Simo Järvenoja’s inheritance: Long-term verification of Hirlam forecasts at the Finnish Meteorological Institute

Kalle Eerola

Finnish Meteorological Institute

1 Introduction

This study was initiated and inspired by a few figures, which we found in the bookshelf of the late Simo Järvenoja. The figures presented field verification results from the Finnish Meteorological Institute’s (FMI) Hirlam model for the period 1990-2006. The parameters in those figures were mean sea level pressure, 500 hPa geopotential and 925 hPa temperature. The first idea was to pick up the values from the figures by hand and compute the more recent verification scores from the Hirlam archive. However, it appeared that Järvenoja’s original verification data is available in FMI archives and after several trials and failures we succeeded in reading and interpreting the data. So we had the whole verification dataset for the period 1990-2006 available. On the other hand, for the period 2003 onwards it was possible to create the corresponding dataset from the Hirlam archives at FMI and ECMWF.

There is a common practice to mention three reasons for verification (Jolliffe and Stephenson, 2003). The general headings for these three reasons are administrative, scientific and economic. From the administrative point of view it is useful to compute a few numbers to describe the quality of the forecasts. They allow the quality of the forecasts to be followed from year to year and to judge the value of operational and research investments. The keyword in scientific verification is understanding. By verifying the forecasts in different ways it may be possible to understand the weaknesses of the models. This can then direct our research to improve the system. For instance, ECMWF continuously follows the systematic errors in their forecasts in order to reveal weak points of their model (see for instance Jung (2005)). The third category, economic verification looks at the quality of forecasts more from the user’s point of view.

This study falls mainly into the administrative category: A long time series of verification scores reveal, how well we have succeeded in the modeling work, if our forecasts have improved during the years and if the Hirlam model has been worth the investments. But there is also the scientific aspect. The improvements in scores after the implementation of new versions can sometimes be traced back to some meteorological changes in the model, which in turn helps us to understand our models better. But first of all, with this work we want to respect the late Simo Järvenoja and his work for Hirlam.

The scope of this study is to study the dynamical part of the Hirlam model and its behavior in the free atmosphere. Thus we do not verify the real weather parameters, like 2-metre temperature, 10-metre wind or precipitation. This is a natural selection, because we use the field verification concept, i.e. verification against numerical analysis. Weather parameter verification requires the verification against observations and different verification methods.

The structure of this report is the following. After this introduction the datasets and their processing are described on section two. Section two also describes the verification scores used in this study. A brief history of the operational Hirlam versions at FMI is given in section three. Then the results are shown in section four by presenting monthly time-series of verification scores followed by seasonal scores as a function of forecast length on selected seasons and years. Finally, summary is given in section five.
2 Datasets and the processing of the data

2.1 The original datasets

The data in this study consist of two datasets. Hereafter they are called as dataset A and dataset B. Dataset A was created by Simo Järvenoja as a part of the monthly monitoring of operational Hirlam systems at FMI during the years 1990 . . . 2006. Dataset B was created from the RCR archive at ECMWF using the standard Hirlam field verification package. In this section the contents and further processing of both datasets are described.

Dataset A was found in the FMI data archive. Simo Järvenoja had collected it as a part of monthly monitoring of operational Hirlam systems during the years 1990 ... 2006. Only parts of it have been published earlier in connection with other studies, and there was hardly any documentation. Earlier discussions with Simo Järvenoja and some Fortran source code lists were the only knowledge about the contents and format of the data. An extra difficulty was that the files had been written with Fortran unformatted write-statements. Thus the file structures were computer-dependent. However the creation times of the files revealed, on which computer the individual files had been created. Finally, after many trials and fails, we succeeded in reading all the data. To help the further processing all the data was first converted into ASCII-format.

Dataset A consisted of monthly sums of the following parameters:

- Analysis
- Analysis squared
- Forecasts every six hours from 6 to 48 hours
- Forecast squared
- Analysis times corresponding forecast every six hours

These sums and the corresponding numbers of cases had been computed and archived in every second grid point. The parameters consisted of mean sea level pressure and geopotential, temperature, relative humidity and wind components on selected constant pressure levels. Thus the data contained enough information for computing the monthly or seasonal verification scores for the period 1990 . . . 2006 for any area or sub-area of the original area.

Dataset B was created using the daily constant pressure level data from the RCR archive at ECMWF. The RCR archive contains all the analyses and forecasts from the FMI operational RCR Hirlam runs starting from February 2004. In addition, some Hirlam data, archived at FMI, was used. In total, the dataset B contains data from the time period from May 2003 to April 2008. These data were processed using the standard Hirlam field verification package. This package produces monthly values of bias and root-mean-square error (rms-error) for every field, which exists in the forecast and analysis files. The bias and rms-error fields are produced separately for forecasts starting from the 00, 06, 12 and 18 UTC analysis. The resulted fields are archived in GRIB-format.

2.2 The verification scores and further processing of data

In this study we concentrate on two verification scores, bias (mean-error) and root-means-square error (rms-error). Generally, the bias is defined as
In this study we concentrate on two verification scores, bias (mean-error) and root-means-square error:

\[ \text{bias} = \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i) \]  

and rms-error as

\[ \text{rms} = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (\hat{x}_i - x_i)^2} \]

where \( \hat{x}_i \) is the forecast in point \( i \) and \( x_i \) the corresponding observation or analysis and \( n \) the number of cases. The forecast and analysis fields are functions of space and time. Our aim is to get single scores for every month or season. Hence we want to take into account both the spatial and temporal variation. For this reason the final monthly or seasonal rms-error is computed as double sum over the time (over month or season) and space (over every grid point):

\[ \text{rms} = \sqrt{\frac{1}{n_s} \sum_{s=1}^{n_s} \frac{1}{n_t} \sum_{t=1}^{n_t} (\hat{x}_{st} - x_{st})^2} \]

where \( n_s \) is the number of grid points the area and \( n_t \) is the number of forecasts in a month or season. For bias or any linear score it does not matter in which order the mean value is computed.

The monthly and seasonal scores were computed for the two areas shown in Figure 1. The ATLEUR area is the area of the first operational Hirlam and it is the largest area, which is common to all operational Hirlam models at FMI. However, such a large area also filters out information, because different weather types are present at the same time in different parts of the area. Therefore the verification scores were also computed to a smaller Scandinavian area (SCANDI), which covers Scandinavia and the surroundings. It is naturally of special interest to Finland, but it also gives knowledge about the behavior of Hirlam in an area, which meteorologically is characterized by varying meteorological conditions. The Scandinavian area is also well covered by surface and upper-air observations. Thus the analysis is supposed to correspond closely to observations.

The two datasets A and B differ from each other in several respects. In dataset A the verification is done against uninitialized analysis, while in dataset B it is done against the initialized analysis. In dataset A only every second grid point is used in the computations, while dataset B contains every grid point. Also the accuracy, on which the results are archived, differs. In dataset A the monthly sums are archived with the accuracy of the floating point accuracy of the computer where the computations were done. On the other hand, in dataset B the bias and rms-error are packed into GRIB using the same accuracy as in the original fields.

On the other hand, in field verification the analysis and forecast values are in the same geographical points and the variables are the same. So the interpolation to the same points or conversion to the same unit are avoided. Also there is no need to consider the quality control of observations. Thus field verification avoids many difficulties, which must be considered when verifying against observations (Yang, 2007).

Fortunately, the period 03/2003 . . . 08/2006 is covered by both datasets making it possible to compare the scores computed based on the two datasets. It appeared that in monthly and seasonal values there are only marginal differences between the scores computed from the two datasets. Therefore we can combine two datasets and compute complete time-series of verification scores.
3 A brief history of Hirlam at FMI

The first Hirlam data assimilation and forecasting system (Hirlam 1) was developed during the years 1985-1988. It consisted of a data assimilation system based on optimal interpolation, a grid point semi-implicit forecast model and necessary pre- and post-processing software. FMI was the first institute to implement this system into operational use. This happened on 2 January 1990. Thereafter there has been a continuous series of Hirlam systems at FMI. Table 1 shows some characteristic features of different synoptic-scale operational Hirlam at FMI. In addition to these versions there has been most of the time a nested higher-resolution model inside the synoptic-scale system.

The left column in Table 1 shows the acronyms of different model versions. They are used throughout this report. The next column shows the time, when the version was in operational use. The updating frequency has increased clearly during the latest years. The first Hirlam version (FIN) was operational for more than four years. Also the version ATA was in use over three years. Recently the updating frequency has been about one new version every year. There has been a change in the working strategy. In early days of Hirlam new versions were tested for a long time at FMI before operational implementation. Modifications were made also on the fly to the operational system. Nowadays the Hirlam project tests new releases much more carefully before the official release of a new version and the operational implementation can take place much quicker. On the other hand, only severe bugs are corrected in an operational system. Minor problems can wait for the next release.

The number of grid points and number of levels reflects the available computer power. The total
number of grid points \((nx \times ny \times no\ levels)\) is now 75 times of that of the first Hirlam version. Version V621, which was introduced into operational use in February 2004, was the first RCR system of Hirlam. The RCR (Regular Cycle with the Reference) concept is an agreement between the Hirlam project and FMI. The main idea is that FMI adopts the latest Hirlam Reference System as the FMI operational synoptic-scale NWP system with minimum modifications. This guarantees that the reference system is tested so well that it can implemented into operational use. The concept has worked well and the Hirlam reference system is nowadays much more ready for operational use than earlier. In addition, all the RCR products are archived at ECMWF and are available for the whole Hirlam community.

4 Results

4.1 Time series of monthly verification scores

In this section the monthly rms-error and bias values for the two areas are presented. For the bigger area (area ATLEUR in Figure 1) the scores of +12, +24 and +48 hours’ forecasts are shown. For the Scandinavian area (area SCANDI in Figure 1) only the +48 hours’ forecasts are presented, but different Hirlam versions are shown to be able to identify the possible effects of model upgrades to the scores. In addition, the 13 months’ moving average is plotted in both figures to reveal the overall improvements by hiding the seasonal fluctuations.

Figure 2 shows the scores of mean sea level pressure for Atlantic-European area (upper panel) and for the Scandinavian area (lower panel). In both areas a prominent feature is the seasonal variation in the rms-error, which is a natural consequence of the stronger general circulation in winter time.

**Figure 1**: The verification areas. ATLEUR is the largest common area for all Hirlam versions at FMI. SCANDI is used to show a more active region.
From the Hirlam model point of view the most important feature is the clear reduction of rms-error. In the ATLEUR area a reduction from an approximate yearly value of 4 hPa to the value of 2 hPa can be seen. Thus, measured by the rms-error, the current +48 hours’ surface pressure forecasts are better than +24 hours’ forecasts in the early 1990s’ and almost as good as +12 hours’ forecasts then.

In the Scandinavian area, the fluctuation from month to month is larger than in the larger area. This is due to the smaller area, where the scores are more dependent on the local weather type. It is generally known that some weather types can be forecasted better than others. Also the reduction of rms-error is slower in the first years: the annual mean value goes below 4 hPa in the beginning of 2000s’, while in the larger area this happens almost ten years earlier. In the most recent years the rms-error in both areas have reached the same level.

Figures 2 also reveals that the reduction of rms-error has been large during the latest two years: introduction of the version 7.0-version (V641) led to a clear jump downwards. The most probable reason for this is the introduction of the large-scale mixing or re-forecast procedure (Yang, 2005). The previous cycle is re-run in order to utilize late-arriving, high-quality ECMWF analyses and forecasts (boundaries) to improve the quality of the large-scale structure of the background fields in the next data assimilation cycle. Another new feature is the implementation of the AMSU-A data in the assimilation. This should improve the analysis especially over the oceans, but the overall effect is smaller.

A well-known weakness in the earlier Hirlam surface pressure forecasts has been the too slow filling of cyclones. In the error fields this was seen as negative bias in Scandinavia and in the northern Atlantic. When introducing the version 6.4 (V637), an artificial turning of the surface stress vector was introduced (Sass and Nielsen, 2004). In the tests the negative bias was clearly reduced in the Norwegian Sea and Scandinavia (Järvenoja (2005) and Eerola (2005)). Figure 2 reveals that the effect is even too strong on both areas: the negative bias changes to positive bias. However, in the later versions the positive bias reduces close to zero.

Figure 3 shows the corresponding curves for the 500 hPa height. The main results are very similar.
to those of surface pressure. The rms-error of 48 hours’ forecasts is now on the same level, about 20 m, as that of 24 hours’ forecasts in the early 1990s’. The positive development during the latest two years is even more striking than in surface pressure. As mentioned earlier, the probable reason for this quick improvement is the introduction of large-scale mixing of the ECMWF data into the first guess of Hirlam analysis via the re-cycle phase. The ECMWF global data assimilation system uses a long cut-off time for observations and utilizes a lot of satellite data. So it supposed to be superior to the Hirlam assimilation system in the area of sparse conventional observation network and in upper atmosphere. Thus the Hirlam system benefits even more in the upper levels from the high-quality ECMWF analysis and boundary fields than in the surface. The positive development of verification scores in the latest years is even more pronounced in 300 hPa scores (not shown).

Figure 4 shows the monthly verification scores for the 850 hPa temperature in the Scandinavian area. There is a large negative bias in the first Hirlam version in the years 1990 . . . 1994. This feature was even more striking in the 925 hPa temperature (not shown). The problem was traced back to the too moist and cold lower troposphere and hence to the problems in the radiation scheme. Several ad hoc corrections were tried, but the introduction of the new radiation scheme (Savijärvi, 1990) in the next Hirlam version (SFI) improved significantly the situation. However, there is a small negative bias up to the most recent Hirlam versions.

During most of the time there have been only small improvements in the skill. However, there seems to be a reduction in the rms-error when implementing the ATX version in 2004. This was the first
version containing the three dimensional variational data assimilation (3DVAR) instead of optimal interpolation. Also it is noticeable that the seasonal variation in the rms-error has decreased in the latest Hirlam versions.

4.2 Forecast error as a function of forecast length
In this section the seasonal verification scores as a function of forecast length on selected years are shown. We concentrate on winter seasons, which means the three months’ period December, January and February.

Figure 5 shows the bias and rms-error of 300 hPa geopotential as a function of forecast length on some winter seasons on the Atlantic-European area. The years are shown in the legends. Note that in selecting years we have given more weight to the most recent years. We see a clear reduction of the rms-error during the years in all forecast lengths. The highest rms-error is clearly in the first winter, winter 1991. However, the winter 1994 shows much lower rms-error, although the basic Hirlam version is the same. This is probably partly due to the different weather conditions in these winters and partly due to the modifications in the model, for instance the implementation of 31 vertical levels. During the following years there is a gradual improvement in the scores in all forecast lengths. Also here we see a remarkable decrease in the rms-error during the latest two winters. Another important thing is that the error growth during the forecast has decreased during the years: the slope of the curves has decreased. This is again most prominent during the latest years. In the behavior of the bias there is no clear trend in the course of the years.

The 500 hPa height and surface pressure scores behave in a very similar way as those of 300 hPa height and are therefore not presented here. Instead, Figure 6 shows the scores of 925 hPa temperature. Here we see much less improvements after the first winter and the latest winters do not by no means have the lowest rms-error. The large negative bias in the first winter accounts for a large part of the exceptional large rms-error.

In the bias there is a slight cooling during the forecast in all shown years expect one. However, the cooling is less than 0.5 deg. in 48 hours. In summer the cooling is stronger than in winter on most years. It is noticeable here that there has been little, if any, improvement during the latest years in the 925 hPa temperature.

Figure 5: Seasonal bias and rms-error scores of 300 hPa height as a function of forecast length on selected winters as shown in the legend. Winter means a three months’ period December-January-February.

Figure 6: Same as Figure 5, but for 925 hPa temperature.
5 Conclusions

The inspiration of this study was the finding that it is possible to create a continuous time-series of monthly field verification scores since the beginning the Hirlam era at FMI, i.e. since the year 1990, thanks to the work by the late Simo Järvenoja. He had collected the monthly data for field verification statistics. On the other hand, the RCR archive at ECMWF made it possible to create the corresponding statistics for the latest years. Despite of the differences in methods, the both data sets give similar results. This could be verified, because the datasets overlap each other several years.

The purpose of this study can be classified to be mainly administrative. The trends in the long time-series of verification scores reveal the overall progress in the operational Hirlam system at FMI. But during the work it was also possible to identify some scientific issues by looking at the possible reasons for the changes in the scores related to upgrading of the Hirlam system. This can help to understand the relevance of the different scientific aspects.

The results show that there has been a substantial improvement in the Hirlam forecasts since the beginning of Hirlam runs in 1990. The rms-error of the mean sea level pressure and 500 hPa geopotential is nowadays about half of the value it was in the first versions of Hirlam in the early 1990s. The yearly rms-error has decreased from about 4 hPa to 2 hPa. The 500 hPa geopotential height rms-error shows a decrease from an approximate yearly value of 40 m to the value of 20 m. This means that two-day forecasts are now better than one-day forecasts in the beginning of 1990s. This development is in line with the experience at ECMWF. Their rule of thumb is that the length of useful forecasts has increased with one day in a decade.

The reduction in rms-error and error growth has been very prominent during the latest years. The most probable reason for this is the re-run concept and large-scale mixing of the ECMWF analysis to improve to first guess in the Hirlam data assimilation.

On the other hand, it was noticed that the same reduction in rms-error is not seen in the lower troposphere (850 hPa and 925 hPa) temperatures, where the improvements could only be seen in the first years of Hirlam.

The purpose of this study was not to verify the real weather parameters, like 2-metre temperature and humidity, 10-metre wind or precipitation, although these are the parameters that the users really need. Instead, in this study the emphasis was in the dynamical behavior of the Hirlam model. To verify the weather parameters it is necessary to do the verification against observations.

References
Eerola, K., 2005: Verification of Hirlam version 6.3.5 against RCR in autumn conditions. Hirlam Newsletter, 47, 72–90.

Järvenoja, S., 2005: Tests with the hirlam 6.3.5 version - comparison to some earlier versions and rcr. Hirlam Newsletter, 47, 41–62.


