COMMISSION FOR BASIC SYSTEMS

MEETING OF EXPERT TEAM TO DEVELOP A VERIFICATION SYSTEM FOR LONG-RANGE FORECASTS

(Montreal, Canada, 22 - 26 April 2002)

FINAL REPORT

WORLD METEOROLOGICAL ORGANIZATION
April 2002

TABLE OF CONTENTS

1. OPENING OF THE MEETING

1.1 At the kind invitation of Canada, the meeting of the Expert Team to develop a verification system for long-range forecasts was held at the Canadian Meteorological Centre in Montreal Canada, 22-26 April 2002. Mr Terry Hart, Chair of the Team opened the meeting, welcomed the participants, particularly the representatives of CCl, CLIVAR and CAS. He outlined the work expected of the meeting.

1.2 Ms Angele Simard, Director of Informatics of the CMC and Chair of the OPAG on Data Processing Forecasting Systems extended a warm welcome to the participants. She informed the meeting the CMC activities in long range forecasting and verification and gave guidance on the expectation of the OPAG on DPFS as to the focus of activities the Team was expected to achieve as deliverables before CBS-Ext. 2002. She noted that although the activities envisaged in the action plan for development of SVS for LRF and experimental exchange of scores including lead centre activities were behind schedule, notable achievements had been realised and some centres were making available verification scores at their web sites at least for their national products. She urged the Team to further review and adjust in the light of experience the SVS and enhance implementation actions to assure appropriate recommendations to CBS-Ext. 2002.

1.3 Ms Angele Simard informed the meeting of progress achieved in the activities of the ICTT on regional climate centres. In particular with regards to assignment of responsibilities where the CBS verification team was to develop and implement verification schemes for SI forecasts in collaboration with CAS while the CCl team provide leadership in the development and implementation of verification of post processed products to end users.

1.4 Mr Morrison Mlaki, Chief Data Processing System Division, on behalf of Professor G.O.P Obasi, Secretary General of WMO, thanked Ms Simard for her warm welcome and Canada for hosting this meeting in the magnificent facilities of the CMC. He noted major tasks: -

- Coordinate experimental exchange of scores and to evaluate results;
- Recommend operational practices on validation results to be obtained on LRF products;
- Consult other Commissions on improvements and future work and recommend operational procedures.

1.5 Mr Mlaki noted that the meeting was to review progress being achieved in the implementation of the plan for exchange of scores and adjust the procedures in response to identified difficulties. It was also to review the scope of the experimental standards and procedures, and recommend definitive standards and procedures for presentation and exchange of verification scores, to be adopted by CBS-Ext-02. He invited the meeting to consider any crosscutting issues, with respect to the responsibility of the CCI Group on Verification, to ensure that the proposed solutions of the SVS for LRF converge with and can be consistently extended or extrapolated to end user verification solutions. He noted that a divergence in verification procedures at the level of LRF producing Centres and at the level of downscaled end user products was not desirable, as it would not facilitate inter-comparison of trends in improvements of forecast skill, as forecast systems improve.

1.6 Mr Mlaki invited the team to consider the issue of the indication of the level of confidence of long-range forecast products for the benefit of RCCs, NMHSs and other users. He noted that there is now ample evidence that the full global, regional and national long-range forecast systems, coupled with a standard verification system and the presentation on the products of a level of confidence of the forecasts, would enhance the utility and economic value of long-range forecasts.

2. ORGANIZATION OF THE MEETING

2.1 The agreed Agenda of the meeting is given in Appendix III.

2.2 The meeting agreed on its working arrangements. There were 11 participants as given in Appendix IV.

3. EXCHANGE OF LONG-RANGE FORECASTS VERIFICATION SCORES

3.1 Background

3.1.1 The meeting noted that a major growth area in meteorological services has been in the provision of long-range forecasts, originating from national meteorological services and other organisations. Several WMO Programmes have an interest in the coordination of these activities in support of NMHSs. In all programmes one recurring topic has been the need for standardised verification of long-range forecast products. Since the number of instances of seasonal forecasts is low, the major focus needs to be on performance in *hindcast* mode rather than ongoing predictions.

3.1. 2 CBS-XI in 1996 gave its Working Group on Data Processing the task of building on adopted guidelines and principles to develop standard verification procedures and to organise intercomparisons of such scores. Around the same time (1997), the CCI and CAS also recognised the need for exchange of quality assessment information as a vital component of long-range prediction activity. The CAS Working Group on Medium and Long Range Weather Forecasting Research considered the matter in detail and made recommendations on specific measures and procedures.

3.1.3 At a meeting of a GDPS team in June 1998 in Washington work was started on developing some procedures for the exchange of verification scores building on the work of CAS with a particular focus on the practical details of producing and exchanging appropriate scores. One session of that meeting was a joint session with representatives of CCl. In that session the CCl/CLIPS representatives confirmed their interest in the exchange of verification results. The meeting noted that there are particular problems to LRF that make the exchange of verification different to verification of NWP. In the latter we have 30 realizations of forecasts over a month, but in LRF we have one realization every month at most. The whole process takes longer to assess systems and the measures used for NWP are not necessarily useful in the context of LRF. The meeting recommended that a pilot exchange of scores be commenced using a limited number of scores and measures, with highest priority being given to the two scores recommended by the CAS group, but with the more comprehensive CBS list as additional measures to be considered. It was noted that a distinction should be made between forecasts (or predictions) that are non-interpreted results from dynamical/ statistical models and value-added forecasts that are resulting from the interpretation of possibly several sources of prediction, adapted to the special needs of well-defined categories of users

3.1. 4 CBS in October 1998 endorsed the approach and as part of the new CBS structure, an Expert Team on a Standardised Verification System (SVS) for Long-range Forecasts was established under the leadership of Francois Lalaurette of ECMWF under the new OPAG on Data-processing and Forecast Systems. This SVS team met in August 1999. It developed specific proposals on the forecast parameters, method of calculation, definition of terms and identified verification data sets. It adopted a template prepared by Richard Verret of Canada and set a timetable for the initiation of an exchange of statistics.

3.1. 5 CBS-XII in November 2000 adopted the recommendations of the team and re-established the team with some changes in membership. The team commenced its work by correspondence prior to this meeting. CBS designated WMC Melbourne as the lead GDPS centre on implementation and operational aspects of the verification system for LRF. It included provision of a centralised web site function.

3.1.6 Significant related activity is taking place under the work of other Commissions. The CAS Working Group on Medium and Long Range Weather Forecasting Research has been disbanded and the activity on long-range forecasting moved under the World Weather Research Programme. The CLIVAR Working Group on Seasonal to Interannual Prediction is very active in research topics concerned with the development of the science of long-range forecasting.

3.1.7 At the end of 2001 CCl established a team on verification as part of the CLIPS program. Its terms of reference were focused on end-user requirements and are complementary to those of the CBS team although there is some overlap. One of the tasks for both teams is to coordinate activities with other relevant groups.

3.1.8 The Inter-Commission Task Team on Regional Climate Centres has been developing a framework for the infrastructure for long-range forecasting. The work of the ICTT has clarified the respective roles of the Commissions and assigned responsibilities. In particular, the roles of CBS are to:

- provide the infrastructure needed for the operational production and exchange of SI forecasts and related services;
- define and implement standards and formats for SI production and exchange;
- develop and implement verification schemes for SI forecasts in collaboration with CAS;
- make verifications available regularly to RCCs and NMHSs;
- monitor scientific progress with a view to improving SI forecast services;
- provide up-to-date information on methodology used by the RSMCs with specialization on long-range forecasting;
- collaborate with CCl and CAS (and representatives of CLIVAR) on matters of common interest.

3.1.9 The ICTT has clarified the question of the “users” whom the SVS is expected to serve. The aim of the SVS is to provide estimates of the skill of the global-scale products in support of NMHSs and RCCs in their use of these products to provide long-range forecasts to end users.

3.1.10 CBS has also appointed an Expert Team on Infrastructure for Long-range Forecasting to progress those aspects relevant to CBS. The team includes representatives from most major modelling centres. The meeting of the team in November 2001 was also attended by the Chair of the ICCTT on RCCs (Stefan Mildner). The meeting refined a list of requirements prepared by the ICTT and proposed a set of products that could be expected from the Global-scale Producing Centres in support of NMHSs and Regional Climate Centres.

3.1.11 These were mainly global-scale products for fields:

- 2 metre temperature over land;
- sea surface temperature;
- precipitation;
- 500 hPa geopotential height; MSLP
- 850 hPa temperature.

as well as some specific index values such as SST for the Nino regions and SOI.

3.1.12 The verification Team therefore used this set of products as the basis for a selection of products from the Global-scale Producing Centres in support of NMHSs and Regional Climate Centres to be verified.

3.1.13 Also relevant to the work of the Team is the report of the meeting of the CBS Expert Team on Ensemble Prediction Systems held in Tokyo in October 2001. The types of ensemble products for short- and medium-range forecasting are very similar in essence to those produced for long-range forecasts. The verification Team noted the report of the EPS team with a view to harmonising its recommendations with those of the EPS team.

3.2 Current status

3.2.1 The meeting evaluated the current status against the planned operations and timetable for implementation and developed specific conclusions. It considered the current status of implementation of the verification system including lead centre role. At this stage there was no coordinated exchange of verification statistics, and the later stages of the timetable set out by the previous team was not being met. Templates and procedures had been prepared by CMC and were available on the WMO web site. Some other tasks of the team have been completed such as definition of the recommended verification data sets.

3.2. 2 The Bureau of Meteorology had not progressed its Lead Centre role to the point where verification results from a range of global centres are available on a single web site. However, the National Climate Centre in Melbourne had a verification site using the recommended templates and based on the SVS procedures. At the time of the report it was a registered user, password-protected site. It could be viewed at:

http://www.bom.gov.au/reguser/by_user/bomw0278/lrfvs/

(For password details contact t.hart@bom.gov.au)

3.2.3 The experience in Melbourne of creating the web site suggested that the templates and SVS procedures were suitable for describing the prediction system and for providing a limited number of statistics, but became cumbersome for the full range of statistics, even using automated procedures to populate the tables. It was found that a more straightforward measure such as *per cent consistent* was more meaningful to end-users. It also allowed results for all forecast categories to be combined. The presentations of results in maps were also desirable by users. Another web site, aimed at users of long-range forecast products was funded by a range of agricultural organisations can be seen on a public web site at:

http://www.bom.gov.au/silo/products/verif/

3.2.4 The Meteorological Service of Canada had previously adopted a similar approach, using maps over Canada and scores such as *per cent correct*. Their products can be seen at:

http://weatheroffice.ec.gc.ca/saisons/index_e.html - verification

3.2.5 These are just two examples of the developing availability of verification results. Other centres (including ECMWF, Met Office UK, Japan and China) are in the process of developing and providing such statistics following the SVS guidelines. The Met Office is expected to complete generation of the SVS standard scores evaluated over a 19 year period (1982-2000) in autumn 2002. In the European DEMETER project verification over a hindcast set (30 years, 1969-1998) is being performed and is expected to be completed in spring 2003.

3.3 Current experimental verification system

3.3.1 The meeting reviewed the current experimental verification system for long-range forecasts and refined the procedures. This focused in particular on the scope and capability of centres to implement them operationally and recommend operational procedures and practices to be followed, in terms of the information on validation results to be attached to the long-range forecasts products. It considered and addressed some concerns over:

(a) The practicality of the proposals particularly in relation to the number of scores to be computed
(b) The relevance and usefulness to forecasters and other users of global-scale long-range forecast products
(c) The impact on differences in verification data sets.

3.3.2 The meeting proposed revised procedures as the basis for operational use. In the proposals the meeting addressed some inconsistencies between the specifications for the experimental SVS and other procedures. These related to the definition of the tropics raised at CBS-XII and the specifications for presenting contingency tables in the proposals from ET on Ensemble Prediction Systems. In presenting the reliability tables the ET on EPS have recommended to present the table according to the number of ensemble members in each class while the current SVS proposal uses probability threshold bins. Some information is lost in this process. However, there is a problem in the ET on EPS proposal when the number of ensemble members is large. In the revised proposal presented here the ET proposal is adopted up to a suggested maximum number of bins. The meeting recommended that the Chair of the ET contact the Chair of the ET on EPS to ensure complete uniformity in the presentation of reliability tables.

3.3.3 The recommendations for revisions to the verification scheme are provided in Section 4. The Team views implementation of the proposals as a developing process that can occur as part of the operational implementation of LRF systems.

4. OPERATIONAL PROCEDURES AND PRACTICES

4.1 Definition of user and user requirements

4.1.1 The meeting recognized three main groups that require verification information of long-range forecasts, and that the requirements of these groups may differ substantially. These three groups can be identified as model developers, forecasters, and end-users. The meeting focused on making recommendations for addressing the requirements of the forecasters for verification information, but recognized that these requirements cannot be considered in isolation of the interests of the other groups.

4.1.2 The primary focus of the model developers (whether dynamical or statistical) is to detect any systematic errors and weaknesses in their models, and to identify reasons for these model imperfections. Their demand is consequently for detailed diagnostics of model performance, which can only be provided by a comprehensive analysis of a range of verification measures that address the multifaceted nature of forecast quality. The focus of the meeting, however, was on the requirements of the forecasters, whose responsibility it is to combine the predictions from the models available to him/her.

4.1.3 In most cases the forecaster will be operating at the level of the RCC or NMHS, and will be responsible for producing an official forecast statement that is to be disseminated to end-users. The objective of the forecaster is to be able to assign some form of weight to each of the available predictions so that they can be combined in an optimal manner. The verification measures should therefore provide information on where, when, and under what conditions, the model predictions can be considered reliable, rather than on the detailed diagnoses that are of interest to the model developers.

4.1.4 The requirements of the third group, the end-users of the forecasts, differ again from those of the other two groups. The end-users are ultimately concerned with questions of forecast value, which is known to bear a highly complex relationship to forecast quality. Forecast quality tends to be the exclusive concern of the model developers, and, to some extent, of the forecasters, and so the verification measures that are of interest to the modellers typically do not directly address the questions raised by the end-users.

4.1.5 In the context of these different interests, the meeting agreed that, where possible, the recommendations should be defined in such a way that the diagnostic requirements of the forecast developers, and the end-users' concerns about forecast value, can be addressed at the same time. To meet this objective it was recognized that scalar measures of forecast quality (such as the ROC area and MSSS), while of some interest, provide incomplete summaries of quality. The meeting therefore aimed to define recommendations for the communication of a minimum set of verification information to communicate forecast quality adequately, whilst avoiding the proliferation of large volumes of data.

4.1.6. It is also desirable that there should be feedback from users to model developers and a mechanism should be established to ensure such feedback.

4.2 Recommendations

4.2.1 Recommendations for scores and verification information first and foremost emphasize the immediate and future needs of RCCs and NMHSs to optimally employ statistical and dynamical model-produced forecasts available to them from global-scale producing centres for seasonal forecasts. However the recommendations are also cognizant of other requirements, ensuring that the raw material to meet end user requirements is exchanged as well as a small suite of basic scores to assess overall model performance.

4.2.2 Recommendations include exchange of scores and tables to accommodate verification of three types of forecasts of three variables: probabilistic and deterministic forecasts of three equally probable categories and continuous deterministic forecasts for (in all cases) two-meter temperatures, precipitation, and sea-surface temperatures.

4.2.3 Further, to ensure that verification information optimally reflects expected performance and the uncertainty in that expectation and that comparisons in performance can be facilitated, it is recommended that common techniques be employed for developing the information and that, to the extent possible, common hindcast verification periods, verification data sets, and update protocols be used. Specifications for these are presented in **Annexes II and IV**.

4.2.4 Requirements for common techniques include rigorous cross-validation of both (1) empirical model hindcasts and (2) climate parameters (like period of verification observed and modelled means, category boundaries, etc.) used in post-processing and verification of dynamical model hindcasts. Common methods should also be used for determination of the uncertainties in verification scores; generally boot-strap techniques will be required because of interdependencies (mostly spatial) in hindcast samples, but for some situations classical statistical methods (like Mann-Whitney tests) will be acceptable. Guidelines for testing significance and developing estimates of uncertainty are provided.

4.2.5 A basic principle for exchanging verification information is that a multi-faceted view of performance be possible. To this end and to accommodate a broad range of needs it is recommended that three levels of information be presented. Summary scores for broad geographic areas, maps of these and additional scores and contingency tables for the two categorical forecast types. Additionally in certain cases for deterministic forecasts, scores utilizing damped persistence as a standard rather than climatology will be more informative. Lastly for all types of forecasts, it will often be highly informative to present separate verification information for moderate to strong El Nino and La Nina episodes

4.2.6 For probabilistic forecasts the recommendation is to produce reliability diagrams (Wilks, 2001) to complement the areas under the ROC curves, recognizing that ROC diagrams are a powerful alternative way to view these forecasts. To this end contingency tables should be exchanged and instructions provided for construction of ROC curves. Global maps of areas under the ROC curves will assist users of model guidance in determining the circumstances under which more extensive examination of the attributes of performance is warranted.

4.2.7 For categorical forecasts in three classes it is recommended to exchange contingency tables rather than summary scores because the former are far more informative. Advice on summary scores can be found in **Annex IV**.

4.2.8 For continuous deterministic forecasts it is recommended that the use of the RMSSS be modified slightly to the use of MSSS. The latter has all the advantages of the former, but importantly allows for a simple direct decomposition into phase, amplitude and bias sources of error. These terms can be mapped, permitting a user of the guidance to make appropriate adjustments in constructing regional and local forecasts. The decomposition should also prove useful as diagnostic information for the producers of the guidance. However, the proposal includes the provision of sufficient fundamental properties to allow other scores (including (RMSSS) to be computed.

4.2.9 It is recognized that RCCs, NMHSs, and end users will want to make increasingly sophisticated use of guidance provided by global-scale producers, including but not limited to downscaling, decision models, etc. Thus it is highly desirable to exchange hindcast forecast histories that includes the model total fields, for use by RCCs and NMHSs in their use of the products.

4.2.10 Finally 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 fields and the corresponding verifying analyses for past forecasts be presented in visual format to the extent possible given restrictions on the availability of verification data.

4.3 Lead Centre role

4.3.1 The team considered that potential users of the verification information would best be served by consistent presentation of the results on a single web site. This would provide a convenient means of assimilating the information on the various prediction systems. The meeting suggested that, in the short term, this could be achieved by the processing at a single centre of digital versions of the verification information. In the longer term, and based on experience, guidelines could be developed for standardised presentation of the information on distributed web sites.

4.3.2 Consequently the meeting concluded that the need for a Lead Centre role was greater than previously envisaged. Revised versions for the roles of such a centre are contained in the Annex to this paragraph. The meeting acknowledged the work done in Melbourne so far and encouraged Melbourne to continue to develop its capabilities. However, the meeting also considered that, in these early stages of the implementation of an operational verification system, it would be beneficial if one or two other centres could assume such a role also. It is suggested that the arrangements be discussed at the ICT of the OPAG on DPFS and at the meeting of the global products producing centres scheduled for late 2002

4.4 Coordination with other Commissions and related working groups

4.4.1 The participation of the representatives of CCl, CLIVAR and CAS was greatly appreciated and the meeting benefited greatly from their active and constructive contributions. The meeting considered that it would be very useful to have broader input from these bodies prior to the presentation to CBS of the recommendations on the verification system for LRF. The next meeting of CBS is scheduled for December 2002.

4.4.2 The meeting recommended that the report of the meeting should be made available for review to CAS, the ET on Verification under the CCl OPAG on Climate Applications, Information and Prediction Services (CLIPS) and the CLIVAR Working Group on Seasonal to Interannual Prediction (WGSIP). Representatives of the two latter teams (Dr Simon Mason, Chair of the CLIPS Team, and Dr Ben Kirtman respectively) were present at the meeting of the team. The meeting invited them to circulate the final version of the report to members of their teams for critical review and comment. Comments would need to be provided before the end of August to allow time for the members of the Team to incorporate comments into the proposals to be submitted to CBS in December.

4.4.3 The meeting also agreed that participation of a Team member in meetings of the CLIPS ET on verification or CLIVAR/WGSIP and an opportunity to comment on any relevant reports would be beneficial. The meeting noted that this aim was well served at present, with the memberships of Drs Mason & Kirtman, and the membership of Dr Shoji Kusunoki in both this CBS Team and the CLIPS ET.

5. FUTURE WORK PROGRAMME

5.1 The meeting developed the following future work programme

(a) Coordinate the provision of long-range forecast verification scores and related information to NMHSs and RCCs, real-time monitoring of forecasts and relevant exchange between participating centres and institutes

(b) Encourage and monitor feed back from NMHSs and RCCs on the usefulness of verification information provided by producing centres under the scheme

(c) Review the effectiveness of the verification scheme in assisting NMHSs and RCCs to use the global-scale products to provide end-user services

(d) Contribute to further development of the activities of lead centre web site with links to producing centres and development and provision of relevant software to NMHSs and RCCs as capacity building measures to access information from producing centres and producing user friendly verification information.

(e) Recommend updates to operational practices to be followed in terms of the information on validation results to be attached to the long-range forecasts products in the light of experience and progress in research on verification activities

(f) Develop relevant standards for representation and presentation of verification information in maps and contingency tables

(g) In consultation with CAS (CLIVAR/WGSIP) and CCl, propose recommendations for improvements to CBS.

6. CLOSURE OF THE MEETING

The meeting was closed on Friday 26 April 2002.

Wilks, D.S. 2001. A skill sore based on economic value for probability forecasts. M*eteorol. Appl. 8*, 209-219.

Annex to paragraph 4.3.2

Role of lead centre

Create and maintain web-site to provide access to the LRF verification information. The web-site will:

(i) Provide access to standardized software for calculating scoring information (ROC curves, areas, contingency table scores, hit rates …).

(ii) preferably, provide consistent graphical displays of the verification results from participating centres through processing of digital versions of the results; alternatively, contain links to the verification information held on the web sites of participating centres;

(iii) contain relevant documentation and links to the web sites of global-scale producing centres;

(iv) provide some means for the collection of feedback from NMHSs and RCCs on the usefulness of the verification information;

(vi) Contain information and, preferably, provide access to available verification data sets;

The centre will also work towards:

(i) Producing monthly verification data sets in common format on 2.5° x 2.5° grid where appropriate;

(ii) developing guidelines for the presentation of verification information including guidelines for consistent graphical displays;

liase with other groups involved in verification (e.g. WGSIP, COLA, etc.) on the effectiveness of the current standardised verification system (SVS) and identify areas for future development and improvement;

provide periodic reports to CBS and other relevant Commissions assessing the effectiveness of the SVS**.**

ANNEX II

VERIFICATION SYSTEMS FOR LONG-RANGE FORECASTS - REVISED EXPERIMENTAL SCORES TO BE EXCHANGED

Formulation

The SVS is formulated in four parts:

1. **Diagnostics**. Three diagnostics are included and are closely defined. Additional diagnostics are suggested but are not incorporated into the Core SVS as yet. Use of the additional diagnostics is optional.

2. **Parameters.** Key variables for initial inclusion are proposed. However products are not limited to these variables. However the list is flexible to ensure that all producers can contribute regardless of the structure of individual forecast systems.

3. **Verification data sets**. Key data sets of observations against which forecasts may be verified are proposed.

4. **System details**. Details of forecast systems employed.

Diagnostics

Three diagnostics are incorporated in the Core SVS - Relative Operating Characteristics, reliability diagrams and accompanying measure of sharpness and Mean Square Skill Scores. These required diagnostics encompass score error estimates These diagnostics permit direct intercomparison of results across different predicted variables, geographical regions, forecast ranges, etc. Both may be applied in verification of most forecasts and it is proposed that, except where inappropriate, all three diagnostics are used on all occasions.

1. **Relative Operating Characteristics:** Calculation details are discussed in Annex I. For deterministic forecasts, the full contingency table should be provided. In addition, values of the Hit and the False Alarm Rates should be supplied. Other contingency measures, as listed in Annex 4, may be added. For probabilistic forecasts, maps providing the standardized area under the curve (such that perfect forecasts, give an area of 1 and a curve lying along the diagonal gives 0.5) should be provided, as a map for gridded data or as a curve for single point/region predictions. Probability values should be labelled on any Relative Operating Characteristics curves.

   A number of contingency table-based diagnostics are listed within Annex IV in addition to Hit and False Alarm Rates, including the Kuiper Score and Percent Correct (both used in assessing deterministic forecasts), and these provide valuable, readily-assimilable information for developers, producers and users of long-range forecasts. They may be considered for inclusion within information supplied to users.

2. **Reliability Diagrams and frequency histograms:** To be completed

   - Mean Square Skill Scores

   - Mean Square Skill Score and its decomposition - To be done.

Parameters

The key list of parameters in the Core SVS is provided below. Any verification for these key parameters, for either the verification history or for real-time forecasts, should be assessed using both Core SVS techniques wherever possible (given exceptions noted above). Many long-range forecasts are produced which do not include parameters in the key list (for example, there are numerous empirical systems that predict seasonal rainfall over part of/or over an entire, country). The Core SVS diagnostics should be used to assess these forecasts also, but full details of the predictions will need to be provided.

1. Diagrams and scores to be calculated for regions

1.1 Atmospheric parameters**.** Predictions for:
T2m Screen Temperature with standard regions:
Tropics 20°N to 20°S
Northern Extratropics >=20°N
Southern Extratropics <=20°S

Precipitation with standard regions:
Tropics 20°N to 20°S
Northern Extratropics >=20°N
Southern Extratropics <=20°S

1.2 Scores and diagrams to be produced for probabilistic forecasts
Reliability diagram
Frequency histograms.
The area under the ROC curve.
Estimations of error (significance) in the scores.

The above scores and diagrams to be produced for equi-probable tercile categories.

1.3 Scores to be used for deterministic forecasts
Mean Square Skill Score (MSSS) with climatology as standard reference forecasts.

1.4 Stratification by season
Four conventional seasons MAM, JJA, SON, DJF

1.5 Lead-time
Preferred minimum: 2 lead-times, one preferably to be 2-weeks or greater, with lead-time not greater than 4 months.

2. Verification to be produced in map format on 2.5^o x 2.5^o grid

2.1 Verification maps to be produced for each of the following variables
T2m
Precipitation
SST

2.2 Scores to be calculated for probabilistic forecasts
ROC area for 3 tercile categories
Significance of the ROC scores should also be calculated, and shown on the ROC area map, or an accompanying map.

2.3 Scores to be calculated for deterministic forecasts

The three terms of the Murphy decomposition of MSSS, produced with climatology as standard reference forecast. As a second, optional, control damped persistence should be used.

Significance estimates for each of the three Murphy terms should be calculated and shown on the relevant map or an accompanying map.

2.4 Stratification by season

Four conventional seasons MAM, JJA, SON, DJF

2.5 Lead-time

Preferred minimum: 2 lead-times, one preferably to be 2-weeks or greater, with lead-time not greater than 4 months.

2.6 Stratification by the state of ENSO should be provided if sufficient ENSO events are within the hindcast period used.

All hindcast seasons

Seasons with El Nino active

Seasons with La Nina active

3. Tabulated information to be exchanged

3.1 Contingency tables

Tabular information to be provided for grid points of a 2.5x2.5 global grid

3.1.1 Contingency tables to be produced each of the following variables

T2m

Precipitation

SST

3.1.2 Tables to be produced for probabilistic forecast verification.

The number of forecasts hits and false alarms to be recorded against each ensemble member or probability bin for each of three equi-probable categories (terciles). It is recommended that the number of bins remain between 9 and 10. The forecast providers can bin according to percentage bins or ensemble members as deemed necessary. No latitude weighting of the numbers of hits and false alarms is to be applied.

The user is encouraged to aggregate the tables over grid points for the region of interest. Methods of assessing statistical significance of the aggregated tables should be supplied by the user.

3.1.3 Tables to be produced for deterministic forecasts

3x3 contingency tables comparing the forecast tercile with the observed tercile, over the hindcast period.

3.1.4. Stratification by season

Four conventional seasons, 12 rolling 3months periods

3.1.5. Lead-time

Preferred minimum: 2 lead-times, one preferably to be 2-weeks or greater, with lead-time not greater than 4 months.

3.1.6 Stratification by the state of ENSO should be provided if sufficient ENSO events are within the hindcast period
All hindcast seasons
Seasons with El Nino active
Seasons with La Nina active

3.2 Grid point data for mapping
The information requested in this section is the digital data required to produce the maps described in Section 2. This will allow maps to be generated by users (or the lead centre) in a consistent format and allow access to maps for twelve rolling seasons rather than the four conventional seasons specified in section 2.

3.2.1 Grid point verification data to be produced for each of the following variables
T2m
Precipitation
SST

3.2.2 Verification to be produced for deterministic verification
Five parameters of the Murphy MSSS decomposition.
Significance estimates for each of the three Murphy terms should be calculated.

3.2.3. Verification to be provided for probability forecasts
ROC area for three tercile categories.
Significance of the ROC scores should also be calculated.

3.2.4. Stratification by season
If available twelve rolling 3-month periods (e.g. MAM, AMJ, MJJ). Otherwise as in 1.5.

3.2.5. Lead-time
Preferred minimum: 2 lead-times, one preferably to be 2-weeks or greater, with lead-time not greater than 4 months.

3.2.6. Stratification by the state of ENSO should be provided if sufficient ENSO events are within the hindcast period.
All hindcast seasons
Seasons with El Nino active
Seasons with La Nina active

4. Verification for indices

4.1 Indices
NINO3.4. Other indices may be added in due course.

4.2 Scores to be calculated for probabilistic forecasts
ROC area for 3 tercile categories
Significance of the ROC scores should also be.

4.3 Scores to be calculated for deterministic forecasts
The three terms of the Murphy decomposition of MSSS, produced with climatology as standard reference forecast. As a second, optional, control damped persistence should be used.
Significance estimates should accompany each of the three terms.

4.4 Lead-time

Preferred minimum: 2 lead-times, one preferably to be 2-weeks or greater, with lead-time not greater than 4 months.

4.4.1 Stratification by the state of ENSO should be provided if sufficient ENSO events are within the hindcast period.
All hindcast seasons
Seasons with El Nino active
Seasons with La Nina active

Verification Data Sets

The key list of data sets to be used in the Core SVS for both climatological and verification information is provided below. The same data should be used for both climatology and verification, although the centre's analysis (where available) and the ECMWF and NCEP/NCAR Reanalyses and subsequent analyses may be used when other data are not available. Many seasonal forecasts are produced that may not use the data in either the key climatology or verification data sets (for example, there are numerous systems which predict seasonal rainfall over part of, or over an entire, country). Appropriate data sets should then be used with full details provided.

1. Sea Surface Temperature

Reynolds OI, (1981 to present). Smith et al (?) (1971-1980)

2. Precipitation

Xie-Arkin; GPCP data

3. T2m Screen Temperature

UKMO/CRU T2m data set.
When gridded data sets are used, a 2.5° by 2.5° grid is recommended.

System Details

Information will be requested for exchange of scores concerning the following details of the forecast system; information labelled * should also be attached to user information:

1. Is the system numerical/hybrid/empirical*?
2. Do the results relate to the verification history or to real-time forecasts*?
3. Is the system deterministic/probabilistic*?
4. List of parameters being assessed. *
5. List of regions for each parameter. *
6. List of forecast ranges (lead times) and periods (e.g. seasonal average) for each parameter. *
7. The number of hindcasts/predictions incorporated in the assessment and the dates of these hindcasts/predictions.
8. Details of climatological and verification data sets used (with details of quality controls when these are not published).
9. If appropriate, resolution of fields used for climatologies and verification.
10. The period over which data are averaged to produce persisted anomalies.
11. Results of significance tests (Monte Carlo tests are recommended) on the historical verification period. *
12. Bias correction.

ADDITIONAL DIAGNOSTICS

1. **Categorical forecasts**

   Linear Error in Categorical Space for Categorical Forecasts (LEPSCAT)
   Bias
   Post Agreement
   Percent Correct
   Kuiper Score

2. **Probability Forecasts of Binary Predictands**

   Brier Score
   Brier Skill Score with respect to Climatology
   Reliability
   Sharpness (measure to be decided)
   Continuous Rank Probability Score

3. **Probability of Multiple-Category Predictands**

   Ranked Probability Score
   Ranked Probability Skill Score with respect to Climatology

4. **Continuous Forecasts in Space**

   Murphy-Epstein Decomposition (phase error, amplitude error, bias error) including the Anomaly Correlation

ANNEX IV

COMMISSION FOR BASIC SYSTEMS

Standardised Verification System (SVS) for Long-Range Forecasts (LRF)

WORLD METEOROLOGICAL ORGANIZATION

Standardised Verification System (SVS) for Long-Range Forecasts (LRF)

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.

Definitions

1.1 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.

1.2 Deterministic Long-Range Forecasts

Deterministic LRF provide details of expected occurrences or non-occurrences of an event (categorical or non-categorical). 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.

1.3 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.

1.4 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.

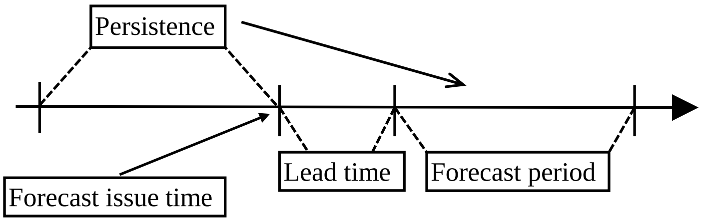

Figure 1: Definition of forecast period, lead time and persistence as applied in a forecast verification framework.

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.

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.

SVS for Long-Range Forecasts

Parameters to be verified

Surface air temperature anomaly at screen level (T2m), Precipitation anomaly and Sea surface temperature (SST) anomaly are the three parameters to be verified.

It is recommended that three types of verification be done:

1. large scale aggregated overall measures of forecast performance.
2. verification at grid points.
3. grid point by grid point contingency tables for more extensive verification.

Both deterministic and probabilistic forecasts are verified if available.

Aggregated verification

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:

1. Tropics: from 20°S to 20°N all inclusive.
2. Northern Extra-Tropics: from 20°N to 90°N, all inclusive.
3. Southern Extra-Tropics: from 20°S to 90°S, all inclusive.

In addition to these three areas, Sea surface temperature anomaly forecast verification will be done over Niño-3.4 region from 170°W to 120°W and from 5°S to 5°N all inclusive.

Grid point verification

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.

Contingency tables

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.

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).

Verification scores

The following verification scores are to be used: Mean Square Skill Score (MSSS) and 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 number of observed occurrences and non-occurrences of an event as a function of the forecast occurrences and non-occurrences of that event (deterministic or probabilistic). The events are defined as binary, which means that only two outcomes are possible, an occurrence or a non-occurrence.

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.

MSSS for 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:

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

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

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 (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:

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

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 on equal area or approximately equal area grids and is equal to cos (θ_i), where θ_i is the latitude at grid point i on latitude-longitude grids.

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

$$RMSSS = 1 - \left(1 - MSSS\right)^{\frac{1}{2}}$$

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

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

where r_{fxj} is the product moment correlation of the forecasts and observations at point or station j.

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

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

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

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).

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.

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

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

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

A recommended skill score for the three by three tables 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

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

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

where

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

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 tables. 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.

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 $\cos(\theta_i)$, 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):

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

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:

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

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 4) are part of the exchange of LRF verification scores. The scaled Hanssen and Kuipers score for deterministic categorical forecasts is also included together with the contingency tables. One contingency table is filled in when the outcome of the LRF system is in two categories; however, one contingency table has to be filled in for each type of possible binary events, when the outcome of the LRF system is in three (or more) categories (for example, for LRF system whose forecasts are in three categories, three contingency tables are filled in, one for each category against the remaining two). When deterministic LRF are generated with an Ensemble Prediction System, the ensemble size should be specified.

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 n^{th} 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 n^{th} 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 and for all probabilities exceeding 0% respectively). The further the curve lies towards the upper left-hand corner (where HR=1 and FAR=0) the better; no-skill forecasts are indicated by a diagonal line (where HR=FAR).

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 part of the exchange of LRF verification scores. The ROC score (area under the ROC curve, normalised to one) for probabilistic forecasts is also included together with the contingency tables. One contingency table is filled in when the outcome of the LRF system is in two categories; however, one contingency table has to be filled in for each type of possible binary events, when the outcome of the LRF system is in three (or more) categories (for example, for LRF system whose forecasts are in three categories, three contingency tables are filled in, one for each category against the remaining two). When LRF are generated with an Ensemble Prediction System, the ensemble size should be specified.

Reliability diagrams 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 instead of 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 probability each probability bin (HR_n) is defined as:

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

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

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, 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 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.

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.

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.

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

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 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.

Status of the verification data sets

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

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

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/

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. Geophysics., 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.

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

Reporting

Information must be provided on the system being verified. This information should include:

1. Type of system (statistical versus dynamical).
2. Model type and resolution.
3. Boundary conditions specifications.
4. Ensemble size.
5. Period of 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.

Section 8 gives an example of template to exchange information on the description of the forecasts system being verified.

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.

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.

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. We have attempted to classify the intensity of each event by focusing on a key region of the tropical Pacific (along the equator from 150°W to the date line). The process of classification was primarily subjective using reanalysed sea surface temperature analyses produced at the National Centres for Environmental Prediction/Climate Prediction Centre and at the United Kingdom Meteorological Office. An objective procedure for classifying intensity is being explored at NCEP/CPC. In the following table, weak periods are designated as C- or W-, moderate strength periods as C or W, strong periods as W+ or C+, and neutral periods as N. An ENSO event is defined as moderate or strong cold or warm events. The data is taken from NOAA/NCEP/CPC at www.cpc.ncep.noaa.gov.

Years	JFM	AMJ	JAS	OND
1950	C	C	C	C
1951	C	N	N	W-
1952	N	N	N	N
1953	N	W-	W-	N
1954	N	N	C-	C
1955	C	C-	C-	C+
1956	C	C	C	C-
1957	N	W-	W-	W
1958	W+	W	W-	W-
1959	W-	N	N	N
1960	N	N	N	N
1961	N	N	N	N
1962	N	N	N	N
1963	N	N	W-	W
1964	N	N	C-	C
1965	C-	N	W	W+
1966	W	W-	W-	N
1967	N	N	N	N
1968	N	N	N	W-
1969	W	W-	W-	W-
1970	W-	N	N	C
1971	C	C-	C-	C-
1972	N	W-	W	W+
1973	W	N	C-	C+
1974	C+	C	C-	C-
1975	C-	C-	C	C+
1976	C	N	N	W-
1977	N	N	N	W-
1978	W-	N	N	N
1979	N	N	N	N
1980	W-	N	N	N
1981	N	N	N	N
1982	N	W-	W	W+
1983	W+	W	N	C-
1984	C-	C-	N	C-
1985	C-	C-	N	N

1986	**N**	**N**	**W-**	**W**
1987	**W**	**W**	**W+**	**W**
1988	**W-**	**N**	**C-**	**C+**
1989	**C+**	**C-**	**N**	**N**
1990	**N**	**N**	**W-**	**W-**
1991	**W-**	**W-**	**W**	**W**
1992	**W+**	**W+**	**W-**	**W-**
1993	**W-**	**W**	**W**	**W-**
1994	**N**	**N**	**W**	**W**
1995	**W**	**N**	**N**	**C-**
1996	**C-**	**N**	**N**	**N**
1997	**N**	**W**	**W+**	**W+**
1998	**W+**	**W**	**C-**	**C**
1999	**C+**	**C**	**C-**	**C**
2000	**C**	**C-**	**N**	**C-**
2001	**C-**	**N**	**N**	**N**

Example of a template for LRF description

Example of a template used for LRF system description:

Identification				
Country:	[1] CANADA			
Meteorological Centre:	[2] CANADIAN METEOROLOGICAL CENTRE			
LRF system identification:	[3] CMC-DYN			
Description of Long-Range Forecast (LRF) System				
Status of LRF system:	**information:**	[4] Yes		
	dissemination:	[5] Disseminated trough the Meteorological Service of Canada regional offices, media and Internet.		
	guidance:	[6] Historical (hindcast) percent correct map is provided with the forecast.		
Type of LRF system:	**numerical:**	[7] Yes	**empirical:**	[8] No
	hybrid:	[9] No	**coupled:**	[10] No
	statistics:	[11] Blend of the model 1000-500 hPa thickness using an arithmetic average and forecast SAT using a Perfect Prog simple linear regression with 1000-500 hPa hPa anomaly as predictor.		
Type of forecasts:	**deterministic:**	[12] Yes	**probabilistic:**	[13] No
LRF output products:	**parameter:**	[14] Surface air temperature anomaly and accumulated precipitation anomaly.		
	categories:	[15] Below (< -0.43 times the inter-annual standard deviation), above (>0.43 times the inter-annual standard deviation) and near normal (otherwise).		
	frequency:	[16] 4 times a year		
	forecast period:	[17] Season (3 months)		

	lead time:	[18] Zero		
	forecast area:	[19] Canada		
Model description:	[20] Two numerical models are used: 1) GCM version 2 (McFarlane et al., 1992) and 2) Global NWP model (Ritchie, 1991).			
Model resolution:	**horizontal:**	[21] GCM=T32 NWP=T63	**vertical:**	[22] GCM=L10 NWP=L23
Bias correction:	[23] No			
Ensemble forecasting:	**ensemble size:**	[24] 12 members (6 for each model)		
	initialisation:	[25] Initial conditions for each run are lagged by 24 hours.		
SST specification:	[26] Anomaly of the month prior to the start of the runs is prescribed through the 3 months forecast period.			
Sea-Ice specification:	[27] Climatology for the GCM and monthly mean of the month prior to the start of the runs is relaxed to climatology during the first 15 days.			
Snow specification:	[28] Initial snow line from observations. The GCM has a prognostic scheme while the NWP model uses a prescribed 10 day snow anomaly for half of the simulation afterwards the system reverts to climatology.			
Soil temperature:	[29] Climatology			
Soil moisture:	[30] Climatology			
Hindcast evaluation:	**hindcast period:**	[31] 1969 to 1994		
	cross-validation:	[32] No		
	particularities:	[33] Initial conditions are lagged by 6 hours (instead of 24 as in the operational set-up).		

ANNEX V

AGENDA

1. OPENING OF THE MEETING

2. ORGANIZATION OF THE MEETING

   2.1 Adoption of the agenda

   2.2 Other organizational questions

3 EXCHANGE OF LONG-RANGE FORECASTS VERIFICATION SCORES

4 OPERATIONAL PROCEDURES AND PRACTICES

5 FUTURE WORK PROGRAMME

6 CLOSURE OF THE MEETING

LIST OF PARTICIPANTS

Mr Terry Hart, *Chair*
Bureau of Meteorology
GPO Box 1289K
MELBOURNE VIC 3001
Australia
Tel: (613) 9669 4030
Fax: (613) 9662 1222
Email: t.hart@bom.gov.au

Dr Laura Ferranti
ECMWF
Shinfield Park
Reading, BERKSHIRE RG2 9AX
United Kingdom
Tel: (44 118) 949 9601
Fax: (44 118) 986 9450
Email: laura.ferranti@ecmwf.int

Dr Robert Livezey
Office of Climate, Water, and Weather Services
National Weather Service, NOAA
SSMC Rm 13228
1325 East-West Highway
CAMP SPRINGS MD 20910-3283
USA
Tel: (1 301) 713 1867 ext. 182
Fax: (1 301) 713 1520
Email: robert.e.livezey@noaa.gov

Dr Richard Graham
Met Office
London Road, Bracknell
BERKSHIRE RG12 2SZ
United Kingdom
Tel: (44 1344) 856 447
Fax: (44 1344) 854 026
Email: rjgraham@meto.gov.uk

Mr Richard Verret
Canadian Meteorological Centre
2121 Trans-Canada Highway
DORVAL QUEBEC
Canada H9P 1J3
Tel: (1 514) 421 4683
Fax: (1 514) 421 4657
Email: richard.verret@ec.gc.ca

Dr Shoji Kusunoki
Japan Meteorological Agency
1-3-4 Otemachi Chiyoda-ku
TOKYO 1OO-8122
Japan
Tel: (81) 3 3211 8406
Fax: (81) 3 3211 8406
Email: s-kusunoki@met.kishou.go.jp

Dr Benjamin Kirtman
COLA Studies
4041 Powder Mill Road,
suite 302 Calverton
MD 20705-3106
U.S.A
Tel.: (1 301) 902-1244
Fax: (1 301) 595-9793
Email: kirtman@cola.iges.org

Dr Simon Mason
IRI,
Scripps Institution of Oceanography,
University of California, San Diego,
0230, La Jolla, CA 92093
U.S.A
Tel.: (1 858) 822-2574
Fax: (1 858) 534-8561
Email: simon@lacosta.ucsd.edu

Dr Peiqun Zhang
National Meteorological Centre
China Meteorological Administration
46 Zhongguancun South Street
BEIJING 100081
China
Tel: (8610) 6840 6403
Fax: (8610) 6840 8508
Email: zhangpg@cma.gov.cn

Mr Normand Gagnon
Canadian Meteorological Centre
2121 Trans-Canada Highway
DORVAL QUEBEC
Canada H9P 1J3
Tel: (1 514) 421 4712
Fax: (1 514) 421 4657
Email: normand.gagnon@ec.gc.ca

WMO Secretariat

Mr Morrison MLAKI
Chief
Data Processing System Division
World Meteorological Organization
7 bis avenue de la Paix
Case postale No.2300
CH-1211 Geneva 2
Switzerland
Tel.: 41(0)22 730 8231
Fax: 41 (0) 22 730 8021
Email: mmlaki@www.wmo.ch and Mlaki_m@gateway.wmo.ch