Earth System Science Organization
Ministry of Earth Sciences

Report by the Working Group on
Forecast verification methods and data sets

07 October 2013

Background

ESSO-MoES constituted a working group to discuss and finalize the forecast verification methods and various verification data sets to be used for verification by the ESSO institutions for medium range, extended range and seasonal forecasts. The working group consists of the following scientists:

Dr. M Rajeevan, Scientist-G, ESSO-MoES, Chairman
Dr A.K.Sahai, Scientist-F, ESSO-IITM
Dr Gopal Iyengar, Scientist-F, ESSO-NCMRWF
Dr D.S.Pai, Scientist E, ESSO-IMD
Dr Pulak Guhathakurta, Scientist-E, ESSO-IMD
Dr A.K.Mitra, Scientist-D, ESSO-IMD

The terms and reference of the working group are:

- To review and recommend region-specific verification methods (metrics) for verification of quantitative forecasts (medium range to seasonal forecasts)
- To review and recommend the observed data sets for verification of forecasts (medium range, extended range and seasonal forecasts).
- To recommend research issues on verification methods and development of data sets for verifications
- To recommend common criteria for defining monsoon onset over Kerala using NWP/Climate forecasts.

Part- I: Extended range to Seasonal Forecasts

In the first part, we recommend verification methods for extended range to seasonal forecasts.

1. Verification methods

In this section, we discuss the methods recommended for verification of forecasts (extended range and seasonal) being issued by the ESSO institutions.

1.1) **Seasonal forecasts**

At present, ESSO-IMD issues long range (seasonal) forecasts of monsoon rainfall as quantitative as well as probabilistic in five categories. The operational forecast of ESSO-IMD is based on statistical models and the table below shows the forecasts being issued.

Table-1

Parameter	Region	Quantitative/Probabilistic
Seasonal Rainfall (JJAS)	All India Four geographical regions (NW India, NE India, Central India and South Peninusla)	Quantitative and probabilistic (5-category for all India and 3-category forecast for four geographical regions of India.
Monthly rainfall July, August, September	All India	Quantitative and probabilistic (3-category)
Monsoon onset	Kerala	Quantitative
NE monsoon Seasonal rainfall (OND)	NE monsoon region	Quantitative and probabilistic
NW India winter precipitation (Jan-March)	NW India	Quantitative and probabilistic

ESSO-IITM also prepares experimental seasonal forecasts for all India using the monsoon mission model, CFS V2.0, both in quantitative and probabilistic way.

Deterministic forecasts provide a single expected value for the forecast variable. Probabilistic forecast provides probabilities of occurrences or non-occurrences of an event or a set of fully inclusive events. The events can be also classified into categories (above/below normal or above/near/below normal for example). Although equi-probable categories are preferred for consistency, other classifications can be used in a similar fashion. The most commonly used equal groupings are terciles, which are three ranges, or intervals, of values of a variable (e.g. rainfall or temperature) that are defined to describe the lower, middle, and upper *thirds* of the climatologically expected distribution of values. ESSO-IMD uses 5-categories of rainfall that are not equal groupings to describe the all India season (June-September) rainfall.

1.1.1 Parameters to be verified

At present the extended range and seasonal forecasts are being prepared/issued for the following two parameters. We recommend that these two parameters may be rigorously verified against observations.

a) Rainfall amount/anomaly

b) Sea surface temperature (SST) anomaly

1.1.2 Verification Strategy

The operational rainfall forecasts of ESSO-IMD are based on statistical models and are one dimensional (time series of area averaged rainfall). On the other hand, the experimental forecasts being generated by ESSO-IITM are based on dynamical models and the rainfall and SST forecasts are two-dimensional. Therefore, the verification strategy for the forecasts of ESSO-IMD and ESSO-IITM needs to be different.

For verification of dynamical model based hindcasts/forecasts, the following strategy is recommended. The strategy is to verify area averaged and large scale performance.

(a) Seasonal/monthly precipitation anomaly averaged over

   a. a larger domain, the south Asian monsoon region (0 – 40^{0}N, 50^{0}E-120^{0}E)

   b. Indian main land region including north-east India (grid points covering all the Indian land mass), which is the same for the ESSO-IMD’s seasonal forecasts and

   c. four large geographical regions of the country as defined by ESSO-IMD

(b) Sea surface temperature (SST) anomaly averaged over the Niño3.4 and the Indian Ocean Dipole Index. For the JJAS period, Indian Ocean Dipole index is not a proper measure. Therefore, SST anomalies from the eastern box may be considered for verification during JJAS period.

(c) Large scale aggregated overall measures of rainfall forecast performance at grid points over

a. South Asian monsoon region (0-40N, 50E-120E)
b. Indian main land region
c. four large geographical regions of the country as defined by IMD

This item is different from the item (a), as this verification is done on each grid point and then aggregating, while the item (a) is done on an area averaged quantity.

(d) Large scale aggregated overall measures of SST anomaly forecast performance at grid points over

a. Tropics (30^0S-30^0N, 0-360^0E)
b. lndian Ocean (25^0S -25^0N, 50^0E-120^0E), and
c. Tropical Pacific (25^0S-25^0N, 120^0E-90^0W)

(e) Large scale monsoon index such as Monsoon Hadley Index (meridional wind difference between 850 hPa and 200 hPa over a broader region 10^0-30^0N, 70^0E-110^0E).

(f) Tropospheric temperature gradient index (vertically integrated temperature difference between 200 and 600 hPa between a north box (10-35N, 30-100E) and south box (15S-10N, 30-100E)).

Large scale verification statistics (c and d above) are required to evaluate the overall skill of the models and ultimately for assessing its improvements. These are bulk numbers calculated by aggregating verification over all grid points within large regions; they will not necessarily reflect skill for any sub-region.

For the operational forecasts of ESSO-IMD, the verification will be done for the forecast issued as per Table-1 given above.

1.1.3. **Verification scores**

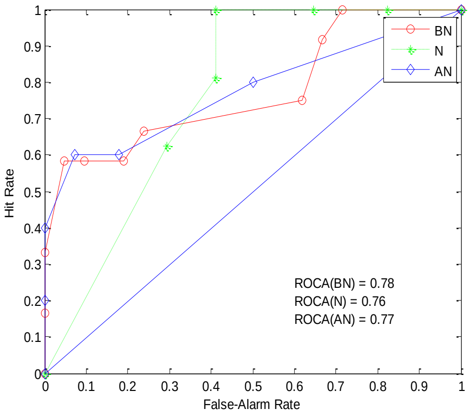
After discussing many verification scores, the committee recommends the following minimum verification scores for verification of extended range and seasonal forecasts. These scores pertain to both deterministic and probabilistic forecasts.

1) Anomaly correlation coefficient (ACC) (for deterministic forecasts).
2) Mean Square Skill Score (MSSS) (for deterministic forecasts)
3) Relative Operating Characteristics (ROC) (both for deterministic and probabilistic forecasts)
4) Reliability diagrams (probabilistic forecasts)
5) Heidke Skill Score (HSS) (Categorical forecasts)

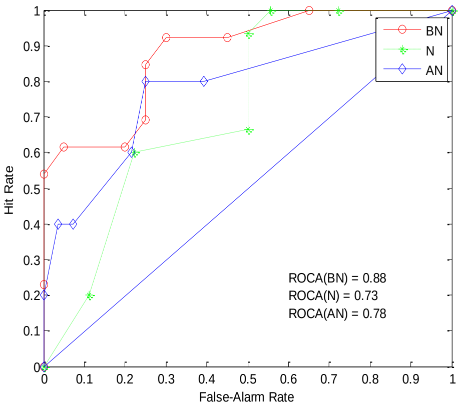
Anomaly correlation coefficient may be used only as a preliminary verification. More rigorous verifications may be done with other 4 scores mentioned above. MSSS is applicable to deterministic forecasts only, while ROC is applicable to both deterministic and probabilistic forecasts. MSSS is applicable to non-categorical forecasts (forecasts of continuous variables), while ROC is applicable to categorical forecasts either deterministic or probabilistic in nature. The verification methodology using ROC is derived from signal detection theory. This methodology is intended to provide information on the characteristics of systems upon which management decisions can be taken. In the case of weather/climate forecasts, the decision might relate to the most appropriate manner in which to use a forecast system for a given purpose. ROC is applicable to both deterministic and probabilistic categorical forecasts and is useful in contrasting characteristics of deterministic and probabilistic systems. The details of the verification scores are given in Annexure-1.

1.1.4 Verification scores for the ESSO- IMD's operational model

ESSO-IMD has been issuing seasonal forecasts using statistical models. The present models based on ensemble statistical methods have been introduced by ESSO-IMD in 2007. Since the real time forecast period is short (only from 2007), we have calculated suitable verifications scores (MSSS and ROC) for the operational models from 1981 onwards. The scores were calculated for both deterministic as well as probabilistic forecasts. The results of verification of ESSO-IMD's operational ensemble prediction system are given below:



ROC plot for April Model



ROC plot for June Model

Relative operating characteristics curves for 3 category (normal, above normal and below normal) forecasts are given below.

Verification Score (1981-2012)	April Model	June Model
Means Squared Skill Score (MSSS)	0.57	0.63
Roc area (Above Normal)	0.77	0.78
Roc area (Normal)	0.76	0.73
Roc area (Below Normal)	0.78	0.88

In general, ROC curve area for below normal category is higher than the other two categories, especially for the June model.

1.2 Extended Range Forecasts

In India, systematic research work on extended range forecasts has started very recently. ESSO-IITM prepares experimental extended range forecasts using the CFS V2.0 model and GFS model with bias corrected SSTs predicted by the CFS model. For last couple of years, IITM is preparing experimental extended range forecasts up to 4 pentads for rainfall averaged over the monsoon core region and four homogenous regions. The forecasts are generated for each pentad for the next 4 pentads and these forecasts are repeated on every 5 days. The forecasts are issued in terms of rainfall anomaly averaged over the monsoon core region as well as probabilities of rainfall in excess, normal and deficient categories (terciles). In addition, the phase evolution of MISO indices is also prepared and issued on experimental basis. From 2013 onwards, using the CFS V 2.0/GFS models, forecast for monsoon onset over Kerala and progress of the monsoon is also being prepared. ESSO-IMD Delhi prepares extended range forecasts as a multi-model ensemble (MME) using three different coupled model forecasts.

1.2.1 Parameters to be verified

At present the extended range forecasts are being prepared for rainfall anomalies and the monsoon onset dates. We recommend that forecasts of rainfall and monsoon onset dates may be verified against observations.

1.2.2. Verification scores

The working group recommends the following scores for verification of extended and long range forecasts. These scores pertain to both deterministic and probabilistic forecasts.

1. Anomaly correlation coefficient (for deterministic forecasts and for preliminary verification)

2. Mean Square Skill Score (MSSS) (for deterministic forecasts)
3. Relative Operating Characteristics (ROC) (both for deterministic and probabilistic forecasts)
4. Reliability diagrams (probabilistic forecasts)
5. Heidke Skill Score (HSS) (Categorical forecasts)

2) VERIFICATION DATA SETS

For verification of forecasts of rainfall and SST, the working group discussed pros and cons of different data sets. The analysis shows that there is a big difference of 20 cm between GPCP rainfall and IMD’s gridded rainfall over the Indian region. As GPCP data is of coarser resolution (2.5 degree) and limited station data from India was used for the analysis, we do not recommend to use the GPCP data for verification of rainfall over the Indian land mass. We recommend IMD's gridded rainfall data for verification of rainfall over the Indian land mass. Since IMD's gridded data are only for the Indian land mass, we recommend use of GPCP data for spatial scale larger than Indian land mass including neighbouring oceans.

The working group recommends the following data sets for verification of rainfall and SST.

2.1 Data Sets for verification of seasonal rainfall forecasts

a) IMD 1x1 Gridded Daily Rainfall Data

For verification of model forecasts, IMD 1x1 gridded daily rainfall data (Rajeevan et al. 2006) prepared based on rain-guage observation will be used. This is applicable to verification of both rainfall averaged over the land area as well as at grid points over the Indian main land. The geographical domain of this data set is 6.5N to 37.5N, 66.5E to 101.5E. The data is available for the period 1951 to 2012. There were some problems in this data set from 2004 onwards due to inadequate network. However, this problem is now solved and the data from 2004 onwards have been updated by including more number of stations for the analysis. It may be mentioned that this data set was prepared using a fixed network of 2140 stations. Therefore, unnecessary bias due to changes in the rainfall network does not arise.

b) NOAA Global Precipitation Climatology Project (GPCP) 2.5X 2.5 Gridded Monthly Precipitation Data:

For verification over the Indian monsoon region (including land and oceans), it is recommended to make use of GPCP monthly precipitation dataset. The GPCP data is available for the period 1979-2012. This data set consists of monthly means of precipitation derived from satellite and gauge measurements derived into 2.5°x2.5° global grids. The spatial coverage of the data is 88.75N - 88.75S, 1.25E - 358.75E. Higher resolution (1 X 1 degree) data set is however available from 1998 onwards.

2.2 Data Sets for verification of extended range rainfall forecasts

For verification of extended range forecasts for rainfall over the monsoon zone and the four homogenous regions of India, the IMD's gridded data of 1X 1 degree data will be used, thus keeping the consistency with the data set being used for seasonal forecast verification. However, to examine the large scale aspects of rainfall over a larger domain (including the oceans), the merged daily gridded data produced by Dr A.K. Mitra (1998 onwards) may be used. This data set uses IMD's rain-gauge data over the Indian land mass and satellite rainfall over the oceans. The data set since 1998 is available at 1 degree resolution. IMD Pune now prepares the merged data set on real time basis using the same algorithm, but with a higher resolution of 0.5 degree. This real-time rainfall product being prepared by IMD Pune can be used for verification of real-time forecasts.

2. 3 Data Sets for verification of seasonal SST forecasts

a) NOAA Monthly Extended Reconstructed Sea Surface Temperature (ERSST.v3b):

For verification of grid point SSTs and SSTs averaged over specified regions(see section 3.1), it is recommended to use the most recent version of the Extended Reconstructed Sea Surface Temperature (ERSST) analysis data of version v3b (Smith et al. 2008). The monthly SST data is available at 2.0 degree latitude x 2.0 degree longitude global grid for the period 1854 to the present. The data were constructed using the most recently available International Comprehensive Ocean-Atmosphere Data Set (ICOADS) SST data and improved statistical methods that allow stable reconstruction using sparse data. The spatial domain of the data is 88.0N - 88.0S, 0.0E - 358.0E.

3. Research on verification methods and data sets

The working group recommends the following aspects for research and development by the ESSO institutions on verification methods and data sets.

1. Monsoon mission may follow these recommendations of verification methods and data sets. The verification of hindcasts (extended range and seasonal) of the monsoon mission models (GFS, CFS and UKMO Unified model) may be carried out as per these recommendations. Even though, under the monsoon mission, more comprehensive verification/assessment of the models is expected, verification as per these recommendations (with minimum verification scores) may be considered as mandatory.
2. Eventhough the working group recommends the verification for extended range and seasonal forecasts of rainfall and temperature parameters only, it is important to diagnose the NWP/Climate model fidelity in total in terms of its dynamics and physics. More diagnostic studies to examine the systematic biases and errors associated with the dynamics (circulation features) and physics (thermodynamics, radiation and clouds etc) should be carried out.
3. The ESSO institutions are also encouraged to do verifications at higher level, i.e., verification at grid points by grid points and for 36 meteorological sub-divisions. At the grid point level, verification is performed with observed and predicted values at each grid point, which is recommended for a regionalised assessment of the skill of the model. The verification of rainfall anomaly is to be done at the same latitude/longitude grid as that of the observed data sets used for verification.
4. For verification data sets, the working group recommends the development of the following data sets:
   a. Merged daily gridded rainfall data set at 0.5 degree resolution from 1 May to 30 October using a fixed network of IMD rain-gauges and satellite derived data (TRMM and INSAT) from 1998 onwards.
   b. Merged monthly gridded rainfall data set at (preferably at 1.0 degree resolution) for all 12 months over the South Asian monsoon region from 1979 onwards. This can be developed using rain-gauge data from India, APROHDITE data (Xie et al. ---) over the land region outside India and

satellite (IR and Microwave) over the oceans. This data set can replace the GPCP data being used for verification of large scale rainfall patterns.

c. Merged hourly rainfall data over the Indian region using IMD rain-gauge data (SRRG, AWS and ARG stations), DWR data and satellite data (TRMM 3G68) from 2006 onwards.

4. Summary of the recommendations

The working group emphasized the need for a suitable verification for the extended range and seasonal forecasts being generated by the ESSO institutions. However, it was decided unanimously to keep minimum verification scores to assess the quality of extended range and seasonal forecasts. We also emphasize the need of reliable data sets for verification purposes to ensure that the models are not penalized due to inaccurate verification data sets.

We summarize the recommendations of verification scores as given below. The following table gives a summary of parameters, validation regions and verification scores.

Parameters	Verification regions	Deterministic forecasts	Probabilistic forecasts
Seasonal Forecasts			
Precipitation amount/ anomaly	(i) Indian Monsoon region (0–40N, 50E-120E) (ii) Indian main land region (iii)Four broad geographical regions of the country iv) Large scale Monsoon Hadley Circulation Index v) Tropospheric temperature gradient	Anomaly correlation coefficient and MSSS for the parameter averaged over the verification regions and aggregated over all grid points within the verification regions.	ROC curves ROC areas Reliability diagrams Heidke Skill Score
Sea Surface Temperature (SST) anomaly	Niño3.4 and IOD regions.	MSSS parameter averaged over the verification regions	ROC curves ROC areas Reliability diagrams Heidke Skill Score
	(i) Tropical lndian Ocean, and (ii) Tropical Pacific	MSSS aggregated over all grid points within the verification regions.	

Extended range forecasts			
Precipitation amount/ anomaly	(i) Monsoon Core zone (ii)Four broad geographical regions of the country iii) Monsoon onset dates over Kerala and interior parts of India	Anomaly correlation coefficient and MSSS for the parameter averaged over the verification regions.	ROC curves ROC areas Reliability diagrams Heidke Skill Score

Annexure-1

Technical Details of the Verification Scores

1. MSSS for non-categorical deterministic forecasts

Let x_{ij} and $f_{ij}(i=1,\dots,n)$ denote time series of observations and continuous deterministic forecasts respectively for a grid point or station j over the period of verification (POV). Then, their averages for the POV, $\bar{x}_j$ and $\bar{f}_j$ and their sample variances s_{xj}^2 and s_{fj}^2 are given by:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ \bar{f}_j = \frac{1}{n}\sum_{i=1}^{n} f_{ij}$$

$$s_{xj}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2,\ s_{fj}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(f_{ij}-\bar{f}_j\right)^2$$

The mean squared error of the forecasts is:

$$MSE_j = \frac{1}{n}\sum_{i=1}^{n}\left(f_{ij}-x_{ij}\right)^2$$

For the case of cross-validated, POV climatology forecasts where forecast/observation pairs are reasonably temporally independent of each other (so that only one year at a time is withheld), the mean squared error of 'climatology' forecasts (Murphy, 1988) is:

$$MSE_{cj} = \left(\frac{n}{n-1}\right)^2 s_{xj}^2$$

The *Mean Squared Skill Score* (MSSS) for j is defined as one minus the ratio of the squared error of the forecasts to the squared error for forecasts of 'climatology':

$$MSSS_j = 1 - \frac{MSE_j}{MSE_{cj}}$$

For the three domains described in Sec. 3.1.1 it is recommended that an overall MSSS be provided. This is computed as:

$$MSSS = 1 - \frac{\sum_j w_j MSE_j}{\sum_j w_j MSE_{cj}}$$

where w_j is unity for verifications at stations and is equal to $\cos(\theta_j)$, where θ_j is the latitude at grid point j on latitude-longitude grids.

For either $MSSS_j$ or MSSS a corresponding *Root Mean Squared Skill Score* (RMSSS) can be obtained easily from

$$RMSSS = 1 - (1 - MSSS)^{1/2}$$

MSSSj for forecasts fully cross-validated (with one year at a time withheld) can be expanded (Murphy, 1988) as

$$MSSS_j = \left\{ 2\frac{s_{fj}}{s_{xj}} r_{fxj} - \left(\frac{s_{fj}}{s_{xj}}\right)^2 - \left(\frac{\left[\bar{f}_j - \bar{x}_j\right]}{s_{xj}}\right)^2 + \frac{2n-1}{(n-1)^2} \right\} \Big/ \left\{ 1 + \frac{2n-1}{(n-1)^2} \right\}$$

wherer$_{fxj}$ is the product moment correlation of the forecasts and observations at point or station j.

$$r_{fxj} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(f_{ij} - \bar{f}_j\right)\left(x_{ij} - \bar{x}_j\right)}{s_{fj} s_{xj}}$$

The first three terms of the decomposition of $MSSS_j$ are related to phase errors (through the correlation), amplitude errors (through the ratio of the forecast to observed variances) and overall bias error, respectively, of the forecasts. These terms provide the opportunity for those wishing to use the forecasts for input into regional and local forecasts to adjust or weight the forecasts as they deem appropriate. The last term takes into account the fact that the ‘climatology’ forecasts are cross-validated as well.

Note that for forecasts with the same amplitude as that of observations (second term unity) and no overall bias (third term zero), $MSSS_j$ will not exceed zero (i.e. the forecasts squared error will not be less than for ‘climatology’) unless r_{fxj} exceeds approximately 0.5.

The SVSLRF requires grid-point values of the correlation, the ratio of the square roots of the variances, and the overall bias i.e.

$$r_{fxj}, \frac{s_{fj}}{s_{xj}}, \left[\bar{f}_j - \bar{x}_j\right].$$

In addition it is recommended that grid-point (j) values of the following quantities are provided:

$$n, \bar{f}_j, \bar{x}_j, s_{fj}, s_{xj}, r_{fxj}, MSE_j, MSE_{cj}, MSSS_j$$

2. Contingency tables and scores for categorical deterministic forecasts

For the categorical deterministic forecasts, full contingency tables are necessary as they constitute the most informative way to evaluate the performance of the forecasts. These contingency tables then form the basis for several skill scores that are useful for comparisons between different deterministic categorical forecast sets (Gerrity, 1992) and between deterministic and probabilistic categorical forecast sets (Hanssen and Kuipers, 1965) respectively.

The contingency tables should be provided for every combination of parameter, lead time, target period (month, season etc.) and ENSO stratification (when appropriate) at every verification point for both the forecasts. If x_i and f_i now denote an observation and corresponding forecast of category i (i = 1,…,3), let n_{ij} be the count of those instances with forecast category i and observed category j. The full contingency table is defined as the nine n_{ij}. Graphically the nine cell counts are usually arranged with the forecasts defining the table rows and the observations the table columns:

Table 1: General three by three contingency table.

Observations					
		Below Normal	Near Normal	Above Normal	
Forecasts	Below Normal	n_{11}	n_{12}	n_{13}	$n_{1\bullet}$
	Near Normal	n_{21}	n_{22}	n_{23}	$n_{2\bullet}$
	Above Normal	n_{31}	n_{32}	n_{33}	$n_{3\bullet}$
		$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	T

In the table (Table-1) above, $n_{i\bullet}$ and $n_{\bullet i}$ represents the sum of the rows and columns respectively; T is the total number of cases. Generally about at least 90 forecast/observation pairs are required to properly estimate a three by three contingency table. Thus it is recommended that the provided tables be aggregated over windows of target periods, like several adjacent months or overlapping three-month periods, or over verification points. In the case of the latter the weights $W_i$should be used in summing n_{ij} over different pointsi. W_i is defined as:

$W_i = 1$when verification is done at stations or at single grid points within a limited geographical region.

$W_i = \cos(\theta_i)$at grid point i, when verification is done on a grid.

$\theta_i =$ the latitude at grid point i.

On a latitude-longitude grid the minimally acceptable sample is easily attained even with a record as short as n = 10 by aggregating over all grid points with a 10 degree box. Or alternatively in this case, an adequate sample can be achieved by aggregation over three adjacent months or overlapping three-month periods and within a 5 degree box. Regardless, scores derived from any contingency table should be accompanied by error bars, confidence intervals or level of significance.

The *relative sample frequencies*p$_{ij}$ are defined as the ratios of the cell counts to the total number of forecast/observation pairs N (n is reserved to denote the length of the POV):

$$p_{ij} = {}^{n_{ij}}\!/_{N}$$

The sample probability distributions of forecasts and observations respectively then become

$$p(f_i) = \sum_{j=1}^{3} p_{ij} = \hat{p}_i \, ; i = 1, ..., 3$$

$$p(x_i) = \sum_{j=1}^{3} p_{ji} = p_i \, ; i = 1, ..., 3$$

A recommended skill score for the three by three table, which has many desirable properties and is easy to compute is the *Gerrity Skill Score*, GSS. The definition of the score uses a scoring matrix s$_{ij}$ (i = 1,…,3), which is a tabulation of the reward or penalty every forecast/observation outcome (represented by the contingency table) will be accorded:

$$GSS = \sum_{i=1}^{3} \sum_{j=1}^{3} p_{ij} s_{ij}$$

The scoring matrix is given by

$$s_{ii} = \frac{1}{2} \left(\sum_{r=1}^{i-1} a_r^{-1} + \sum_{r=i}^{2} a_r \right)$$

$$s_{ij} = \frac{1}{2} \left[\sum_{r=1}^{i-1} a_r^{-1} - (j-i) + \sum_{r=j}^{2} a_r \right]; 1 \le i < 3, i < j \le 3$$

where

$$a_i = \frac{1 - \sum_{r=1}^{i} p_r}{\sum_{r=1}^{i} p_r}$$

Note that GSS is computed using the sample probabilities, not those on which the original categorisations were based (i.e. 0.33, 0.33, 0.33).

The GSS can be alternatively computed by the numerical average of two of the three possible two-category, unscaled Hanssen and Kuipers scores (introduced below) that can be computed from the three by three table. The two are computed from the two two-category contingency tables formed by combining categories on either side of the partitions between consecutive categories: (1) above normal and a combined near and below normal category and (2) below normal and a combined near and above normal category.

The GSS's ease of construction ensures its consistency from categorization to categorization and with underlying linear correlations. The score is likewise equitable, does not depend on the forecast distribution, does not reward conservatism, utilizes off diagonal information in the contingency table, and penalizes larger errors more. For a limited subset of forecast situations it can be manipulated by a forecaster to his/her advantage (Mason and Mimmack 2002), but this is not a problem for objective forecast models that have not been trained to take advantage of this weakness. For all these reasons it is the recommended score.

An alternative score to the GSS for consideration is LEPSCAT (Potts et al.1996)

Table-2 shows the general form for the three possible two by two contingency tables referred to above (the third is the table for the near normal category and the combined above and below normal category). In Table-2, T is the grand sum of all the proper weights applied on each occurrence and non-occurrence of the events.

Table-2: General ROC contingency table for deterministic forecasts.

		Observations		
		occurrences	**non-occurrences**	
forecasts	**occurrences**	O_1	NO_1	O_1+ NO_1
	non-occurrences	O_2	NO_2	O_2+ NO_2
		O_1+ O_2	NO_1+ NO_2	T

The 2X2 table in Table-2 may be constructed from the 3X3 table described in Table-1 by summing the appropriate rows and columns.

In Table-2, O_1 represents the correct forecasts or hits:

$$O_1 = \sum W_i \left(OF\right)_i$$

(OF) being 1 when the event occurrence is observed and forecast; 0 otherwise. The summation is over all grid points or stations.

NO_1 represents the false alarms:

$$NO_1 = \sum W_i \left(NOF\right)_i$$

(NOF) being 1 when the event occurrence is not observed but was forecast; 0 otherwise. The summation is over all grid points or stations.

O_2 represents the misses:

$$O_2 = \sum W_i \left(ONF\right)_i$$

(ONF) being 1 when the event occurrence is observed but not forecast; 0 otherwise. The summation is over all grid points or stations.

NO_2 represents the correct rejections:

$$NO_2 = \sum W_i \left(NONF\right)_i$$

(NONF) being 1 when the event occurrence is not observed and not forecast; 0 otherwise. The summation is over all grid points or stations.

$W_i = 1$ when verification is done at stations or at single grid points.

$W_i = \cos(\theta_i)$ at grid point i, when verification is done on a grid.

$\theta_i =$ the latitude at grid point i.

When verification is done at stations, the weighting factor is one. Consequently, the number of occurrences and non-occurrences of the event are entered in the contingency table of Table -2.

However, when verification is done on a grid, the weighting factor is cos(θ_i), where θ_i is the latitude at grid point i. Consequently, each number entered in the contingency table of Table-3, is, in fact, a summation of the weights properly assigned.

Using stratification by observations (rather than by forecast), the Hit Rate (HR) is defined as (referring to Table-2):

$$HR = O_1 \Big/ \left(O_1 + O_2\right)$$

The range of values for HR goes from 0 to 1, the latter value being desirable. An HR of one means that all occurrences of the event were correctly forecast.

The False Alarm Rate (FAR) is defined as:

$$FAR = NO_1 \Big/ (NO_1 + NO_2)$$

The range of values for FAR goes from 0 to 1, the former value being desirable. A FAR of zero means that in the verification sample, no non-occurrences of the event were forecast to occur.

Hanssen and Kuipers score (Stanski et al 1989) is calculated for deterministic forecasts. Hanssen and Kuipers score (KS) is defined as:

$$\begin{aligned} KS &= HR - FAR \\ &= \frac{O_1 NO_2 - O_2 NO_1}{(O_1 + O_2)(NO_1 + NO_2)} \end{aligned}$$

The range of KS goes from -1 to +1, the latter value corresponding to perfect forecasts (HR being 1 and FAR being 0). KS can be scaled so that the range of possible values goes from 0 to 1 (1 being for perfect forecasts):

$$KS_{scaled} = \frac{KS + 1}{2}$$

The advantage of scaling KS is that it becomes comparable to the area under the ROC curve for probabilistic forecasts (see section 3.33) where a perfect forecast system has an area of one and a forecast system with no information has an area of 0.5 (HR being equal to FAR).

⇒ Contingency tables for deterministic categorical forecasts (such as in Table-2) are mandatory for level 2 verification. These contingency tables can provide the basis for the calculation of several scores and indices such as the Gerrity Skill Score, the LEPSCAT or the scaled Hanssen and Kuipers score and others, which are not essential in the recommended minimum verification frame of work but can be calculated as the additional information.

3. ROC for probabilistic forecasts

Tables-3 and 4 show contingency tables (similar to Table 2) that can be built for probabilistic forecasts of binary events.

Table-3: General ROC contingency table for probabilistic forecasts of binary events with definitions of the different parameters. This contingency table applies when probability thresholds are used to define the different probability bins.

bin number	forecast probabilities	observed occurrences	observed non-occurrences
1	0-P_2 (%)	O_1	NO_1
2	P_2-P_3 (%)	O_2	NO_2
3	P_3-P_4 (%)	O_3	NO_3
●●●	●●●	●●●	●●●
N	P_n-P_{n+1} (%)	O_n	NO_n
●●●	●●●	●●●	●●●
N	P_N-100 (%)	O_N	NO_N

In Table -3 above,

n = number of the nth probability interval or bin n; n goes from 1 to N.
P_n= lower probability limit for bin n.
P_{n+1} = upper probability limit for bin n.
N = number of probability intervals or bins.

$$O_n = \sum W_i \left(O\right)_i$$

(O) being 1 when an event corresponding to a forecast in bin n, is observed as an occurrence; 0 otherwise. The summation is over all forecasts in bin n, at all grid points or stations.

$$NO_n = \sum W_i \left(NO\right)_i$$

(NO) being 1 when an event corresponding to a forecast in bin n, is not observed; 0 otherwise. The summation is over all forecasts in bin n, at all grid points i or stations i

$W_i = 1$ when verification is done at stations or at single grid points.

$W_i = \cos\left(\theta_i\right)$ at grid point i, when verification is done on a grid.

$\theta_i =$ the latitude at grid point i.

Table-4: General ROC contingency table for probabilistic forecasts of binary events with definitions of the different parameters. This contingency table applies when the different probability bins are defined as function of the number of members in the ensemble.

bin number	member distribution	observed occurrences	observed non-occurrences
1	F=0, NF=M	O_1	NO_1
2	F=1, NF=M-1	O_2	NO_2
3	F=2, NF=M-2	O_3	NO_3
●●●		●●●	●●●

	N	F=n-1, NF=M-n+1	O_n	NO_n	
	●●●		●●●	●●●	
	N	F=M, NF=0	O_N	NO_N	

In Table-4 above,

M = number of members in the ensemble

n = number of the nth bin; n goes from 1 toN=M+1.
F = the number of members forecasting occurrence of the event.
NF = the number of members forecasting non occurrence of the event.

The bins may be aggregated.

$$O_n = \sum W_i (O)_i$$

(O) being 1 when an event corresponding to a forecast in bin n, is observed as an occurrence; 0 otherwise. The summation is over all forecasts in bin n, at all grid points i or stations i.

$$NO_n = \sum W_i (NO)_i$$

(NO) being 1 when an event corresponding to a forecast in bin n, is not observed; 0 otherwise. The summation is over all forecasts in bin n, at all grid points i or stations i.

$W_i = 1$ when verification is done at stations or at single grid points.

$W_i = \cos(\theta_i)$ at grid point i, when verification is done on a grid.

$\theta_i =$ the latitude at grid point i.

To build the contingency table in Table-4, probability forecasts of the binary event are grouped in categories or bins in ascending order, from 1 to N, with probabilities in bin n-1 lower than those in bin n (n goes from 1 to N). The lower probability limit for bin n is P_n and the upper limit is P_{n+1}. The lower probability limit for bin 1 is 0%, while the upper limit in bin N is 100%. The summation of the weights on the observed occurrences and non-occurrences of the event corresponding to each forecast in a given probability interval (bin n for example) is entered in the contingency table.

Tables-3 and 4 outline typical contingency tables. It is recommended that the number of probability bins remain between 10 and 20. The forecast providers can bin according to percent thresholds (Table-3) or ensemble members (Table-4) as deemed necessary. Table-5 gives an example of a table based on ensemble members.

Hit rate and false alarm rate are calculated for each probability threshold P_n (see Tables-3 and 5). The hit rate for probability threshold P_n (HR_n) is defined as (referring to Tables-4 and 5):

$$HR_n = \sum_{i=n}^{N} O_i \Big/ \sum_{i=1}^{N} O_i$$

and the false alarm rate (FAR$_n$) is defined as:

$$FAR_n = \sum_{i=n}^{N} NO_i \Big/ \sum_{i=1}^{N} NO_i$$

where n goes from 1 to N. The range of values for HR$_n$ goes from 0 to 1, the latter value being desirable. The range of values for FAR$_n$ goes from 0 to 1, zero being desirable. Frequent practice is for probability intervals of 10% (10 bins, or N=10) to be used. However the number of bins (N) should be consistent with the number of members in the ensemble prediction system (EPS) used to calculate the forecast probabilities. For example, intervals of 33% for a nine-member ensemble system could be more appropriate.

Hit rate (HR) and false alarm rate (FAR) are calculated for each probability threshold P$_n$, giving N points on a graph of HR (vertical axis) against FAR (horizontal axis) to form the Relative Operating Characteristics (ROC) curve. This curve, by definition, must pass through the points (0,0) and (1,1) (for events being predicted only with >100% probabilities (never occurs) and for all probabilities exceeding 0% respectively). No-skill forecasts are indicated by a diagonal line (where HR=FAR); the further the curve lies towards the upper left-hand corner (where HR=1 and FAR=0) the better

The area under the ROC curve is a commonly used summary statistics representing the skill of the forecast system. The area is standardised against the total area of the figure such that a perfect forecast system has an area of one and a curve lying along the diagonal (no information) has an area of 0.5. The normalised ROC area has become known as the ROC score. Not only can the areas be used to contrast different curves, but they are also a basis for Monte Carlo significance tests. It is proposed that Monte Carlo testing should be done within the forecast data set itself. For the verification of LRF, the area under the ROC curve should be calculated using the Trapezium rule (Other techniques are available to calculate the ROC score (see Mason, 1982).)

4. Reliability diagrams and frequency histograms for probabilistic forecasts

It is recommended that the construction of reliability curves (including frequency histograms to provide indications of sharpness) be done for the large-sampled probability forecasts. Given frequency histograms, the reliability curves are sufficient for the ROC curve, and have the advantage of indicating the reliability of the forecasts, which is a deficiency of the ROC. It is acknowledged that the ROC curve is frequently the more appropriate measure of forecast quality than the reliability diagram in the context of verification of long-range forecasts because of the sensitivity of the reliability diagram to small sample sizes. However, because measures of forecast reliability are important for modellers, forecasters, and end-users, it is recommended that in the exceptional cases of the forecasts being spatially aggregated over the tropics and over the two extratropical hemispheres, reliability diagrams be constructed in addition to ROC curves.

The technique for constructing the reliability diagram is somewhat similar to that for the ROC. Instead of plotting the hit rate against the false alarm rate for the accumulated probability bins, the hit rate is calculated only from the sets of forecasts for each probability bin separately, and is plotted against the corresponding forecast probabilities. The hit rate for each probability bin (HR_n) is defined as:

$$HR_n = \frac{O_n}{O_n + NO_n}$$

This equation should be contrasted with the hit rate used in constructing the ROC diagram.

Frequency histograms are constructed similarly from the same contingency tables as those used to produce reliability diagrams. Frequency histograms show the frequency of forecasts as a function of the probability bin. The frequency of forecasts (F_n) for probability bin n is defined as:

$$F_n = \frac{O_n + NO_n}{T}$$

where T is the total number of forecasts (and $T = \sum_{n=1}^{N}(O_n + NO_n)$).

5. Heidke Skill Score (HSS) (for categorical forecasts)

The Heidke skill score tallies the number of hits (cases in which the category having the highest forecast probability matches the later observed category) and compares this number to that expected by chance alone. The HSS is a scaled measure of the percentage improvement in skill relative to a set of random forecasts or climatology (equal chances) forecasts and is defined by

HSS = (c-e)*100/ (t-e)

where c is the number of cases (here, grid points) withhits, t is the total number of grid points in the outlook, and e is number of grid points expected to be correct bychance (and equals t/3 for the tercile based categorization). In the case of more than one category sharing thehighest forecast probability, a hit is divided if one of them is later observed. Hence, in the case of the ‘‘equalchances’’ forecast (1/ 3 probability for each category), there is a three-way tie and a 1/ 3 hit is tallied, equalling the hit rate expected by chance and contributing to a HSS of 0. For a two-way tie,2 a ½ hit is tallied, contributing to positive skill more weakly than a non-tied hit.

6. Hindcasts

In contrast to short- and medium-range dynamical Numerical Weather Prediction (NWP) forecasts, seasonal forecasts are produced relatively few times a year (for example, one forecast for each season or one forecast for the following 90-day period, issued every month). Therefore the verification sampling for seasonal forecasts may be limited, possibly to the point where the validity and significance of the verification results may be questionable. Providing verification for a few seasons or even over a few years only may be misleading and may not give a fair assessment of the skill of any seasonal forecasting system.

Seasonal forecasting systems should be verified over as long a period as possible in hindcast mode. Although there are limitations on the availability of verification data sets and in spite of the fact that validating numerical forecast systems in hindcast mode requires large computer resources, the hindcast period should be as long as possible.

References

Brier, G.W., 1950, Verification of forecasts expressed in terms of probability. *M on.Weather Rev*., **78**, 1–3.

Gerrity, J. P. Jr., 1992, A note on Gandin and Murphy’s equitable skill score. *Monthly Weather Review*, 120, pp. 2707-2712.

Jollife, I.T., and D.B. Stephenson, (Eds), 2011, Forecast verification: a practitioner’s guide in atmospheric science, John Wiley Sons Limited. P240.

Mason,I., 1982, A model for assessment of weather forecasts. Aust. Met. Mag, 30:291-303.

Mason, S. J., and G. M. Mimmack, 2002: Comparison of some statistical methods of probabilistic forecasting of ENSO.*J. Climate*, 15, pp. 8-29.

Mitra, Ashis K., et al., 2009, Daily Indian precipitation analysis formed from a merge of rain-gauge data with the TRMM TMPA satellite-derived rainfall estimates, Journal of the Meteorological Society of Japan 87, 265-279.

Murphy, A. H., 1988, Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116, pp. 2417-2424.

Potts J. M., C. K. Folland, I. T, Jolliffe and D. Sexton, 1996, Revised “LEPS” scores for assessing climate model simulations andlong-range forecasts, *J. Climate*, 9, pp. 34-53.

Rajeevan, M., J. Bhate, K. D. Kale, and B. Lal, 2006, High resolution daily gridded rainfall data for the Indian region: Analysis of break and active monsoon spells, Curr. Sci., 91, 296–306.

Stanski H. R., L. J. Wilson and W. R. Burrows, 1989, Survey of common verification methods in meteorology. *World Weather Watch Technical Report No. 8*, WMO/TD-No.358, 114 pp.

Smith, T.M., R.W. Reynolds, Thomas C. Peterson, and Jay Lawrimore, 2008, Improvements to NOAA's Historical Merged Land-Ocean Surface Temperature Analysis (1880-2006). *Journal of Climate*,**21**, 2283-2296.

Wilks, D.S. 1995, *Statistical Methods in the Atmospheric Sciences: An Introduction*. San Diego: Academic Press.

Part-II: Medium Range Forecasts

In the first part of the working group report, the verification methods for extended range and seasonal forecasts were proposed along with the suitable data sets to be used for verification.

In the second part here, we discuss the verification methods and data sets suitable for verification for medium range forecasts. In India, medium range forecasts are being prepared by ESSO-NCMRWF and ESSO-IMD. While at ESSO-NCMRWF, the GFS and UKMO models are being used for medium range forecasts, at ESSO-IMD, the GFS model is being used for generating medium range forecasts (up to 7 days). The GFS model is run at 22 km resolution, the UKMO model is run at 25 km resolution.

1) Parameters to be verified

a) Day-1to Day-7 forecasts of rainy day (rain or no rain i.e., rainfall >= 0.5 cm at each grid over land regions)

b) Day-1to Day-7 forecasts of 24 hr accumulated rainfall (mm/day) (rainfall exceeding different thresholds)

c) Day-1 to Day-7 forecasts of surface air temperatures (minimum and maximum temperatures)

2) Verification scores

The working group recommends the following scores for verification of medium range forecasts. These scores pertain to both deterministic and probabilistic forecasts. The scores are based on the categorical contingency table whereby an event (“yes”) is defined by rainfall greater than or equal to the specified threshold; otherwise it is a non-event (“no”). Further the verification statistics are presented for thresholds of rainfall excess of 0.1, 0.5, 1.0, 3.0, 5.0, 7.0 and 9.0 cm in 24 hours.

(2.a)Standard Verification Scores for Rainfall Forecasts (deterministic)

(a)Bias Score (BIAS)
(b)Probability of Detection (POD)
(c)False Alarm Ratio (FAR)
(d)Equitable threat score (Gilbert skill score)
(e)Root Mean Squared Error (RMSE)
(f)Mean Error
(g)Pattern Correlation.

(2.b)Spatial Verification of Rainfall Forecasts (deterministic)

(a) Contiguous Rain Areas (CRA): The Contiguous Rain Area (CRA) method is a features based approach that evaluates the properties of forecast rain system, namely, their location, intensity and pattern.

(2.c)Verification Scores for ensemble probabilistic forecasts of rain categories

(a)Brier Score (BS)
(b)Brier Skill Score (BSS)
(c)Continuous Ranked Probability Score (CRPS)
(d)Ranked Probability Skill Score (RPSS)

To put the skill of the model forecasts into context, model verification statistics should also be compared to verification statistics for persistence or climatology. The persistence forecast is simply the observed rainfall from the previous forecast period and is the most commonly used unskilled forecast when evaluating QPFs. Climatology can be either some long-term mean rainfall or the most probable value. The model forecasts can be considered useful only if they outperform these two simple forecasts.

3) Data sets for verification of medium range forecasts

For rainfall verification, the India Meteorological Department (IMD) and NCMRWF merged rainfall data at 0.5□ grid resolution (NSGM) may be used. At present, the data set is being prepared at 0.5 degree resolution. However, ESSO-IMD plans to start the development work at 0.25 degree very soon. The NMSG rainfall analysis is obtained by merging satellite (TRMM) and gauge (IMD) estimates of rainfall over India (Mitra et al. 2009).

For temperature verifications, the daily gridded temperature data (Srivastava et al. 2009) may be used. This data set makes use maximum and minimum temperature data of more than 300 surface observatories on real time basis. The data set is developed at a 1 X 1 degree resolution.

The summary of the verification scores for precipitation and temperature is given below.

Parameters	Verification regions	Deterministic forecasts	Probabilistic forecasts
Precipitation amount	(i) Indian Monsoon region (0–40N, 65E-100E) (ii) Indian main land region	(i)Standard verification scores (ii)Spatial verification using CRA method.	BS, BSS ROC curves Reliability diagram, CRPS and RPSS
Temperatures	(i) Indian Monsoon region (0–40N, 65E-100E) (ii) Indian main land region	(i)Standard verification scores	BS, BSS ROC curves

4. ENSEMBLE PREDICTION SYSTEM (EPS)

Ensemble forecasting (EF) has been embraced as a practical way of estimating the uncertainty of a weather forecast. Since 1960s, it has been recognized that perfect numerical weather forecasts will always be unattainable; even the smallest of errors in the initial conditions will grow inexorably, eventually rendering any single deterministic forecast useless. Rather than pinning unrealistic hopes upon the accuracy of a single numerical forecast, EF adopts an alternative approach: generate multiple, individual numerical forecasts from different initial conditions and/or different numerical model configurations. Probabilistic forecasts of the weather may then be generated from the relative frequencies of events in the ensemble. Medium range ensemble forecasts have been produced operationally at several centers around the world. Each of the centers produces the ensembles using different forecast models and different ensemble construction techniques.

ESSO-NCMRWF recently implemented an ensemble prediction system using the NCEP GFS model. With the improved computing facility being acquired by ESSO-MOES, the ensemble prediction system will be made operational soon at ESSO-IMD. Therefore, we need to develop/propose suitable verification scores meant for the ensemble prediction system (EPS).

A single number is usually inadequate for evaluating all the desired information about the performance of an analysis/forecast system. Each verification measure provides unique information on performance. For that reason, the utility of EPS forecasts will be evaluated by several measures. Ensemble forecast verification is required to enable better decisions than could be made using alternate information (persistence, climatology, deterministic forecast), and to determine how do forecast systems differ in performance. It will assist in bias removal and downscaling and pinpoint sources of error in ensemble forecast system. Ensemble method is used to generate probability distributions with quality characterized by skill and spread. The skill represents accuracy to assess whether the forecasts are close to the observed.

The statistical problem is to determine whether an accuracy for the EPS forecasts is significantly better than a baseline forecast (e.g., climatology), or whether changes in EPS configuration from one year to the next, or between the EPS forecasts and idealized perfect model configurations, produce a significant change in the accuracy or skill measures. The significance of differences between rmse's, Brier skill scores, areas under relative operating characteristic (ROC) curves and rank frequencies is judged by nonparametric resampling. The use of bootstrap techniques to estimate statistical significance has clear advantages over parametric tests such as the Student's *t*-test (e.g., Livezey and Chen 1983; Wilks 1995, 145–150, and references within). Some advantages of resampling techniques are that: (1) an a priori assumption of the background distribution of the sample populations (e.g., equal variances) is not required, (2) an a priori estimate of the number of degrees of freedom is not required, and (3) differences between *any metric* can be tested.

The consensus on the verification of EPS among the prediction centers was that, a minimum, the following metrics should be used to evaluate the EPS:

1) Probabilistic skill score

Score measures such as the Brier Score, Brier Skill Score, Ranked Probability Score, or Ranked Probability Skill Score can provide an overall single number metric for judging the quality of probabilistic forecasts. Their very simplicity also prohibits them from being very informative about the nature of probabilistic forecast errors. However, the Brier score can be decomposed into reliability, resolution and uncertainty terms. A similar decomposition for a continuous ranked probability score also is available.

2) Reliability Diagrams

The reliability diagrams plotted together with a decomposition of the Brier score and information on the distribution of forecasts issued (the sharpness). Reliability diagrams can provide information on conditional biases of ensemble forecasts. However, if they are not based on large set of cases, they can be noisy and uninformative.

3) The Relative Operating Characteristic (ROC)

The ROC curve graphs probabilities of incorrect null and alternative hypotheses as each sorted ensemble member is used as a decision making threshold. The ROC curve is based on stratification by observations, it is independent of reliability and instead provides a measure of resolution. It is particularly valuable for comparing the performance of ensemble systems against single deterministic forecasts at higher resolution and the more general resource issue of ensemble size/configuration versus model resolution.

4) Rank Histograms

The Rank histograms diagnose the ability of the ensemble to sample from the correct probability distribution. Model bias and under or over variability of the ensemble can be detected from the shape of the rank histogram.

Other evaluative techniques (spread/skill relationships, cluster analysis, etc.) may prove useful depending on the research issue in question.

5. Verification of forecasts for extreme /rare events

Extreme weather events such as high wind speeds, heavy precipitation, or high temperatures can have severe impacts on society. Improving predictions of such events therefore has a high priority in national weather services. An important part of this activity is to determine whether or not prediction quality is improved when prediction systems are updated. Assessing the quality of predictions of extreme weather events, however, is complicated by the fact that measures of forecast quality typically degenerate to trivial values as the rarity of the predicted event increases. The drive to improve predictions of extreme events and the associated difficulties of measuring the quality of such predictions has generated a growing interest in better ways of verifying forecasts of extreme events.

The verification scores like hit rate, false alarm ratio and odds ratio cannot be used for verifying extreme weather events, as these score typically degenerate to trivial values as the definition of the event is changed to become increasingly rare (Ferro and Stephenson, 2011). Two new measures are proposed by Ferro and Stephenson (2011) for verifying deterministic forecasts of rare binary events. The scores for verification of extreme weather events are a) Extremal Dependence Index (EDI) and b) the Symmetric extremal dependence Index (SEDI). These two scores are defined as follows:

$$EDI = \frac{logF - logH}{\log F + logH}$$

$$SEDI = \frac{logF - logH - \log(1-F) + \log(1-H)}{logF + logH + \log(1-\mathbf{F}) + \log(1-\mathbf{H})}$$

Where, H is hit rate and F is false alarm ratio.

Both EDI and SEDU are base-rate independent because they are functions of H and F only. However, it is recommended that forecasts must be calibrated before computing these measures if a clear understanding of forecast performance of rare events is desired.

More details of these indices can be obtained from Ferro and Stephenson (2011).

APPENDIX-2

(i)Standard Verification Scores for Rainfall Forecasts (deterministic)

(a)Bias Score (BIAS): Also called the frequency bias, it is the ratio of forecast rain frequency to observed rain frequency.

$$BIAS = {(a+b)}/{(a+c)}$$

(b)Probability of Detection (POD): It is the fraction of observed events that were correctly predicted.

$$POD = {a}/{(a+b)}$$

(c)False Alarm Ratio (FAR): It is the fraction of forecast events that were observed to be non-event.

$$FAR = {b}/{(a+b)}$$

(d)Equitable threat score (Gilbert skill score): It is the fraction of all events forecast and.or observed that were correctly diagnosed, accounting for the hits that would occur purely due to random chance.

$$ETS = {(a - arandom)}/{(a+b+c- \ arandom)}$$

where – $arandom = \dfrac{1}{n[(a+b)(a+c)]}$

(e)Root Mean Squared Error (RMSE): RMSE measures the average error magnitude but gives greater weight to larger errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_i - O_i)^2}$$

Additionally mean squared error (MSE) and pattern correlation are also computed. The MSE can be decomposed into component error sources as shown in next section.

(ii)Spatial Verification of Rainfall Forecasts (deterministic)

(a) Contiguous Rain Areas (CRA): In the CRA method a weather system is defined as a region bounded by a user specified isopleth of precipitation in the union of the forecast and observed rain fields. The horizontal displacement of the forecast is determined by translating the forecast rain field until the total squared difference (*pattern correlation*) between the observed and forecast fields is minimized (*maximized*). This allows a decomposition of total error into components due to: (a) location; (b) rain volume and (c) pattern.

The CRA method is an object-oriented verification procedure suitable for gridded quantitative precipitation forecasts (QPFs). For each entity (user defined isopleth) that can be identified in forecast and observations, the CRA method uses pattern matching techniques to determine the location error, as well as errors in area, mean and maximum intensity, and spatial pattern. The total error can be decomposed into components due to location, volume, and pattern error. This is a useful property for model developers who need such information to improve the numerical weather prediction models.

CRA method was developed for estimating the systematic errors in the rainfall forecasts (Ebert and McBride 2000; Ebert and Gallus 2009). It was one of the first methods to measure errors in predicted location and to separate the total error into components due to errors in location, volume and pattern. The steps involved in CRA technique are described in Ebert and Gallus (2009). A brief summary of the procedure is given here.

Firstly a CRA is defined for an observation/Forecast pair based on a user-specified isohyet (rain rate contour) in the forecast and/or the observations. It is the *union* of the forecast and observed entities (blobs) as illustrated in the Figure shown below. The forecast and observed entities need not overlap, but they must be associated with each other. The two entities must be nearby and also associated with common synoptic situation. During the monsoon season large parts of India regularly receive rainfall in the range up to 10 mm/day. It was found that choice of 1, 2 and 5 mm/day contours spread the CRA across large geographical areas. CRAs defined by higher thresholds of 10, 20, 40 and 80 mm/day were usec

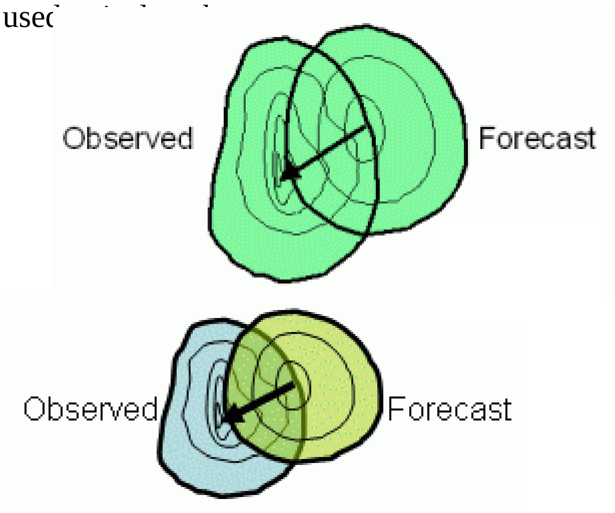


Figure Typical CRA entity based on user defined contour.

In the next step pattern matching technique is used for estimating the location error. Here the forecast field is horizontally translated over the observed field until the best match is obtained. The location error is then simply the vector displacement of the forecast.

The best match between the two entities can be determined either: (a) by maximizing the correlation coefficient, (b) by minimizing the total squared error, (c) by maximizing the overlap of the two entities, or (d) by overlaying the centers of gravity of the two entities. For a good forecast all of the methods will give very similar location errors. In the present study the best match is determined by maximizing the correlation. The mean squared error (MSE) and its decomposition (location error, volume error and pattern error) are done as shown below.

$$MSE_{Total} = (\boldsymbol{F} - \boldsymbol{O})^2 + (\mathbf{s_O} - r\mathbf{s_F})^2 + (1- r^2)\mathbf{s_F}^{\,2} \qquad (1)$$

-where $\boldsymbol{F}$ ($\mathbf{s_F}$) and $\boldsymbol{O}$ ($\mathbf{s_O}$) are the mean (standard deviation) values of the forecast and observed precipitation respectively before obtaining the best match via shifting the forecast. The spatial correlation between the original forecast and observed features (r) increases to an optimum value (r_{OPT}) in the process of correcting the location via pattern matching. The contribution to total error due to displacement, volume and pattern errors are estimated as-

$$MSE_{Displacement} = 2\mathbf{s_F}\mathbf{s_O}\,(r_{OPT} - r),$$

$$MSE_{Volume} = (\boldsymbol{F'} - \boldsymbol{O'}), \text{ and} \qquad (2)$$

$$MSE_{Pattern} = 2\mathbf{s_F}\mathbf{s_O}\,(1 - r_{OPT}) + (\mathbf{s_F} - \mathbf{s_O})^2$$

-where the $\boldsymbol{F}$ ‘and $\boldsymbol{O'}$ are the mean values after shifting.

(iii)Verification Scores for Probability forecasts of rain meeting or exceeding specified thresholds

For verification of the forecast rain probability (P_i) at any location i the observed occurrence O_i (0 or 1) is considered.

(a)Brier Score (BS): It measures the mean squared error in probability space.

$$BS = \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}$$

(b)Brier Skill Score (BSS): BSS expresses the value of BS for the forecast with reference to climatology.

$$BSS = 1 - \frac{BS_f}{BS_c}$$

(c)Rank Probability Score (RPS), Continuous Ranked Probability Score (CRPS): RPS tells how well did probability forecast predict the category that the observation fell into. CRPS forms the RPS for continuous

(d)Ranked Probability Skill Score (RPSS):RPSS shows the relative improvement of the probability forecast relative to climatology.

References

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

Ebert E. E, 2008: Fuzzy verification of high-resolution gridded forecasts; A review of proposed framework. Meteorol. Appl. 15, 51-64.

Ebert, E. E and W. A. Gallus Jr, 2009: Towards better understanding of Contiguous Rain Areas (CRA) method of spatial verification, Weather and Forecasting, 24, 1401-1415

Ferro, C.A.T., and D.B. Stephenson, 2011, Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events, Weather and Forecasting, DOI: 10.1175/WAF-D-10-05030.1, 699-713.

Hamill, T.M. et al, 2000, Meeting Summary, Bull Amer. Met Society, 81, 2653-2664. (the references therein for EPS)

Hamill, T.M., and S.J.Colucci, 1997, Verification of Eta-RSM short range ensemble forecasts, Mon Wea Rev, 125, 1312-1327.

Livezey, Robert E., W. Y. Chen, 1983: Statistical Field Significance and its Determination by Monte Carlo Techniques. *Mon. Wea. Rev.*, **111**, 46–59.

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

Srivastava A. K. M Rajeevan, S.R Kshirsagar, 2009: Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region. Atmospheric Science Letters (2009) http://dx.doi.org/10.1002/asl.232

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. Environment Canada Research Rep. 89-5, 114 pp.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability,* Reading, United Kingdom, ECMWF, 1–26.

Wilks, D.S. 1995, *Statistical Methods in the Atmospheric Sciences: An Introduction*. San Diego: Academic Press.

Acknowledgements

We thank Dr Shailesh Nayak, Chairman, ESSO and the Secretary, MoES for his constant encouragement, support and guidance for preparing this report. We also thank ESSO-MoES scientists for their comments and feedback, which helped us to improve the report.
