MAUSAM, **66**, 3 (July 2015), 387-402

551.509.5 : 551.515.2 : 551.577

Spatial verification of rainfall forecasts for very severe cyclonic storm 'Phailin'

SAJI MOHANDAS and HARVIR SINGH

NCMRWF, A-50, Sector-62, Noida, U. P. – 201 309, India

e mail : saji.md@nic.in

सार – इस शोध पत्र में मॉडल पूर्वानुमानों के व्यापक मूल्यांकन के लिए पारंपरिक और स्थानिक नैदानिक दोनों प्रकार की तकनीक को साधन के रूप में किए गए उपयोग को दर्शाया गया है। इसका मूलभूत कार्य भौतिक प्रक्रियाओं के संबंध में विशेष रूप से उच्च विभेदन मॉडल्स तथा प्रेक्षणों के लिए मॉडल की अन्य कमियों और अच्छाईयों को उजागर करना है। पारंपरिक निष्कर्ष भी तथाकथित 'डबल पेनल्टी' मद से बाधित होता है और इस प्रकार यह अकेले पूर्वानुमान तथा प्रेक्षित वर्षा के पैटर्न्स के बीच स्थानिक एवं कालिक आधार पर वर्षा के परिमाण उपलब्ध नहीं करा सकता है। ऑब्जेक्ट-बेस्ड डॉयग्नोस्टिक इवैल्यूएशन की विधि एक प्रकार से विस्थापन विधियों की श्रेणी की एक स्थानिक (spatial) सत्यापन तकनीक है जबकि वेवलेट विश्लेषण स्थानिक सत्यापन के फिल्टरिंग टाइप से किया जाता है। इसमें पहले वाला विशेषताओं पर आधारित सत्यापन तकनीक है जबकि बाद वाला स्केल सेपरेशन सिद्धान्त पर आधारित है। इस शोध पत्र में अति प्रचंड उष्णकटिबंधीय चक्रवात 'फैलीन' की स्थिति को अध्ययन के लिए लिया गया है और भूमंडलीय पूर्वानुमान प्रणाली से वर्षा के पूर्वानुमान लिए गए है और राष्ट्रीय मध्य अवधि मौसम पूर्वानुमान केंद्र में इस एकीकृत मॉडल को चलाया गया है तथा इसका सत्यापन उपग्रह सह वर्षामापी मिश्रित वर्षा विश्लेषण के साथ किया गया है। सुनिश्चित और अनवरत परिमाणों का उपयोग करते हुए परंपरागत सत्यापन स्कोर्स की गणना की गई है तथा विभिन्न अवसीमाओं से स्थानिक सत्यापन स्कोर्स के साथ गणना की गई है। वर्षा पूर्वानुमान के संबंध में दोनों भूमंडलीय मॉडल्स के समग्र निष्पादन के परिणामों को यहाँ संक्षेप में प्रस्तुत किया गया है।

ABSTRACT. The current study demonstrates the utilisation of a tool for the comprehensive evaluation of model forecasts using both traditional and spatial diagnostic techniques. The fundamental idea is to provide additional and meaningful insight into the model weaknesses and strengths in terms of underlying physical processes especially for very high resolution models and observations. The traditional scores also suffer from the so called "double penalty" issue and hence alone cannot provide a measure of spatial and temporal match between the forecast and observed rainfall patterns. Method for Object-based Diagnostics Evaluation is a spatial verification technique in the category of displacement methods while wavelet analysis comes into filtering type of spatial verification. Former is a features based verification technique while the latter is based on scale-separation principle. The case of Very Severe Tropical Cyclone 'Phailin' is taken up for the study and the rainfall forecasts from Global Forecast System and Unified Model run at National Centre for Medium Range Weather Forecasting are verified against gridded satellite-cum-raingauge-merged rainfall analysis. The traditional verification scores were computed using categorical and continuous measures and the spatial verification scores were computed against various thresholds. The results are presented to summarise the overall performance of both the global models with respect to the rainfall prediction.

Key words – Model evaluation tool, Categorical verification scores, Object-based diagnostics, Intensity-scale analysis.

1. Introduction

National Centre for Medium Range Weather Forecasting (NCMRWF) has a mandate to constantly improve upon the numerical weather prediction models for the prediction of weather over India and its neighbourhood by adopting the latest developments in the modelling community. The modelling systems and its year-to-year improvement should be hand-in-hand with the performance evaluation of the available systems at the centre and its mutual comparisons. Day-to-day weather forecasts over the regions predicted by the operational model and other experimental models should be constantly monitored and the statistical measures of the different aspects of the various model-generated prognostic and diagnostic variables should be produced and archived to look into the various properties from all possible angles (For information on general framework of verification see Murphy and Winkler, 1987; Jolliffe and Stephenson, 2003; Stansky *et al*., 1989; Wilks, 2006 and Ebert, 2008). Not only the simple and direct properties of the model in terms of the traditional parameters like anomaly correlation and RMSEs, but also the measures related to the spatially coherent features of the model

should be investigated simultaneously to understand the performance and to diagnose the limits of the skills of the model and for comparison between the modelling systems (Davis *et al*., 2006 & 2009; Brown *et al*., 2007; Gilleland, 2013; Gilleland *et al*. 2009; 2010a & 2010b; Casati, 2010; Casati *et al*., 2004; 2008; Ebert, 2008 & 2009, Gallus, 2010; Ebert and McBride, 2000; Ebert and Gallus, 2009; Ahijevych, *et al*., 2009; Mittermaier and Roberts, 2010). The day-to-day statistics can be aggregated to estimate the overall performance of each of the episodes of synoptic scale and mesoscale phenomena occurring on a very regular basis in different types of weather regimes. The daily statistics of the episodes can be aggregated and summarised for every month to assess the monthly performance of the models, which in turn can be again aggregated over a season or year to condense the huge amount of information into very few quantitative figures. This will allow the year-to-year comparison of performance of multiple modelling systems or year-to-year variability for a single modelling system.

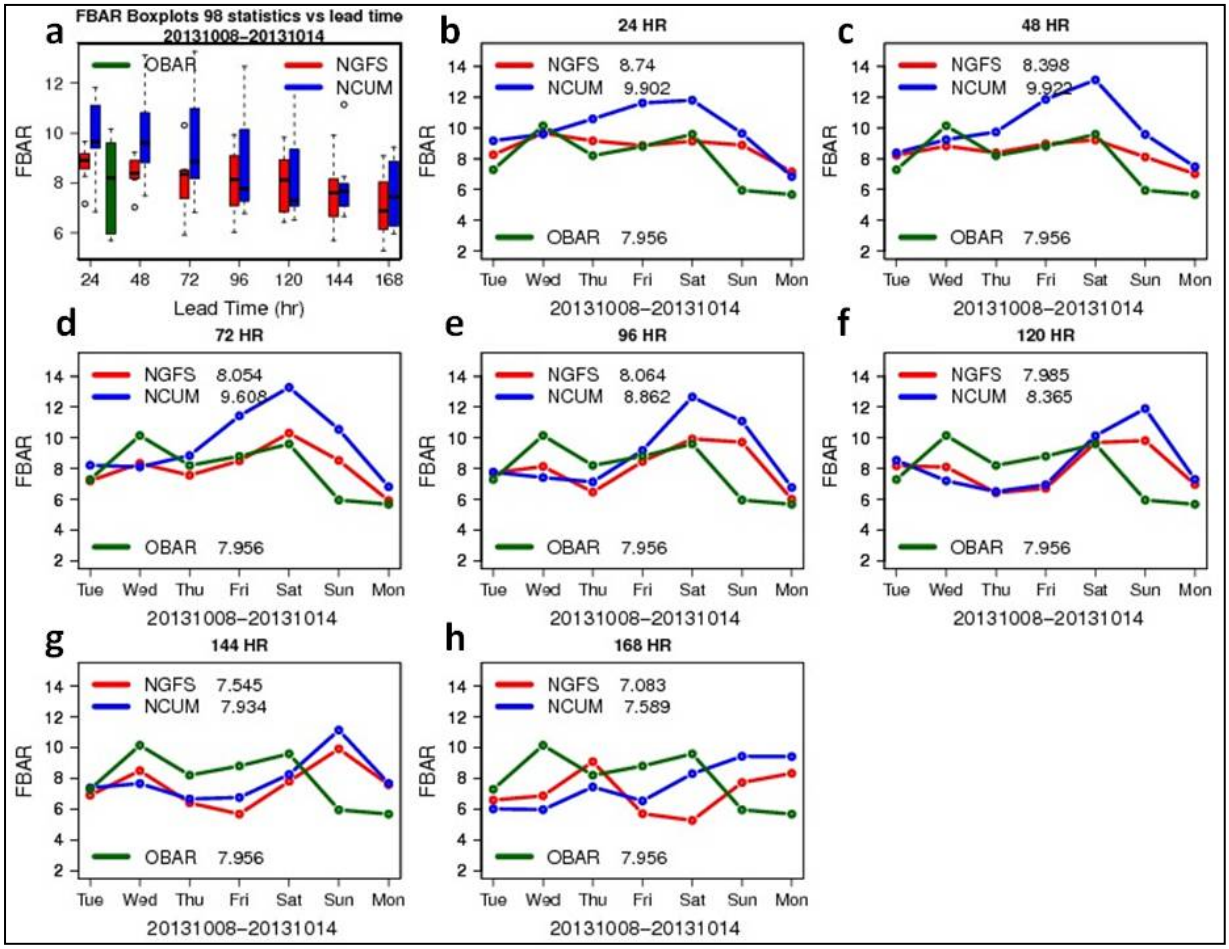
This paper focuses on the verification and evaluation of the rainfall predictions for a recent tropical cyclone event, namely the Very Severe Cyclonic System (VSCS) 'Phailin'. The current study evaluates the performance of the forecast of this system by two global models at NCMRWF adopted from National Centres for Environmental Prediction (NCEP) and United Kingdom Met. Office (UKMO), namely, NGFS and NCUM respectively (Prasad *et al*., 2011 & 2013; Rajagopal *et al*., 2012). The forecast of the tropical cyclone (TC) 'Phailin' was very successful with more or less accurate prediction of track, intensity and landfall and the gradual decay after the landfall. The current study is a demonstration of Model Evaluation Tools (MET) implemented on IBM Power 6 High Performance System at NCMRWF. MET is a tool for comprehensive performance evaluation between different models, of any variable with a forecast and with any corresponding observation or analysis. It incorporates both traditional scores as well as spatial verification scores like, Method for Object-based Diagnostics Evaluation (MODE) and wavelet analysis. It has been implemented for models like, Weather Research and Forecasting (WRF), Global Forecast System (GFS), Unified Model (UM) and the regional versions of UM.

The traditional scores suffer from the problem of the so-called double-penalty issue. This is because, traditional grid-point verification methods penalise a minor shift in the location twice, once for missing grid points where the precipitation event occurred, and also for predicting false alarms at some other grid points. MODE is an effective alternative to provide additional diagnostic information and an objective assessment of location, size and intensity errors of the synoptic systems which is otherwise impossible through traditional approaches. Wavelet analysis uses scale-decomposition approach to identify the scale at which the skill is maximised. It is applied to forecast and observation fields to obtain spatial scale components and to compute the bias, error and skill of forecast on each spatial scale. It provides the information on the ability of the model in reproducing the observed scale structure and scale dependency of error and skill. The current study is an attempt to diagnose the overall performance of the two global models for TC 'Phailin' in all angles using traditional and spatial verification techniques. This study is a preliminary attempt to formulate and design a set of standard diagnostic measures for the routine monitoring and objective assessment of the overall performance of the numerical models in rainfall prediction and for the comparison between different modelling systems. The following sections deal with the data and methodology, results and discussions followed by conclusions.

2. Data and methodology

TC 'Phailin' originated from a depression over north Andaman Sea on 8th October, 2013 near (12° N, 96° E) and moved west-northwestwards intensifying into deep depression on 9th and crossed Odisha and adjoining north Andhra Pradesh coast near Gopalpur at 2230 hrs IST of 12th October, 2013 as a Very Severe Cyclonic Storm. Sustained maximum surface wind speed reported was 215 kmph with estimated central pressure of 940 hPa as per estimates by India Meteorological Department (IMD). Maximum rainfall was over north-east sector at the time of landfall (38 cm at Banki in Cuttack district). All the model runs starting with initial analyses of 0000 UTC 8-14 October, 2013 and 7 days of forecasts were considered for the current study, The predicted 24-hour accumulated precipitation is compared between the two global models namely NGFS and NCUM. The rainfall forecasts valid for these 7 days are examined in detail to assess the overall performance with respect to traditional verification scores and the features-based verification procedures. The domain of study is (75-100° E, 5-30° N) which covers the TC system during the period of study. The resolution of NGFS is T574L64 global spectral corresponding to an average resolution of about 23 km near equatorial regions. The resolution of NCUM is N512L70 corresponding to an average resolution of around 30 km near equatorial latitudes.

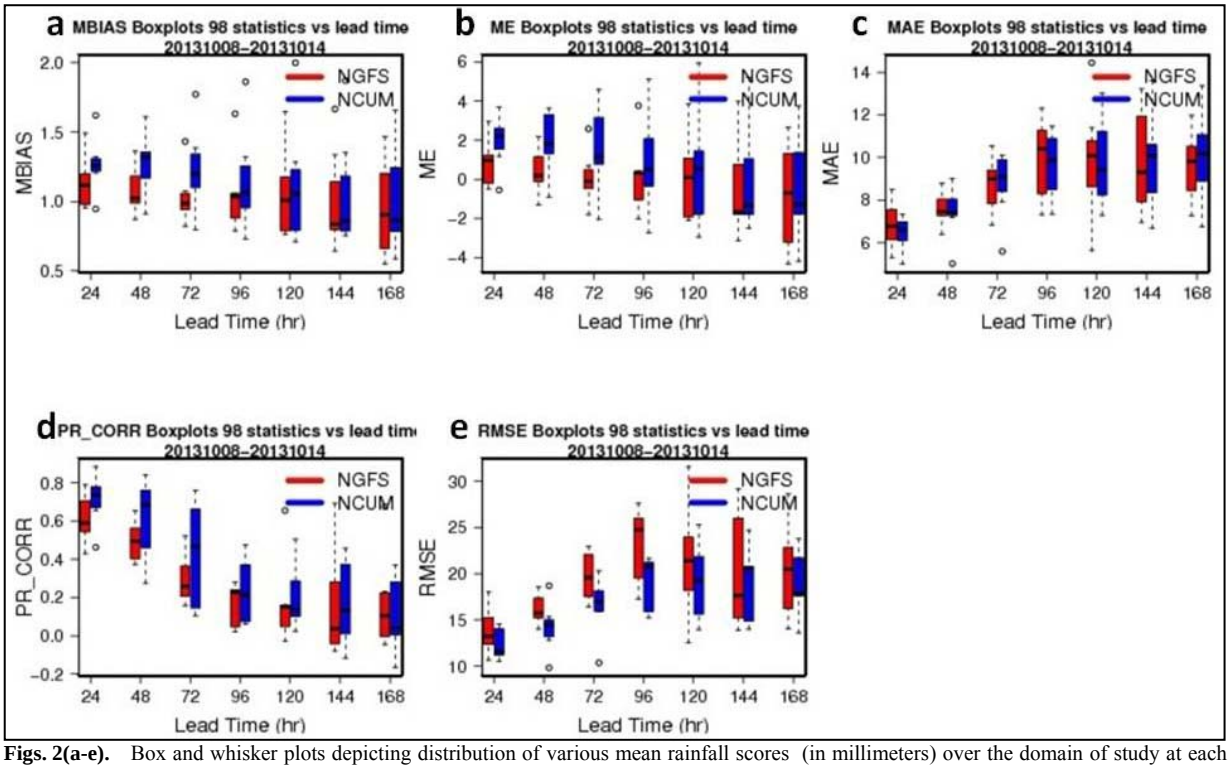
MET provides four major tools to estimate various kinds of verification statistics, namely, Point-stat, Grid-stat, MODE and Wavelet. Point-stat is the standard verification measure computed at station points and Grid-stat is the same computed at some common regular grid



Figs. 1(a-h). Box and whisker plots depicting distribution of mean rainfall (FBAR) (in millimeters) over the domain of study at each forecast lead times of 1-7 days from NGFS and NCUM runs vs. mean observed gridded rainfall analyses (OBAR) along with the corrresponding time series at each forecast lead times for the period 8-14 October, 2013. The mean value for all the 7 days period is also written along with the legends

points. Before the computation of the statistics, both observation and forecast matched pairs need to be generated at a common grid by some of the most popular re-gridding techniques suitably selected for the variable under study. The resolution of the current study is restricted by the resolution of the gridded rainfall analysis which is 50 km. The model rainfall is regridded to 50 km bi-linearly using copygb utility. Traditional scores were computed for both continuous and categorical measures using Point-stat and Grid-stat. Continuous measures are basically based on the difference between forecast and observed rainfall, whereas categorical measures are based on the 2 × 2 categorisation of 'yes' or 'no' of rainfall values at different rainfall thresholds by generating a contingency table for each of the threshold. As the focus is on Tropical Cyclone and the number of land raingauge stations reporting the associated rainfall is very less, the gauge-based metrics are not shown in the current paper. Grid-stat results are presented for traditional scores, which are computed against the IMD-NCMRWF gridded satellite-rainguage merged rainfall analysis (Mitra *et al*., 2003 & 2009).

Spatial verification of rainfall comes into at least four types - neighbourhood, object-based, scale-separation and deformation. But the two categories being used here are scale separation (a filtering approach) and object-based (a displacement approach). For object-based verification, the MODE was used and for scale separation, wavelet stat tool was used both of which are part of the Model Evaluation Tools (MET). Individual days of scores were averaged across the days and the forecast lead times to assess all the aspects of the verification and overall summary scores. Appendix-I gives the brief description of MODE and the settings adopted for the current study (Brown *et al*., 2007 and Davis *et al*., 2006). Wavelet stat tool decomposes the forecasts and observations according to intensity and scale, by thresholding the same to convert



Figs. 2(a-e). Box and whisker plots depicting distribution of various mean rainfall scores (in millimeters) over the domain of study at each forecast lead times of 1-7 days from NGFS and NCUM runs computed against observed gridded rainfall analyses for the period 8-14 October, 2013. The scores are Multiplicative Bias (MBIAS), Mean Error (ME), Mean Absolute Error (MAE), Pearson Correlation (PR_CORR) and Root Mean Squared Error (RSME)

into binary fields and decomposing into sum of components of different scales. Casati *et al*. (2004) describes the methodology in detail. A 2-dimensional Haar wavelet filter is used. Discrete wavelet transforms are usually performed on square domains of $2^n \times 2^n$ grid points. Automated tiling method is adopted here which figures out the maximum tile of dimension $2^n \times 2^n$ that fits within the domain and places the tile at the centre of the domain. For each threshold and for each scale component of binary forecast and observation, mean squared error (MSE) is evaluated. The largest error is typically associated with smallest scale and highest threshold. For each threshold and scale component, the intensity-scale skill (ISS) score based on the MSE of binary forecast and observation scale components is evaluated taking random chance as reference forecast (Casati *et al*., 2004; Jolliffe and Stephenson, 2003; Wilks, 2006). For each threshold (t) and scale component (j), the MSE for random binary forecast is equipartitioned on the n + 1 scales to evaluate the ISS.

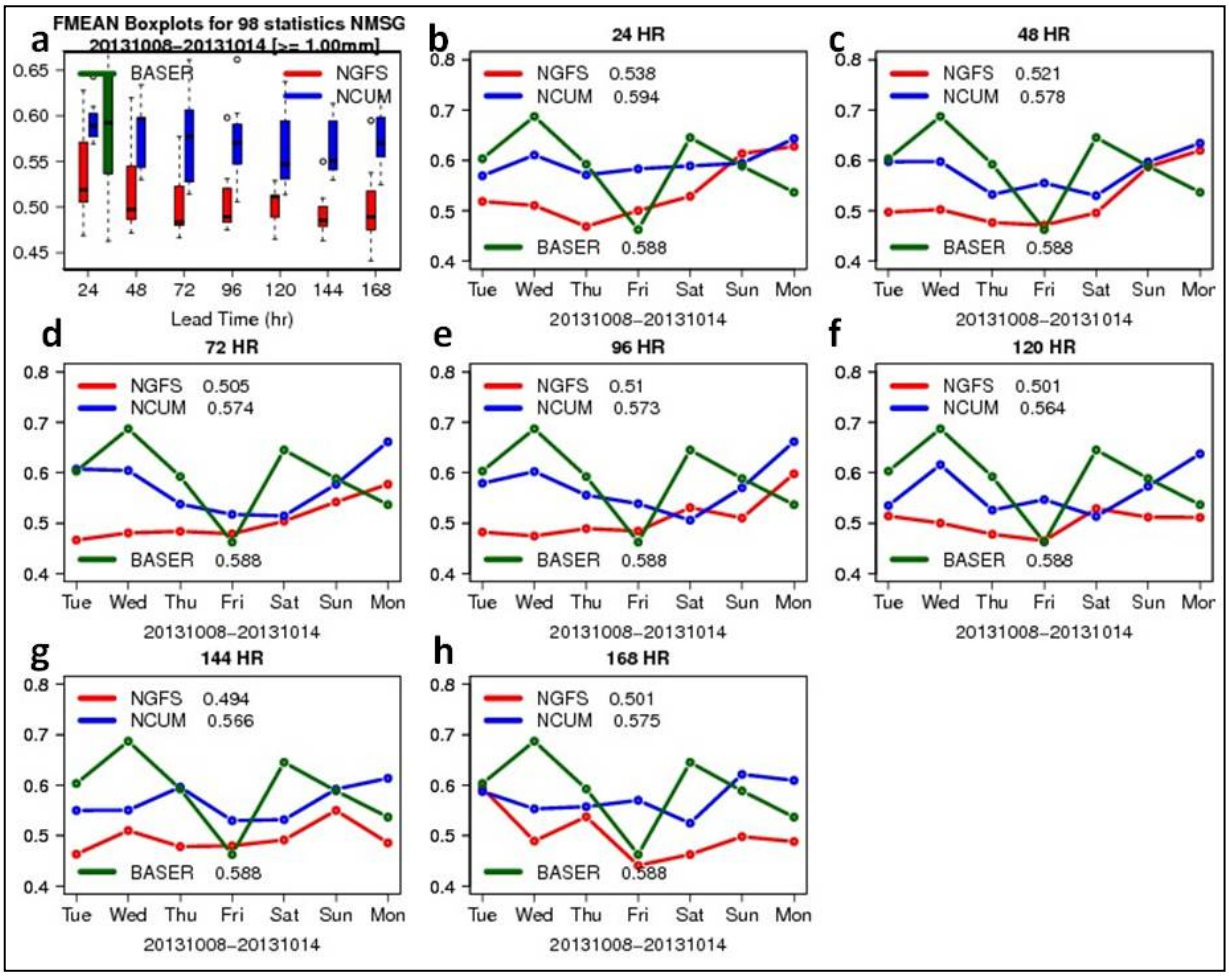
$$ISS(t, j) = [MSE(t)_{random} - MSE(t, j)*(n+1))]/MSE(t)_{random}$$

The detailed analysis of the mature stage of the tropical cyclone around the period of landfall of the system was carried out to demonstrate the capabilities of the MODE and wavelet stat tool and finally, the overall performance was objectively assessed.

3. Results and discussion

3.1. *Traditional verification scores*

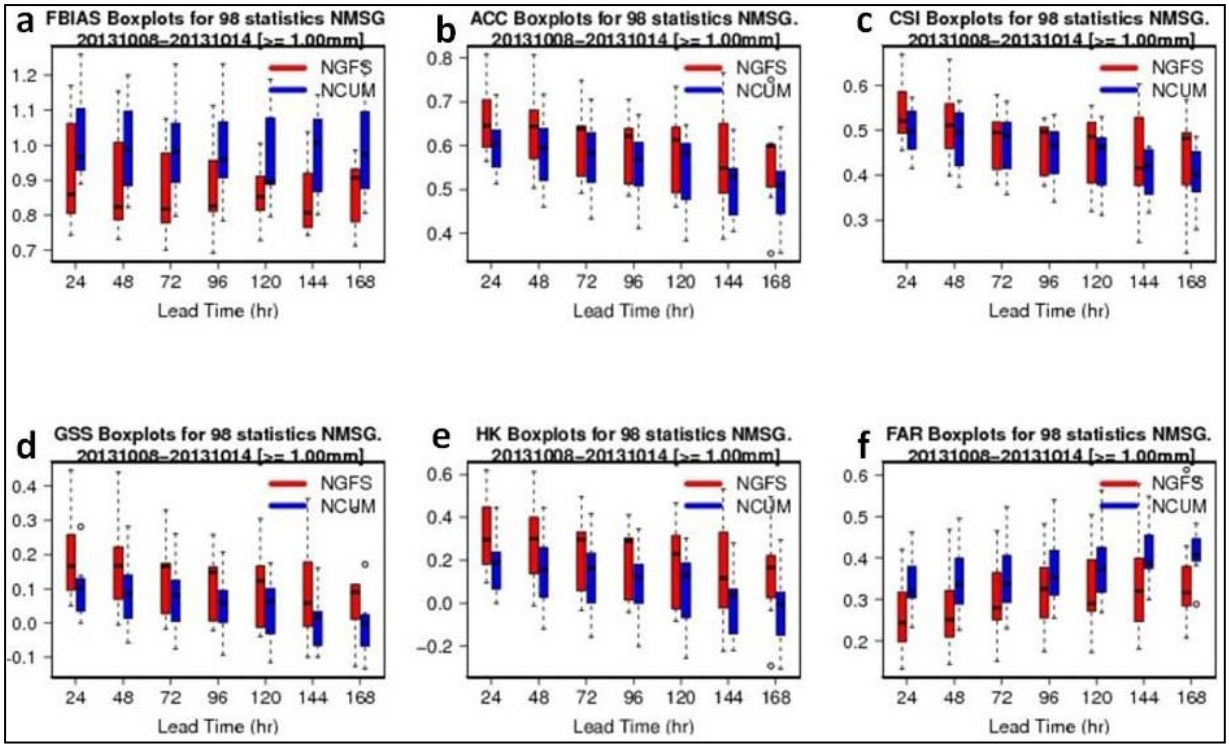
The current section deals with the traditional verification of NGFS and NCUM forecasts of 'Phailin' tropical cyclone against IMD-NCMRWF satellite-cum-raingauge-merged gridded rainfall analysis as mentioned in the previous section. Two types of metrics are generated, the first with the continuous variables and the second with categorical variables. For continuous variables, the verification methods are consistent with the general framework of verification outlined by Murphy and Winkler (1987). The domain mean of the forecast and the observation computed over the forecast-observation pairs (FBAR and OBAR) is only one of the many important aspects of performance of the models. FBAR and OBAR are plotted together in Figs. 1(a-h), with colours of red (NGFS) and blue (NCUM) and dark green (OBAR) for the seven days period of 8-14 October, 2013. Fig. 1(a) depicts the boxes denoting the first, second and third quartiles of the mean rainfall on each forecast day of the period. The whiskers represent the maximum-minimum,



Figs. 3(a-h). Similar to Fig. 1, except for categorical rainfall metrics for the rainfall threshold of ≥ 1 mm. The categorical rainfall score (in millimeters) shown is Mean Forecast (FMEAN) alongwith Base Rate (BASER) with its distribution according to the forecast lead times as well as the time series at each of the lead times 1-7

if there are no outliers, but the minimum fencing when there are suspected outliers represented by open circles. The mean values of individual forecast days are plotted as times series with the mean of all the days is shown along with the legends on each of the time series panels.

The observed domain mean rainfall could be seen to be ranging from around 6 mm to 10 mm with the median value of near around 8 mm. The mean value is also nearly equal to the median value for the 7 days period of 8-14 October, 2013, in the case of observed rainfall. The observed rainfall reached around 10 mm on 9th and 12th and after 12th there was a reduction in the domain mean rainfall after the landfall. It could be clearly seen that NCUM had a tendency to generally over predict the domain mean rainfall compared to NGFS, during the first four days of forecasts. There was a clear-cut tendency to over predict the associated rainfall for the tropical cyclone by NCUM up to day-5 forecasts [also seen in the figures of geographical plots of regridded rainfall of Figs. 6(a&b)], mainly owing to the larger spread of heavy or rather heavy rainfall contours. Overall, up to day-3 forecasts the domain mean rainfall by NGFS is more or less closer to the observed contours in the time series panels up to day-3, beyond which NGFS rainfall showed a tendency to under predict the same almost on all days. Both models were consistent in predicting maximum rainfall on 12th October, 2013 up to day-4, after which there was a general tendency to shift the maximum towards the later period. NCUM was able to predict the reduction in domain mean rainfall after 12th up to day-4 forecasts, while in the case of NGFS, up to day-3. Beyond day-5, there was found a general reduction in the skill of both models. The domain mean rainfall decreased from 8.74 on day-1 to 7.083 on day-7 for NGFS, while for NCUM there was a reduction from 9.902 to 7.589. The rainfall dipped below observed after day-5 for NGFS and after day-6 for NCUM.

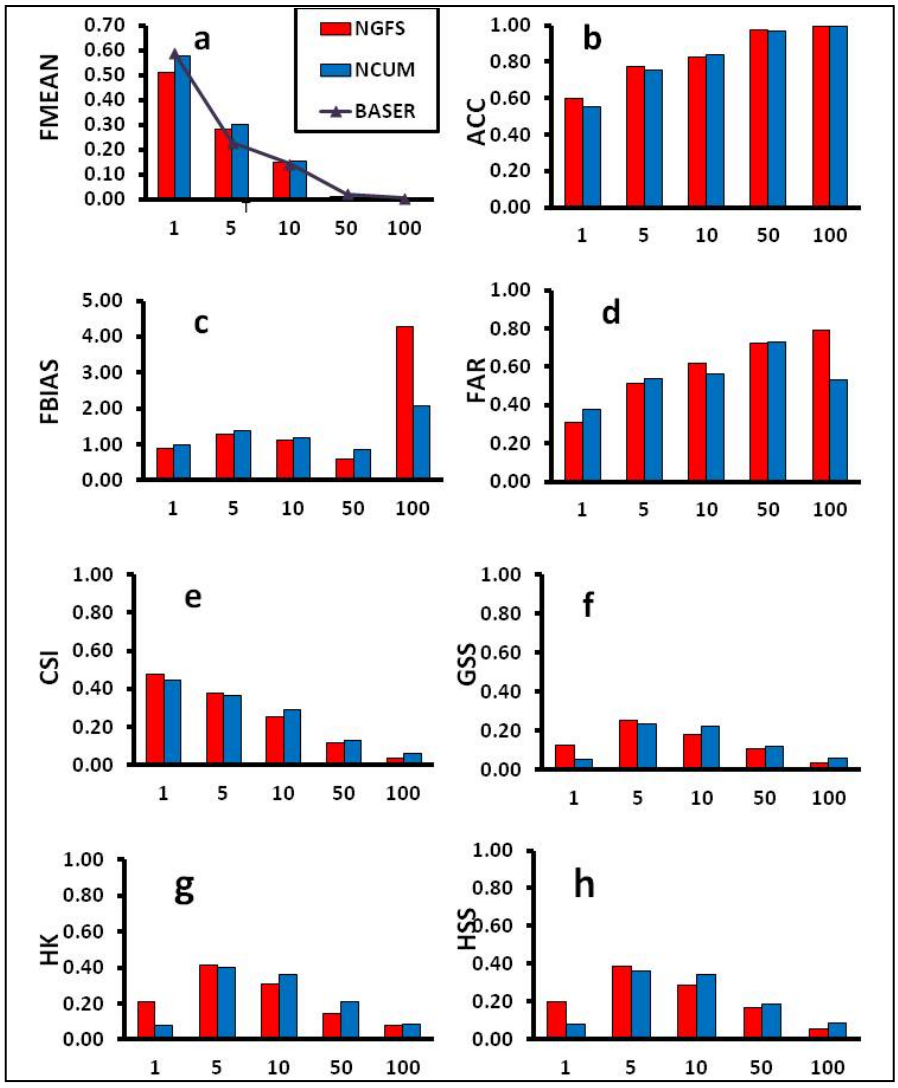


Figs. 4(a-f). Similar to Fig. 2, except for the categorical rainfall metrics for a rainfall threshold of ≥ 1 mm. The categorical scores (in millimeters) shown are Frequancy Bias (FBIAS), Accuracy (ACC), Critical Success Index (CSI), Gilbert Skill Score (GSS), Hanssen_kuipers Discriminant (HK) and False Alarm Ratio (FAR)

For other scores for continuous variables the box and whisker plots are shown in Figs. 2(a-e). The measures shown are multiplicative bias (MBIAS), Mean Error (ME), Mean Absolute Error (MAE), Pearson Correlation (PR_CORR) and Root Mean Squared Error (RMSE). MBIAS also gave the similar conclusions as FBAR/ OBAR that both NGFS and NCUM were slightly over predicting till day-4 and after that hovering around a value of 1. But NCUM was showing more over prediction in the first four days. From day-5, the day-to-day variability was larger in general for both the models. Mean Error (ME) decides the direction of bias. It was showing positive bias till day-4, and after that the bias was showing large variability around 0. Mean Absolute Error (MAE) indicated the order of error. It was showing a sharp increase in error from day-1 to day-4, from around 6 mm to 10 mm and after that the bias was more for NCUM during the first four days as well as RMSE. However, on a positive note, NCUM was also having better correlation between forecast and observation throughout the forecast period, compared to NGFS. The order of RMSE values was approximately between 10 and 25 mm, whereas the order of MAE was mostly between 6 & 12 mm. As RMSE imposes high penalty on large errors than does the MAE, this is very likely with a small sample size. Also RMSE was slightly higher for NCUM compared to NGFS.

For categorical measures, the daily rainfall is divided into 5 categories with thresholds of 1, 5, 10, 50 and 100 mm and various metrics were computed. Figs. 3(a-h) shows statistics for lower threshold value of 1mm. At 1 mm threshold, the observed values of Base Rate (BASER) are ranging from 0.45 to 0.67. The first quartile is near 0.56 and third quartile is near 0.65 with the median value of nearly 0.60. The mean value of Base Rate is 0.588 which is also closer to the median value. NCUM shows the mean forecast (FMEAN) of the same order as that of the observed throughout the forecast period. NGFS shows under prediction of rainy areas of 1mm and above at all lead times with most of the individual days being predicted to be lower and lower with lead time. NCUM does not show much reduction in the area extend of rainy grid points at the light rainfall threshold. However, NCUM values of FMEAN are closer to the Base Rate throughout the forecast period compared to NGFS. In short, it can be seen that NCUM predicts large areas of light rainfall during most of the forecast period, while NGFS showing much lesser area coverage of light rainfall category.

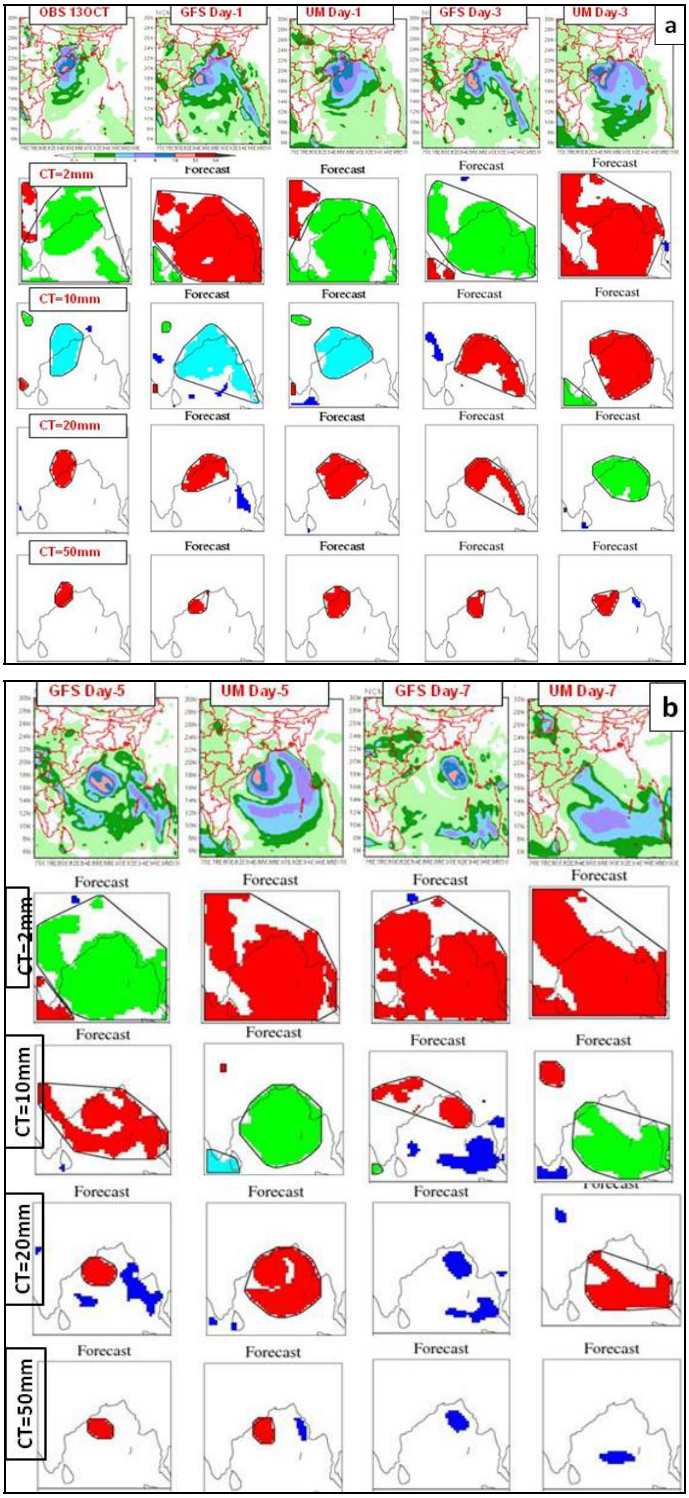
Similar picture arises when looking at all the other metrics in Figs. 4(a-f) for the same threshold. Figs. 4(a-f) shows the scores of important categorical measures,



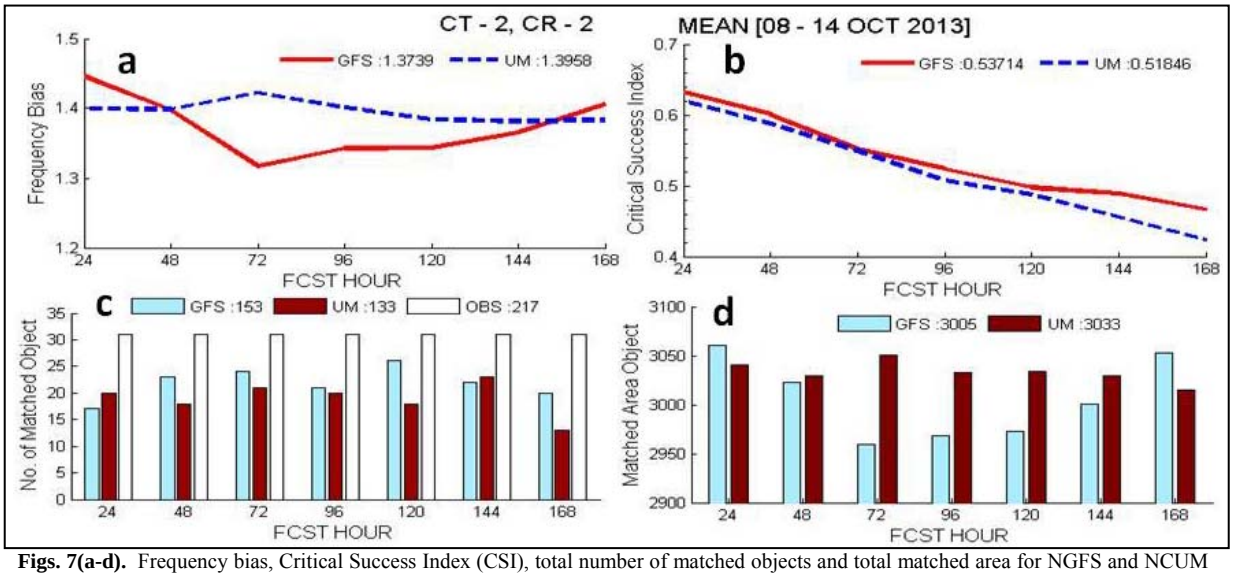
Figs. 5(a-h). Continuous verification statistics for daily rainfall (millimeters) against the gridded rainfall analyses averaged for the period 8-14 October, 2013 for NGFS and NCUM models and over each forecast lead times from day-1 to day-7; (a) mean forecast (FMEAN) (b) Accuracy (ACC) (c) Frequency Bias (FBIAS) (d) False Alarm Ratio (FAR) (e) Critical Success Index (CSI) (f) Gilbert Skill Score (GSS) (g) Hanssen-Kuipers Discriminant (HK) and (h) Heidke Skill Score (HSS). Thresholds used are 1 mm, 5 mm, 10 mm, 50 mm and 10 mm

Frequency Bias (FBIAS), Accuracy (ACC), Critical Success Index (CSI), Gilbert Skill Score (GSS), Hanssen-Kuipers discriminant (HK) and False Alarm Ratio (FAR). FBIAS is nearly 1 for NCUM for all lead times indicating that the frequency of 'yes' events are matching very well with that of the observed 'yes' events on all lead times. In contrast, NGFS under predicts the area throughout the forecast period for 1mm threshold. Accuracy is a measure of proportion of correct forecasts to the total number of events. Throughout most of the lead time, NGFS shows better performance with Accuracy always being on the higher side of 0.5 (with a range of about 0.6 to 0.7) for day-1 which gradually decreases to a range of (0.5-0.6) at day-7. NCUM is mostly showing values less than 0.5 after day-5, indicating poor performance beyond day-5. It can be seen that FAR is mostly on the higher side for NCUM on individual days at any lead time compared to NGFS. All the skill scores (CSI, GSS, and HK) shows slight upper hand for NGFS at all lead times. To summarise, it can be seen that at lower thresholds of 1mm, NGFS is clearly showing better skill than NCUM, though the area covering lower rainfall threshold is too less.

Looking at only one threshold will not give any conclusive measure of the performance of model rainfall forecast. So one needs to look at a range of thresholds,



Figs. 6(a&b). Observed and forecast rainfall regridded to 50 km resolution along with the simple objects and clusters captured at convolution thresholds 2, 10, 20, 50 mm. NGFS and NCUM forecasts are shown for (a) day-1 and day-3 and (b) day-5 and day-7 lead times. Rainfall contours are coloured at intervals 0.1, 1, 2, 4, 8, 16 and 32 cm and the objects of the same cluster are of single colour in a field but can have different colors in different fields. The blue objects are always un-clustered ones



Figs. 7(a-d). Frequency bias, Critical Success Index (CSI), total number of matched objects and total matched area for NGFS and NCUM for the period 8-14 October, 2013 as a time series of forecast lead time 24-168 hours for daily rainfall (mm). The mean value is also indicated in each panel. (Convolution threshold = 2 mm, Convolution radius = 2 grid sizes)

which are carried out in Figs. 5(a-h). It gives an overall overview of the performance of the models at thresholds of 1, 5, 10, 50 and 100 mm. As the number of days being verified are only 7 the sample size dwindles too fast as one goes from lower range to heavy to rather heavy rainfall category. So at thresholds of 50 and 100 mm, the comparison becomes statistically irrelevant, where the model mean rainfall (FMEAN) and the sample climatology (BASER) both converges to nearly zeroes. At 1mm threshold, the average FMEAN value is nearly 0.5 for NGFS and 0.6 for NCUM, which was also seen in Figs. 3(a-h). In between these thresholds, the BASER is nearly 0.2 at 5 mm and 0.15 at 10 mm. At 1 mm threshold, NCUM value of FMEAN is closer to BASER. At 5 mm, both models are slightly over predicting the frequency of rainy areas while NGFS is closer to the observed. At 10mm, both models are showing almost comparable performances and matching very well with the BASER.

In the case of Accuracy, at lower thresholds (1 and 5 mm), NGFS is showing higher values compared to NCUM. NGFS shows comparatively less frequency bias (FBIAS) especially in lower thresholds. Only in 100 mm threshold, NGFS is showing extremely high value of bias compared to NCUM. This may be mostly due to the difference in the intensity of the system in day-7 prediction by both the models. NGFS is mostly able to predict the intensity of the system up to day-7 whereas NCUM shows poor performance in that range. NCUM mostly fails to predict the cyclonic system in day-7, while NGFS is able to predict the same, but in a completely wrong location (as shown by the geographical plots of the mean rainfall which is not shown here). FAR is also less for NGFS in lower thresholds of 1 and 5 mm, while at higher thresholds (especially at 100 mm) NGFS values are high. This also supports the conclusions that NGFS is able to predict better intensity but at a wrong location beyond day-5 compared to NCUM, so that the relative small sample size of 100 mm threshold can generate very high value of FAR. All the skill scores (CSI ,GSS and HK) shown in Figs. 5(a-h) also in general shows that at lower thresholds of 1 mm and 5 mm NGFS shows better performance while at higher thresholds, NCUM shows better performance in the 'rather heavy' to 'heavy' category. Here is the relevance of verification methods which account for the performance of the models in terms of the location error and to qualify the models according to the question, 'which model has shown more location error and how far?'. The next section is devoted to identify the spatially coherent features of rainfall predicted by the models which is most likely to mimic the observed pattern and to arrive at an objective quantification of the spatial match or error.

3.2. *Method for object-based diagnostics evaluation (MODE)*

Method for Object-based Diagnostics Evaluation (MODE) is a spatial verification method, the details of which can be obtained from Davies *et al*. (2004). This is more suitable for mesoscale model verification of highly discontinuous fields like precipitation and cloudiness. Davies *et al*. (2004) used this for WRF model outputs at 4 km resolution. The current study may perhaps be the first attempt to apply it to the global model forecasts at a relatively coarse resolution of 50 km. Hence some tuning

TABLE 1

Median of Maximum Interest (MMI) for NGFS (G) and NCUM (U) for 24, 72, 120 and 168 hour forecasts valid for 13 October, 2013 at thresholds 1, 5, 10, 20 and 50 mm

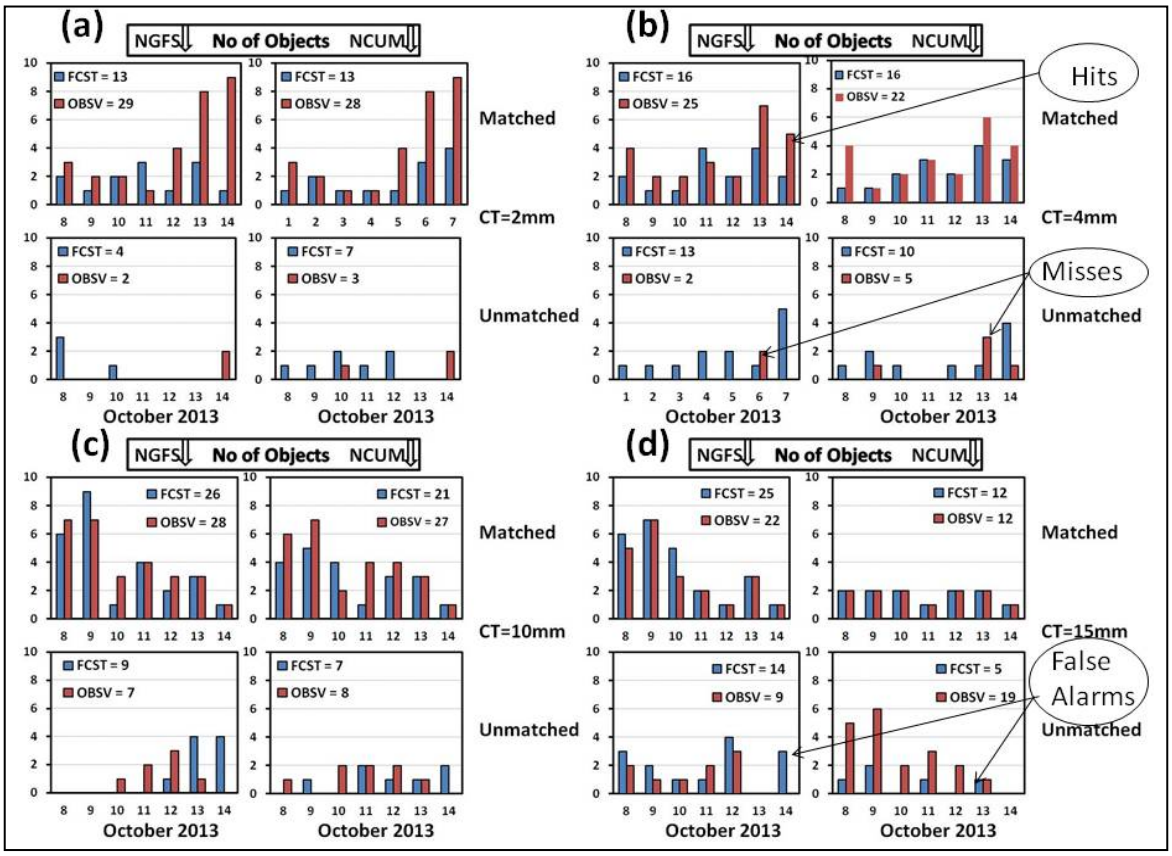
CT (mm)	24 hr		72 hr		120 hr		168 hr	
	NGFS	NCUM	NGFS	NCUM	NGFS	NCUM	NGFS	NCUM
1	0.861	0.8569	0.8398	0.8467	0.8322	0.842	0.8291	0.8383
5	0.8572	0.7822	0.712	0.8605	0.7513	0.8524	0.803	0.8577
10	0.8772	0.9383	0.6381	0.7841	0.6128	0.7969	0.6411	0.8213
20	0.7414	0.9163	0.7772	0.7676	0.5772	0.8712	0.6032	0.5137
50	0.921	0.9514	0.9625	0.9833	0.8004	0.9404	0.6564	0.5426

of the interest maps was required in the parameter settings rather than using the same settings given by Davies *et al*. (2004). Hence a brief description of the methodology with the inclusion of these modified parameter settings are given in Appendix - I. Also we are using only one convolution radius (2 grid spaces) with a range of convolution thresholds rather than many convolution radii, as at higher convolution radii the field is getting too smooth and fails to capture any simple objects.

Figs. 6(a&b) shows the simple objects generated by the MODE analysis tool for gridded rainfall analysis, and day-1, day-3, day-5 and day-7 forecasts by NGFS and NCUM at various convolution thresholds, 2 mm, 4 mm, 10 mm, 20 mm and 50 mm valid for 13 October, 2013, the most intense period at landfall. At 2 mm threshold, the objects cover maximum area and are clustered together with the Fuzzy logic over a region covering most of the domain. The observation object cluster contains 3-4 objects of the same colour which are compared against the day-1, day-3 and day-5 forecast clusters of NGFS and NCUM. Most matching clusters in different fields may have different colours and hence a strict colour matching should not be attempted to when comparing clusters between different fields. In all panels one common feature is that unmatched objects are all colored in blue and all other colours are clusters, pairs of which are all matched between the observation and the forecast fields. In general it could be noted that the clusters occupy larger areas in the forecasts compared to the observation and the total interest computed will be the maximum for the lowest threshold. As it goes to higher and higher thresholds, the objects areas and the cluster sizes decrease and the total interest also diminishes very fast. At highest threshold of 50mm, the objects are very less in both number and area coverage and at longer lead times of day-5 and beyond the cluster itself is not formed often due to the very low total interest (< 0.7) between the objects in the same fields and hence the match is not being made.

Median of Maximum Interest (MMI) can be considered as a single objective score to assess the general agreement of all the forecast objects in the entire domain with the observed objects. This is because, MMI accounts for all the attributes of the forecast-observation pairs characteristic of the errors in location, orientation and intensity distribution of the simple objects (Appendix - I). Table 1 lists the Median of Maximum Interest (MMI) values for day-1, day-3, day-5 and day-7 forecasts valid for 13 October, 2013 for thresholds 1, 5, 10, 20 and 50 mm. In general it can be seen that NCUM features better performance with higher number of MMIs while at day-7 NCUM fails to produce any strong system and hence has poor MMI at higher thresholds of 20 mm and above. Though NGFS produced the Tropical Cyclone in day-7 forecasts, it probably failed to produce Total Interests for the simple objects at higher rainfall range so as to exceed the threshold value of 0.7 to make a cluster. This may be due to the higher weights given to the centroid distance and other parameters being taken into consideration for the computation of Total Interest as the location not the intensity is in more error in this case. Thus in general, except at longer lead times beyond day-5, the MMI values are above a useful threshold value of 0.7 and can be considered as a measure of better performance. Also it can be seen that NCUM produced higher values of MMIs compared to NGFS almost at all lead times and on all rainfall thresholds, except on day-7 at higher rainfall ranges, in which case, NGFS scores are predictably higher owing to the better intensity forecasts.

Traditional verification scores applied to the model output rainfall computed by defining matched observed objects to be hits, unmatched observed objects to be misses and unmatched forecast objects to be false alarms, weighted by object area [Figs. 7(a-d)] shows that at lower convolution thresholds of 2 mm and convolution radius of 2 grid sizes, NCUM features slightly higher frequency bias and lower CSI compared to NGFS, when averaged

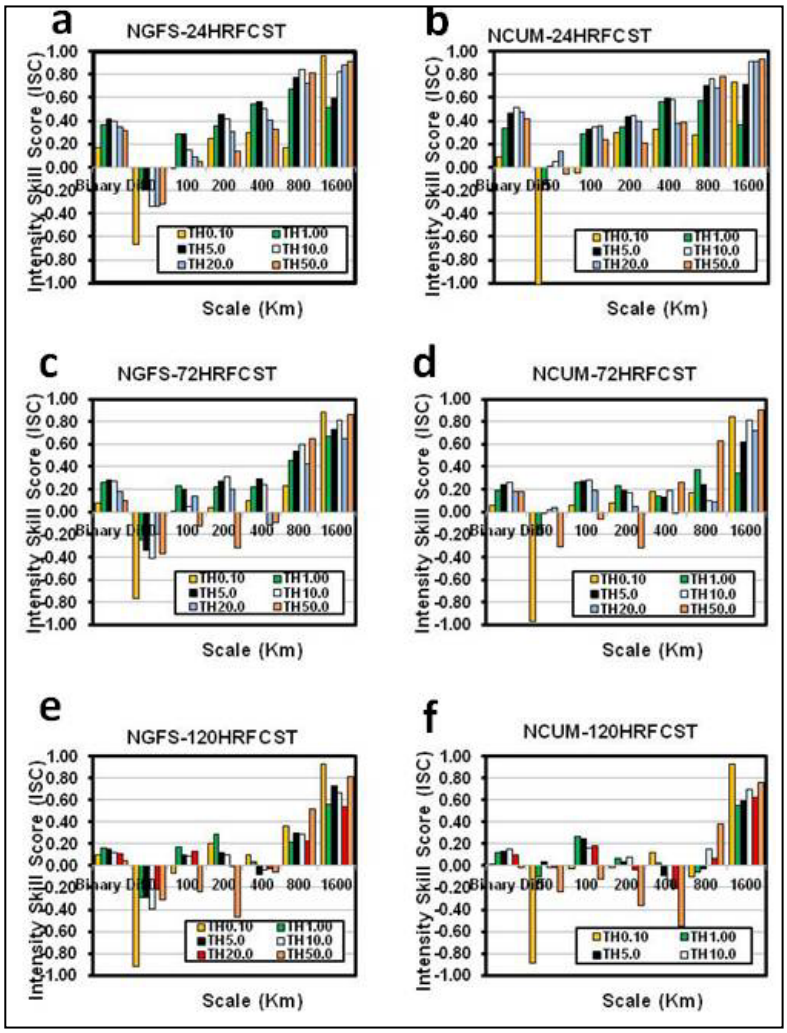


Figs. 8(a-d). The bars denote the time series of the number of matched observed (blue) and forecst (red) objects (top two) and unmatched observed (blue) and forecast (red) objects (bottom two) for 24 hour rainfall forecasts for the period 8-14 October, 2013 by NGFS and NCUM respectively for convolution thresholds (a) 2 mm (top left 2×2 square panels) (b) 4 mm (top right 2×2 square panels) (c) 10 mm (bottom left 2×2 square panels) and (d) 15 mm (bottom right 2×2 square panels). Matched number of observation objects are represented as 'hits' and unmatched number of observation objects as 'misses', while the unmatched number of forecast objects are represented as false alarms

over 7 days of lead time. NGFS is having better number of matched objects, while features lesser matched area except for day-1 and day-7. This can be an indication of more number of forecast objects of small sizes for NGFS and lesser number of large contiguous areas for NCUM as model-specific characteristics. Figs. 8(a-d) gives statistics of time variation of total number of matched and unmatched objects captured during the period of TC Phailin for 24-hour rainfall forecasts for four thresholds, 2, 4, 10 and 15 mm. In general it can be stated that, for the case of number of matched objects, 24-hour rainfall features more number of hits for NGFS for the entire episode. Also the number of misses and false alarms are less for NGFS compared to NCUM.

3.3. *Wavelet analysis*

Wavelet analysis evaluates the forecast skill as a function of the precipitation intensity and the spatial scale of error. Casati *et al*. (2004) states that, the loss of forecast skill of a mesoscale model mainly owes to small spatial scale errors at larger precipitation thresholds. Different scales are associated with different physical processes. For example, small scales are associated with convective showers and mesoscale events and large scales are associated with frontal systems and other large scale synoptic systems. Any weather phenomena can be considered as consisting of all range of scales from micro to the maximum size of the event. The wavelet analysis carried out at a finite number of scales will quantify the performance of the rainfall forecasts for these scales and will give an idea about the scales of maximum and minimum average displacement error at each threshold. Different categorical scores are computed for each particular scale component, like Intensity-scale Skill score (ISS) which is based on the mean squared error (MSE). Thus, this approach enables the user to assess the skill of the model in simulating these scales and hence the associated physical processes.



Figs. 9(a-f). Intensity skill scores of rainfall at different thresholds (bars) of 0.1, 1, 5, 10, 20 and 50 mm, plotted at different scales 50, 100, 200, 400, 800 and 1600 km along with the binary difference averaged for 8-14 October, 2013 for forecast lead times 24, 72 and 120 hours by NGFS and NCUM

Here, as the grid resolution of study is 50 km, the mesoscale features cannot be resolved, the scales of the range of (50-1600 km) are considered. Figs. 9(a-f) shows intensity skill score plotted against the scale in kilometres as bar diagrams for day-1, day-3 and day-5 forecasts by NGFS and NCUM. Different colour bar for each scale denotes different thresholds of 0.1, 1, 5, 10, 20 and 50mm. At day-5, both the models show considerable degradation in the skill to simulate the scales even up to 800 km. NCUM shows poorer skill at 800 km scale compared to NGFS. In general, both the models show better capability to simulate scales of 800 km and 1600 km at all thresholds and NGFS on average, shows relatively better skill in 800 km scale compared to NCUM. This may be partly due to the difference in the resolution of the models, as NGFS runs at comparatively higher resolution. At 50 km scale, both models show the least skill perhaps due to more displacement error. Averaged over the entire episode (not shown here), NGFS shows lower skill at higher thresholds compared to NCUM at 50 km scale.

4. Summary and conclusions

The study evaluates the overall performance of the rainfall forecasts by NGFS and NCUM global models for the case of TC 'Phailin'. There are 7 days of forecasts taken into account during the period of 8-14 October, 2013, from the time when the system is formed over the southeast Bay of Bengal and till the land fall has occurred. The verification is carried out against gridded rainfall analysis to assess the spatial pattern and the performance of the predicted rainfall.

The traditional verification scores are giving mixed results for continuous variables with mean value of the rainfall predicted closer to that observed by NGFS as compared to NCUM, in which case there is always an over prediction up to day-4. NCUM suffers from more bias and poor skill scores during the first four days of forecasts. The categorised rainfall evaluation with a lower threshold value of 1mm shows that the relative frequency of occurrence by NCUM is better matching with the frequency of the observed events (Base Rate) throughout the seven days forecast period. NGFS predicts lesser frequency of occurrence of the event compared to the observed and also less variability. Though NCUM shows less frequency bias in the lower threshold of 1 mm, Accuracy and all other skill scores are comparatively better for NGFS and False Alarm Ratio relatively less. Frequency bias is very high for NGFS compared to NCUM at 100 mm threshold. For FAR and all skill scores, it can be concluded that at lower thresholds of 1 and 5 mm, NGFS performs better than NCUM, while for higher thresholds, NCUM is superior to NGFS.

MODE analysis shows that NCUM is having marginally better scores in terms of total interests. In general, the wavelet analysis for both models show better capability to simulate synoptic scales of order of 800 km and 1600 km. The difference in the resolution of the models may have some impact on the scale of best performance. At 50 km scales both the models show the worst skill and NGFS is worse than NCUM in this respect. The NCUM forecasts are found to be superior to NGFS up to day-5 forecasts though NCUM is unable to predict any system on day-7 forecasts.

It can be noted that MODE analysis is carried out only at one Convolution Radius (CR = 2 mm) and over a number of Convolution Thresholds (CT). Thus the Total Interest indicates an overall match between the smooth rainfall forecasts against the gridded rainfall analysis. As the effective grid resolution is taken as 50 km, which can be considered as quiet coarse, the use of higher values of CR over-smooth the pattern and will not survive the 'Convolution-thresholding' process at higher CT's. So it is most apt for rainfall verification of higher resolution forecasts against higher resolution observations. The entire study is conducted at 50 km resolution where as the models it selves are having higher resolution. We are actually limited by the resolution of the observations. The experiment however proves that we are ready with the verification tool whenever the rainfall forecasts by mesoscale models are made available along with comparable or high resolution observations.

The current study demonstrates the new ways of model performance evaluation and more comprehensive analysis techniques. These types of standard scores are useful in assessing the overall quality of the forecasts for the kind of extreme weather events which last for at least about a week. However, as the sample size is relatively small, the scores cannot be generalised or a definite statement of the performance of a particular model cannot be arrived at. The object-based scores are useful in day-to-day assessment of the agreement between forecast and observed rainfall patterns and in-depth analysis of the performance in the simulation of various physical processes. A long period forecast experiment or a large sample size can be used to assess the strength and weakness of the models. So an ensemble of scores of large set of extreme weather events or accumulation of scores through a longer period like a month or season can help in assessing the performance of the models in different scenarios or convective environments. One aspect of the verification which is out of scope of the current study is the issue of 'if the errors are within the acceptable limits or not'. As this is a one-off case study of a tropical cyclone, the sample size is too inadequate to decide whether the forecast biases are within acceptable limits or whether the forecast biases reflect the deviation occurring within the verification sample data. For that, we need a large set of cases of tropical cyclone predictions by the same formulations of models over the region to generate the climatology and to evaluate separately the forecast and observation biases against the climatology.

Acknowledgements

The authors are grateful to Director, NCMRWF for the support and encouragement for the current study. The editorial comments and the two anonymous reviewers helped in improving the quality of the paper.

References

Ahijevych, D., Gilleland, E., Brown, B. and Ebert, E., 2009, "Application of spatial forecast verification methods to gridded precipitation forecasts", *Wea. Forecasting*, **24**, 1485-1497.

Brown, B. G., Bullock, R., Gotway, J. H., Ahijevych, D., Davis, C., Gilleland, E. and Holland, L., 2007, "Application of the MODEL object-based verification tool for the evaluation of model precipitation fields", AMS 22nd Conference on Weather Analysis and Forecasting and 18th Conference on Numerical Weather Prediction, 25-29 June, Park City, Utah, American Meteorological Society (Boston) (Available at http://ams/confex.com/ams/pdfpapers/124856.pdf).

Casati, B., 2010, "New developments of the intensity-scale technique within the spatial verification methods intercomparison project", *Wea. Forecasting*, **25**, 113-143.

Casati, B., Ross, G. and Stephenson, D. B., 2004, "A new intensity-scale approach for the verification of spatial precipitation forecasts", *Meteor. Appl.*, **11**, 141-154.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown B. G. and Mason, S, 2008, "Forecast verification: Current status and future directions", *Meteor. Appl*., **15**, 3-18.

Davis, C. A., Brown, B. G. and Bullock, R. G., 2006, "Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas", *Mon. Wea. Rev*., **134**, 1772-1784.

Davis, C. A., Brown, B. G., Bullock, R. G. and Gotway, J. H., 2009, "The method for object-based diagnostic evaluation (MODE) applied to WRF forecasts from the 2005 NSSL/SPC Spring Program", *Wea. Forecasting*, **24**, 1252-1267.

Ebert, E. E., 2008, "Fuzzy verification of high resolution gridded forecasts: A review and proposed framework", *Meteor. Appl*., **15**, 51-64.

Ebert, E. E., 2009, "Neighbourhood verification: A strategy for rewarding close forecasts", *Wea. Forecasting*, **24**, 1498-1510.

Ebert, E. E. and Gallus, W. A. Jr., 2009, "Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification", *Wea. Forecasting*, **24**, 1401-1415.

Ebert, E. E. and McBride, J. L., 2000, "Verification of precipitation in weather systems: Determination of systematic errors", *J. Hydrol.*, **239**, 179-202.

Gallus, W. A. Jr., 2010, "Application of object-oriented verification techniques to ensemble precipitation forecasts", *Wea. Forecasting*, **25**, 144-158.

Gilleland, E., 2013, "Testing competing precipitation forecasts accurately and efficiently: The Spatial prediction comparison test", *Mon. Wea. Rev*., **141**, 1, 340-355.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B. and Ebert, E. E., 2009, "Intercomparison of spatial forecast verification methods", *Wea. Forecasting*, **24**, 1416-1430.

Gilleland, E., Ahijevych, D. A., Brown, B. G. and Ebert, E. E., 2010a, "Verifying Forecasts Spatially", *Bull. Amer. Meteor. Soc*., **91**, 1365-1373.

Gilleland, E. D., Lindström, J. and Lindgren, F., 2010b, "Analyzing the image warp forecast verification method on precipitation fields from the ICP", *Wea. Forecasting*, **25**, 1249-1262.

Jolliffe, I. T and Stephenson, D. B., 2003, "Forecast verification: A practitioners' guide in atmospheric science", Wiley and Sons Ltd., p240.

Murphy, A. H. and Winkler, R. L., 1987, "A general framework for forecast verification", *Mon. Wea. Rev*., **115**, 1330-1338.

Mitra, A. K., Das Gupta, M., Singh, S. V. and Krishnamurti, T. N., 2003, "Daily rainfall for Indian monsoon region from merged satellite and raingauge values: Large-scale analysis from real-time data", *J. Hydrometeorol*., **4**, 5, 769-781.

Mitra, A. K., Bohra, A. K., Rajeevan, M. N. and Krishnamurti, T. N., 2009, "Daily Indian precipitation analyses formed from a merge of rain-gauge with TRMM TMPA satellite derived rainfall estimates", *J. Meteor. Soc. Japan*, **87A**, 265-279.

Mittermaier, M. and Roberts, N., 2010, "Intercomparison of spatial forecast verification methods: identifying skilful spatial scales using the fractions skill score", *Wea. Forecasting*, **25**, 343-354.

Prasad, V. S., Mohandas, S., Das Gupta, M., Rajagopal, E. N. and Datta, S. K., 2011, "Implementation of upgraded Global Forecasting Systems (T382L64 and T574L64) at NCMRWF", NCMRWF Technical Report No. NCMR/TR/5/2011, National Centre for Medium Range Weather Forecasting, Min. of Earth Sciences, A-50, Sector-62, Noida, May 2011, p72.

Prasad, V. S., Mohandas, S., Dutta, S. K., Das Gupta, M., Iyengar, G. R., Rajagopal, E. N. and Basu, S., 2013, "Improvements in Medium Range Weather Forecasting System of India", *Journal of Earth System Science*, **123**, 2, March, 2013, 347-258.

Rajagopal, E. N., Iyengar, G, R., George, J. P., Das Gupta, M., Mohandas, S., Siddharth, R., Gupta, A., Chourasia, M., Prasad, V. S., Aditi, Sharma, K. and Ashish, A., 2012, "Implementation of Unified Model based Analysis-Forecast System at NCMRWF", NCMRWF Technical Report No. NMRF/TR/2/ 2012, NCMRWF, Ministry of Earth Sciences, A-50, Sector-62, Noida, UP-201309, May 2014, p45.

Stansky, H. R., Wilson, L. J. and Burrows, W. R, 1989, "Survey of common verification methods in meteorology", World Weather Watch Tech. Rept. No. 8, WMO/TD No. 358, WMO Geneva, p114.

Wilks, D., 2006, "Statistical methods in the atmospheric sciences", Elsevier, San Diego, Second Edition, International Geophysics Series, **91**, p630.

APPENDIX - I

Method for Object-based Diagnostics Evaluation (MODE)

This is a displacement technique of spatial verification methods which provides information which is not otherwise possible to obtain using traditional grid-point based verification methods (Davis *et al*., 2004). It objectively identifies simple objects in rainfall fields at different thresholds, which would mimic what humans call as "regions of interest". This process is a multistep one which is called the 'convolution-thresholding' technique. It basically involves application of a simple circular filter which in terms is a function of convolution radius (CR). Once the filter is applied, the convolved field is thresholded using a convolution threshold (CT) to generate a mask field. These simple objects are the connected regions of "1" in the mask field. Finally, the actual data is restored inside the mask regions of object interiors to obtain the object field. Thus these objects are a function of CR and CT.

Once simple objects are generated in the rainfall field, various object attributes are computed and compared to merge the objects in the same field and match the objects between the two different fields, say forecast and observation. The summary statistics can be computed based on the single object statistics as well as statistics of the pairs of objects. As an example, Area is an attribute which is simply the count of the number of grid squares an area of object occupies. Axis angle gives the tilt of the object as curvature implies the curviness. Aspect ratio is the ratio of the width and length of the rectangle which is aligned so as to have the same axis angle as the object and for which the length and width are chosen so as to just enclose the object. Complexity is defined by comparing the area of an object to the area of its convex hull. Similarly pair attributes are defined such as centroid distance, angle difference, union area, intersection area and symmetric difference.

Matching and merging of the objects are achieved by various techniques and "Fuzzy engine" logic is applied for the current study. This involves assigning "interest maps", "confidence maps" and weights for the attributes (α) which are taken in to consideration. Interest maps [I(α)] range from zero to one and are applied to each attributes to convert it into interest values. 1 indicates high interest and 0 indicates no interest and there will be some attributes featuring intermediate interests. Confidence maps [C(α)] also range from zero to one, but is a function of the entire set of attributes to indicate the relative confidence of one field in terms of other fields thus is dependent of other parameters also. By default if the attribute is independent of any other attribute, the confidence map is defined as 1. The scalar weights (ω) are to be assigned to each attribute giving preference to which attribute the user assign maximum weightage. Finally a single number called total interest [T(α)] is computed using all the three maps by a formula as given below.

$$T(\alpha)=\frac{\sum_i w_i C_i(\alpha) I_i(\alpha)}{\sum_i w_i C_i(\alpha)}$$

The total interest is then thresholded and the pairs of objects that are having total interest more than the threshold are merged if they are in the same field and matched if they are in the different fields. MODE outputs the statistics of single as well as cluster of objects. The scores can be summarised as Median of Maximum Interest (MMI), which is an example of a useful single measure of the general agreement between forecast and observation for the entire domain (See Davis *et al.*, 2009). Median is taken instead of Mean to avoid the effect of outliers. The details of the MODE configurations and the definitions of various interest maps and confidence maps are as given below.

General

Grid_res = 50 km

Convolution thresholds = 1, 2, 5, 10, 20 and 50 mm

Convolution radius = 2 (grid spaces)

Forecast_merge_flag = 2 (Fuzzy Engine merging method)

max_centroid_dist = 200

total_interest_thresh = 0.7

Interest functions and piecewise linear functions

Centroid Distance	Interest
0.0	1.0
100.0/grid_res	1.0
1000.0/grid_res	0.0

Boundary Distance	Interest
0.0	1.0
500.0/grid_res	1.0
2000.0/grid_res	0.0

Convex Hull Distance	Interest
0.0	1.0
500.0/grid_res	1.0
2000.0/grid_res	0.0

Angle Difference	Interest
0.0	1.0
30.0	1.0
90.0	0.0

Area Ratio	Interest
0.0	0.0
1.0	1.0

Intersecting area ratio	Interest
0.00	0.00
0.10	0.50
0.25	1.00
1.00	1.00

Confidence functions

aspect_ratio_conf(t) = ((t - 1)**2/(t**2 + 1))**0.3;

area_ratio_conf(t) = t

Weights

centroid_dist_weight = 2.0

boundary_dist_weight = 4.0

convex_hull_dist_weight = 0.0

angle_diff_weight = 1.0

area_ratio_weight = 1.0

int_area_ratio_weight = 2.0

complexity_ratio_weight = 0.0

intensity_ratio_weight = 0.0