

Towards a common verification for operational HIRLAM

Xiaohua Yang, DMI

1 Introduction

Verification of model results against observation or analysis is an important part of NWP research and operational applications. Due to historical reasons, different approaches and softwares have been applied, within HIRLAM community, for observation verification. Presently, for routine observation verification in HIRLAM operational centers, the "reference verification package" is only applied in KNMI, FMI and INM, whereas different local softwares are used in DMI, Met Eireann, met.no and SMHI. In addition, there exists a verification package in the HIRLAM data assimilation community (HIRVDA), an extension of which has now become part of the multi-purpose utility packages "gl", which is used for daily monitoring and verification in the HIRLAM meso-scale modeling ("HARMONIE") community for forecasts made with HIRLAM-ALADIN meso-scale system.

In this notes, we present some evidence of incompatibility in verification scores due to diversity in methodology and data selection, describe the recently proposed approaches within the HIRLAM-A programme to harmonize on verification approaches for model inter-comparison and impact studies, and the efforts towards a common verification for operational HIRLAM forecasts.

2 Incompatibility of verification results

HIRLAM forecast systems as implemented in operational centers differ, sometimes radically. In the past, centers used to be asked to provide, periodically, to the international HIRLAM community with certain selected observation verification scores for inter-comparison and for monitoring of long term quality trend of the HIRLAM forecast system. The usefulness of such a collection of verification scores from different sources has, however, been recognized to be limited, sometimes even misleading, due to the substantial diversity in verification algorithm, selection of verifying data sets including the criteria for quality control.

Figure 1 and 2 demonstrate such an example. Shown in Figure 1 are the root mean square (rms) and bias errors of temperature at 850 hPa by the operational HIRLAM models (including that for FMI-RCR) for January 2006, as computed by operational centers using their own routine verification softwares. Different colors corresponds to scores for different models, (the information is omitted here). In Figure 2, the results for one of these model results, FMI-RCR, calculated with reference and DMI verification packages, respectively, are plotted. Comparing the two figures, the difference between FMI-RCR rms scores as shown in Figure 2, which is purely due to use of different verification packages, is strikingly large in comparison to the relative differences between the scores for operational models collected from operational centers, as shown in Figure 1. Thus it would be unreliable to derive from the kind of results in Figure 1, which is based on a simple collection of verification scores produced at different centers, relative skills between different models.

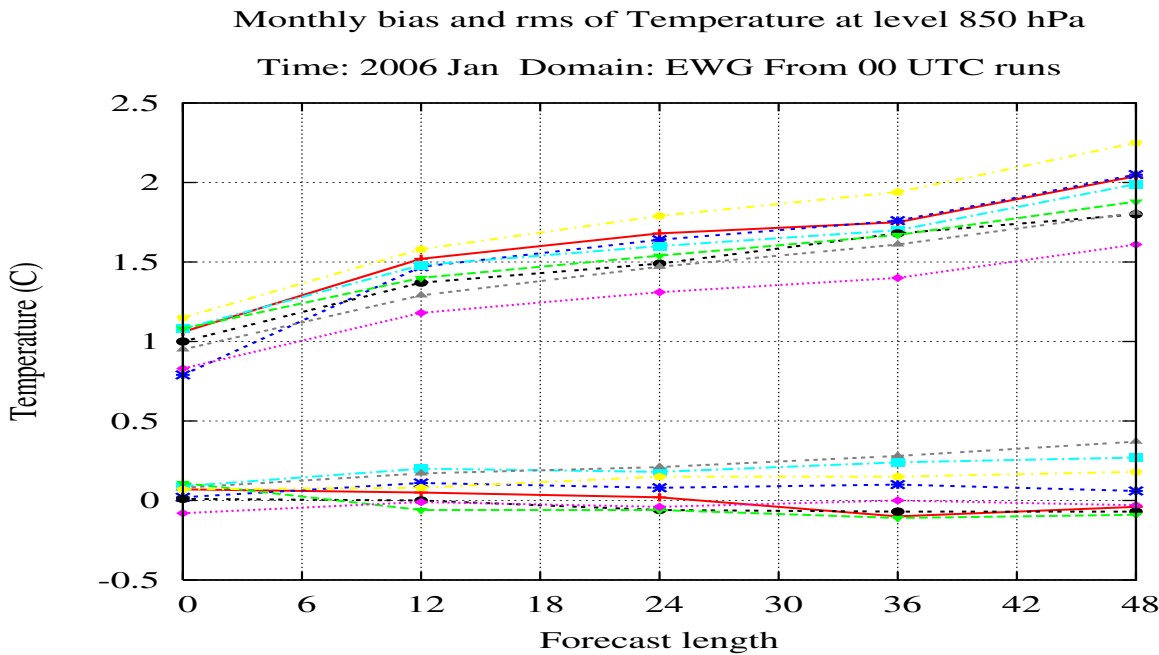


Figure 1: Averaged bias and rms for HIRLAM 850 hPa temperature forecasts along forecast lead-time, validated against EWGLAM data for Jan 2006. Colors denote results from different HIRLAM centers, as computed with routine verification softwares used at individual services. Courtesy of Kalle Eerola.

3 Sensitivity of verification scores

To illustrate further sensitivity of results to verification procedure, we conduct here some sensitivity experiments taking into account a number of known contributing factors that may influence verification results. For generality the selection of forecast data here has been arbitrary. The chosen forecast results are from a month-long numerical experiment using Hirlam 7.1 beta 2 for January, 2006. Observation verification scores in rms and bias for EWGLAM-data are calculated with the following 5 different verifications:

- Control verification (CTL) as produced using HIRVDA verification package. With this software, experiment results on model levels for up to 48 h with 6 h interval are first interpolated bi-linearly to EWGLAM stations. The verifying observation data is extracted from the ECMWF MARS archive. The data is first checked against analysis of the same experiment to discard those with large departure, according to the criteria specified in HIRVDA verification.
- Alternative verification package (REF). In this case the same experiment results are verified using the reference verification package, which, in addition to a different procedure for verification, differs from CTL primarily in terms of quality control aspect and in different algorithm for diagnosing MSLP and geopotential height.
- Alternative source of verifying observation (ROB). With everything else being equal to CTL, the FMI-RCR archived observation is used here instead of data extracted from MARS. Note that although the observation data in both ROB and CTL are originated from GTS, the RCR archive has a 2 h cutoff whereas MARS archive has a longer one. As such, the data amount from MARS is higher than that from RCR.
- Alternative analysis for quality control (RQC). With everything else being equal to CTL, the HIRLAM analysis used for quality control of verifying observations is that from FMI-RCR instead of native ones.

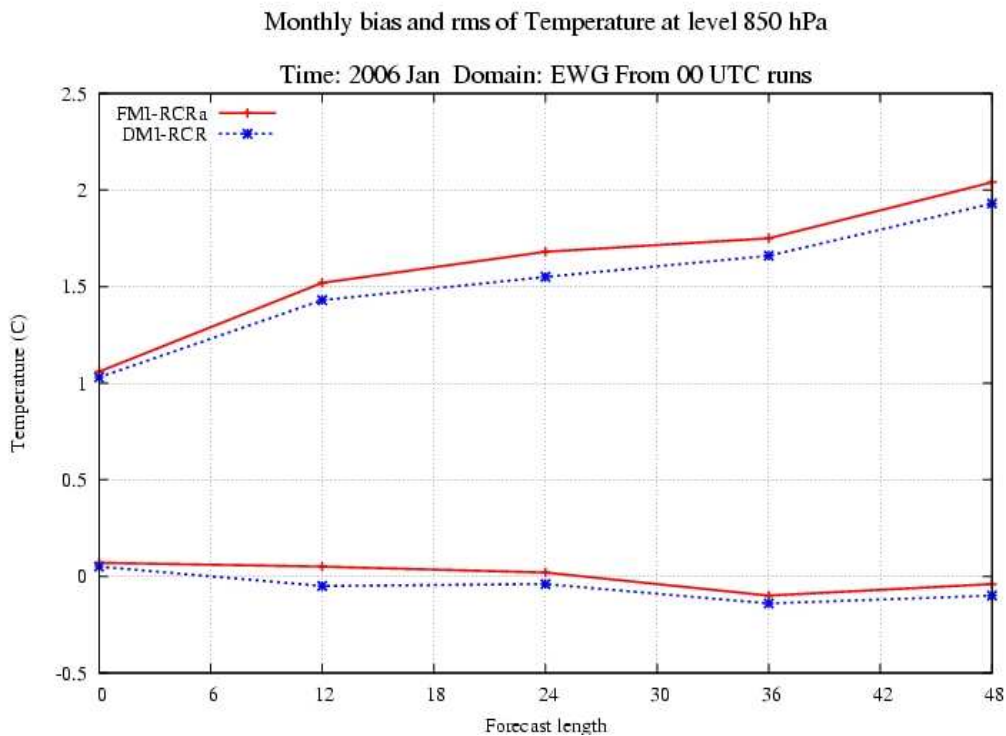


Figure 2: same as Figure 1 but with verification results for FMI RCR calculated with reference (in red) and DMI (in blue) verification packages, respectively.

- Alternative quality control criteria (QRA). With everything else being equal to CTL, the quality control criteria used to discard 'questionable observations' are those similar to the default ones used in the reference verification.

Figure 3 shows the key verification scores computed for the near surface and upper air (850 and 500 hPa) parameters with CTL and REF. It is seen clearly that use of different verification packages (HIRVDA and reference) results in rather significantly different verification scores for most parameters. In particular, the differences in humidity parameter and in bias features for T2m are alarmingly large.

Figure 4 shows comparison of scores for upper air parameters using same verification packages (HIRVDA) but with different verifying observation data sets (CTL vs ROB), different verifying analyses for quality control (CTL vs RQC) and different error tolerance criteria in observation data quality control (QRA). Noticeable differences in resulting verification scores are seen in all these cases, most remarkably in use of different observation data set for verification. The impact of differences in observation data selection procedure is seen to affect mainly the humidity parameters, but less significant for other parameters. Overall, the corresponding sensitivity for surface parameters to these differences seem to be less significant, presumably due to a much larger sample size (not shown here)

It should be added that there are a few additional factors that make the inter-comparisons difficult:

- Difference in source of model data for verification calculation. In reference package the model data used are those from 've' files (selected pressure levels + surface and accumulated parameters), which are post-processed during model integration. In some other packages (e.g., HIRVDA, DMI, gl) verification is done with model output from the so-called history file. The diagnosis/interpolation used to derive MSLP and geopotential height may thus be inconsistent from those implemented in the integration model. The difference caused by this factor has been shown, e.g., to be significant in some earlier studies at DMI. The difference due to inconsistency in selection of use for model level or 've' files are especially important at

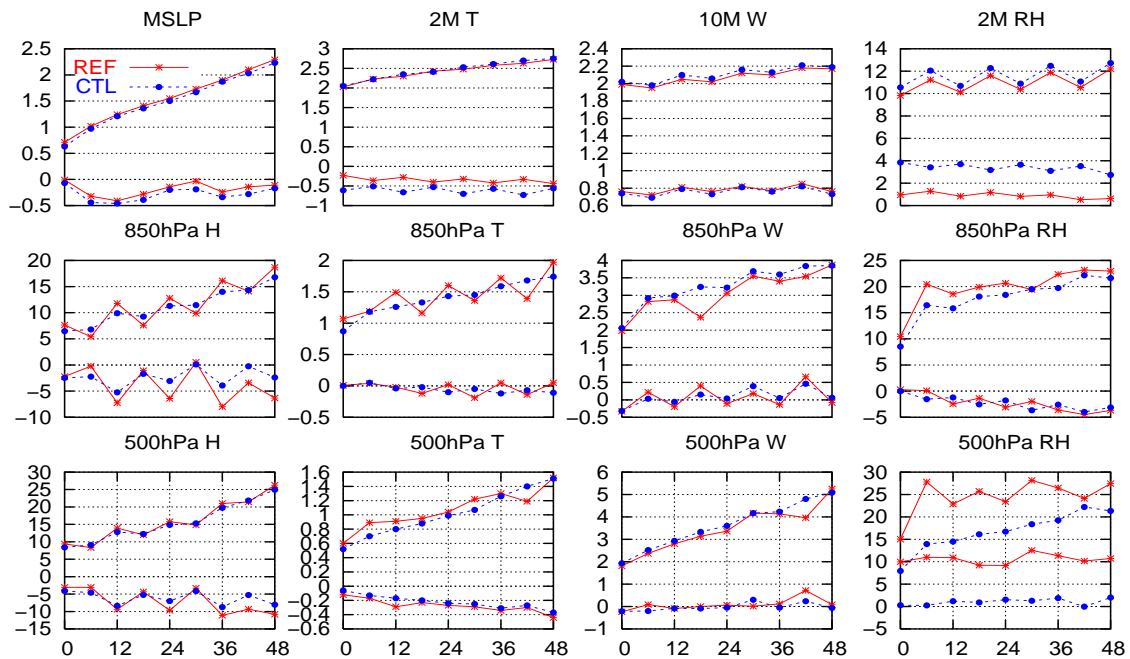


Figure 3: EWGLAM verification scores in standard deviation and bias for key parameters calculated with CTL and REF, respectively.

analysis time, since the analysis data is used for quality control to reject questionable observations. In the model level file, the surface temperature are from that of surface analysis, whereas the value appearing in the corresponding 've' files are the diagnosed ones via HIRLAM's SAPP (stand-alone post-processing) procedure. Unfortunately, the latter has not been updated to be consistent with the forecast model since the introduction of ISBA surface scheme.

- Difference in station-list definition. The most common verification in model inter-comparison is done against EWGLAM-station list. For centers using the online approach with reference verification package, due to the sharing of observation data for assimilation and for verification, the number of EWGLAM stations included in score calculation is incomplete due to short cut-off. Similar shortcoming is not shared in other centers where the observation data used for verification is usually from history archive. For Met Eireann, the EWGLAM score is only nominal due to the fact that the model domain does not cover full EWGLAM area. Another potential difference in score calculation can arise due to differences in specified observation station locations and heights (due to station re-location).
- Different verification parameters such as those for humidity data and for precipitation. For the latter, different categories for contingency table have been defined, which makes it difficult for a direct score comparison.

In summary, based on the above, it is concluded that a simple collection of verification scores calculated with inconsistent methodologies from operational centers can not be expected to provide reliable and trustworthy verification information. On the contrary, results may be misleading. Secondly, even if same verification package is used, inconsistencies in implementations such as selection of verifying analysis, observation data set, data quality screen procedure etc. can potentially result in significant uncertainties. This is especially true for upper air verification parameters, for which the ratio between "noise" (the error margins due to uncertainties in methodology) and "signal" (the difference in forecast qualities between model systems) may be too large to offer useful guidances. For usual surface parameters such as MSLP, T2m and W10m, the uncertainties due to methodology and observation data selection may be relatively less, thanks to a larger sample size.

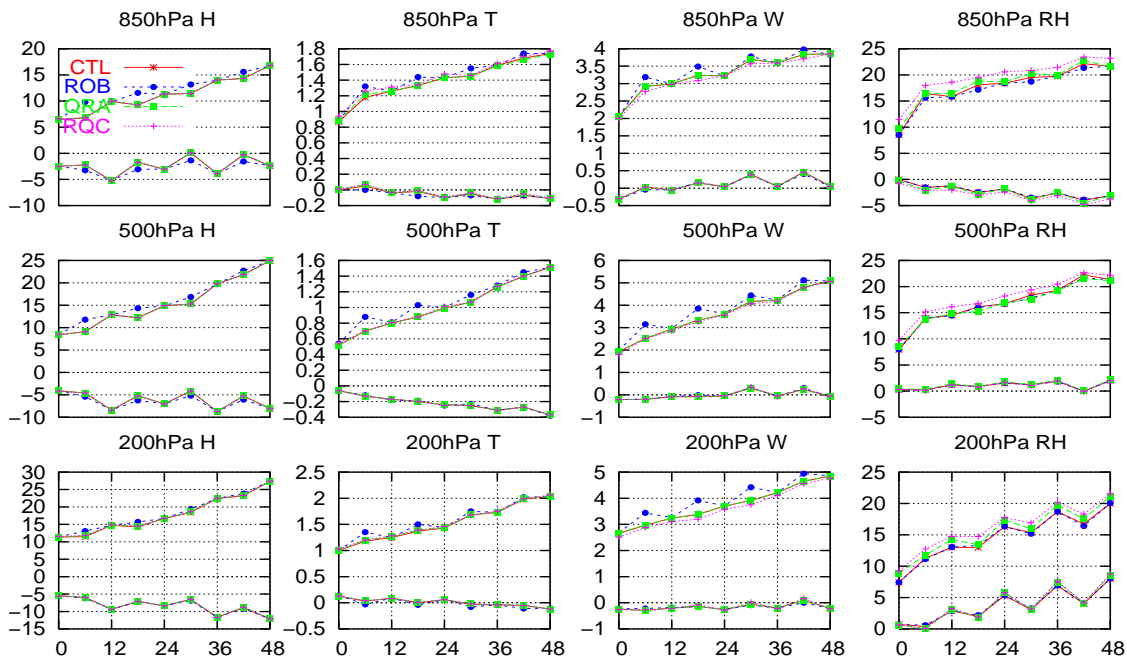


Figure 4: EWGLAM verification scores for key parameters calculated using HIRVDA package with different setups. CTL: standard HIRVDA package with verifying observation data set from MARS archive; ROB: using RCR observation data as verifying observation; RQC: using quality control criteria as in the reference verification package.

It is hence concluded, from the above discussion, that a reliable verification for model inter-comparison requires a common framework in which same verification software, same analysis or observation data and same quality control are applied to all model outputs.

4 Towards a common verification framework

Within the HIRLAM-A programme, a clear mission has been defined to improve quality of the HIRLAM forecast system and to demonstrate such with a measurable way. In concrete, a systematic mechanism has been requested by the HIRLAM council to monitor performance of the operational HIRLAM forecasts via regular model inter-comparison, both inside and outside of the HIRLAM consortia. Against that background, a reliable, functional common verification framework has been a target in the HIRLAM-A application project.

During the HIRLAM system workshop in Madrid, Oct 2006, the existing verification packages in HIRLAM communities were reviewed. In view of practical difficulties to harmonize, in near future, the verification softwares among member services, a consensus was reached among participants to aim for, a coordinated, regular model inter-comparison with central data management. It was proposed that within such a framework, operational centers will produce in real time a special output stream containing minimum set of model data which is suitable to be collected centrally for automatic, unsophisticated monitoring and comparison. For that purpose, a common verification software shall be adopted to generate verification products centrally and in real time.

As a follow-up to the Madrid system workshop, a special verification working meeting was held in Copenhagen, March 2007 with participation of key staffs working on verification issues. The meeting started with inventory on the currently available verification packages, followed by a discussion on verification packages suitable to be used for common verification and general model inter-comparison.

Main differences in verification methods

Between the reference package, on one hand, and other reviewed packages as used in DMI, met.no, SMHI,

HIRVDA, GL, on the other hand, there exists a clear difference in procedure for calculating observation verification statistics. In the former, verification scores are calculated per cycle, and the sum for a chosen episode/area/station-list are then aggregated using per-cycle statistics. In the latter, interpolated forecast data to a station-list is archived first, and calculation of verification scores is done in final step. The aggregation method in the reference package has the advantage that only very limited information needs to be saved per cycle to make final aggregation. The drawback is that many factors that affect final scores, such as definition of station-list, observation and analysis used for data selection have to be available when doing calculation per cycle. This limits flexibility in later use of such data for aggregation and implies differences between final scores generated on-line and off-line. As a contrast, by archiving only extracted model data to a pre-defined station-list, other verification packages have a flexibility to treat forecast data generated at the different time and place consistently, because it allows comparative verification for the extracted model data against a common set of observation that is screened by common analyses data set.

Towards convergence of verification methods

Based on analysis of the main differences between reference verification package and those used in other centers, a consensus was reached in the Copenhagen meeting to work out a modified reference verification package for model inter-comparison. The modification involves mainly addition of one output step, in the reference procedure, to archive, at each cycle, an intermediate (GRIB or ASCII) file which contains interpolated model data to a pre-defined, sufficiently extensive observation list. Such data is then used in final calculation of verification scores, as is done in other verification packages. In practice, the intended intermediate file contains information that is already generated in the reference verification procedure, thus the real change lies only in the final algorithm to aggregate statistics. Instead of using archived per-cycle statistics for aggregation, it now reads, similar to the practice in other verification packages, extracted model data per cycle to calculate final scores. In principle, such a change to the reference system will bring most of the available verification packages into convergence. In addition, this extension can be added as an extra procedure, so that the current reference package can retain all of its current functionalities for centers that have been using it, for sake of operational consistency in these centers.

Minimum data for common verification of operational HIRLAM

While the proposed modification to the reference package will enable it to be used in general purpose model inter-comparison such as impact studies, it is considered necessary to pursue with the idea for centralize model inter-comparison for operational HIRLAM forecasts. For such purpose, it is proposed that following minimum data set should be produced and collected in real time from operational systems:

- field data at 3 hour interval, from 0 to 48h, for 00 and 12 UTC runs, for pmsl, t2m, u10m, v10m, rh2m
- field data at 6 hour interval, from 0 to 48h, for 00 and 12 UTC runs:
 - acc.-precip., visibility, tot. cloud cover;
 - u, v, T, rh, Z for pressure levels 850, 500, 250 hPa;
- mast data for Cabauw, Sodankula, Valladolid, with the parameters in accordance with the available observation data in these towers.

With these data set collected, some other simple, real-time monitoring products can be produced, such as forecast charts for European area, meteogram for EWGLAM stations, mast profiles, etc. With the limited data volume as defined in such way, the Hirlam data server hirlam.org should be able to host such an action.

Acknowledgment The reported work has benefited from many discussions with HIRLAM colleagues. In particular, Kalle Eerola, Jose A. G. Moya, Ben Wichers Schreur, Ulf Andrae, John B Bremnes, Kai Sattler and Bjarne Amstrup are acknowledged for many stimulating discussions in connection with the the HIRLAM verification working meeting in Copenhagen, March 2007.