Twenty-One Years of Verification from the HIRLAM NWP System

KALLE EEROLA

Finnish Meteorological Institute, Helsinki, Finland

(Manuscript received 5 July 2012, in final form 17 September 2012)

ABSTRACT

The High-Resolution Limited-Area Model (HIRLAM) international research program maintains a synoptic-scale NWP system. At the Finnish Meteorological Institute, the HIRLAM system has been run operationally since 1990. The HIRLAM forecasts from 1990 to 2012 have been verified against the numerical analysis. In 2-day forecasts, the monthly rms error of the mean sea level pressure has decreased from about 4 to about 2 hPa; that is, the error is now about half of the value it was in the early 1990s. Similar reduction is seen in the 500-hPa height. The negative bias has decreased significantly. In addition, the dependence on the weather regime, measured as the correlation between the North Atlantic Oscillation (NAO) index and rms error, has decreased. The reason for these improvements can often be attributed to changes in the HIRLAM system. A single improvement, improving most significantly the forecast skill, is the rerun concept, which improves the HIRLAM first guess by utilizing the high-quality ECMWF analysis. Verifying against observations or against the initial analysis gives similar results for a 48-h forecast. For a 6-h forecast, however, the field verification gives lower rms error values and lower bias values. In summary, the results indicate that the goal of the HIRLAM program has been fulfilled: to develop and maintain an up-to-date NWP system for 1- and 2-day forecasts on a limited domain.

1. Introduction

The High-Resolution Limited-Area Model (HIRLAM) international research program is a cooperative research effort among 10 European meteorological institutes. The aim of the program is to develop and maintain numerical short-range weather forecasting systems for operational use by the participating meteorological institutes. The program was initiated in 1985 and is on going. The first numerical forecasting tool developed by the program is called the HIRLAM system. It is intended to be used at horizontal scales down to 5–10-km horizontal grid spacing. It is a complete numerical weather prediction (NWP) system, containing a data assimilation system, a limited-area forecasting model with a comprehensive set of physical parameterizations, and pre- and postprocessing of observations and forecasts. The first HIRLAM version (HIRLAM 1) was developed during 1985–88. The Finnish Meteorological Institute (FMI) was the first institute to introduce it into operational use in January 1990. Since then, the HIRLAM system has been run at FMI on a routine basis. The verification system against the HIRLAM numerical analysis, hereafter called field verification, was implemented in June 1990. These verification data up to February 2012 form the basis of this study.

Jolliffe and Stephenson (2003) mention three reasons for verification. Their general headings are administrative, scientific, and economic. First, administrative verification monitors the overall quality of the forecasts and is normally done operationally on a regular basis. This information is used, for instance, to judge the needed or proposed human or financial investments. Normally, the information is condensed down to a small number of scores or numbers describing the essential features of the forecast quality. Second, in scientific verification, the keyword is understanding. By verifying the forecasts in versatile ways, the weaknesses and deficiencies of the forecasting system can be detected and understood. This will then direct the research and possibly improve the forecasting system. Normally, a large variety of tools and methods are needed and they may vary from case to case. However, certain standard scores are used on a regular basis. For instance, the European Centre for Medium-Range Weather Forecasts (ECMWF) monitors
the systematic errors in their forecasting system regularly (Jung 2005). Statistical verification scores, like those in this study, are one tool in a toolbox of scientific verification. They give a measure and geographical distribution for the error, but they do not reveal the reasons for the errors. Thus, they are the first step in the process of revealing and understanding weaknesses and deficiencies in the forecasts. The third category, economic verification, looks at the quality of forecasts more from the customer’s point of view and is not discussed in this study. This study focuses mainly to the first category, but also partly covers aspects belonging to the second category.

Research in all areas of NWP has been active during the last 20 years, both inside the HIRLAM program and worldwide. At the same time, advances in computer science have made it possible to run more sophisticated and higher-resolution models. A natural question then arises: How are these advances reflected in the quality of the numerical forecasts? A long time series of verification scores, computed in a uniform way, reveals the overall advance in the quality of the forecasts, thus helping to answer the following question: Have the investments to research and operations of NWP been worth doing? More broadly, has the HIRLAM program succeeded in one of its main goal to develop and maintain an up-to-date synoptic-scale NWP system? Since 2004, FMI has been the so-called Regular Cycle of the Reference model (RCR) center within the HIRLAM program, meaning that FMI runs the latest well-tested and accepted version of the HIRLAM system (reference system) with operational status. This gives more weight to the current results in judging the success of the HIRLAM program.

The verification dataset of this study consists of monthly field verification scores from the FMI HIRLAM forecasts from June 1990 to February 2012. Field verification limits the meteorological parameters to be verified to those analyzed by the data assimilation system. Important weather elements, such as precipitation, cloudiness, and 2-m temperature, cannot directly be verified with this concept. This study concentrates on mean sea level pressure (MSLP) and geopotential and temperature on constant pressure levels. The verification scores are the two conventional scores: the root-mean-square error (rms error) and the mean error (bias). Although every meteorological center running a NWP system also verifies their forecasts, there are not many published studies about the long-term behavior of the forecast quality.

The structure of this article is the following. Section 2 describes the history of the HIRLAM system by highlighting the major scientific developments during the latest 20 years. The data in this study consist of two datasets. These datasets and their further processing are described in section 3. Special emphasis is given to ensure that the two datasets together form a continuous and homogeneous time series. This section also contains a description and a discussion of the verification scores. Section 4 describes the results. Section 5 discusses the role of developments in improving the forecasts and also the role of weather type in the forecast skill. Also, a comparison of field verification and observation verification is discussed. Finally, a summary is given in section 6.

2. A brief history of the HIRLAM forecasting system

The HIRLAM program maintains a standard version of the HIRLAM forecasting system, called the HIRLAM reference system. It is a well-tested complete NWP forecasting system, containing everything that is needed for its operation: code, scripts, libraries, and tools. The FMI HIRLAM system follows the developments in the HIRLAM reference system and is regularly upgraded, when a new reference system is released. In addition to the major developments described in this section, numerous smaller upgrades, tunings, and corrections have been made.

Table 1 gives the characteristic features of the different HIRLAM versions at FMI, including the horizontal and vertical resolutions, the number of grid points, and the times they became operational. The first column gives the acronyms used throughout this study. In February 2004, FMI became the RCR center in the HIRLAM program and since then it has been running the latest reference system with minimal changes as its operational NWP system. Thus, the results since 2004 reflect the quality of the HIRLAM reference system.

The first version (FIN), implemented operationally at FMI in January 1990, consisted of a limited-area version of the ECMWF optimal interpolation data assimilation method (Lorenc 1981), a nonlinear normal mode initialization method (Machenhauer 1977), and a limited-area version of the ECMWF hydrostatic primitive equation gridpoint forecast model (Burridge and Haseler 1977). The physical parameterizations contained horizontal and vertical diffusion, radiation, surface processes, and convective and stratiform precipitation. All of them, except radiation, were based on the first ECMWF global model (Tiedtke et al. 1979).

The development of three- and four-dimensional variational data assimilation methods (called 3DVAR and 4DVAR, respectively) started in the 1990s. In March 2003, the 3DVAR system (Gustafsson et al. 2001; Lindskog et al. 2001) was ready to be implemented at FMI as a part of the ATX version. The first step toward 4DVAR data assimilation (taking time into account) was the
inversion approach (STRACO; Undén et al. 2002), which was for a long time based on the Soft Transition Con-

vention scheme developed by Savijärvi (1990) replaced it in 1994 in the SFI version.

The conventional in situ observations [synoptic ob-
servations (SYNOP), ship reports (SHIP), radiosonde soundings (TEMP), pilot reports (PILOT), aircraft re-
ports (AIREP), aircraft to satellite data relay (ASDAR), and drifting buoy (DRIBU)] are still the main observation types used in the FMI HIRLAM. So far only Advanced Microwave Sounding Unit (AMSU-A)/Ad-
anced TIROS (Television and Infrared Observational Satellite) Operational Vertical Sounder (ATOVS) sat-
ellite observations (Schyberg et al. 2003) are used oper-
ationally at FMI.

An alternative initialization method for normal mode initialization, digital filter initialization, was developed by Lynch and Huang (1992) and further generalized by Huang and Lynch (1993). The incremental form of digi-
tal filtering (Lynch and Huang 1994), using the Dolph–
Chebyshev window (Lynch 1997), was implemented into HIRLAM. At FMI, it was introduced into operational use in February 2004 in version V621.

In the dynamics, the trend in the 1990s was toward semi-Lagrangian advection schemes, which allowed longer time steps that were computationally more efficient. The two-time-level semi-Lagrangian advection in HIRLAM (McDonald and Haugen 1992, 1993) was implemented at FMI in March 2003 in version ATX.

The original HIRLAM radiation scheme had a ten-
dency to cool the lower atmosphere too strongly. There-
fore, a radiation scheme developed by Savijärvi (1990) replaced it in 1994 in the SFI version.

The stratiform and convective precipitation schemes were for a long time based on the Soft Transition Con-
densation approach (STRACO; Undén et al. 2002), which takes care of both large-scale and convective precipitation and puts special emphasis on achieving a gradual transition between both regimes. The stratiform precip-
sitation scheme developed by Rasch and Kristjánsson (1998) and the convection scheme of Kain and Fritsch (1993) were also tested in HIRLAM with good results. Their main advantage was to produce more realistic small precipitation amounts. At FMI, these schemes replaced the STRACO scheme in version V72 during September 2008.

The original surface scheme was the same as used in the first ECMWF global model (Tiedtke et al. 1979). Avisar and Pielke (1989) suggested a method based on the idea of mosaic tiles. The total vertical flux within a grid box is an area-weighted mean of fluxes in differ-
ent types of surfaces. Five surface types were defined: sea/lake water, ice, bare land, forest, and agricultural terrain/low vegetation. For the three land surface types, a two-layer Interactions between Surface–Biosphere–
Atmosphere (ISBA) scheme (Noilhan and Planton 1989; Noilhan and Mahfouf 1996) was implemented in version ATX during March 2003. In version V73, an extension to the previous scheme was implemented by introducing a more realistic snow scheme and a completely new forest formulation (Gollvik and Samuelsson 2010). For lake temperature and ice-conditions, the lake parameterization scheme FLake (Mironov et al. 2010) was introduced in version V74.

The original turbulent mixing scheme in HIRLAM was based on the first ECMWF model (Tiedtke et al. 1979). Later, the Cuxart–Bougeault–Redelsperger (CBR; Cuxart et al. 2000), based on a prognostic equation for the turbulent kinetic energy (TKE), was implemented into HIRLAM. At FMI, it was implemented in the ATA version during November 1999. The latest version, moist CBR, where the liquid water potential temperature is used instead of the dry potential temperature (Tijm and

### Table 1. Some characteristic features of the different synoptic-scale HIRLAM systems at FMI. Here, $n_x$ and $n_y$ are the number of grid points in the x and y directions, respectively, and $dx$ is the horizontal resolution (’). For details, see section 2.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Period</th>
<th>$n_x \times n_y$</th>
<th>$dx$</th>
<th>Levels</th>
<th>Version</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIN</td>
<td>Jan 1990–May 1994</td>
<td>$130 \times 100$</td>
<td>0.5</td>
<td>16</td>
<td>HIRLAM 1</td>
<td>31 levels in 1992, boundaries from two daily runs</td>
</tr>
<tr>
<td>SFI</td>
<td>Jun 1994–Aug 1996</td>
<td>$130 \times 100$</td>
<td>0.5</td>
<td>31</td>
<td>HIRLAM 2</td>
<td>Savijärvi radiation, new physiography</td>
</tr>
<tr>
<td>NSF</td>
<td>Sep 1996–Nov 1997</td>
<td>$194 \times 140$</td>
<td>0.4</td>
<td>31</td>
<td>HIRLAM 2.1</td>
<td></td>
</tr>
<tr>
<td>ATL</td>
<td>Dec 1997–Oct 1999</td>
<td>$194 \times 140$</td>
<td>0.4</td>
<td>31</td>
<td>HIRLAM 2.5</td>
<td></td>
</tr>
<tr>
<td>ATA</td>
<td>Nov 1999–Feb 2003</td>
<td>$194 \times 140$</td>
<td>0.4</td>
<td>31</td>
<td>HIRLAM 4.6.2</td>
<td>CBR, ECMWF boundaries 4 times day$^{-1}$</td>
</tr>
<tr>
<td>ATX</td>
<td>Mar 2003–Jan 2004</td>
<td>$256 \times 186$</td>
<td>0.3</td>
<td>40</td>
<td>HIRLAM 5.1.4</td>
<td>3DVAR, ISBA, semi-Lagrangian advection</td>
</tr>
<tr>
<td>V621</td>
<td>Feb 2004–May 2005</td>
<td>$436 \times 336$</td>
<td>0.2</td>
<td>40</td>
<td>HIRLAM 6.3</td>
<td>FGAT, first RCR, digital filter</td>
</tr>
<tr>
<td>V637</td>
<td>Jun 2005–May 2006</td>
<td>$436 \times 336$</td>
<td>0.2</td>
<td>40</td>
<td>HIRLAM 6.4</td>
<td></td>
</tr>
<tr>
<td>V641</td>
<td>Jun 2006–Mar 2007</td>
<td>$436 \times 336$</td>
<td>0.2</td>
<td>40</td>
<td>HIRLAM 7.0</td>
<td>Rerun concept</td>
</tr>
<tr>
<td>V71</td>
<td>Apr 2007–Aug 2008</td>
<td>$583 \times 448$</td>
<td>0.15</td>
<td>60</td>
<td>HIRLAM 7.1</td>
<td></td>
</tr>
<tr>
<td>V72</td>
<td>Sep 2008–Sep 2010</td>
<td>$583 \times 448$</td>
<td>0.15</td>
<td>60</td>
<td>HIRLAM 7.2</td>
<td>4DVAR, Kain–Fritsch</td>
</tr>
<tr>
<td>V73</td>
<td>Oct 2010–Feb 2012</td>
<td>$583 \times 448$</td>
<td>0.15</td>
<td>60</td>
<td>HIRLAM 7.3</td>
<td>Improved surface scheme</td>
</tr>
<tr>
<td>V74</td>
<td>Mar 2012–present</td>
<td>$1030 \times 816$</td>
<td>0.07</td>
<td>65</td>
<td>HIRLAM 7.4</td>
<td>Flake lake model</td>
</tr>
</tbody>
</table>
more than 100,000 times greater than that in 1990. The power available to NWP at FMI during spring 2012 is in the first HIRLAM version in 1990. The computing directions, respectively) in version V74 is 262 times that (nx ny) where nx and ny are number of grid points in x and y directions, respectively) in version V74 is 262 times that in the first HIRLAM version in 1990. The computing power available to NWP at FMI during spring 2012 is more than 100,000 times greater than that in 1990.

3. Data and processing of data

a. The verification datasets

The data in this study consist of two datasets, hereafter called dataset A and dataset B. Dataset A has been created as a part of the monthly monitoring of the operational HIRLAM system at FMI during the years 1990–2006 by using software developed at FMI. It contains the monthly values of the following quantities in every grid point: sum of the forecasts (\( \sum_{i=1}^{n} \tilde{x}_i \)), sum of the forecasts squared (\( \sum_{i=1}^{n} \tilde{x}_i^2 \)), sum of the analysis (\( \sum_{i=1}^{n} x_i \)), sum of the analysis squared (\( \sum_{i=1}^{n} x_i^2 \)), and sum of the product of the forecast and verifying analysis (\( \sum_{i=1}^{n} \tilde{x}_i x_i \)), where \( \tilde{x}_i \) is the forecast, \( x_i \) the verifying analysis, and \( n \) the number of cases in a month. Bias and rms error can be computed from these quantities. The meteorological parameters consist of MSLP and geopotential height, temperature, relative humidity, and wind components on several constant pressure levels.

Dataset B has been created from the archived daily data using the standard HIRLAM field verification package. It contains geographical fields of monthly mean, bias, and rms error values for the same meteorological parameters as dataset A. This dataset contains data from May 2003 onward.

The two datasets differ from each other in several respects. In dataset A, the verification is done against the uninitialized analysis, whereas, in dataset B, it is done against the initialized analysis. In dataset A, the scores are computed in every second grid point (only 25% of the total number of points), whereas dataset B contains every grid point. In addition, dataset A has been archived as floating point values of the originating computer, differing from computer to computer. Data- set B has been packed in gridded binary (GRIB) format with fixed accuracy. Fortunately, the two datasets overlap each other from March 2003 to June 2006. Figure 1 shows the comparison of the rms error of MSLP for different forecast lengths from these two datasets. The values are almost identical, except for one month, May 2003. Closer examination revealed that dataset B during this month was incorrectly created, containing only part of the month. The final time series of verification scores were constructed using dataset A until January 2004 and dataset B from February 2004 onward.

b. The verification scores

Bias and rms error are two commonly used verification scores. Generally, bias is defined as

\[
\text{bias} = \frac{1}{n} \sum_{i=1}^{n} (\tilde{x}_i - x_i)
\]
and rms error as

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

(2)

where \( \bar{x}_i \) is the forecast at point \( i \), \( x_i \) is the verifying observation or analysis, and \( n \) is the number of cases. The forecast and analysis fields are functions of space and time. To get a single monthly or seasonal rms error and bias value, both the spatial and temporal variations are taken into account. The final monthly or seasonal rms error is computed as a double sum over time (month or season) and space (grid points):

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

(3)

where \( n_s \) is the number of grid points in the area and \( n_t \) is the number of forecasts in a month or season. Correspondingly, a double sum is used to calculate bias, as well.

The temporal and spatial structures of bias are important parameters in verifying NWP models (e.g., Jung 2005; Elmore et al. 2006a,b). Biased forecasts lead to an incorrect model climate, which is especially dangerous for general circulation models, but the growth of bias with forecast length in short-range forecasts is a sign of a problem in the model.

The rms error is not independent of bias, because it contains a contribution both from the bias and standard deviation of the forecast error. In rms error, the forecast error is squared, meaning that it gives larger weight to larger errors but does not show the direction of the error. It is flow dependent, and a large contribution to rms error comes from the variability of the forecasts and verifying analyses, not only from their difference. Therefore, rms error is larger in winter when the variability of the atmosphere is larger and the general circulation stronger.

Baldwin et al. (2001) demonstrate, in a theoretical example, how rms error favors smooth fields and punishes especially detailed forecasts with phase errors. This is called the double-penalty problem: rms error increases twice, because a detailed pattern is not in the right place and is in a wrong place.

The monthly verification scores have been computed for several areas, the characteristics of which differ from each other concerning the observational network and typical weather conditions. In this study, the results are shown for two areas (Fig. 2). The Atlantic–Europe

---

**Fig. 2.** The three verification areas. ATLEUR is the largest common area for all HIRLAM versions at FMI. SCANDI contains Scandinavia with its surroundings. EWGLAM is the area with a good observation network and is used when comparing observation and field verification scores. The entire map shows the integration area of version V73.
(ATLEUR) area is the largest area, common to all synoptic-scale HIRLAM implementations at FMI. It covers the northern Atlantic Ocean and Europe. We also show the verification scores over a smaller Scandinavian area (SCANDI), which covers Scandinavia and its surroundings, and is well covered by surface and upper-air observations. This region is the most important target area for the FMI HIRLAM forecasts. The EWGLAM area is used to compare the verification scores computed against observations and numerical analyses.

In addition to rms error, the growth of rms error as a function of forecast length (called error growth in this study) is also discussed. The error growth is computed as a difference in the rms error of two different forecast lengths for three forecast length intervals, from +12 to +24 h, from +24 to +36 h, and from +36 to +48 h.

4. Results

a. Time series of monthly verification scores

Time series of monthly rms error and bias of MSLP and 500-hPa height are shown in Figs. 3 and 4. A characteristic feature in rms error is the seasonal fluctuation, which is due to the stronger general circulation in winter. The variation in bias from month to month is more irregular. Negative bias, increasing with the forecast length, is prominent in winter, especially in the 1990s. In the SCANDI area, this feature is stronger and is also visible in recent years.

The rms errors of both MSLP and 500-hPa height have decreased over the years in both areas. For MSLP, the reduction is from about 4 hPa in the early 1990s to about 2 hPa in the most recent years in the 48-h forecasts (Fig. 3). The corresponding numbers for 500-hPa height are 40 and 20 m (Fig. 4). Measured by the rms.
error, the current HIRLAM 2-day forecasts in the ATLEUR area are better than 1-day forecasts in the early 1990s and almost as good as 12-h forecasts then. In the SCANDI area, the rms errors in 36- and 48-h forecasts are larger in the 1990s compared to the ATLEUR area, but reach the same error level after the model revision in 2006. Because the error is downward limited by the analysis error, the gap between 48- and 12-h forecasts has decreased. The variation from month to month is larger in the early years, partly due to larger bias. The decrease in rms error after the system upgrade in 2006 is worth noting. After that, the quality has been relatively stable, except during the winters of 2010/11 and 2011/12, when there is a slight increase in the rms error of the MSLP, but not in the 500-hPa height. During the first two winters (1990/91 and 1991/92), the rms error in the ATLEUR area (Figs. 3a and 4a) was larger than in the subsequent years even if the model version had not changed. However, during these years, several improvements were made, including an increase in the number of vertical levels from 16 to 31, introduction of real sea surface temperatures instead of climatology, and fresh lateral boundaries from ECMWF twice a day. In 2004, 2005, and partly 2006, there is an increase in rms error compared to the previous years, especially in the SCANDI area. In 12-h forecasts, rms error and its seasonal fluctuation are reduced after the introduction of the ATX version in March 2003. The reduction seems to be larger in the shorter, 6-h forecasts, and higher at the 300-hPa level (not shown). This is obviously related to the introduction of 3DVAR during March 2003.

The seasonal variation in the rms error has decreased especially after the model upgrade in 2006. In the 1990s, the seasonal variation in MSLP is about 2 hPa in the ATLEUR area and 3 hPa in the SCANDI area. In recent years, the corresponding values are of the order 1 hPa. In the SCANDI area, the reduction is partly due to the decrease in wintertime bias. In 2-day forecasts, this was typically −2 hPa in the 1990s, whereas the rms error was about 6 hPa, meaning that over 10% of the

![Fig. 4. As in Fig. 3, but for 500-hPa height (m).](image-url)
total error variance was due to the bias. In the ATLEUR area, this effect is smaller.

The variation in bias of both MSLP and 500-hPa height from month to month is more irregular than in rms error. In the SCANDI area, the bias is negative in most of the winter months, increasing with the increasing forecast length. The bias is very prominent in the 1990s, with a maximum monthly value of $-4.1$ hPa for MSLP and $-31$ m for 500-hPa height in 48-h forecasts in December 1993. The negative bias in the winter months exists even in the latest versions. In particular, the peaks in December 2010 and February 2012 are prominent. Despite several attempts to understand the bias in the case of December 2010, the reason for the large negative bias could not be found. In February 2012, the variance of the analysis was very large in Scandinavia, showing that the weather was quite very variable (not shown). In winters 2010/11 and 2011/12, the negative bias increasing with lead time also appears in the ATLEUR scores. The question of whether the feature is also present in the current operational version (V74) remains open and needs attention in the coming winter of 2012/13.

The bias is larger, more variable, and irregular in the small SCANDI area than in the large ATLEUR area. Partly this is due to the compensation of positive and negative values in a large geographical area. However, comparison of the bias over the Atlantic Ocean and over the European continent reveals that the bias over the Atlantic is smaller and less variable than over the continent (not shown). This suggests that over the ocean, with a limited number of observations, the changes are slower, because the verifying analysis is based mainly on the short forecast from the previous cycle.

Figure 5 shows the geographical distribution of the monthly bias of MSLP in 2-day forecasts during February 2012. Although the magnitude of the bias is extreme in February (see Fig. 3), the geographical structure is typical throughout the most recent 10 years during the winter: negative bias occurs in the eastern and southeastern parts of Europe and positive bias occurs in northern Canada, but the strength varies from year to year and from month to month. In the 1990s, the structure was more variable, depending on the prevailing weather regime.

Figure 6 shows the monthly scores of the 925-hPa temperature. From 1990 to 1993, a prominent feature is a large negative bias. This pattern is also reflected in the rms error. The bias is normally largest in spring and is reflected also into the screen-level temperatures, which were too cold in spring (not shown). Closer examination revealed that the lower troposphere below 850 hPa was too moist and cold, thus causing low stratus cloud cover. Several corrections were made in the early years: increasing the number of levels from 16 to 31, introducing analyzed sea surface temperatures, and using fresh boundaries from ECMWF twice a day. Importantly, the introduction of higher vertical resolution in 1992...
decreased the cold bias, along with the introduction of a new radiation scheme (Savijärvi 1990) in 1994. Even later, the negative bias continued to be a prevailing feature with a peak often in spring. There is less improvement in the rms error until the introduction of the 3DVAR assimilation system in 2003 in the ATX version. The improvement in temperature then is seen at all levels and in both areas (not shown). Especially in the ATLEUR area, the change is clear. After 2004, there is a slight increase in rms error in the shorter, +12-h, forecasts, which lasts to 2008 and remains stable after that. If we look the time period from 2004 to 2008 separately for the continental area and sea area, the increase is concentrated only on the sea area and there it is seen at the 925- and 850-hPa levels, but disappears at higher levels (not shown).

b. Forecast error as a function of forecast length

Figure 7 shows the growth of the rms error (referred as error growth) for MSLP, computed as the growth of the rms error in 12 h, for three different forecast lengths. The first 12 h are not considered because of the possible spinup problems.

First, there is the same seasonal variation with higher values in winter as in the rms error itself. The trend of decreasing values is similar, as well. In the ATLEUR area (Fig. 7a), the seasonal cycle is regular, and the improvements occur mainly when introducing a new version. After 2006 the annual cycle becomes more regular. In the SCANDI area (Fig. 7b), the annual cycle is larger and the variability around the annual cycle is larger and more irregular, especially during 1990–2000. Normally, there is no significant variability in the available observations from month to month. So, in case the model version has not changed, the explanation for the differences between months likely lies in the prevailing weather regime. It has long been known that the predictability of NWP models varies in different weather regimes (e.g., Branstator 1986; Palmer 1988).

Based on Fig. 7b, the whole dataset can be divided into four periods. The first period consists of 1990–2000, when the error growth has large variability, especially in
winter. The rms error decreases steadily, but the variability from month to month remains large and the annual cycle is strong. During the second period, from 2000 to 2003, the error growth is much reduced and has a more regular annual cycle. The decreased error growth and reduced annual cycle mean reduced rms error and reduced annual cycle, as seen in Figs. 3 and 4. The ATA version was operational for the whole period. During the third period, from 2004 to 2006, the error growth increases, especially in winter. During this period, there were three different model versions in use. In winters 2003/04 and 2004/05, very high error growth values can be seen in the SCANDI area between the 12- and 24-h forecasts. Large error growth is reflected as higher rms error values. Different from MSLP and 500-hPa height, the rms error of 925-hPa temperature decreases, which is due to the introduction of 3DVAR data assimilation (Fig. 6). The fourth period occurs from 2007 onward. The amplitude of the error growth is small and the annual cycle is quite regular, especially in the ATLEUR area (see also Figs. 3, 4, and 6).

In summary, the HIRLAM forecasts have improved significantly during the last two decades. The rms error has decreased so that the rms errors of MSLP and 500-hPa height in 2-day forecasts are now lower than the corresponding values in 1-day forecasts 20 yr ago. At the same time, the seasonal cycle of the rms error has decreased by about half. In bias, the negative values in winter have been particularly reduced. The improvements are concentrated on the implementation times of new model versions, showing the benefits of the new versions. In the first two years, there were improvements in both scores (Figs. 3, 4, and 6), even if the model version did not change. This was due to many modifications in the first version of the HIRLAM system.

5. Discussion

a. The importance of system upgrades

The results from the previous section indicate that there have been remarkable improvements in the quality of the HIRLAM forecasts during the past 21 yr. In this
section, we link Table 1, showing significant system upgrades, to the quality of forecasts discussed in the previous section.

During 1990–2012, conventional in situ observations were the backbone of the observational data in the HIRLAM data assimilation system at FMI. The only satellite data being used were the AMSU-A/ATOVS data since 2006. The amount of conventional upper-air data has not increased, except for aircraft observations. For surface data, the number of observing stations has not increased, although the frequency of the observations has increased. Hence, the improvements are probably not due to better observational data, but are due to improvements in the data assimilation and forecast models, and due to better lateral boundaries.

Both in MSLP and 500-hPa height (Figs. 3, 4 and 7), the reduction of rms error and error growth has been remarkable after introducing version V641 in June 2006. In this version there were two major changes: the introduction of AMSU-A/ATOVS observations and the introduction of the rerun concept (often also called large-scale mixing). According to Eerola (2005), the impact of AMSU-A/ATOVS data was only marginally positive. Some improvements could be seen in the longer forecasts (up to +48 h) and in the upper troposphere. Geographically, the improvements were concentrated over the Atlantic Ocean and western coast of Europe. Therefore, the main reason for improvements was the introduction of the rerun concept (Yang 2005), which utilizes the latest ECMWF analysis. The ECMWF global data assimilation system utilizes more satellite observations than HIRLAM, has a long cutoff time for observation collection, and is global. All of these features improve the analysis, especially over the oceans and other areas where the conventional observational network is sparse. Therefore, it is supposed to be superior to HIRLAM analysis in these areas and in the upper atmosphere. However, it is not available at the time of HIRLAM forecast runs. To utilize the high quality ECMWF analysis in HIRLAM in the tight daily schedule, the previous, 6-h-old, HIRLAM data assimilation cycle is repeated just before a new cycle starts. By this time, the ECMWF analysis for that earlier cycle is available. The initial analysis for the HIRLAM recycle is a combination of the ECMWF analysis and the HIRLAM analysis. This analysis mixes the two sources using the digital filter approach, utilizing the high quality, large-scale structure of the ECMWF analysis, while preserving the small-scale structures of the HIRLAM analysis. A short forecast then gives the best possible background field for the HIRLAM data assimilation system.

The period 2004–05 shows degradation in the quality of MSLP and 500-hPa geopotential forecasts, especially in the SCANDI area (Figs. 3, 4, and 7). Of note in the SCANDI area is that the bias is large. The geographical structure of the bias in winter 2004/05 resembles that of February 2012 (Fig. 5) and, similar to it, the variability is large in Scandinavia. January 2004 looks different in both respects. The reason for the degradation is either the weather conditions or that the ATX, V621, and V637 versions were not behaving as well as the previous ATA version. Note that in those years many new developments became operational and the model versions changed frequently. Perhaps the combined effect of these new features and their interactions were not tested thoroughly enough, as indicated by Järvenoja (2004) while testing V621. However, in 925-hPa temperature forecasts, the signal is not clear: there are improvements in SCANDI area, but the opposite is found for the ATLEUR area (Fig. 6).

Figure 6 reveals that the 925-hPa temperature forecasts were improved after the introduction of the ATX version in March 2003. This feature could be seen also in temperature scores at other levels (not shown). The seasonal oscillation in the rms error was reduced, and the negative bias value was also reduced. The rms error decreased by more than 0.5 K. However, the seasonal oscillation in the bias remained, with negative values in winter and positive values in spring. There were three important new features in this version: 3DVAR, semi-Lagrangian advection, and an ISBA surface scheme. The improvements were largest over the Atlantic Ocean and more modest over the European continent (not shown). The semi-Lagrangian advection scheme mainly improved the efficiency, not the quality. So the most probable reason for improvement was 3DVAR.

In spite of the improvements, there have also been occasional disappointments when introducing new model versions. An example is version V637, implemented in 2005. In that version, an artificial turning of the surface stress vector (Sass and Nielsen 2004) was implemented in order to correct the weakness of filling cyclones too slowly. Figures 3 and 4 reveal that the effect was perhaps too strong: the negative bias changed to positive. Later, this modification was replaced with a more physically based approach by modifying the CBR scheme (Tijm and Lenderink 2003).

b. Effect of the weather type on verification scores

Cassou et al. (2004) found that there are four weather regimes on the monthly scale over the North Atlantic–European area. Two of them are related to the positive and negative phases of the North Atlantic Oscillation (NAO). The NAO index (e.g., Hurrell and Deser 2009) is widely used to classify the strength of the jet stream in the North Atlantic and Europe. A positive NAO index
means westerly flow over the Atlantic, and a negative index means blocking. It is well known that NWP models have large temporal variability in skill. We speculate that strong westerly flow and strong cyclone activity in North Atlantic affect the skill of the HIRLAM forecasts in Europe, especially earlier in the dataset, when fewer observations were available over the Atlantic. This hypothesis was tested by computing the correlations between the NAO index and the rms error and bias of the 48-h forecasts in the SCANDI area. The correlations between rms error and NAO for December–February were 0.37 for 500-hPa geopotential and 0.36 for MSLP, respectively. Taking separately the winters 1990–99 and 2000–12, the correlations between the NAO index and rms error were, respectively, 0.41 and 0.42 for 500-hPa geopotential and for MSLP over the first period and 0.25 and 0.21 over the second period. Although the correlations were not very high, HIRLAM was more sensitive to the monthly weather type defined by the NAO index in the early years. At that time, the integration area was smaller, the western boundary being partly over the Atlantic Ocean. The boundaries were received only once or twice a day, the rerun concept using ECMWF analysis was not used, and no satellite observations were used.

As an example of different prevailing weather regimes and their influence on the bias in the early years of HIRLAM, Fig. 8 shows the monthly mean MSLP and related 48-h bias for January 1992 and December 1993. In January 1992, high pressure was situated over western Europe, and the jet stream took a northern route. In December 1993, the Icelandic low was quite deep and zonal flow over the Atlantic was very strong. The monthly NAO indices for these months were −0.66 and 1.36, respectively. The geographical error maps for these months reveal a clear difference in the structure of bias. In January 1992, the bias was positive almost over the whole European continent. On the other hand, during December 1993, very large negative bias covered the whole of western Europe. This is an indication that the cyclones were too deep in the areas of interest for European weather.

c. Comparison of verification against observations and analysis and comparison to other models

So far our discussion has concentrated on the field verification scores, where the forecasts are compared to the numerical analysis. However, Casati et al. (2008) argue that this method gives scores that are too optimistic because the background for the analysis is a short-range forecast and the background field is used to quality control observations. This filters out small-scale features present in the observations. On the other hand, Simmons and Hollingsworth (2002) argue that this should not be a problem, at least for the medium-range forecasts in the Northern Hemisphere midlatitude troposphere.

The longest time series of observation verification scores of HIRLAM forecasts at FMI extends to the year 1995. The verification scores computed against numerical analysis and against observations were compared over the European continent (the EWGLAM area in Fig. 2), where the observational network is good both for surface and upper-air data.

Figure 9 compares the verification scores of MSLP for 6- and 48-h forecasts. For 6-h forecasts (Fig. 9a), the field verification gives systematically lower rms error values than the observation verification, and the difference increases in the recent years. In bias, the observations verification gives slightly positive bias, whereas the field verification is almost unbiased in the most recent years. In 3DVAR and especially in 4DVAR, data assimilation and forecast models are more tightly coupled together than earlier in the optimal interpolation. Thus, the previous cycles have a larger effect on short forecasts via the background field.

For 48-h forecasts (Fig. 9b), the rms errors given by the two verification scores are similar. In the first ten years of the period, the observation verification seems to give even slightly lower values in winter.

After all the improvements discussed earlier, is the HIRLAM system comparable to other limited-area models (LAMs) and can the products be used safely in everyday forecasting? In the EUMETNET SRNWP-V program, five different European LAMs and the ECMWF global model are compared in a unified way, using the same methods, the same verifying observations, and the same areas (Wilson and Mittermaier 2011). According to these results, no model has a significant advantage over the others. For some parameter, some models are better but for other parameters other models perform better. Compared to ECMWF scores, the LAMs show their value in the near-surface parameters, especially in the 10-m winds.

6. Summary

The HIRLAM system is a complete NWP system developed and maintained by the HIRLAM consortium. It has been operational at FMI since January 1990. More than 10 major and numerous minor upgrades have been made since then, the horizontal and vertical resolutions have improved, and the integration area has been enlarged. The forecasts have been verified against analyses since summer 1990. The focus of this study is on the general quality of the HIRLAM forecasts in light of this verification dataset. Thus, the purpose can be classified
FIG. 8. Monthly mean MSLP (hPa) and bias in 48-h forecasts during January 1992 and December 1993.
mainly as administrative. The trends in the long time series of verification scores, compiled in a uniform way, reveal the overall success or failure of the HIRLAM system and are important information for decision makers. In addition, the purpose is to see if the changes in the scores could be explained by the changes in the data assimilation or forecast system. This aspect is more scientific in nature. It reveals which changes have had the largest impact on the improvements in skill.

The data were drawn from two different datasets. The first covered 1990–2006, and the second 2004–12. By comparing the overlapping period, it was possible to show that they were comparable to each other and could be used as one homogeneous time series.

There has been a substantial improvement in the HIRLAM forecasts since the beginning of HIRLAM runs in 1990. In 2-day forecasts, the rms errors of MSLP and 500-hPa height are currently about half of the values of the first version of HIRLAM. The average yearly rms error of MSLP has decreased from about 4 to 2 hPa. Similarly, the rms error of 500-hPa height has decreased from an approximate yearly value of 40 m to the value of 20 m. This means that 2-day forecasts are now equal to or better than 1-day forecasts in the beginning of the 1990s. Also, the negative bias that was very prominent especially in winter in the early years has almost vanished. These developments are in line with the experiences in other centers running NWP systems. The often-used rule of thumb is that the forecasts have in the past improved by 1 day in a decade.

The conventional in situ observations have been the backbone of the observation data in the HIRLAM system at FMI during the whole period, and little has happened in the quality or amount of observations. Therefore, the improvements have come from the improvements in the HIRLAM system itself (data assimilation, model dynamics, and parameterizations of physical processes) and from the quality of lateral boundaries. Normally, the improvements have proceeded with small steps. However, an example of a big improvement is the introduction

![Graph](image-url)
of the rerun concept, where the ECMWF high quality analysis is used to improve the large-scale structure of the HIRLAM first guess of the data assimilation. In the temperature scores of the lower troposphere, the introduction of 3DVAR seems to be a milestone. Its influence can be seen both in the rms error and in the bias. In the very first version, the large negative bias was corrected by several steps (increasing the vertical levels, new radiation scheme, etc.). On the other hand, there are few cases where the introduction of a new model version diminishes the scores.

From everyday experience, we know that the weather type also affects the forecast quality (i.e., some weather types are more difficult to predict than others). This was studied by computing the correlation between the NAO index and rms error in the SCANDI area. In the first 10 yr, there seemed to be some positive correlation, indicating that, in westerly flow with increased cyclone activity in the Atlantic toward Europe, (i.e., positive NAO index), the forecasts are worse than in the case of a negative NAO index. In recent years, this relation has vanished.

In 6-h forecasts, the verification against observations gives higher rms error values than verification against numerical analysis. This is due to using the short forecast as a background for the analysis. In 48-h forecasts, both verification methods give essentially similar results.

As a summary, the quality of the HIRLAM system has improved during the years and the quality of the forecasts indicates that the goal of the HIRLAM projects has been fulfilled: to develop and maintain an up-to-date NWP system for 1- and 2-day forecasts on a limited domain.

Acknowledgments. The late Simo Järvenoja collected the dataset A during the years 1990–2005. I am also grateful for him for many discussions in the early years of HIRLAM. I am especially grateful to Prof. David M. Schultz for encouragement, suggestions, and patience during the long way with the manuscript. Drs. Carl Fortelius and Laura Rontu are acknowledged for comments and discussions during different phases of the manuscript. The comments of the two anonymous reviewers helped to improve the manuscript.

REFERENCES


Kain, J. S., and J. M. Fritsch, 1993: Convective parameterization for mesoscale models: The Kain–Fritsch scheme. The


