Ministério da Ciência e Tecnologia - MCT
Instituto Nacional de Pesquisas Espaciais - INPE
Centro de Previsão de Tempo e Estudos Climáticos - CPTEC

Forecast verification methods

Chou Sin Chan

chou@cptec.inpe.br
+55-12-3186-8424

Fatores que afetam a qualidade das previsões

- O modelo numérico
- Validação e verificação do modelo
- Posição geográfica, tamanho e resolução do dominio
- Maior esforço na interface previsor/modelo
- Topografia, percentagem de oceanos e continentes
- Nos tropicos predomina a representação dos processos físicos como turbulencia, convecção, radiação, processos de superfície, etc. Parametros empiricos,
- Ensemble de baixa qualidade.

Reasons for verification of forecasts:

-To monitor forecast quality over time,
- To compare the quality of different forecast systems, and
- To improve forecast quality through better understanding of forecast errors.

Verification should be done both against:
(a)**gridded observations (model-oriented verification) on a common 0.5° latitude/longitude grid**

NCEP reanalyses, ECMWF reanalyes, ERA-Interim, GPCP, CRU, satellite obs,

(b) station observations (user-oriented verification)

Nature of forecast:	Example(s)	Verification methods
deterministic (non-probabilistic)	quantitative precipitation forecast	visual, dichotomous, multi-category, continuous, spatial
probabilistic	probability of precipitation, ensemble forecast	visual, probabilistic, ensemble
qualitative (worded)	5-day outlook	visual, dichotomous, multi-category

Space-time domain:		
time series	daily maximum temperature forecasts for a city	visual, dichotomous, multi-category, continuous, probabilistic
spatial distribution	map of geopotential height, rainfall chart	visual, dichotomous, multi-category, continuous, probabilistic, spatial, ensemble
pooled space and time	monthly average global temperature anomaly	dichotomous, multi-category, continuous, probabilistic, ensemble

Specificity of forecast:		
dichotomous (yes/no)	occurrence of fog	visual, dichotomous, probabilistic, spatial, ensemble
multi-category	cold, normal, or warm conditions	visual, multi-category, probabilistic, spatial, ensemble
continuous	maximum temperature	visual, continuous, probabilistic, spatial, ensemble
object- or event-oriented	tropical cyclone motion and intensity	visual, dichotomous, multi-category, continuous, probabilistic, spatial

9 attributes that contribute to quality of forecasts

Bias - the correspondence between the mean forecast and mean observation.

Association - the strength of the linear relationship between the forecasts and observations (for example, the correlation coefficient measures this linear relationship)

Accuracy - the level of agreement between the forecast and the truth (as represented by observations). The difference between the forecast and the observation is the *error*. The lower the errors, the greater the accuracy.

Skill - the relative accuracy of the forecast over some reference forecast. The reference forecast is generally an unskilled forecast such as random chance, persistence (defined as the most recent set of observations, "persistence" implies no change in condition), or climatology. Skill refers to the increase in accuracy due purely to the "smarts" of the forecast system. Weather forecasts may be more accurate simply because the weather is easier to forecast -- skill takes this into account.

Reliability - the average agreement between the forecast values and the observed values. If all forecasts are considered together, then the *overall reliability* is the same as the *bias*. If the forecasts are stratified into different ranges or categories, then the reliability is the same as the *conditional bias*, i.e., it has a different value for each category.

Resolution - the ability of the forecast to sort or resolve the set of events into subsets with different frequency distributions. This means that the distribution of outcomes when "A" was forecast is different from the distribution of outcomes when "B" is forecast. Even if the forecasts are wrong, the forecast system has resolution if it can successfully separate one type of outcome from another.

Sharpness - the tendency of the forecast to predict extreme values. To use a counter-example, a forecast of "climatology" has no sharpness. Sharpness is a property of the forecast only, and like resolution, a forecast can have this attribute even if it's wrong (in this case it would have poor reliability).

Discrimination - ability of the forecast to discriminate among observations, that is, to have a higher prediction frequency for an outcome whenever that outcome occurs.

Uncertainty - the variability of the observations. The greater the uncertainty, the more difficult the forecast will tend to be.

What is "truth" when verifying a forecast?

The "truth" data that we use to verify a forecasts generally comes from observational data. These could be rain gauge measurements, temperature observations, satellite-derived cloud cover, geopotential height analyses, and so on.

In many cases it is difficult to know the exact truth because there are errors in the observations. Sources of uncertainty include random and bias errors in the measurements themselves, sampling error and other errors of representativeness, and analysis error when the observational data are analyzed or otherwise altered to match the scale of the forecast.

Rightly or wrongly, most of the time we ignore the errors in the observational data. We can get away with this if the errors in the observations are much smaller than the expected error in the forecast (high signal to noise ratio). Even skewed or under-sampled verification data can give us a good idea of which forecast products are better than others when intercomparing different forecast methods. Methods to account for errors in the verification data currently being researched.

Deterministic forecasts

The *mean value is useful for putting the forecast errors into perspective*

$$\overline{O} = \frac{1}{N}\sum_{i=1}^{N} O_i \qquad \overline{F} = \frac{1}{N}\sum_{i=1}^{N} F_i$$

The *sample variance (s2) describes the rainfall variability*

$$s_O^2 = \frac{1}{N-1}\sum_{i=1}^{N}(O_i - \overline{O})^2 \qquad s_F^2 = \frac{1}{N-1}\sum_{i=1}^{N}(F_i - \overline{F})^2$$

The *mean error (ME) measures the average difference between the forecast and observed* values

$$ME = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i) = \overline{F} - \overline{O}$$

The *mean absolute error (MAE) measures the average magnitude of the error.*

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|F_i - O_i\right|$$

The *mean square error (MSE) measures the average squared error magnitude, and is often used* in the construction of skill scores.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2$$

The *root mean square error (RMSE) measures the average error magnitude but gives greater* weight to the larger errors.

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

The *(product moment) correlation coefficient (r) measures the degree of linear association* between the forecast and observed values, independent of absolute or conditional bias. As this score is highly sensitive to large errors it benefits from the square root transformation of the rain amounts

$$r = \frac{\sum_{i=1}^{N}(F_i - \overline{F})(O_i - \overline{O})}{\sqrt{\sum_{i=1}^{N}(F_i - \overline{F})^2}\sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2}} = \frac{s_{FO}}{s_F s_O}$$

The *skill score measures the fractional* improvement of the forecast system over a reference forecast. The most frequently used scores are the *MAE and the MSE. The reference estimate is persistence for forecasts of 24h or less, and* climatology for longer forecasts.

$$MAE\_SS = \frac{MAE_{forecast} - MAE_{reference}}{MAE_{perfect} - MAE_{reference}} = 1 - \frac{MAE_{forecast}}{MAE_{reference}}$$

$$MSE\_SS = \frac{MSE_{forecast} - MSE_{reference}}{MSE_{perfect} - MSE_{reference}} = 1 - \frac{MSE_{forecast}}{MSE_{reference}}$$

The MSSS is essentially the Mean Square Error (MSE) of the forecasts compared to the MSE of climatology for a station or grid point.

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

where

$$MSE_{cj} = \frac{n-1}{n} s_{xj}^2$$

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

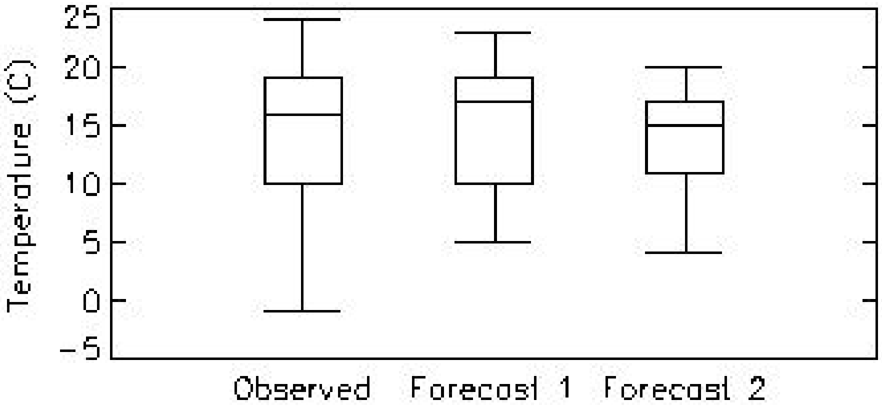
The Mean Square Skill Score (MSSS) is applicable to deterministic forecasts only

Percent improvement in *MSE* (mean square error) over a climatological forecast

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

The first three terms of the decomposition of MSSSj 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.

Box plot - Plot boxes to show the range of data falling between the 25th and 75th percentiles, horizontal line inside the box showing the median value, and the whiskers showing the complete range of the data.



Contingency Table – for categorical forecasts

		Observed yes	Observed no		
Forecast	yes	*hits* a	*false alarms* b	*forecast yes*	a+b
	no	*misses* c	*correct rejections* d	*forecast no*	c+d
		observed yes a+c	*observed no* b+d	*N = total*	

The *frequency bias (BIAS) gives the ratio of the forecast rain frequency to the observed rain frequency.*

$$BIAS = \frac{hits + false\ alarms}{hits + misses}$$

The *probability of detection (POD (HR)) measures the fraction of observed events that were correctly* forecast.

$$POD = \frac{hits}{hits + misses}$$

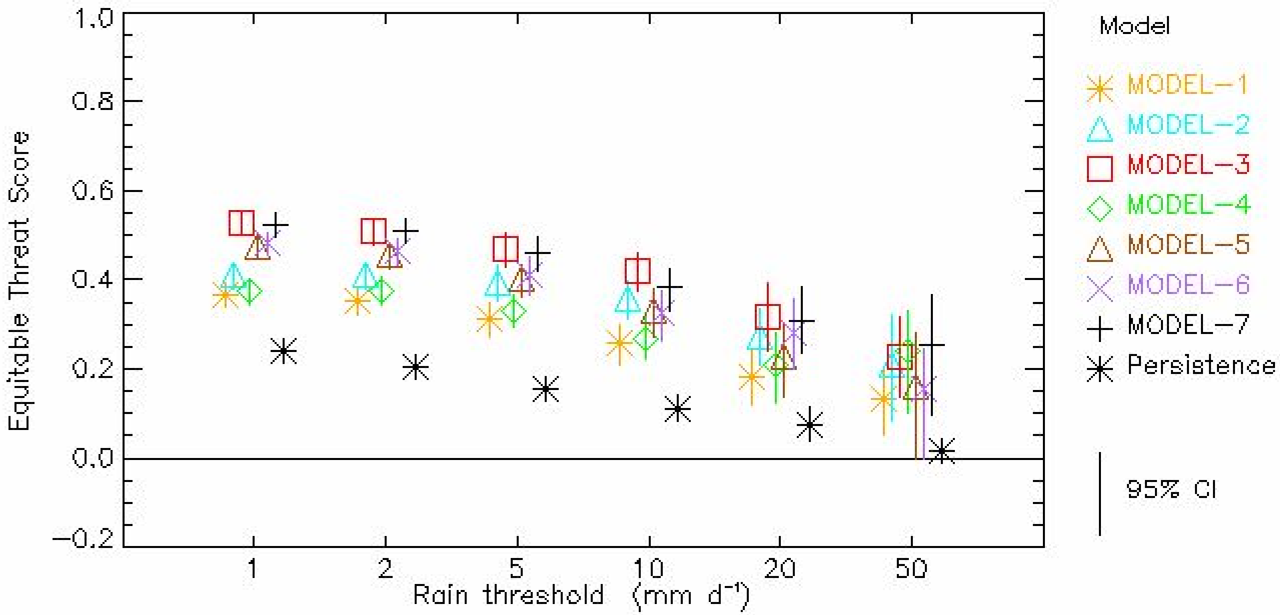
The *false alarm ratio (FAR) gives the fraction of forecast events that were observed to be nonevents.*

$$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

The *equitable threat score (ETS) measures 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

$$ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}$$

$$hits_{random} = \frac{1}{N}(observed\ yes \times forecast\ yes)$$



For probabilistic forecasts

An accurate probability forecast system has:

reliability - agreement between forecast probability and mean observed frequency.

sharpness - tendency to forecast probabilities near 0 or 1, as opposed to values clustered around the mean.

resolution - ability of the forecast to resolve the set of sample events into subsets with characteristically different outcomes

Brier score - $$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 = \frac{1}{N}\sum_{k=1}^{K} n_k (p_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})$$

(1) (2) (3)

Brier score: measure the mean squared probability error . It can be partitioned into three terms: (1) reliability, (2) resolution, and (3) uncertainty.

Range: 0 to 1. Perfect score: 0.

Relative Operating Characteristics (ROC- for probabilistic forecasts)

Plot hit rate (PODy) vs false alarm rate (POFn), using a set of increasing probability thresholds (for example, 0.05, 0.15, 0.25, etc.) to make the yes/no decision.

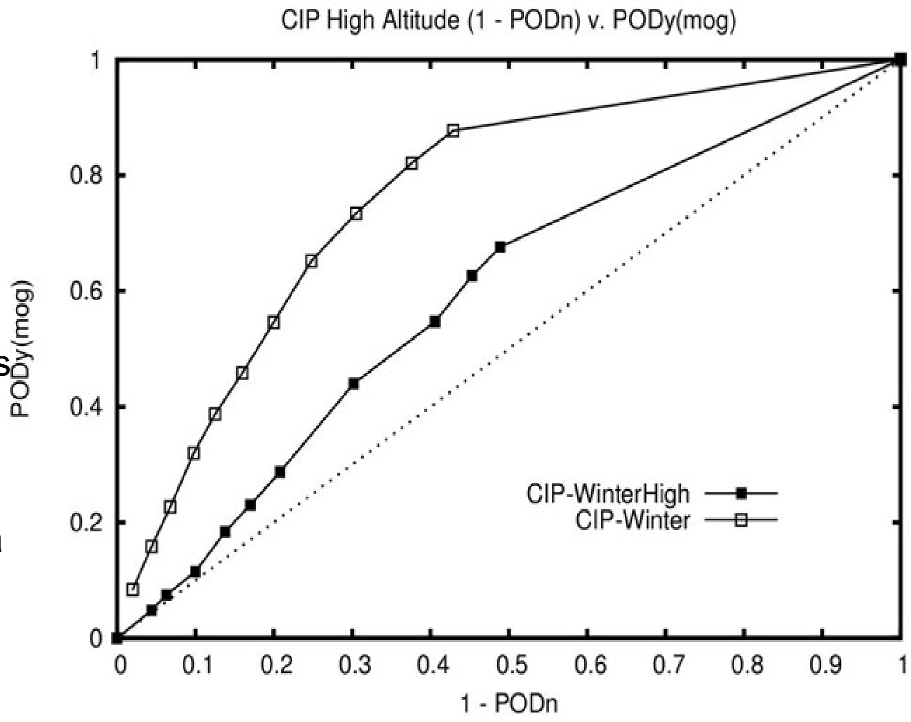
The area under the ROC curve is frequently used as a score.

Perfect: Curve travels from bottom left to top left of diagram, then across to top right of diagram. Diagonal line indicates no skill.

Range: 0 to 1, 0.5 indicates no skill. Perfect score: 1

What is the ability of the forecast to discriminate between events and non-events?

ROC measures the ability of the forecast to discriminate between two alternative outcomes, thus measuring resolution. It is not sensitive to bias in the forecast, so says nothing about reliability. A biased forecast may still have good resolution and produce a good ROC curve, which means that it may be possible to improve the forecast through calibration. The ROC can thus be considered as a measure of potential usefulness.

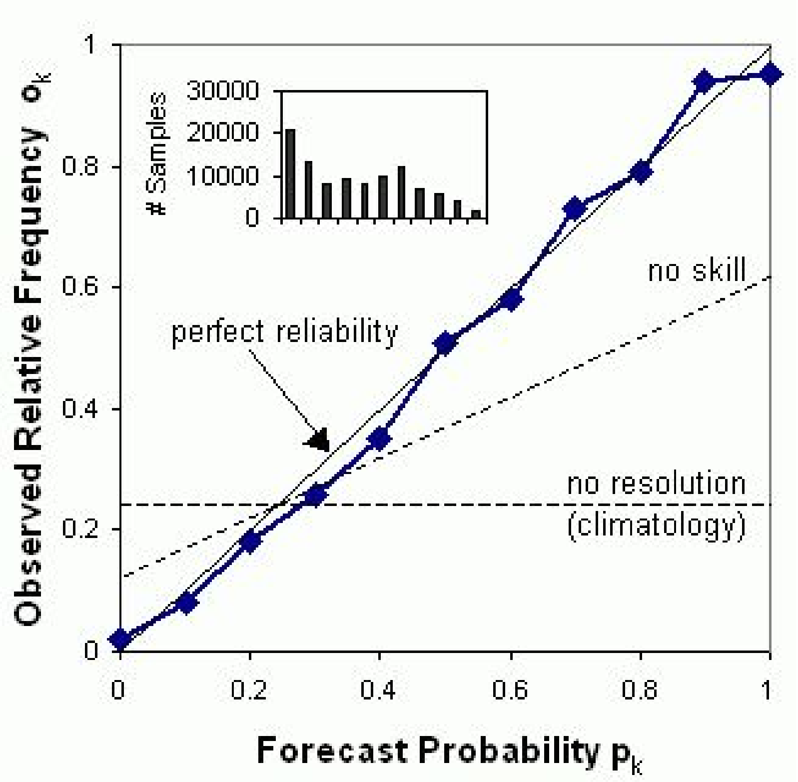


RELIABILITY DIAGRAM (obs freq X fcst probability)

The range of forecast probabilities is divided into K bins (0-5%, 5-15%, 15-25%, etc.).

The sample size in each bin is often included as a histogram or values beside the data points.

How well do the predicted probabilities of an event correspond to their observed frequencies?



The deviation from the diagonal gives the conditional bias. If the curve lies below the line, this indicates overforecasting (probabilities too high); points above the line indicate underforecasting (probabilities too low).

The flatter the curve in the reliability diagram, the less resolution it has. A forecast of climatology does not discriminate at all between events and non-events, and thus has no resolution.

The frequency of forecasts in each probability bin (histogram) shows the sharpness of the forecast.

The reliability diagram is conditioned on the forecasts (i.e., given that X was predicted, what was the outcome?), and can be expected to give information on the real meaning of the forecast.

It is a good partner to the ROC, which is conditioned on the observations.

Talagrand Diagram for probabilistic forecasts

How well does the ensemble spread of the forecast represent the true variability (uncertainty) of the observations?

Rank histogram (Talagrand et al. 1997; Hamill, 2001)

The diagram checks where the verifying observation usually falls with respect to the ensemble forecast data, which is arranged in increasing order at each grid point. In an ensemble with perfect spread, each member represents an equally likely scenario, so the observation is equally likely to fall between any two members.

To construct a rank histogram, do the following:
1. At every observation (or analysis) point rank the N ensemble members from lowest to highest. This represents N+1 possible bins that the observation could fit into, including the two extremes
2. Identify which bin the observation falls into at each point
3. Tally over many observations to create a histogram of rank.

Interpretation:
Flat - ensemble spread about right to represent forecast uncertainty
U-shaped - ensemble spread too small, many observations falling outside the extremes of the ensemble
Dome-shaped - ensemble spread too large, most observations falling near the center of the ensemble
Asymmetric - ensemble contains bias

Note: A flat rank histogram does not necessarily indicate a good forecast, it only measures whether the observed probability distribution is well represented by theensemble.

Taylor Diagram

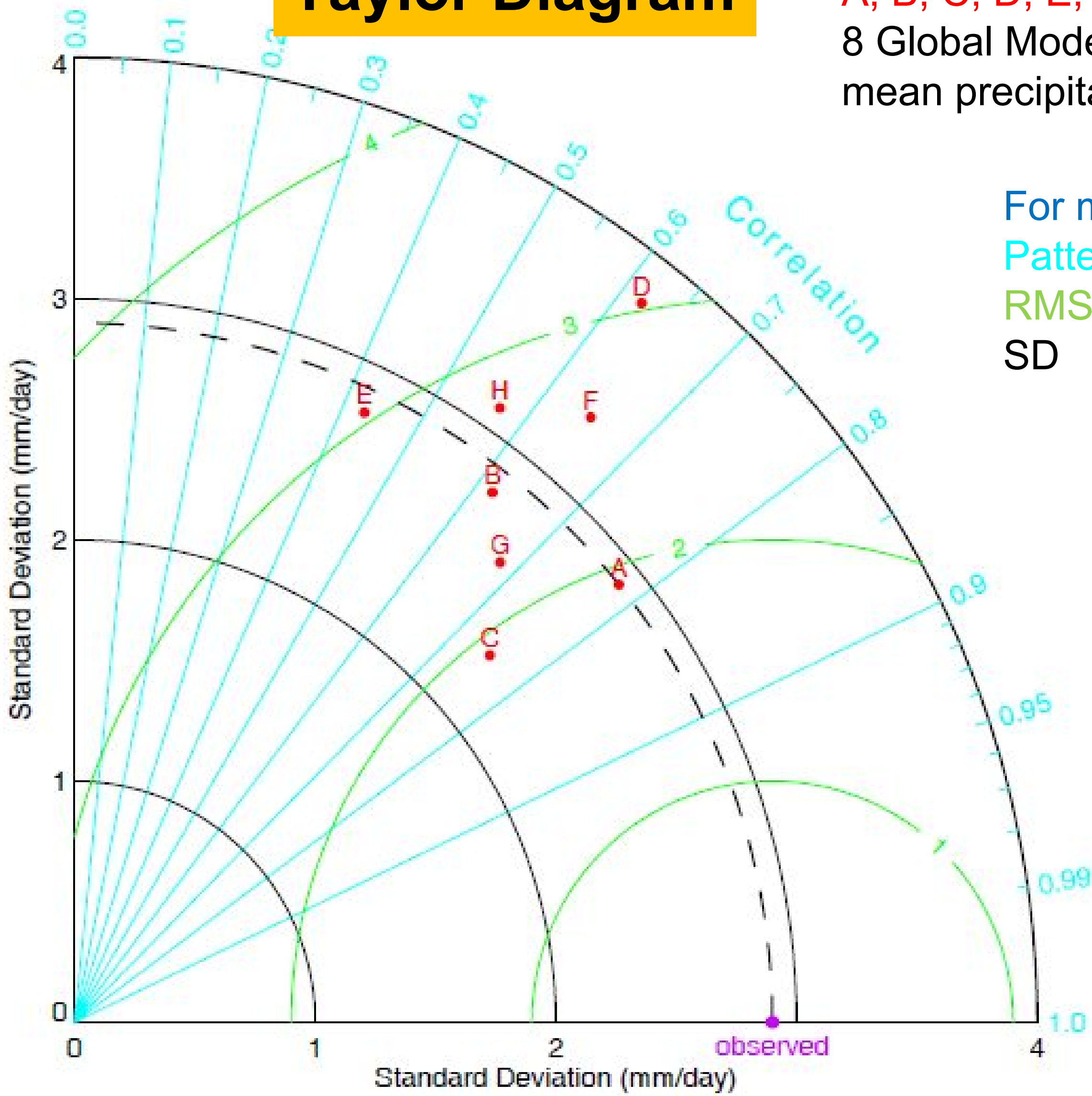
Taylor diagrams (Taylor, 2001) provide a way of graphically summarizing how closely a pattern (or a set of patterns) matches observations.

The similarity between two patterns is quantified in terms of
-their correlation,
-their centered root-mean-square difference and
-the amplitude of their variations (represented by their standard deviations).

These diagrams are especially useful in evaluating multiple aspects of complex models or in gauging the relative skill of many different models (e.g., IPCC, 2001).

-**Estatísticas são referidas como "padrão estatístico"**
-**Em geral :**
-**Caracteriza as relações estatísticas entre as saídas de modelos e observações-O diagrama não fornece informação sobre todos os Vieses-Caracteriza somente o erro padrão centrado -> variabilidade climática**

Taylor Diagram



A, B, C, D, E, F, G, H :
8 Global Models spatial pattern of annual mean precipitation

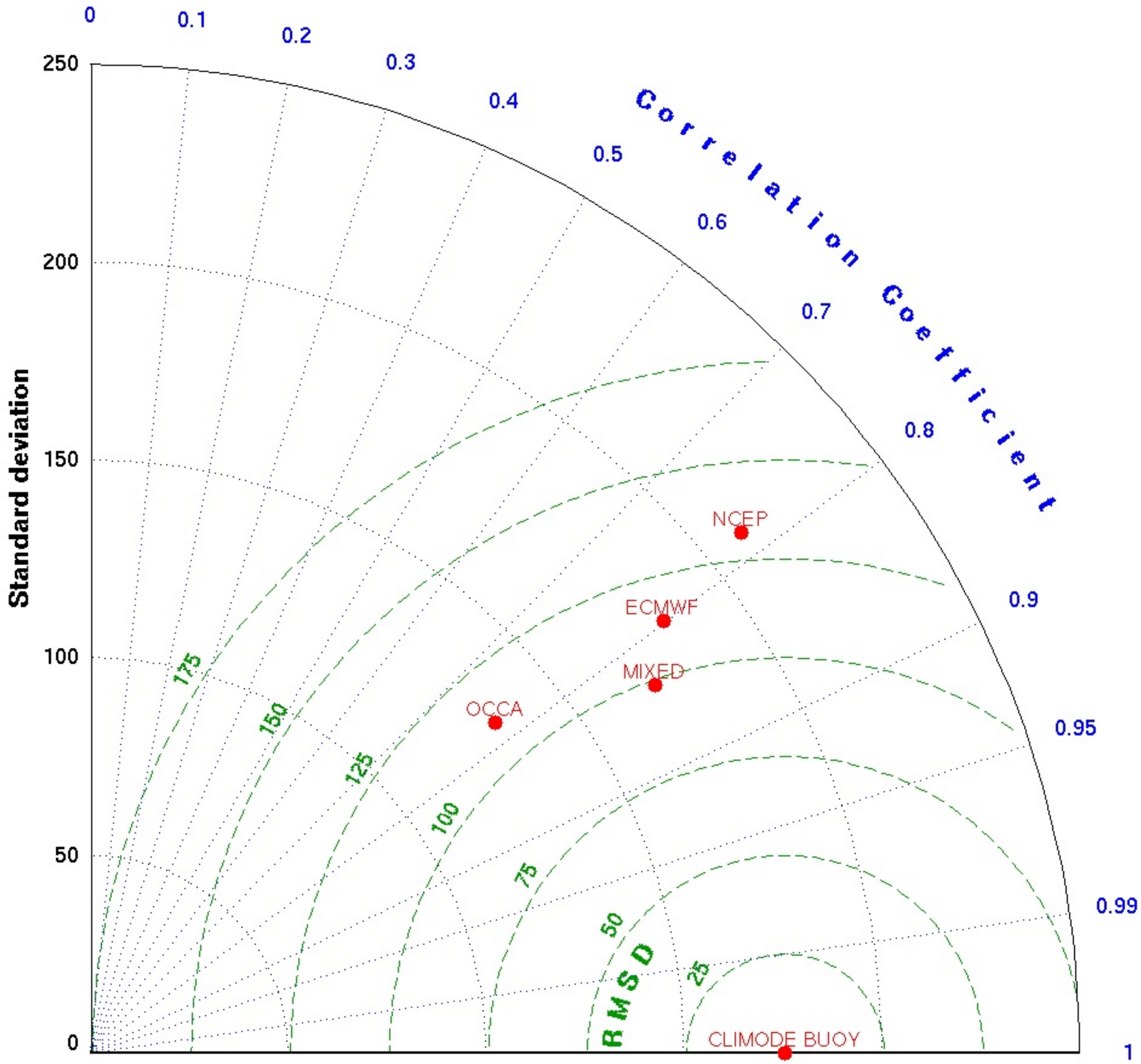
For model F:
Pattern correlation: : 0.65
RMS : 2.6 mm/d
SD : 3.3 mm/d

The means of the fields are subtracted out before computing these statistics,

so the diagram does not provide information about overall biases,

but solely characterizes the *centered pattern error.*

.: Sample Taylor diagram displaying a statistical comparison with observations of eight model estimates of the global pattern of annual mean precipitation.



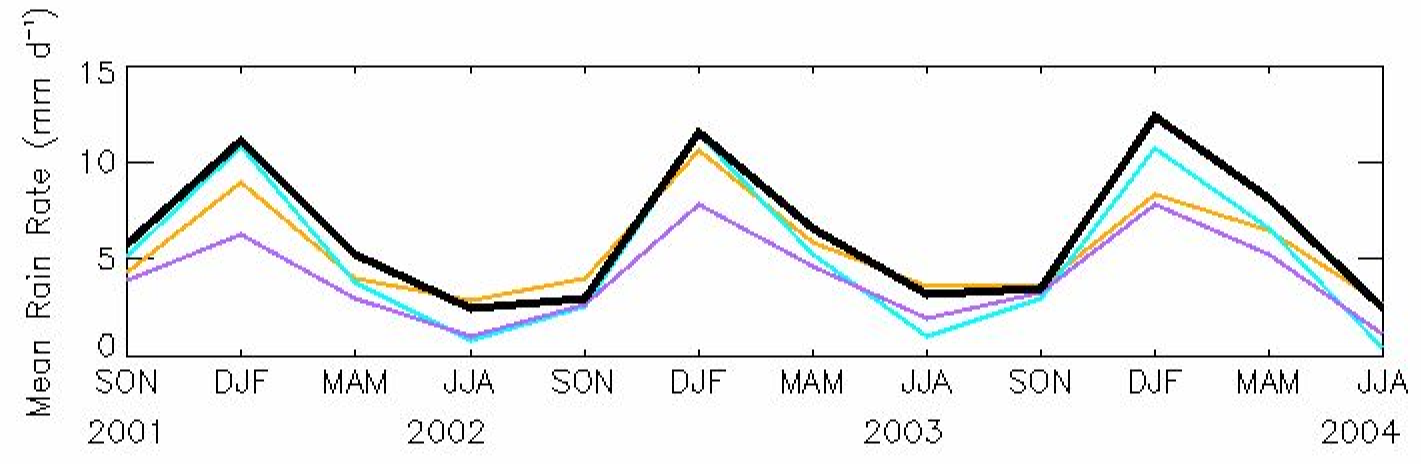
Correlation Coefficient
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.99
1
Standard deviation
250
200
150
100
50
0
RMSD
175
150
125
100
75
50
25
NCEP
ECMWF
MIXED
OCCA
CLIMODE BUOY

Simple diagnostic methods

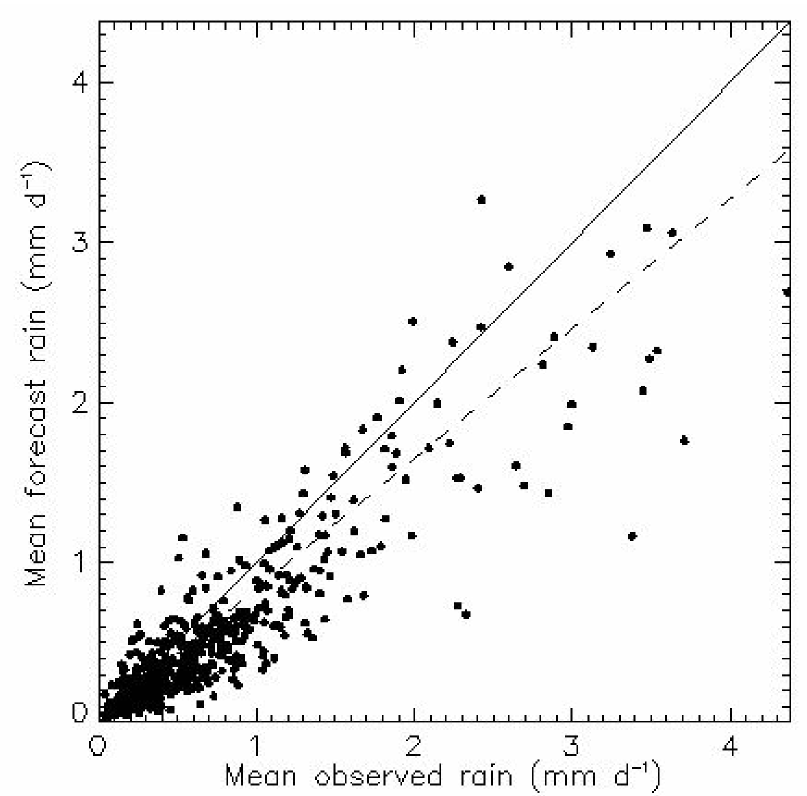
Maps of seasonal mean rainfall are highly recommended. Maps of the frequency of rainfall exceeding certain thresholds (for example, 1 mm d-1 and 10 mm d-1) are recommended.

Time series of observed and forecast domain mean rainfall allow us to see how well the temporal patterns are simulated by the model. **Time series of seasonal mean rainfall are highly recommended. Time series of mean rainfall for shorter time series are recommended. Time series of the seasonal frequency of rainfall exceeding certain thresholds (for example, 1 mm d-1 and 10 mm d-1) are recommended.**

A *scatter plot simply plots the forecast values against the observed values to show their correspondence.* The results can be plotted as individual points, or if there are a very large number, as a contour plot. **Scatter plots of forecast versus observed rain are highly recommended. Scatter plots of forecast error versus observed rainfall are recommended.**

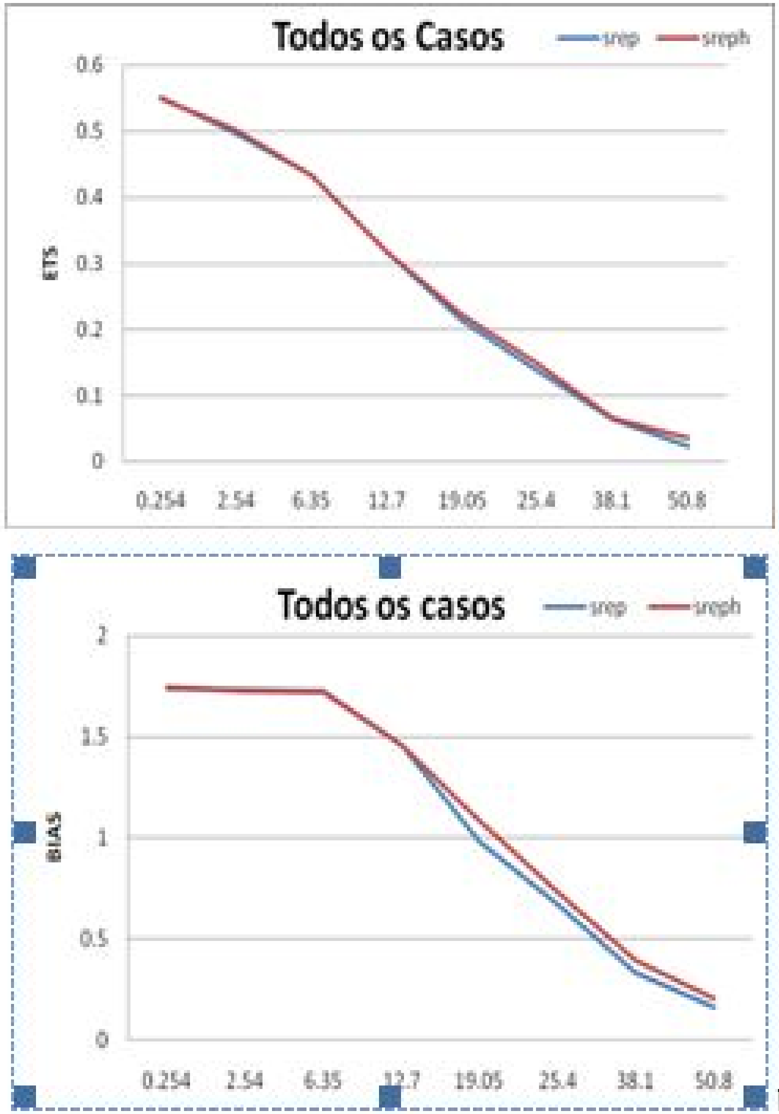


Seasonal time series of forecast and observed mean rainfall



Scatter plot of forecast versus observed rainfall. The dashed line shows the best fit to the data when normalized using a square root transformation

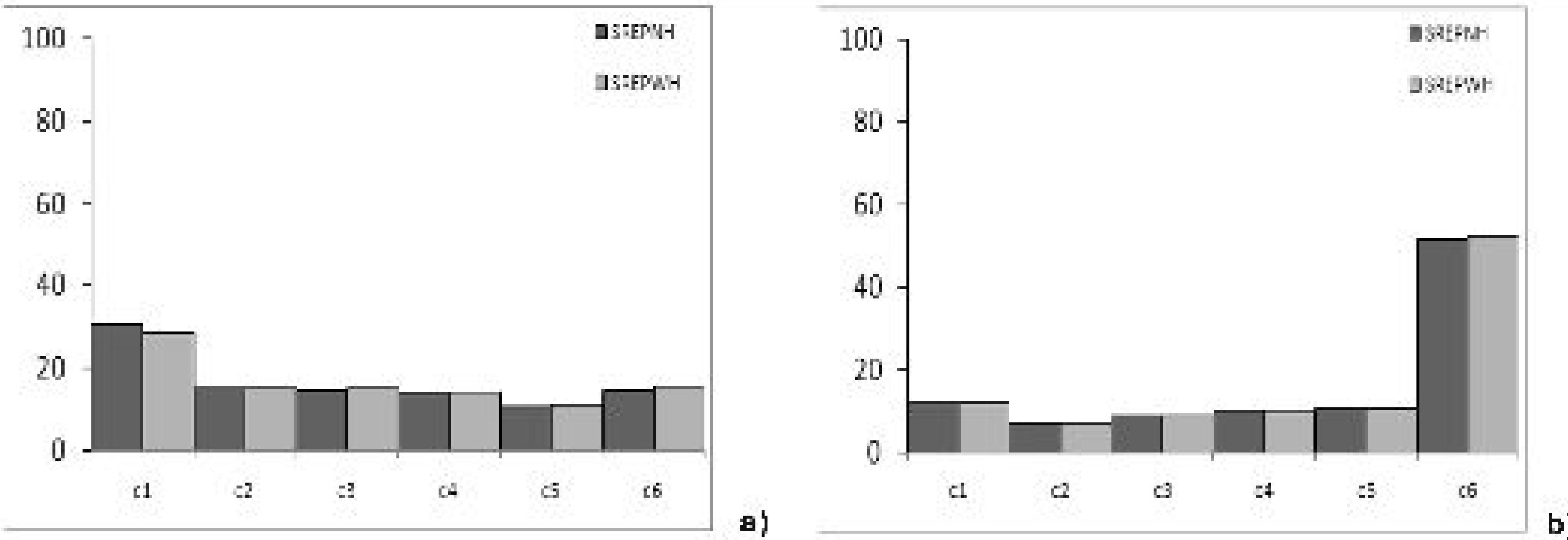
Bustamante e Chou,JGR, 2009



a)
b)

Figure 18: 24-hour accumulated precipitation from ensemble mean: a) ETS; b) BIAS.

Bustamante e Chou,JGR, 2009



ure 12: 850-hpa temperature Talagrand diagrams: a) 24-hour forecast; b) 144-hour forecast.

Table 1. Specifications of hindcast

Model type	Two-tiered method is used. The atmospheric model is TL95L40 version of the Global Spectral model used for a short- and medium-range forecast in JMA
Boundary conditions	SST: Combination of persisted anomaly, climate and prediction with the El Nino prediction model (atmosphere-ocean coupled model; CGCM) in JMA
Ensemble size and ensemble method	11 members. Singular vectors are used for atmospheric initial perturbation.
Training period	22 years from 1984 to 2005. Initial date is 10th of every month.
Forecast range	120 days

Mostrar pagina CPTEC –

Avaliacoes de modelo tempo: http://avaliacaodemodelos.cptec.inpe.br/

Previsoes sazonais: http://clima1.cptec.inpe.br/gpc/

Previsoes e verificacao

http://eurobrisa.cptec.inpe.br

WMO LRFSVS (Long-Range Forecast Standardised Verification System

http://www.bom.gov.au/cgi-bin/climate/wmo.cgi

http://www.wmo.int/pages/prog/www/DPS/LRF/ATTACHII-8SVSfrom%20WMO_485_Vol_I.pdf

References

WMO, 2002: Standardised Verification System (SVS) for Long-Range Forecasts (LRF). New attachment II-9 to the *Manual on the GPDS (WMO-No.485), Volume 1. Available on the internet at* http://www.wmo.ch/web/www/DPS/LRF-standardised-verif-sys-2002.doc

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences. An Introduction. Academic Press,* San Diego, 467 pp.

Taylor, K.E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* ***106, 7183-7192.***

JWGV, 2004: Forecast verification – Issues, methods, and FAQ. http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

Bougeault, P., 2002: WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. *CAS/JSC WGNE Report No. 18, Appendix C. Available on the* internet at http://www.wmo.ch/web/wcrp/documents/wgne18rpt.pdf.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlin. Proc. Geophys.,* ***8, 401-417.***