New Attachment II-9 to the
***Manual on the GDPS* (WMO-No. 485), Volume I**

Standardised Verification System (SVS)

for

Long-Range Forecasts (LRF)

Table of contents

Standardised Verification System (SVS)

for Long-Range Forecasts (LRF)

1. Introduction

The Commission for Basic Systems (CBS) of the World Meteorological Organisation (WMO) noted that there has been considerable progress in the development of long-range forecasting activities but that no comprehensive documentation of skill levels measured according to a common standard was available. It was noted that assessments of the scientific quality of long-range forecasts were not generally made available to users, apart from simple measures of skill and warning provided along with Internet products from some issuing Centres/Institutes.

Long-range forecasts are being issued from several Centres/Institutes and are being made available in the public domain. Forecasts for specific locations may differ substantially at times, due to the inherent limited skill of long-range forecast systems. The Commission acknowledged the scientific merit of those differences and encouraged the various approaches as a means to spur progress on the research front. However, concerns were raised that this situation tended to lead to confusion amongst users, and ultimately was reflecting back on the science behind long-range forecasts.

There was agreement on the need to have a more coherent approach to verification of long-range forecasts. The Commission agreed that its role was to develop procedures for the exchange of verification results, with a particular focus on the practical details of producing and exchanging appropriate verification scores.

This document presents the detailed specifications for the development of a Standardised Verification System (SVS) for Long-Range Forecasts (LRF) within the framework of a WMO exchange of verification scores. The SVS for LRF described herein constitutes the basis for long-range forecast evaluation and validation, and for exchange of verification scores. It will grow as more requirements are adopted.

2. Definitions

2.1 Long-Range Forecasts

LRF extend from thirty (30) days up to two (2) years and are defined in Table 1.

Table 1: Definition of long-range forecasts.

Monthly outlook:	Description of averaged weather parameters expressed as departures from climate values for that month.
Three-month or 90-day outlook:	Description of averaged weather parameters expressed as departures from climate values for that three-month or 90-day period.
Seasonal outlook:	Description of averaged weather parameters expressed as departures from climate values for that season.

Seasons have been loosely defined in the Northern Hemisphere as December-January-February (DJF) for Winter (Summer in the Southern Hemisphere), March-April-May (MAM) for Spring (Fall in the Southern Hemisphere), June-July-August (JJA) for Summer (Winter in the Southern Hemisphere) and September-October-November (SON) for Fall (Spring in the Southern Hemisphere). In the Tropical areas, seasons may have different definitions. Outlooks over longer periods such as multi-seasonal outlooks or tropical rainy season outlooks may be provided.

It is recognised that in some countries long-range forecasts are considered to be climate products.

This document is mostly concerned with the three-month or 90-day outlooks and the seasonal outlooks.

2.2 Deterministic Long-Range Forecasts

Deterministic LRF provide a single expected value for the forecast variable. The forecast may be presented in terms of an expected category (referred to as categorical forecasts, e.g. equiprobable terciles) or may take predictions of the continuous variable (non-categorical forecasts). Deterministic LRF can be produced from a single run of a Numerical Weather Prediction (NWP) model or a General Circulation Model (GCM), or can be produced from the grand mean of the members of an Ensemble Prediction System (EPS), or can be based on an empirical model.

The forecasts are either objective numerical values such as departure from normal of a given parameter or expected occurrences (or non-occurrences) of events classified into categories (above/below normal or above/near/below normal for example). Although equi-probable categories is preferred for consistency, other classifications can be used in a similar fashion.

2.3 Probabilistic Long-Range Forecasts

Probabilistic LRF provide probabilities of occurrences or non-occurrences of an event or a set of fully inclusive events. Probabilistic LRF can be generated from an empirical model, or produced from an Ensemble Prediction System (EPS).

The events can be classified into categories (above/below normal or above/near/below normal for example). Although equi-probable categories is preferred for consistency, other classifications can be used in a similar fashion.

2.4 Terminology

There is no universally accepted definition of forecast period and forecast lead time. However, the definition in Table 2 will be used in this document.

Table 2: Definitions of forecast period and lead time.

Forecast period:	Forecast period is the validity period of a forecast. For example, long-range forecasts may be valid for a 90-day period or a season.
Lead time:	Lead time refers to the period of time between the issue time of the forecast and the beginning of the forecast validity period. Long-range forecasts based on all data up to the beginning of the forecast validity period are said to be of lead zero. The period of time between the issue time and the beginning of the validity period will categorise the lead. For example, a Winter seasonal forecast issued at the end of the preceding Summer season is said to be of one season lead. A seasonal forecast issued one month before the beginning of the validity period is said to be of one month lead.

Figure 1 presents the definitions of Table 2 in graphical format.

Forecast range determines how far into the future LRF are provided. Forecast range is thus the summation of lead time and forecast period.

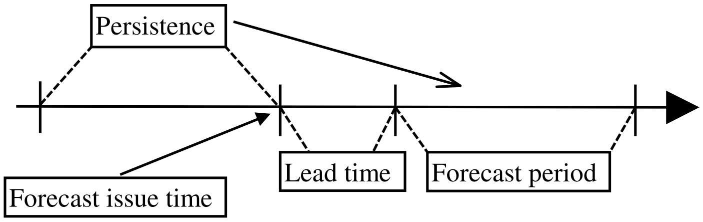

Figure 1: Definition of forecast period, lead time and persistence as applied in a forecast verification framework.

Persistence, for a given parameter, stands for persisting the anomaly which has been observed over the period of time with the same length as the forecast period and immediately prior to the LRF issue time (see Figure 1). It is important to realise that only the anomaly of any given parameter can be persisted. The persisted anomaly is added to the background climatology to retrieve the persisted parameter. Climatology is equivalent to persisting a uniform anomaly of zero.

3. SVS for Long-Range Forecasts

3.1 Parameters to be verified

The following parameters are to be verified:

a) Surface air temperature (T2m) anomaly at screen level
b) Precipitation anomaly
c) Sea surface temperature (SST) anomaly.

In addition to these three parameters, the Niño3.4 Index, defined as the mean SST anomaly over the Niño-3.4 region from 170°W to 120°W and from 5°S to 5°N all inclusive is also to be verified.

It is recommended that three levels of verification be done:

a) level 1: large scale aggregated overall measures of forecast performance (see section 3.1.1).
b) level 2: verification at grid points (see section 3.1.2).
c) level 3: grid point by grid point contingency tables for more extensive verification (see section 3.1.3).

Both deterministic and probabilistic forecasts are verified if available. Level 1 is applicable to T2m anomaly, Precipitation anomaly and Niño3.4 Index. Levels 2 and 3 are applicable to T2m anomaly, Precipitation anomaly and SST anomaly.

3.1.1 Aggregated verification (level 1)

Large scale verification statistics are required in order to evaluate the overall skill of the models and ultimately for assessing their improvements. These are bulk numbers calculated by aggregating verification at grid points and should not be used to assess regionalised skill. This aggregated verification is performed over three regions:

a) Tropics: from 20°S to 20°N all inclusive.
b) Northern Extra-Tropics: from 20°N to 90°N, all inclusive.
c) Southern Extra-Tropics: from 20°S to 90°S, all inclusive.

The verification of Niño3.4 Index is also part of level 1 verification.

3.1.2 Grid point verification (level 2)

The grid point verification is recommended for a regionalised assessment of the skill of the model. The appropriate way to make these verifications available is by visual rendering. The verification latitude/longitude grid is recommended as being 2.5° by 2.5°, with origin at 0°N, 0°E.

3.1.3 Contingency tables (level 3)

It is recommended to make available the raw verification material used for the grid point verification in section 3.1.2. This data is provided in contingency tables to allow users to perform more detailed verifications and generate statistics that are relevant for localised regions. The contingency tables are defined in sections 3.3.2 and 3.3.3. It is recommended to code all contingency tables at all grid points into a single file. Forecasts producers are required to provide a complete description of the format to ensure proper decoding of these contingency table files.

3.1.4 Summary of the Core SVS

The following gives a summary of what is part of the core SVS:

Level 1			
Parameters	**Verification regions**	**Deterministic forecasts**	**Probabilistic forecasts**
T2m anomaly Precipitation anomaly	Tropics Northern Extra-Tropics Southern Extra-Tropics (section 3.1.1)	MSSS (bulk number) (section 3.3.1)	ROC curves ROC areas Reliability diagrams Frequency histograms (sections 3.3.3 and 3.3.4)
Niño3.4 Index	N/A	MSSS (bulk number) (section 3.3.1)	ROC curves ROC areas Reliability diagrams Frequency histograms (sections 3.3.3 and 3.3.4)
Level 2			
Parameters	**Verification regions**	**Deterministic forecasts**	**Probabilistic forecasts**
T2m anomaly Precipitation anomaly SST anomaly	grid point verification on a 2.5° by 2.5° grid (section 3.1.2)	MSSS and its three term decomposition at each grid point in graphic representation number of forecast-observation pairs mean of observations and forecasts variance of observations and forecasts correlation of forecasts and observations (section 3.3.1)	ROC areas at each grid point in graphic representation (section 3.3.3)
Level 3			
Parameters	**Verification regions**	**Deterministic forecasts**	**Probabilistic forecasts**
T2m anomaly	grid point verification on	3 by 3 contingency tables	ROC reliability tables at

Precipitation anomaly SST anomaly	a 2.5° by 2.5° grid (section 3.1.2)	at each grid point (section 3.3.2)	each grid point (section 3.3.3)

The number of realisations of LRF is far smaller than in the case of short term numerical weather prediction forecasts. Consequently it is mandatory as part of the core SVS, to calculate and report error bars and level of significance (see section 3.3.5).

In order to ease implementation, participating LRF producers may stage the introduction of the core SVS according to the following priorities:

a) Verification at levels 1 and 2 in the first year of implementation
b) Verification at level 3 by the middle of the year following implementation of levels 1 and 2
c) Level of significance by the end of the year following implementation of levels 1 and 2.

Other parameters and indices to be verified as well as other verification scores can be added to the core SVS in future versions.

3.2 Verification strategy

LRF verification should be done on a latitude/longitude grid, and at individual stations or groups of stations representing grid boxes or local areas as defined in section 3.1.1. Verification on a latitude/longitude grid is performed separately from the one done at stations.

The verification latitude/longitude grid is recommended as being 2.5° by 2.5°, with origin at 0°N, 0°E. Both forecasts and the gridded verifying data sets are to be interpolated onto the same 2.5° by 2.5° grid.

In order to handle spatial forecasts, predictions for each point within the verification grid should be treated as individual forecasts but with all results combined into the final outcome. The same approach is applied when verification is done at stations. Categorical forecast verification can be performed for each category separately.

Similarly, all forecasts are treated as independent and combined together into the final outcome, when verification is done over a long period of time (several years for example).

Stratification of the verification data is based on forecast period, lead time and verification area. For example, seasonal forecast verification should be stratified according to season, meaning that verification results for different seasons should not be mixed. Forecasts with different lead times are similarly to be verified separately. It is also recommended to stratify verification according to warm and cold ENSO events (see Section 7 for definitions).

3.3 Verification scores

The following verification scores are to be used:

- Mean Square Skill Score (MSSS)
- Relative Operating Characteristics (ROC).

MSSS is applicable to deterministic forecasts only, while ROC is applicable to both deterministic and probabilistic forecasts. MSSS is applicable to non-categorical forecasts (or to forecasts of continuous variables), while ROC is applicable to categorical forecasts either deterministic or probabilistic in nature.

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 derivation of ROC is based on contingency tables giving the hit rate and false alarm rate for deterministic or probabilistic forecasts. The events are defined as binary, which means that only two outcomes are possible, an occurrence or a non-occurrence. It is recognised that ROC as applied to deterministic forecasts is equivalent to the Hanssen and Kuipers score (see section 3.3.2).

The binary event can be defined as the occurrence of one of two possible categories when the outcome of the LRF system is in two categories. When the outcome of the LRF system is in three (or more) categories, the binary event is defined in terms of occurrences of one category against the remaining ones. In those circumstances, ROC has to be calculated for each possible category.

3.3.1 MSSS for non-categorical deterministic forecasts

Let x_{ij} and f_{ij} (i=1,…,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:

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

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

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 (see section 3.4) 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:

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

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

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

where r_{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.

It is recommended that maps of the correlation, the ratio of the square roots of the variances, and the overall bias be produced for all forecast parameters and leads for each of the conventional seasons:

$Map: r_{fxj}, \frac{s_{fj}}{s_{xj}}, \left[\bar{f}_j - \bar{x}_j\right]$, all parameters, leads, and target months and seasons.

In addition to the bulk measures of MSSS and the maps of the three quantities just described, it is recommended that a table be produced for every parameter, lead, and target containing for every station or grid point j the following quantities:

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

As an additional standard against which to measure forecast set performance, cross-validated *damped persistence* (defined below) should be considered for certain forecast sets. A forecast of *ordinary persistence*, for a given parameter and target period, stands for the persisted anomaly (departure from cross-validated climatology) from a period immediately preceding the start of the lead time for the forecast period (see Figure 1). This period must have the same length as the forecast period. For example, the ordinary persistence forecast for a 90-day period made 15 days in advance would be the anomaly of the 90-day period beginning 105 days before the target forecast period and ending 16 days before. Ordinary persistence forecasts are never recommended as a standard against which to measure other forecasts if the performance or skill measures are based on squared error, like herein. This is because persistence is easy to beat in this framework.

Damped persistence is the optimal persistence forecast in a least squared error sense. Even *damped persistence should not be used in the case of extratropical seasonal forecasts*, because the nature of the interannual variability of seasonal means changes considerably from one season to the next in the extratropics. For all other cases damped persistence forecasts can be made in a cross-validated mode (Section 3.4) and the skill and performance diagnostics based on the squared error described above (bulk measures, maps, and tables) can be computed and presented for these forecasts.

Damped persistence is the ordinary persistence anomaly $x_{ij}(t-\Delta t)-\bar{x}_{ij}^{m}(t-\Delta t)$ damped (multiplied) towards climatology by the cross-validated, lagged product moment correlation between the period being persisted and the target forecast period.

Damped persistence forecast: $r_{\Delta,j}^{m}\left[x_{ij}(t-\Delta t)-\bar{x}_{ij}^{m}(t-\Delta t)\right]$

$$r_{\Delta,j}^{m}=\frac{\frac{1}{m}\sum_{m}\left[x_{ij}(t-\Delta t)-\bar{x}_{ij}^{m}(t-\Delta t)\right]\left[x_{ij}(t)-\bar{x}_{ij}^{m}(t)\right]}{s_{xj}^{m}(t-\Delta t)\,s_{xj}^{m}(t)}$$

where t is the target forecast period, t-Δt the persisted period (preceding the lead time), and m denotes summation (for $r_{\Delta,j}^{m}$, $\bar{x}_{ij}^{m}$, s_{xj}^{m}) at each stage of the cross-validation over all i except those being currently withheld (Section 3.4).

⇒ MSSS, provided as a single bulk number, is mandatory for level 1 verification in the core SVS. MSSS together with its three term decomposition are also mandatory for level 2 verification in the core SVS.

3.3.2 Contingency tables and scores for categorical deterministic forecasts

For two- or three-category deterministic forecasts it is recommended that full contingency tables be provided (digitally not graphically), because it is recognized that 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 month or season, and ENSO stratification (when appropriate) at every verification point for both the forecasts and (when appropriate) damped persistence. The definition of ENSO events is provided in Section 7.

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 3: 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 Table 3, $n_{i\bullet}$ 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 by users 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 points i (see discussion on Table 4). W_i is defined as:

$W_i = 1$ when verification is done at stations or at single grid points within a 10 degree box.

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

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

On a 2.5 degree 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.

Contingency tables such as the one in Table 3 are mandatory for level 3 verification in the core SVS.

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):

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

The sample probability distributions of forecasts and observations respectively then become

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

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

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

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

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

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 Hannssen 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 4 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 4, T is the grand sum of all the proper weights applied on each occurrence and non-occurrence of the events.

Table 4: 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 4 may be constructed from the 3X3 table described in Table 3 by summing the appropriate rows and columns.

In Table 4, O_1 represents the correct forecasts or hits:

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

(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:

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

(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:

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

(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:

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

(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 4.

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 5, 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 4):

$$HR = O_1 \big/ (O_1 + O_2)$$

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 (see Hanssen and Kuipers, 1965 and 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 3.3.2.2) 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 3) are mandatory for level 3 verification in the core SVS. 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.

3.3.3 ROC for probabilistic forecasts

Tables 5 and 6 show contingency tables (similar to Table 4) that can be built for probabilistic forecasts of binary events.

Table 5: 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 5,

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.

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

(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 or stations.

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

(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.

Table 6: 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=N	O_1	NO_1
2	F=1, NF=N-1	O_2	NO_2
3	F=2, NF=N-2	O_3	NO_3
•••		•••	•••
n	F=n-1, NF=N-n+1	O_n	NO_n
•••		•••	•••
N+1	F=N, NF=0	O_{N+1}	NO_{N+1}

In Table 6,

n = number of the nth bin; n goes from 1 to N+1.
N = number of members in the ensemble.
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 6, 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-1} and the upper limit is P_n. 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 5 and 6 outline typical contingency tables. It is recommended that the number of probability bins remain between 9 and 20. The forecast providers can bin according to percent thresholds (Table 5) or ensemble members (Table 6) as deemed necessary. Table 6 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 5 and 6). The hit rate for probability threshold P_n (HR_n) is defined as (referring to Tables 5 and 6):

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

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. The area under the ROC curve can be calculated using the Trapezium rule. Although simple to apply, the Trapezium rule renders the ROC score dependent on the number of points on the ROC curve, and care should be taken in interpreting the results. Other techniques are available to calculate the ROC score (see Mason, 1982).

⇒ Contingency tables for probabilistic forecasts (such as in Tables 5 and 6) are mandatory for level 3 verification in the core SVS. ROC curves and ROC areas are mandatory for level 1 verification in the core SVS while ROC areas only are mandatory for level 2 verification in the core SVS.

3.3.4 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 aggregated over the tropics and, separately, the two extratropical hemispheres. 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.

⇒ Reliability diagrams and frequency histograms are mandatory for level 1 verification in the core SVS.

3.3.5 Level of significance

Because of the increasing uncertainty in verification statistics with decreasing sample size, significance levels and error bars should be calculated for all verification statistics. Recommended procedures for estimating these uncertainties are detailed below.

ROC area

In certain special cases the statistical significance of the ROC area can be obtained from its relationship to the Mann–Whitney U-statistic. The distribution properties of the U-statistic can be used only if the samples are independent. This assumption of independence will be invalid when the ROC is constructed from forecasts sampled in space because of the strong spatial (cross) correlation between forecasts (and observations) at nearby grid-points or stations. However, because of the weakness of serial correlation of seasonal climate anomalies from one year to the next, an assumption of sequential independence may frequently be valid for long-range forecasts, and so Mann–Whitney U-statistic may be used for calculating the significance of the ROC area for a set of forecasts from a single point in space. An additional assumption for using the Mann–Whitney U-test is that the variance of the forecast probabilities (not that of the individual ensemble predictions per se) for when non-events occurred is the same as those for when events occurred. The Mann–Whitney U-test is, however, reasonably robust to violations of homoscedasticity which means that the variance of the error term is constant across the range of the variable, and so significance tests in cases of unequal variance are likely to be only slightly conservative.

If the assumptions for the Mann–Whitney U-test cannot be held, the significance of the ROC area should be calculated using randomisation procedures. Because the assumptions of permutation procedures are the same as those of the Mann–Whitney U-test, and because standard bootstrap procedures assume independence of samples, alternative procedures such as moving block bootstrap procedures (Wilks, 1997) should be conducted to ensure that the cross- and/or serial-correlation structure of the data is retained.

ROC curves

Confidence bands for the ROC curve should be indicated, and can be obtained either by appropriate bootstrap procedures, as discussed above, or, if the assumption of independent forecasts is valid, from confidence bands derived from a two-sample Kolmogorov-Smirnov test comparing the empirical ROC with the diagonal.

MSSS

Appropriate significance tests for the MSSS and the individual components of the decomposition again depend upon the validity of the assumption of independent forecasts. If the assumption is valid, significance tests could be conducted using standard procedures (namely the F-ratio for the correlation and for the variance ratio, and the t-test for the difference in means), otherwise bootstrap procedures are recommended.

⇒ Level of significance is mandatory in the core SVS. A phased in introduction of level of significance in the SVS may be used (see section 3.1.4).

3.4 Hindcasts

In contrast to short- and medium-range dynamical Numerical Weather Prediction (NWP) forecasts, LRF 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 LRF 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 LRF system. LRF 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. Because of verification data availability, it is recommended to do hindcast over the period from 1981 to present. If data is available, it is recommended to extend the period back to 1971.

Verification in hindcast mode should be achieved in a form as close as possible to the real time operating mode in terms of resolution, ensemble size and parameters. In particular dynamical/empirical models must not make any use of future data. Validation of empirical models, dynamical models with postprocessors (including bias corrections), and calculation of period of verification means, standard deviations, class limits, etc. must be done in a cross-validation framework. Cross-validation allows the entire sample to be used for validation (assessing performance, developing confidence intervals, etc.) and almost the entire sample for model and post-processor building and for estimation of period of verification climatology. Cross-validation proceeds as follows:

1. Delete 1, 3, 5, or more years from the complete sample;
2. Build the statistical model or compute the climatology;
3. Apply the model (e.g. make statistical forecasts or postprocess the dynamical forecasts) or the climatology for one (usually the middle) year of those deleted and verify;
4. Replace the deleted years and repeat 1-3 for a different group of years;
5. Repeat 4 until the hindcast verification sample is exhausted.

Ground rules for cross–validation are that every detail of the statistical calculations be repeated, including redefinition of climatology and anomalies, and that the forecast year predictors and predictands are not serially correlated with their counterparts in the years reserved for model building. For example, if adjacent years are correlated but every other year is effectively not, three years must be set aside and forecasts made only on the middle year (see Livezey, 1999, for estimation of the reserved window width).

The hindcast verification statistics should be updated once a year based on accumulated forecasts.

⇒ Verification results over the hindcast period are mandatory for the exchange of LRF verification scores.

3.5 Real-time monitoring of forecasts

It is recommended that there be regular monitoring of the real time long range forecasts. It is acknowledged that this real-time monitoring is neither as rigorous nor as sophisticated as the hindcast verification; nevertheless it is necessary for forecast production and dissemination. It is also acknowledged that the sample size for this real-time monitoring may be too small to assess the overall skill of the models. However, it is recommended that the forecast and the observed verification for the previous forecast period be presented in visual format to the extent possible given the restrictions on availability of verification data.

4. Verification data sets

The same data should be used to generate both climatology and verification data sets, although the forecasts issuing Centres/Institutes own analyses or ECMWF reanalyses and subsequent operational analyses may be used when other data are not available. Use of NCEP reanalysis data is also another option.

Many LRF are produced that are applicable to limited or local areas. It may not be possible to use the data in either the recommended climatology or verification data sets for validation or verification purposes in these cases. Appropriate data sets should then be used with full details provided.

It is recommended to use:

1. UKMO/CRU for Surface air temperature anomaly at screen level (T2m).
2. Xie-Arkin and/or GPCP for Precipitation anomaly.
3. Reynolds OI for Sea surface temperature (SST) anomaly. Prior to 1981, the reconstructed SST database using EOF of Smith et al, 1996 can be used.

4.1 Status of the verification data sets

The following paragraphs give the status of the various proposed verification data sets:

4.1.1 Xie-Arkin

Availability:	• NOAA
Period:	• 1979-1998.
Type:	• Rain gauges, satellites and model precipitation amount values. • Choice of grids with missing values in the polar regions or completed with model data. • Monthly means.
Grid:	• 2.5° by 2.5°
Update frequency:	• Every 3 to 6 months.
Climatology:	• None.
Reference:	• Xie, Pingping, Phillip A. Arkin, 1997: Global Precipitation: A 17-Year Monthly Analysis Based on Gauge Observations, Satellite Estimates, and Numerical

	Model Outputs. Bulletin of the American Meteorological Society: Vol. 78, No. 11, 2539–2558.
Web site:	• http://www.cdc.noaa.gov/cdc/data.cmap.html

4.1.2 GPCP

Availability:	• NASA
Period:	• 1987-1999.
Type:	• Similar to Xie-Arkin data.
Grid:	• 2.5° by 2.5°
Update frequency:	• Unknown.
Climatology:	• None.
Reference:	• Huffman, George J., Robert F. Adler, Philip Arkin, Alfred Chang, Ralph Ferraro, Arnold Gruber, John Janowiak, Alan McNab, Bruno Rudolf, Udo Schneider, 1997: The Global Precipitation Climatology Project (GPCP) Combined Precipitation Dataset. Bulletin of the American Meteorological Society: Vol. 78, No. 1, 5–20.
Web site:	• http://orbit-net.nesdis.noaa.gov/arad/gpcp/

4.1.3 UKMO/CRU

Availability:	• UKMO/Hadley Centre
Period:	• 1851-1998.
Type:	• Monthly surface air temperature (T2m) anomalies from 1961-1990 climate.
Grid:	• 5° by 5°
Update frequency:	• Monthly.
Climatology	• 1961-1990.
Reference:	• Jones, P. D., M. New, D. E. Parker, S. Martin and I. G. Rigor, 1999: Surface air temperature and its changes over the past 150 years. Rev. Geophys., 37, 173-199.
Web site:	• http://www.cru.uea.ac.uk/cru/data/temperature/

These data sets are available for use in scientific research upon the signing of a short license agreement.

4.1.4 Reynolds OI

Availability:	• NOAA/CDC
Period:	• 1981-1998.
Type:	• Weekly or monthly sea surface temperature (SST) means.
Grid:	• 1° by 1° • 2° by 2°
Update frequency:	• 2-4 times a year.
Climatology:	• None.
Reference:	• Reynolds, R. W. and T. M. Smith, 1994: Improved global sea surface temperature analyses using optimum interpolation. J. Climate, 7, 929-948. • Smith M. T., R. W. Reynolds, R. E. Livezey and D. C. Stokes, 1996: Reconstruction of Historical Sea Surface Temperatures Using Empirical Orthogonal Functions, Journal of Climate, 1403-1420.
Web site:	• http://www.cdc.noaa.gov/cdc/data.reynolds_sst.html

5. System details

Information must be provided on the system being verified. This information should include (but is not restricted to):

1. Whether the system numerical, empirical or hybrid.
2. Whether the system is deterministic or probabilistic
3. Model type and resolution.
4. Ensemble size.
5. Boundary conditions specifications.
6. List of parameters being assessed.
7. List of regions for each parameter.
8. List of forecast ranges (lead times) and periods for each parameter.
9. Period of verification.
10. The number of hindcasts or predictions incorporated in the assessment and the dates of these hindcasts or predictions.
11. Details of climatological and verification data sets used (with details on quality control when these are not published).
12. If appropriate, resolution of fields used for climatologies and verification.

Verification data for the aggregated statistics and the grid point data should be provided on the web. The contingency tables should be made available by the web or anonymous FTP. The Lead Centre will take responsibility for defining a common format for displaying the verification scores. Real-time monitoring should be done as soon as possible and made available on the web.

6. References

Gerrity, J. P. Jr., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev*., 120, 2707-2712.

Hanssen A. J. and W. J. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Koninklijk Nederlands Meteorologist Institua Meded*. Verhand, 81-2-15.

Livezey, R. E., 1999: Chapter 9: Field intercomparison. *Analysis of Climate Variability: Applications of Statistical Techniques*, H. von Storch and A. Navarra, Eds, Springer, pps. 176-177.

Mason I., 1982: A model for assessment of weather forecast. *Australian Met. Magazine*, 30, 291-303.

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

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **16**, 2417-2424.

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

Smith M. T., R. W. Reynolds, R. E. Livezey and D. C. Stokes, 1996: Reconstruction of Historical Sea Surface Temperatures Using Empirical Orthogonal Functions, *J. Climate*, 1403-1420.

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 358, 114 pp.

Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate*, 10, 65-92.

7. Definition of ENSO events.

The following table gives the definition of the ENSO events. The following list of cold (La Niña) and warm (El Niño) episodes has been compiled to provide a season-by-season breakdown of conditions in the tropical Pacific. The data underlying the following table have been taken from NOAA/NCEP/CPC at www.cpc.ncep.noaa.gov and have been subjectively interpolated to fit the conventional seasons DJF, MMA etc.

Years	DJF	MAM	JJA	SON
1950	C	C	C	C
1951	C	N	N	N
1952	N	N	N	N
1953	N	N	N	N
1954	N	N	N	C
1955	C	N	N	C
1956	C	C	C	N
1957	N	N	N	W
1958	W	W	N	N
1959	N	N	N	N
1960	N	N	N	N
1961	N	N	N	N
1962	N	N	N	N
1963	N	N	N	W
1964	N	N	N	C
1965	N	N	W	W
1966	W	N	N	N
1967	N	N	N	N
1968	N	N	N	N
1969	W	N	N	N
1970	N	N	N	C
1971	C	N	N	N
1972	N	N	W	W
1973	W	N	N	C
1974	C	C	N	N
1975	N	N	C	C
1976	C	N	N	N
1977	N	N	N	N
1978	N	N	N	N
1979	N	N	N	N
1980	N	N	N	N
1981	N	N	N	N
1982	N	N	W	W
1983	W	W	N	N
1984	N	N	N	N
1985	N	N	N	N
1986	N	N	N	W
1987	W	W	W	W
1988	N	N	N	C
1989	C	N	N	N
1990	N	N	N	N

1991	**N**	**N**	**W**	**W**
1992	**W**	**W**	**N**	**N**
1993	**N**	**W**	**W**	**N**
1994	**N**	**N**	**W**	**W**
1995	**W**	**N**	**N**	**N**
1996	**N**	**N**	**N**	**N**
1997	**N**	**W**	**W**	**W**
1998	**W**	**W**	**N**	**C**
1999	**C**	**C**	**N**	**C**
2000	**C**	**N**	**N**	**N**
2001	**N**	**N**	**N**	**N**