WORLD METEOROLOGICAL ORGANIZATION

COMMISSION FOR BASIC SYSTEMS
OPAG on DPFS

MEETING OF THE CBS (DPFS) EXPERT TEAM ON OPERATIONAL WEATHER AND FORECASTING PROCESS AND SUPPORT

GENEVA, SWITZERLAND
22-24 OCTOBER 2014

DPFS/ET-OWFPS/Doc. 4.1_v3

(21.X.2014)

Agenda item : 4.1

ENGLISH ONLY

DRAFT DOCUMENT ON STANDARDIZED SURFACE VERIFICATION OF DETERMINISTIC NWP PRODUCTS

(T. Haiden, H. Kabelwa, M. Mittermaier, A. Okagaki, T. Robinson)

Summary and purpose of document

This document summarizes the standardized surface verification of deterministic NWP products for the exchange of scores between global centres, to be included in the new CBS Manual WMO-No.485. It is based on the work of the Task Team for Surface Verification (TT-SV) and has been approved by the Expert Team on the Operational Weather Forecasting Process and Support (ET-OWFPS) on 22 Oct 2014 in Geneva.

Action Proposed

Global NWP centres are invited to take note of, and comment on, the proposal.

1. Introduction

Detailed procedures are presented for the production and exchange of a standard set of verification scores for deterministic global NWP forecasts of surface fields produced by participating centres. The goal is to provide consistent verification information on the NWP surface products of participating centres for forecasters in the NMHSs and to help the centres compare and improve their forecasts.

The term “deterministic NWP” refers to single integrations of NWP models providing products defining single future states of the atmosphere (as distinct from ensemble prediction systems where multiple integrations provide a range of future states).

The standardized verification should provide relevant information appropriate to the state-of-the-art in NWP, while being as simple and as easy to implement as possible. It should ensure a consistent implementation across participating centres with regard to the generation of forecast-observation pairs, scores, and temporal aggregation.

2. Verification statistics

The following subsections define two sets of verification statistics. A mandatory set shall be provided by all participating centres. A set of additional recommended statistics is also defined which all centres should provide if possible. The detailed procedures are required to ensure it is possible to compare results from the different participating centres in a scientifically valid manner.

3. Parameters

Mandatory

- 2-m temperature
- 10-m wind speed
- 10-m wind direction
- 24-h precipitation

Additional recommended

- Total cloud cover
- 6-h precipitation
- 2-m relative humidity
- 2-m dewpoint

For 2-m temperature, a simple height-correction using a constant lapse rate of 0.0065 K/m shall be applied to the forecast.

4. Forecast times

Scores shall be computed daily for forecasts initialized at 00 UTC and 12 UTC separately. For those centres not running forecasts from either 00 UTC or 12 UTC, scores may be provided for forecasts initiated at other times and must be labelled as such.

5. Forecast steps

Mandatory:

6-hourly up to T+72, 12-hourly up to T+240 or end of the forecast

For 24-h precipitation: 24-hourly up to T+240 or end of the forecast

Additional recommended:

3-hourly up to T+72, 6-hourly up to T+240 or end of the forecast (for improved representation of diurnal cycle)

For 6-h precipitation: 6-hourly up to T+240 or end of the forecast

6. Grid and interpolation

Verification shall be based on the native model grid using the grid point nearest to the observation location.

7. Observations

Verification is carried out for a set of surface stations (SYNOP, METAR, fixed marine observing stations) which should be chosen based on high availability and reliability, in order to reduce the effect of observation errors and ensure consistency over time. Each participating centre should aim to include as many stations as possible to ensure good global coverage. The list of stations used in the verification is allowed to differ between centres. This is made possible by the fact that scores for individual stations will be exchanged (see item 9).

Centres are encouraged to make use of the quality control procedures available to them to reduce the effect of observation errors on scores. This includes removal of occasional unphysical values as well as data at individual stations which has been systematically rejected over a certain time period.

8. Scores

Scores are computed for each station individually. A station for which scores are computed should have at least 90% data availability during the verification period.

For 2-m temperature, 2-m relative humidity, 2-m dewpoint, 10-m wind speed, 10-m wind direction, and total cloud cover the following error scores are computed:

- Mean error (ME)
- Mean absolute error (MAE)
- Root mean square error (RMSE)

10-m wind direction is verified only when the observed wind speed is ≥3 m/s. For 10-m wind direction the equivalence of 360 and 0 degrees needs to be taken into account (cyclic continuation).

For 10-m wind speed, precipitation, and total cloud cover, contingency-tables for the following thresholds are provided:

- 10-m wind speed: 5, 10, and 15 m/s
- 24-h precipitation: 1, 10, and 50 mm
- 6-h precipitation: 1, 5, and 25 mm
- Total cloud cover: ≤ 2 okta, ≥ 6 okta

For total cloud cover, the model output should be rounded to the nearest okta prior to verification (for the contingency tables only).

9. Exchange of scores

On a monthly basis, in a common format, where the station information (lat/lon, station height and model height, station ID) is contained within the file. This means that no supplementary files are required, and the exchanged data is fully self-contained.

10. Temporal and spatial aggregation

For any given 1-month period, error scores and contingency tables are computed for each station individually. It forms the basis for aggregation by users of the exchanged verification data, both in time and space.

Spatial aggregation is not part of the exchange, and is left to user discretion. Exchanging scores in this way allows forecast users to get detailed information on model performance for individual stations. It also ensures a high level of transparency and flexibility for model inter-comparison studies. Furthermore, it removes the requirement of coordinating, circulating, and updating whitelists of surface stations for verification. For model intercomparison studies the intersection of the different sets of stations used by global modelling centres would be used for comparison (‘smallest common denominator’).

If users would like to aggregate the exchanged scores, they can refer to Annex A which provides guidelines for the choice of aggregation areas. Compared to upper-air verification, more emphasis needs to be put on aggregating over climatologically relatively homogeneous areas (since absolute thresholds are used for the contingency tables).

Annex A - Guidelines for use of the exchanged scores and metrics

Areas for aggregation

If users want to aggregate the exchanged scores, areas need to be defined in such a way that areal means are statistically meaningful. Areas for spatial aggregation should be relatively homogeneous climatologically. With respect to the regions currently defined by CBS for upper-air verification, only the tropics and polar regions would be considered suitable. For mid-latitudes a further subdivision based on latitude and degree of continentality is recommended.

Contingency-table based scores

The exchange of contingency tables allows users to compute a wide range of scores and metrics. Some suggested scores and metrics are

- Base rate (BR)
- Frequency bias (FB)
- Hit rate (HR)
- False alarm rate (F)
- Equitable Threat Score (ETS)
- Peirce Skill Score (PSS)

It should be noted that for some of the higher thresholds, such as 50 mm for 24-h precipitation, the number of correctly forecast non-events in the contingency tables will be much larger than the other three. In such cases of rare events, many of the usual skill scores such as PSS or ETS may give misleading results. In these cases scores specifically designed for rare events, such as the Symmetric Extremal Dependence Index (SEDI) should be used.

Confidence Intervals

Confidence intervals quantify the uncertainty and facilitate the interpretation of verification scores. The length of the verification period for which data is exchanged (1 month) is considered too short for statistically meaningful bootstrapping. It is therefore recommended to apply bootstrapping techniques in the spatial aggregation of scores based on the exchanged data, in order to derive confidence intervals (e.g. 95%).

As described by Candille et al. (2007), the bootstrap technique involves recomputing scores numerous times after randomly extracting samples from the data set and then replacing them, again randomly, from the original data set.

Reference

Candille, G., C. Côté, P. L. Houtekamer and G. Pellerin, 2007: Verification of an ensemble prediction system against observations, *Mon. Wea. Rev.*, **135,** 2688–2699.

Additional information

The above recommendations define the basic verification of standard NWP surface fields. Additional evaluations of surface fields using different types of observations are encouraged. For example, for a more comprehensive verification of cloudiness, further information can be found in the WMO publication WWRP2012-1 ‘Recommended Methods for Evaluating Cloud and Related Parameters’ which can be downloaded from
http://www.wmo.int/pages/prog/arep/wwrp/new/Forecast_Verification.html

Annex B – Background on some of the recommendations

3. Parameters

The list of mandatory parameters is consistent with the directly weather-related surface parameters listed in Appendix A.II.2.1.1-a (MINIMUM LIST OF GLOBAL DETERMINISTIC NWP PRODUCTS TO BE MADE AVAILABLE ON THE WIS) of the revised Manual on the Global Data Processing and Forecasting Process (GDPFS, WMO-No.485), the only difference being the use of wind speed and direction instead of u and v components.

5. Forecast steps

The mandatory steps for which the forecasts are to be verified is consistent with the directly weather-related surface parameters listed in Appendix A.II.2.1.1-a (MINIMUM LIST OF GLOBAL DETERMINISTIC NWP PRODUCTS TO BE MADE AVAILABLE ON THE WIS) of the revised Manual on the GDPFS (WMO-No.485).

6. Grid and interpolation

The method of using the nearest model gridpoint rather than interpolating the model field to the exact station location has the advantage of not introducing any artificial smoothing. This is especially relevant for heavy precipitation and has been adopted here as a uniform method of forecast/observation matching for all parameters. Using the nearest land point rather than the nearest gridpoint was not found to provide a substantial benefit in terms of forecast skill, partly because models are already taking into account fractional land/sea coverage near coastlines. It was found that use of the nearest land point did not systematically reduce forecast errors of 2-m temperature and 10-m wind speed.

Using only the nearest model gridpoint for verification may become problematic for some parameters as the horizontal resolution of global models increases further in the future, possibly down to convection-permitting scales (5 km and smaller). This would especially affect total cloud cover, which tends to show increasing 0/1 behaviour at higher resolution, and where already now there is a representativeness mismatch between the size of a model grid-box and the area over which a human observer estimates cloud amount.

8. Scores

The thresholds for 10-m wind speed and 24-h precipitation are consistent with, and a subset of, the thresholds listed in Appendix A.II.2.1.3-a (MINIMUM LIST OF GLOBAL EPS PRODUCTS TO BE MADE AVAILABLE ON THE WIS) of the revised Manual on the GDPFS (WMO-No.485).

Reference

A draft (July 2014) of the revised Manual on the GDPFS (WMO-No.485) can be found at
http://www.wmo.int/pages/prog/www/DPS/linkedfiles/Revised-Manual-July2014.zip