The Weather Research and Forecasting (WRF) Model is a next-generation mesoscale numerical weather prediction system designed to serve both atmospheric research and operational forecasting needs as explained in Wikepedia and the official website by "The WRF model". The WRF model features two dynamical cores, a data assimilation system, and a software architecture facilitating parallel computation and system extensibility. The model serves a wide range of meteorological applications across scales from tens of meters to thousands of kilometers. The effort to develop WRF began in the latter part of the 1990's and was a collaborative partnership principally among the National Center for Atmospheric Research (NCAR), the National Oceanic and Atmospheric Administration (represented by

TABLE 1

Calculation of convergence based on 36 hour forecast of *v* and *u* wind velocities at 300 hPa, valid for 1200 GMT 16 June, 2013 (*Source* : as a result of pre-processing WRF output file in .grib format provided by IMD)

Point no. (w. r. t. Fig. 3)	Latitude (°N)	Longitude (°E)	*v* (m/s)	*u* (m/s)	Convergence ($\times 10^{-5}$ per second)
4	30.88	79.67	9	16	-40.42
3	30.95	79.59	9	16	-40.00
0	30.95	79.67	9	16	-40.42
1	30.95	79.75	8	17	-47.08
	31.03	79.59	5	14	-66.25
2	31.03	79.67	4	13	-72.92
	31.1	79.67	-1	-7	-20.42

the National Centers for Environmental Prediction (NCEP) and the (then) Forecast Systems Laboratory (FSL)), the Air Force Weather Agency (AFWA), the Naval Research Laboratory, the University of Oklahoma, and the Federal Aviation Administration (FAA).

The WRF model under study produces forecasts at a resolution of 0.08 degrees latitude and longitude, *i.e*., approximately 9 km.

4.2. *Datasets under consideration*

The datasets produced as forecast by WRF model are in GRIdded Binary (GRIB) format which is a mathematically concise data format commonly used in meteorology to store historical and forecast weather data. It is standardized by the World Meteorological Organization's Commission for Basic Systems. The forecast datasets of the model include values for 74variables (including all atmospheric pressure levels), for latitude 5.673° N to 37.553° N and longitude 63.769° E to 98.659° E at a grid spacing of 0.08°. There are 431 points on meridian (Ny) and 431 points on a parallel (Nx), so making it equal to 431 × 431 grid points i.e. forecast of 74 variables at 1,85,761 grid points. Finally it becomes a huge datasets of 1,37,46,314 (approx. 13 million) values for just one forecast valid for a particular time.

For the purpose of analysis of this case of cloudburst, the WRF output products have been provided by IMD, Delhi. 12 hr, 24 hr, 36 hr, 48 hr, 60 hr, 72 hr forecasts made with initial conditions of 0000 GMT 14 June, 2013, 0000 GMT 15 June, 2013 and 0000 GMT 16 June, 2013 have been provided. The forecasts valid for 1200 GMT 16 June, 2013, 0000 GMT 17 June, 2013 and 1200 GMT 17 June, 2013 have been pre-processed as given in section 4.3 to derive convergence and mined using the k-means clustering technique of data mining as discussed in section 4.4.

4.3. *Pre-processing of WRF model outputs*

The GRIB files have been converted to (.csv) format by using National Digital Forecast Database - NDFD GRIB2 decoder program of NOAA downloaded from Internet. The author has derived Convergence, *i.e*., indirect vertical wind motion by using meridian (*v*) and zonal (*u*) component of wind as forecasted by model. For analysis, convergence at atmospheric pressure level lower upto 500 hPa has been considered. The *v* and *u* wind velocity at 500 hPa, 400 hPa and 300 hPa levels have been used to derive convergence / vertical wind motion which is the most important ingredient for cloudburst formation. The formula used is as follows:-

$$\text{Convergence formula} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}$$

where,

v denotes meridian wind flow in m/s,

u denotes zonal wind flow in m/s,

x denotes longitude and

y denotes latitude.

A snapshot of the grid points used for calculations is shown in Fig. 3. Here 0.08° corresponds to approximately 9 km distance. The expressions used for the calculation of

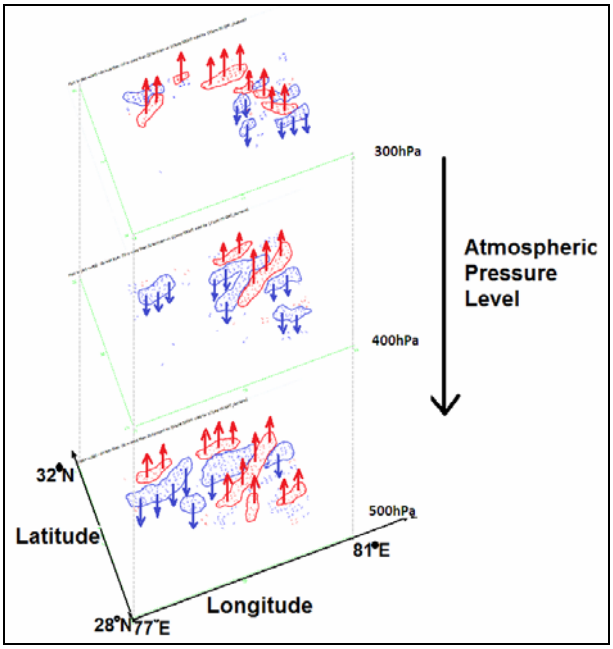


Fig. 5. 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 0000GMT 17 June 2013 (Red points and arrows correspond to convergence, blue points and arrows correspond to divergence)

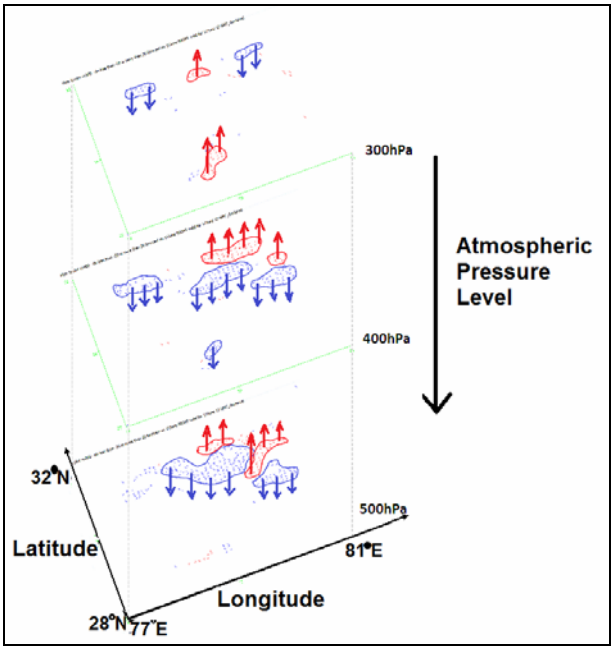


Fig. 6. 3-dimensional visualization of forecast of convergence (arrows represent vertical motion field), valid for 1200GMT 17 June, 2013 (Red points and arrows correspond to convergence, blue points and arrows correspond to divergence)

convergence are based on the forecast of fundamental variables, *viz.*, *v* wind-velocity and *u* wind-velocity at grid points of WRF model, are mentioned below :

Convergence at grid-point named 0

$$= \frac{(u_1 - u_3)}{18\,\text{km}} + \frac{(v_2 - v_4)}{18\,\text{km}}$$

A sample of the calculations of convergence corresponding to gridded location no. 0 (79.67° E, 30.95° N at 300 hPa) in Fig. 3, for 36 hour forecast valid for 1200 GMT 16 June, 2013, is depicted in Table 1.

These calculations have been done for an approximate area of 2.5° × 2.5° window surrounding the location of cloudburst, *i.e.*, Rudraprayag, Uttarakhand (the latitude of 30.28° N and the longitude of 78.98° E) and hence area under consideration is 28.73° N, 77.64° E to 31.33° N, 80.48° E.

4.4. *Application of clustering technique*

We have tried to pick up zones of updraft of air mass that is an early signal of formation of cloudburst, represented by convergence in the pre-processed dataset.

Here, k-means clustering technique offered by WEKA tool as explained by official website of Weka and by Witten and Frank, 2005, has been applied. WEKA is a collection of machine learning algorithms for data mining. It is an open Source Machine Learning Software in Java. It offers many techniques for mining of datasets. We have used Clustering technique of WEKA to group a set of objects (on the basis of value of convergence) into clusters so that the objects within a cluster have a high similarity in comparison to one another, but are dissimilar to objects in other clusters.

The tool provides various options to select/ reject the attributes on the basis of which the clustering should be done. The groups that are identified are exclusive so that an instance belongs to only one group. We have generated clusters on the basis of the feature *viz.* convergence. This field has positive and negative values depicting downdraft and updraft of air mass respectively. So, the number of clusters has been taken as equal to two.

The k-means method first chooses k points at random as cluster centers. All instances are assigned to their closest cluster centre according to the ordinary Euclidean distance metric. Next the centroid/mean of the instances in each cluster is calculated. These centroids are taken to be new center values for their respective clusters.

Finally, the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and will remain same forever.

In the pre-processed datasets, we have taken k = 2 as the idea is to observe the patterns signifying convergence and divergence. The k-means algorithm has been applied three different sets of forecasts, *i.e*., forecast with initial conditions of 0000GMT 15 June, 2013 valid for 1200 GMT 16 June, 2013, forecast with initial conditions of 0000 GMT 15 June, 2013 valid for 0000 GMT 17 June, 2013 and forecast with initial conditions of 0000 GMT 15 June, 2013 valid for 1200 GMT 17 June, 2013.

Out of these three forecasts, the forecast valid for 1200 GMT 16 June, 2013 and 0000 GMT 17 June, 2013 are near the time of occurrence of cloudburst. This data mining technique has hence generated two clusters for each forecast and the data corresponding to downdraft and updraft areas are visualized across the longitude and latitude. This technique has been applied for the datasets at atmospheric pressure levels of 300 hPa, 400 hPa and 500 hPa separately.

The 3-dimensional visualization of the convergence and divergence at the mentioned atmospheric pressure levels for forecast valid for 1200 GMT 16 June, 2013, valid for 0000 GMT 17 June, 2013 and valid for 1200 GMT 17 June, 2013, is being plotted in Figs. 4, 5 and 6 respectively.

5. Results and discussion

There is a very strong vertical motion field in the clusters of ensemble of forecast, shown in Figs. 4, 5 and 6. If we compare these patterns of convergence with that shown in the conceptual model of cloudburst given in Fig. 2(b), the formation of cloudburst signals is clearly indicated especially in Figs. 4 and 5 which are more close to the time of occurrence of cloudbursts. It is observed that this very large region of convergence is an early signal of formation of cloudburst.

Location of Uttarakhand is approximately at 550 hPa so only the clusters at 300 hPa, 400 hPa and 500 hPa are important indicators for observation. Here, the convergence fields at levels of 1000 hPa, 925 hPa, 850 hPa and 700 hPa are virtual areas because of high altitude of Uttarakhand and hence not being analyzed in this particular study.

The study of four real life cases of cloudburst over Indian region and over Dhaka was also taken up earlier by Pabreja and Datta, 2012 that used the Multidimensional Data Model for storage and retrieval of huge forecast data files. This study was based on forecast by European Center for Medium-range Weather Forecasting (ECMWF) model and provided strong evidence to support the cloudburst case of coastal region, *i.e*., Dhaka. Here, by using WRF model output products where grid size is more precise, it has been validated that there is formation of signals that contribute for the occurrence of cloudburst even 3 days in advance.

Hence, it is concluded that there is presence of advance signal of formation of cloudburst that is indicated through ensemble of different temporal forecasts of convergence produced by the model and this technique can supplement the traditional technique using MOS for the forecast of Sub-Grid scale disastrous weather phenomenon.

Acknowledgement

The author would like to express gratitude to Dr. S. K. Roy Bhowmik, DGM (Numerical Weather Prediction), IMD, Delhi for providing co-operation to carry out this study.

References

About The WRF model, downloaded from http://www.wrf-model.org/.

Cloud Burst over Uttarakhand A report by India today.

Das, S., Arshit, R. and Moncrief, M. W., 2006, "Simulation of a Himalayan cloudburst event", *Journal of Earth System Science*, **115**, 3, 299-313.

Heidorn, K. C., 2009, "The weather doctor, Exploring the Science and Poetry of Our Weather and Atmosphere (online book)", Accessed 2 Jan 2009. Available from: http://www.islandnet.com/~see/ weather/ doctor.htm.

NDFD GRIB2 decoder program of NOAA from Internet Available from: www.nws.noaa.gov/mdl/degrib/download.php.

Neiley, P. P. and Hanson, K. A., 2004, "Are model output statistics still needed?", 84th AMS Annual Meeting (session 6), Seattle, Washington, January 11-15, 1-5. Available from: http://ams.confex.com/ams/pdfpapers/73333.pdf.

Pabreja, K. and Datta, R. K.,2012, "A data warehousing and data mining approach for analysis and forecast of cloudburst events using OLAP-based data hypercube", International Journal of Data Analysis Techniques and Strategies, Inderscience Publishers, **4**, 1, 57-82.

Premium Weather, "The ultimate personal weather service", Accessed 10 February, 2009, Available from: http://www.tornadochaser.net/ tornado.html.

Short, N. M.., 2005, "The Remote Sensing Tutorial [CD-ROM]", Federation of American Scientists, Washington. Accessed 20 January 2010.Available from: http://www.fas.org/irp/imint/docs/rst/Sect14/Sect14_1d.html.

Weather Research and Forecasting (WRF), Model, available from Wikipedia, the free encyclopedia.

Weka 3- Data mining with open source machine learning software, downloaded from Internet. Available from:http://www.cs.waikato.ac.nz/ml/weka/.

Witten, I. H. and Frank, E., 2005, "Data Mining Practical Machine Learning Tools and Techniques", Second edition, Morgan Kaufmann Publishers.