

HIRLAM 4D-VAR : where we are and where we go

Xiang-Yu Huang^{1*}, Nils Gustafsson² and Per Undén²

¹Danish Meteorological Institute, Denmark

²Swedish Meteorological and Hydrological Institute, Sweden

November 13, 2003

1 Introduction

A review and a critical discussion of the present status and development of HIRLAM 4D-VAR are presented in this paper. The intention is to provide background information for decisions regarding future HIRLAM research activities in data assimilation and also to provide an analysis of the computational resources needed for operational application of HIRLAM 4D-VAR.

The starting point for our review and discussion is the successful operational implementation of 4D-VAR by several NWP centers (section 2), in particular by the ECMWF and the JMA (Japanese Meteorological Agency). From this and from experiences with 4D-VAR reported in the literature, we then discuss advantages and disadvantages of the 4D-VAR approach for meteorological data assimilation (section 3). The historical development (section 4) and the current status (section 5) of the HIRLAM 4D-VAR are then briefly reviewed. More details can be found in Huang *et al.* (2002). After this review and this discussion we are prepared to provide our vision for the future development and operational implementation of HIRLAM 4D-VAR (section 6). Our vision forms the basis for our plan for further developments of the HIRLAM 4D-VAR (section 7), a plan that covers the nearest future (2003-3004) as well as the more remote future after the end of the HIRLAM 6 project. The long term plan is of course very uncertain, in particular with regard to recent discussions on possible future collaboration between the HIRLAM community and other modelling groups for the development of a high resolution forecasting system, including a non-hydrostatic model. The required computer resources for running HIRLAM 4D-VAR operationally are discussed in section 8 and a brief summary is given in section 9.

Further references to 4D-VAR theory and formulations could be found in, for example, Huang *et al.* (2002) and are therefore not repeated in this paper. A recent review paper by Park and Županski (2003) gives good overview of almost all the methodology developed for 4D-VAR during the years.

* *Corresponding address:* Danish Meteorological Institute, Lyngbyvej 100, DK-2100 Copenhagen Ø, DENMARK. Email: xyh@dmi.dk

2 Successful operational implementations of 4D-VAR

2.1 ECMWF

4D-VAR was introduced operationally at ECMWF in 1997 after several years of parallel testing and comparison with 3D-VAR. Before the operational introduction of 4D-VAR, the positive impact of 4D-VAR versus 3D-VAR on the forecast verification scores was clearly demonstrated in extensive parallel tests. It is not completely clear what aspects of 4D-VAR provides the largest contribution to these improved forecast verification scores and to what extent a 3D-VAR with “First Guess at Appropriate Time” (FGAT) would have given a similarly good impact. One important advantage of 4D-VAR versus 3D-VAR/FGAT is of course the use of high temporal resolution observations, for example the dynamical effect of the implicit use of pressure tendencies via high temporal resolution surface pressure observations in 4D-VAR. The other limitation of 3D-VAR/FGAT is that the analysis increments are applied only at the center of the time window, usually wrong time for asynoptic observations.

An exception to the general improvements provided by 4D-VAR was that the forecast quality in tropical areas was slightly degraded by introduction of the first version of operational 4D-VAR at ECMWF. This problem is related to the basic assumption of a perfect forecast model in 4D-VAR and the limitations of the simplified physics package of the first version of operational 4D-VAR at the ECMWF. (It was a very simple physics package, essentially surface friction, applied in the first inner minimization loops, followed by the more complete simplified physics in the second). The other problem was how to balance the increments in the tropics, since incremental NMI (adiabatic) was then used. Afterwards the incremental DFI soft constraint has been introduced.

Since the operational introduction of 4D-VAR at ECMWF, there has been a continuous and steady improvement of forecast verification scores. It is not clear to what extent improvements in the 4D-VAR algorithms (12 hour assimilation window, improved physics in inner loops, increased number of outer loops) have contributed to these improvements.

Further experience at ECMWF actually indicates that one might have to revert to a 6 hour window as the resolution increases and the (moist) physics becomes ‘more realistic’. There has also been a lot of effort to assimilate precipitation estimates from SSM/I data with particular effort into the moist physics in 4D-VAR. Very impressive results have been shown for particular tropical areas and cyclones. Non-linear effects of 4D-VAR like transpositions of cyclone centers and precipitation bands have been shown.

2.2 Météo-France

4D-VAR has also been introduced in the operational global forecast model ARPEGE at Météo-France. Météo-France also reported about improved forecast verification scores with 4D-VAR as compared to forecasts based on 3D-VAR. The limited area ALADIN system has also a 4D-VAR option, although this has only been used for research. The ALADIN operations are for relatively small areas and mainly for dynamical adaptation at high local resolution. Only a few members run data assimilation so far (3D-VAR in Morocco is operational).

2.3 Japan Meteorological Agency

An operational 4D-VAR coupled to a spectral mesoscale model was introduced operationally the Japan Meteorological Agency in March 2002. The JMA 4D-VAR has many characteristics in common with the HIRLAM 4D-VAR: the spectral model formulation; the assimilation control vector includes unbalanced wind components (with a scale dependent decoupling of geostrophy, however); use of the NMC method for background error statistics.

Important for the successful application of the JMA mesoscale 4D-VAR is the inclusion of important physical parameterization in the inner loop tangent linear and adjoint model integrations: large scale condensation; evaporation of precipitation; moist convective adjustment; vertical diffusion; radiation; water loading.

The JMA 4D-VAR is applied with a model resolution of 10 km for outer non-linear minimization loops and with a model resolution of 20 km for inner minimization loops. Only 20 “iterations” (gradient evaluations) are applied in the inner minimization loops for computational economy reasons.

An important aspect of the JMA mesoscale 4D-VAR is the control of the lateral boundary conditions via assimilation increments on the lateral boundaries and a background constraint for these lateral boundary increments.

Important for the JMA mesoscale 4D-VAR is also the use of 1 hour precipitation observations. Very impressive improvements in precipitation forecast verification scores have been reported, the reference in this case was a statistical interpolation based assimilation system.

3 Discussion about the advantages of 4D-VAR, problem areas and alternatives

From a theoretical perspective there are strong arguments in favour of 4-dimensional variational data assimilation in comparison with more empirically based assimilation techniques like nudging or 3-dimensional methods like 3D-VAR even with FGAT:

- The initial model state is determined through a stringent optimization process, taking all available observations and all relevant a priori knowledge into account.
- The iterative solution of the optimization problem allows for non-linear relations between observed quantities and model state variables - this is very important for the utilization of many types of remote sensing observations.
- The iterative solution can also provide non-linear analysis impacts like transferring a whole vorticity or precipitation pattern from one location to another instead of just distorting it like a 3D method would do.
- The time dimension is taken into account through the forecast model as a strong constraint in the optimization process. This provides us with sequence of model states, consistent with the model equations, and being as close as possible to available observations over the assimilation time window.
- Unobserved phenomena may be analyzed through the model history as demonstrated in idealized studies. E.g. lower troposphere observations to analyze the

upper levels and for analyzing cyclone centers with the aid of only more peripheral observations.

- Inclusion of the forecast model into data assimilation process will make it possible to assimilate not only forecast model state variables, but also diagnostic quantities available through the time integration of the model. Important examples are surface pressure tendency and precipitation intensity.
- It will be made possible to assimilate efficiently observations with a high resolution in time. Examples are surface observations, radar observations and ground-based GPS observations.
- 4D-VAR will provide us with implicit flow-dependent assimilation structure functions, describing, for example, tilting baroclinic structures.
- It can also provide implicit physiographically realistic structure functions due to orography of land/sea contrasts.
- At a mesoscale (a few km) resolution the model generated structure functions are probably much more important than any explicit balance constraints specified for the background term. There are no obvious balance constraints known that can be applied for e.g. the storm scales, and thus 3D methods will be very deficient in the mesoscale.
- 4D-VAR will provide us with a possibility to avoid an external explicit initialization at the start of the forecast to be run from the analyses. It can also be argued that 4D-VAR may reduce completely the spin-up of physical processes during the first hours of the forecast. The first effect is made possible through the application of weak “gravity-wave constraint” in the 4D-VAR procedure. The second effect necessitates, in addition, the application of reasonably realistic physical parameterizations in the 4D-VAR procedure.

The success of the JMA operational application of a mesoscale 4D-VAR for assimilation of precipitation observations to improve precipitation forecasts points to a key issue for further development of HIRLAM 4D-VAR: **The need to introduce moist physics in the inner loop minimizations.** It is needed for the use of new data-types like precipitation intensity from radars and satellites. It is also needed to reduce the spin-up of moist processes.

The introduction of moist physics in the inner loop minimizations as well as the introduction of several outer loop minimization cycles in the HIRLAM 4D-VAR will most likely reduce the spin-up problems. Taking a somewhat longer time perspective, this will make it possible to use NWP model and 4D-VAR also for nowcasting and very short-range forecasting purposes. During a recent workshop at EUMETSAT, discussing the needs of observations for NWP and Nowcasting with a 2015-2025 perspective, it was strongly argued by Brian Golding (Met Office), that NWP models and 4D-VAR are the only way forward for nowcasting and short-range forecasting of precipitation. (Of course it was not complete consensus among the workshop participants on this issue.)

From a theoretical point of view, these arguments are of course very appealing. It should be pointed out, however, that there are also theoretical arguments against 4D-VAR. Some of these theoretical arguments against 4D-VAR may be considered less significant, taking the obvious advantages of 4D-VAR into account. There are also possible

improvements to be introduced into the 4D-VAR algorithm that would reduce these theoretical disadvantages. More important, however, is the likely huge development effort needed to develop, test and validate a fully efficient 4D-VAR, in particular for a mesoscale model, an area of research that is still fairly unexplored.

- One basic weakness of 4D-VAR, at least as it is applied operationally at present, is the assumption of a perfect forecast model over the assimilation window - the forecast model is applied as a strong optimization constraint. Several leading scientists (Bennet, Evensen) in the forefront of data assimilation research have taken this argument seriously, rejected the idea of 4D-VAR and started developments in other directions, for example the application of the model as a weak constraint (the representer method) and the ensemble Kalman filtering (EnKF). However, it is not believed that the EnKF is any better in treating the model errors, since the pdf is around the ensemble mean which is affected by model systematic errors. There exist possibilities to partly compensate for this basic weakness of 4D-VAR, however. The original idea of Lewis and Derber (1985) to use some quantity representing model errors, for example tendency bias, in the assimilation control vector, has recently received re-newed interest. There also exist possibilities to manipulate observations error statistics in such a way to optimize the fit to observations at the end of the assimilation window, thus to be as close as possible to the observations at the time of the initial state for the forecast is to be issued. Zupanski *et al.* (pers. comm.) have successfully included a model error in the control variable and recovered a model error in the regional ETA 4D-VAR system.
- The incremental 4D-VAR approach is based on linearization of the forecast model equations around a model trajectory being close to the true development of the real atmosphere. It may be argued that such linearizations are impossible, or very difficult, taking strongly non-linear processes like convection into account. The EnKF avoids, at least explicitly, this difficulty. The full non-linear model is used and there is no need for TL and AD model code. (Although the real difficulties in strongly non-linear processes does not go away). To solving the heavily non-linear problem of 4D-VAR for a mesoscale model, one needs to introduce the different spatial scales and the different processes stepwise into the minimization, starting with quasi-linear near-adiabatic synoptic scale processes and introducing smaller scales and physical processes gradually. For this purpose it is needed to have access to a range of regularized and simplified tangent linear and adjoint physical parameterization schemes to be applied during different phases of the minimization. The efforts needed to develop and maintain such physics packages for 4D-VAR should not be under-estimated.
- 4D-VAR does not provide an error estimation, and even less, any error covariances. The perhaps biggest argument for EnKF is its associated output of an estimated pdf of the forecast errors that can be used naturally and directly in an ensemble prediction system (EPS), which we are convinced will be necessary at all resolutions.

Ensemble Kalman Filtering has been mentioned under all the problem points above and deserves to be discussed further. EnKF may well be seen as a viable alternative to 4D-VAR and can probably provide many of the advantages of 4D-VAR while also addressing the problem areas of 4D-VAR. The cost of running EnKF is estimated to be similar to

4D-VAR given that one needs at least 50 members and 4D-VAR may converge in 20 iterations (while the TL+AD models are 3 times as expensive as the normal non-linear model). The development cost is much less, as the TL/AD modelling is not needed. Hence has the method received a lot of attention and is a popular research area on the American continent. There are however quite large practical problems in controlling the covariances and to make the “localization” effective, to both avoid spurious long distance correlations and to make the analysis problem tractable (cf the OI method or the PSAS method from the NASA DA Office, both relying on inverting localized matrices). Furthermore, the ensembles may need to, from time to time, be “re-aligned” to remove possible insignificant members. Also covariances will have to be ensured to have the properties of covariances, as at least to be positive definite matrices.

4 The development of HIRLAM 4D-VAR

The development of HIRLAM 4D-VAR started in 1995 following the discussion and suggestions by Gustafsson *et al.* (1997). The adiabatic Eulerian tangent linear model (TLM) and adjoint model (ADM) were developed in 1995—1996. This development was documented in connection with studies on sensitivity of forecast errors to initial states and lateral boundaries (Gustafsson and Huang, 1996; Gustafsson *et al.*, 1998). The ADM was also used to improve the Optimal Interpolation based HIRLAM data assimilation system at that time (Huang *et al.*, 1997).

Further development of the HIRLAM 4D-VAR was on the TLM and ADM of the parameterization schemes of physical processes. With the help of the tangent linear and adjoint model compiler (Giering, 1999), the TLM and ADM of the full HIRLAM physics package were developed during 1997—1999 (Yang, 1998).

In 2000, the HIRLAM 4D-VAR was put together for the first time and a demonstration was made over a 4-day period with $110 \times 100 \times 16$ model grid points. The results were encouraging, but the high demand on computing time made more experimentation difficult.

In the following two years, we made the following two major developments: the incremental formulation (Gustafsson *et al.*, 2001) and the implementation of the simplified physics packages from ECMWF (Buizza, 1994) and Météo-France (Janisková *et al.*, 1997). An overview of the physics packages available in HIRLAM 4D-VAR, including TLM and ADM, was given by Yang (2002).

In 2002, we made our first extensive data assimilation experiments with realistic model configurations. The results from these experiments are reported in a HIRLAM technical report (Huang *et al.*, 2002).

5 The current status

The status HIRLAM 4D-VAR in July 2003 could be summarized as follows (Huang *et al.*, 2002):

- After many experiments, a HIRLAM 4D-VAR configuration has been chosen for a feasibility and impact study, Briefly speaking, $1/3 - 1/2$ resolution for inner loops and the minimum physics package were selected.

- The HIRLAM 4D-VAR is a complete system, although further refinements should improve its performance (to be discussed later).
- From the few storm cases in December 1999, 4D-VAR analyses (using the default configuration) produce deeper and more realistic lows than 3D-VAR. The overall meteorological impact from 4D-VAR, according to the observation verification scores, is marginally positive for the selected periods.
- The cost of the default 4D-VAR Eulerian configuration is about 7 times the cost of 3D-VAR or about the same as the cost of a 72h Eulerian forecast. This is with Eulerian integration; the semi-Lagrangian scheme is being implemented and the final cost is not yet known, but is estimated to be reduced by a factor 0.2 - 0.5 (and more if more advanced physics is used).

All data assimilation experiments were carried out with the default configuration described above and the conclusions were mainly valid with this configuration. There were problems related to radiosonde (and pilot) report time stamps, which may have lead to degradation of 4D-VAR in comparison with 3D-VAR. The observation verification scores and a number of maps may give an incomplete picture of the meteorological aspects. Some work should be placed on examining, *e.g.*, the spin-up problems and precipitation forecasts.

A very low (≈ 135 km) inner loop resolution (the outer loop resolution is ≈ 45 km) and the minimum physics package were chosen due to computational considerations.

6 The Vision

In order to form a vision for the development and operational utilization of HIRLAM 4D-VAR, we first of all need to set a scenario for the future development of operational numerical weather prediction in Europe. At present, the ECMWF is responsible for medium range NWP while the weather services take care of nowcasting and short range forecasting. However, in order to produce the best medium range forecasts, ECMWF also produces the best short range forecasts. For political reasons, the operational distribution of these short range forecasts is not permitted. We are convinced that this will change within a time range of 5-10 years. By Year 2010 it is likely that ECMWF will produce short range forecasts 4/day with a model resolution at approximately 10 km. From this we may conclude that weather services in a longer term perspective should concentrate on mesoscale modelling to be applied, in particular, for nowcasting, very short range forecasting and forecasting for severe weather. For the nearest 1-5 years or so, the weather services need also to carry out NWP on the Atlantic scale for operational short range forecasting with model resolutions in the range 10-40 km. For these models we need 4D-VAR and the developments of the 4D-VAR for the Atlantic scale can also be seen as a preparation for the mesoscale 4D-VAR. Furthermore, it is not yet clear to what extent an Atlantic scale NWP model with similar physics as the mesoscale model is needed to produce consistent lateral boundary condition for the mesoscale model.

In 2003 and 2004 a number of planned improvements to the 4D-VAR algorithm will be made, several of them already in good progress (see section 7 for details). In 2004 we will make real efforts to demonstrate real time 4D-VAR on the synoptic scale for a 20 km resolution Reference Atlantic area. The data usage will include ATOVS radiances, radar winds and scatterometer data in addition to the conventional data. Forecasts shall

be provided (or demonstrated to be possible) 3 hours after observation time, with an observation cut off time of 2 hours and with a production time of 1 hour. Possibly there will be ensemble forecasting as well, but that is another issue.

One can ask what is provided above what ECMWF will provide at 25 km at this time. First of all, the ECMWF products will not be provided in time for short range forecasts. There is something like a 12 h delay as opposed to a 3 hour delay (after observation time main hour) for HIRLAM. Furthermore, a local system is needed to provide a comprehensive data base of all levels and parameters for all forecasting and model applications. HIRLAM forecasts utilizing HIRLAM 4D-VAR will be available at least 4/day and the HIRLAM forecast quality is expected to gain from the use of regional observing systems like radar and ground-based GPS data also at forecast model resolutions in the range 10-20 km.

While the push towards full operational 4D-VAR continues, there should be research effort into developing the 4D-VAR system further to the mesoscale application in 2005. Even though the computer resources will not be anywhere close to sufficient for operational application over any complete national area at e.g. 3 km at this time, there should be research at high resolution with a small area but with high resolution observational input. Radar radial winds is one such input and also radar reflectivity will be worked on. Imagery data from METEOSAT is another source with high resolution both in time and space. Significant developments of the moist physics will be necessary and experiences at ECMWF used as far as possible, even though the resolutions are quite far apart.

Most **importantly**, 4D-VAR is the only assimilation technique that has proven to be able to give a consistent and sustained impact from, for example, precipitation observations to the dynamical model variables (temperature and wind) that survives model integration longer than a few hours. This is strongly needed for nowcasting and very short range forecasting.

7 Recent improvements and planned work for the next few years

There are many steps we can take which may lead to possible improvements of HIRLAM 4D-VAR both in meteorological aspects and in computational aspects. Some steps are easy and straightforward. Some steps need in-depth research and substantial code developments.

7.1 Year 2003

The main theme of Year 2003 activities is to finish the main development:

- Complete and implement the semi-Lagrangian time integration in HIRVDA. This will lead to significant computational savings. A nonlinear semi-Lagrangian code (Gustafsson and McDonald, 1996) was already available in the spectral HIRLAM. This code has now been inserted successfully in the HIRLAM 4D-VAR framework and the corresponding tangent linear and adjoint codes have been developed and tested. This had the effect that the timestep used in 4D-VAR could be increased by a factor 3-4, but the computational savings were reduced to a factor 1.5-2 since this

semi-Lagrangian scheme is quite computationally involved. During 2003, the SETT-S semi-Lagrangian integration scheme (Temperton *et al.*, 2001; Hortal, 2002), applied at the ECMWF, is being implemented in HIRLAM 4D-VAR. Since SETT-S is a more clean two-time-level scheme, it is computationally less demanding, and the computational savings are expected to be of a factor 3-4 for a 3-4 times longer timestep as compared to the Eulerian scheme. The physics interface has been updated, but not to the recent physics package.

The work on the tangent linear and adjoint code has to be completed and tested. More exact computational savings will be estimated in connection with 4D-VAR experiments.

- Enable the outer-loop and develop the multi-incremental formulation. The former is necessary to treat highly nonlinear observation operators properly. The latter will speed up the minimization convergence, by starting with a very low resolution and then to gradually increasing it during the iterations. It will lead to gradually higher accuracy both for innovation vectors for the observations and for the full physics. Problems experienced at Météo-France and ECMWF and solutions to those will be taken on board.

The outer loop and multi-incremental formulation has been introduced into HIRLAM 4D-VAR during summer 2003. The use and configuration of the tangent linear, simplified, physics will be further investigated in connection with the multi-incremental approach as well as when using new data sources like radar reflectivities.

- Use gridpoint HIRLAM for the 4D-VAR background. The reference HIRLAM is formulated in gridpoint space while the assimilation model in HIRLAM 4D-VAR is a spectral model. One step to bridge the gap is to use the gridpoint model to produce the nonlinear trajectories.

This option has been made available during summer 2003 and has been under further testing since then.

- Further test the tangent linear and adjoint physics. The meteorological impact of the Météo-France simplified physics package should be assessed. Special attention will be paid to moist variables and physics related fields.

Until now, all studies of the impact of HIRLAM 4D-VAR have been run with very simplified physics without any condensation processes, for example, in the inner loop minimizations. Some meteorological testing of the impact of Météo-France physics needs to be carried out.

- Remove inconsistencies like the timing of radiosonde data from the current HIRLAM 4D-VAR and re-assess the meteorological impact of 4D-VAR.

At the time of the writing of the HIRLAM 4D-VAR (Huang *et al.*, 2002) it was found, but never understood, that HIRLAM 4D-VAR gave a positive impact compared with HIRLAM 3D-VAR for forecasts starting at 0600 UTC and 1800 UTC, while the impact at 0000 UTC and 1200 UTC was neutral or weakly negative. Recently a possible explanation has been found. Radiosonde reports at the main observational hours 0000 UTC and 1200 UTC are assimilated at the time of the radiosonde launch (mostly 2300 UTC and 1100 UTC) in HIRLAM 4D-VAR, while these data are

assimilated at nominal observation time (0000 UTC and 1200 UTC) (at which time they are in the mid troposphere) in HIRLAM 3D-VAR. This is the time when the sondes are most likely to be in the mid troposphere and this is also the timing of the use of radiosonde data for verification purposes. This gives an unfair advantage for 3D-VAR. At 0600 UTC and 1800 UTC, the positive impact of HIRLAM 4D-VAR most likely is related to the dominance of non-synoptic data like AIREP. Even more recently (May 2003) it became obvious that the same problem occurred with PILOT data (thanks to a comprehensive observation data base made and shown by Ole Vignes). The above problems have since been corrected for both FGAT (First Guess at Appropriate Time) and 4D-VAR.

- To improve the moisture analysis, the moist part of the current 4D-VAR will be re-formulated - a new moisture assimilation control variable with a near Gaussian statistical distribution will be introduced.

7.2 Year 2004

After Year 2003, the main focus will be on extensive data assimilation experiments to tune the HIRLAM 4D-VAR. The resolution will be increased to the operationally interesting ones of around 20 km (with increments at 40 km or initially 60 km) at least for some of the studies:

- Interesting cases will be selected.
- Long runs for different seasons.
- Data impact studies using 4D-VAR. In particular, wind profiler data, radar data and ground based GPS data are expected to give larger impact in 4D-VAR than in 3D-VAR.

Furthermore, a few refinements will also be made and tested:

- The weak digital filter constraint for HIRLAM 4D-VAR will be refined and its meteorological impact will be assessed. This has been shown, at ECMWF and Météo-France, not only to reduce the noise of the 4D-VAR analyses but also to improve the meteorological aspect in the forecasts.
- To take the lateral boundaries into account, a weak lateral boundary constraint will be formulated, probably following the JMA approach.

The lateral boundary values can be modified by the HIRLAM 3D-VAR at present and these modifications of the lateral boundaries also have the chance to penetrate into the inner model area during the course of the forecast (the initial data are used as the first lateral boundary data set). At present, lateral boundaries in HIRLAM 4D-VAR are not modified - HIRLAM 4D-VAR provides zero increments on the lateral boundaries. A control of the lateral boundaries, including a background constraint for the lateral boundaries, has been introduced in the JMA regional model 4D-VAR. The JMA approach needs to be tested also in HIRLAM 4D-VAR .

- The HIRLAM 4D-VAR will be prepared to assimilate precipitation data.

Finally, issues related to operational implementations will be addressed:

- The optimal ratio between the time spent on data assimilation and the time for running deterministic or even ensemble forecasts.
- Possible combinations of 3D-VAR (FGAT) and 4D-VAR.

7.3 After Year 2004

After Year 2004, there should be real effort put into the mesoscale assimilation. It will be in parallel to HIRLAM mesoscale model development, but the actual perturbation model in 4D-VAR may be simpler than the full model (although it will probably include some non-hydrostatics and particularly a good moist physics package). The computational requirements will be very demanding but it is important to gain some experience at very high resolution.

The plan of work after Year 2004 will of course depend very much on decisions on possible collaboration with other modelling groups for the development of the high resolution forecasting system. It will probably be a good strategy for HIRLAM to keep research and development of 4D-VAR assimilation techniques as a “HIRLAM profile” in such possible collaboration projects.

8 Time estimations and optimization of HIRLAM 4D-VAR

The major concern with 4D-VAR seems to be the expense in computational time vs the demonstrated and expected benefits. It is argued above that many of the expected benefits and future needs of 4D-VAR have not yet been realized, so it is necessary to take a long term strategic view. As a background, some time estimations in relationship to the forecast model are given below, for different resolutions.

8.1 Computational requirements estimated for HIRLAM 4D-VAR

The measurements of CPU-time for several 4D-VAR and non-linear model configurations were carried out on the ECMWF Fujitsu VPP5000 computer and reported in Huang *et al.* (2002).

Here we are reporting our recent estimations of the computational requirements for the semi-Lagrangian version of the HIRLAM 4D-VAR, based on some timings of the Eulerian version of HIRLAM 4D-VAR on a SGI 3000 computer. We have roughly estimated the computer time for running the semi-Lagrangian version of the HIRLAM 4D-VAR on the same computer with 64 processors available for operational NWP calculations. The background for this estimate is the (brave) SMHI intention to run HIRLAM 4D-VAR over an Atlantic area with 20 km horizontal resolution and with 40 vertical levels on a LINUX-based PC-cluster with 64 processors in 2005 (the processing speed of each processor will be 2-4 times that of SGI 3000).

The results of the timing on SGI 3000 for 438×310 horizontal gridpoints, with 40 vertical levels and with 22 km horizontal resolution are included in Table 1. The non-linear forecast model with a 1.5 min timestep carried out a +3 h forecast in 425 seconds on

31 SGI 3000 processors. For the 4D-VAR minimization with the simplified Buizza vertical diffusion, 20 minimization iterations (gradient calculations) with a horizontal resolution of 44 km for the increments was carried out in 7474 seconds.

Table 1: The computational costs (in seconds) for 20 iterations with the Eulerian version of the HIRLAM 4D-VAR and for a 3h forecast on a SGI 3000 computer at the National Supercomputer Centre (NSC) in Linköping, Sweden. The experiments were performed for data from 0600 UTC 1 December 1999. The resolution for minimization, mini-res, is in degree. The computational area consisted of 438×310 horizontal points with a grid resolution of 0.2 degree and with 40 vertical levels. The number of processors used for the 4D-VAR experiment was 20, while 8 and 31 processors were used for the forecast experiments.

run	proc-no	NL-physics	TL/AD-physics	mini-res	CPU (s)	mini-no
4D-VAR	20	HIRLAM	Adiabatic	0.4	6396	20
4D-VAR	20	HIRLAM	Buizza	0.4	7474	20
03hFC	8	HIRLAM	-	-	1469	-
03hFC	31	HIRLAM	-	-	425	-

Recently a timing on the SGI 3000 computer was also carried out for the non-linear spectral HIRLAM model with the Eulerian as well as the SETTLS semi-Lagrangian time integration scheme. This timing was for a forecast area with a 44 km horizontal resolution. The cost per timestep was approximately the same for the Eulerian and the semi-Lagrangian integrations, while it was possible to increase the timestep from 3 min for the Eulerian run to 10 min (and even 15 min) for the semi-Lagrangian run. This will allow us to use a cost reduction factor of 3 for estimation of the computational cost for 4D-VAR with semi-Lagrangian time integration.

The intention of SMHI is to run the operational HIRLAM over an area with 306×306 horizontal gridpoints, 40 vertical levels and a 22 km horizontal resolution from 2004. Estimates of the computational cost for running 4D-VAR on this area with a SGI 3000 computer are given in Table 2:

Table 2: Estimated computational costs (in minutes) for 60 iterations with the semi-Lagrangian version of the HIRLAM 4D-VAR and for a 48h forecast on a SGI 3000 computer. The resolution for minimization, mini-res, is in degree. The computational area consists of 306×306 horizontal points with a grid resolution of 0.2 degree and with 40 vertical levels. The number of processors is 64.

run	proc-no	NL-physics	TL/AD-physics	mini-res	CPU (min)	mini-no
4D-VAR	64	HIRLAM	Buizza	0.4	30	60
048hFC	64	HIRLAM	-	-	13	-

Taking into account that the PC-cluster processors are 2-4 times faster than the SGI 3000 processors and the possibilities for further code optimizations, it seems quite safe to state the 4D-VAR with the required model configuration could be run on the planned SMHI PC-cluster with a production time in the range 20-40 minutes.

It has been estimated that a 64 processors SGI 3000 has a sustained performance of about 5 GFLOPS for HIRLAM applications. To be on the safe side, we can assume that a computer with a sustained performance of 10 GFLOPS is needed for running an Atlantic scale HIRLAM 4D-VAR operationally at 20 km resolution. From this we may also give rough estimates of the required computer performance for future operational model scenarios, see Table 3:

Table 3: Rough estimates of required sustained computer calculation performance for running the semi-Lagrangian version of the HIRLAM 4D-VAR in a number of current and future model configurations.

Model version	Resolution	Number of gridpoints	Assimilation window	Computer performance
Atlantic scale	40 km	$150 \times 150 \times 40$	6 h	1.5 GFLOPS
Atlantic scale	20 km	$300 \times 300 \times 40$	6 h	10 GFLOPS
Atlantic scale	10 km	$600 \times 600 \times 60$	6 h	140 GFLOPS
Mesoscale	3 km	$500 \times 500 \times 100$	3 h	500 GFLOPS
Mesoscale	1 km	$1500 \times 1500 \times 100$	3 h	13000 GFLOPS

Besides the requirements on computational capacity, HIRLAM 4D-VAR also put strong requirements on memory size. This is not a problem for a computer of PC-cluster type, but it may be very expensive on a super-computer of type NEC.

8.2 Semi-Lagrangian time integration

The semi-Lagrangian version of the HIRLAM 4D-VAR has been discussed in detail above, and it is estimated that this version will run 3-5 times faster than the Eulerian version of the HIRLAM 4D-VAR. There is certainly also room for further code optimizations, for example, in the treatment of Fourier transforms and in inter-processor communications for the semi-Lagrangian interpolations.

Furthermore, the current Reference physics should be introduced in the spectral HIRLAM, which is far from negligible work. When this is done one should also consider the improved coupling of physics with the semi-Lagrangian scheme following Wedi at ECMWF and Martinez in HIRLAM.

8.3 The multi-incremental approach

The introduction of outer loops in HIRLAM 4D-VAR will give us full freedom to try also the multi-incremental approach to the HIRLAM 4D-VAR. Different spatial resolutions, different semi-Lagrangian interpolation techniques and different physical packages can be used in different sets of inner loop minimizations. This will probably results in improved assimilation quality but will above all provide us with a reduction of the computer time that we need.

Special caution has to be taken as to the degree of resolution differences and how particularly moist physics and turbulence are represented at different resolutions. Météo-France has reported about difficulties when the resolutions were vastly different, and

consequently the physical parametrization worked very differently at the lowest resolution compared with the highest.

8.4 Quasi-continuous assimilation

The above mentioned multi-incremental approach will also be useful for implementing the idea of quasi-continuous of Järvinen *et al.* (1996). The 4D-VAR assimilation can be run with an early subset of observations at a lower resolution and with the result saved and used for the subsequent final analysis(es) at higher resolution. The idea is that the starting point for the final analysis is already quite close to the converged state and the final analysis will need less computation.

8.5 Combinations of 3D-VAR and 4D-VAR

A RUC 3D-VAR will give us possibility to introduce more observed data closer to the actual starting time of a forecast.

Maybe it is better to use a slightly longer forecast based on a consistent 4D-VAR with less spin-up problems than a shorter forecast based on less consistent 3D-VAR initial data, with more severe spin-up problems. This is what is most often stressed by forecasters and staff who are responsible for production of forecast products.

There may be room for both sorts of products; a higher quality consistent 4D-VAR produced probably with a long cut-off time on one hand, and a short cut-off RUC 3D-VAR cycle spun off the 4D-VAR main cycle as often as needed, in order to e.g. include the latest observations for severe weather (mainly conventional ones, N.B.; the more advanced ones involving precipitation and clouds will still need 4D-VAR for the most reasonable benefit).

A combination of 3D-VAR and 4D-VAR was also suggested by Bouttier (2002) and would use 4D-VAR only for the last part of the time window to reduce the cost.

Another, even more defensive option, would be to use 4D-VAR only in the non-time critical re-analysis cycles (repeat cycles run with late incoming observations).

9 Summary

We have strongly argued in this paper for a continued development of the HIRLAM 4D-VAR. Our arguments have been based on scientific reasoning, on the successful implementation of 4D-VAR at several major NWP centers and on a vision for the development of operational numerical weather prediction in Europe.

Main scientific argument in favour of 4D-VAR are:

- The use of a time sequence of observations.
- The much reduced model spin-up.
- The use of time integrated data (precipitation, tendencies).
- The flow dependency of (implicit) structure functions.
- The use of non-linear observation operators (this is also an argument for 3D-VAR).

- The possibility to move “weather patterns” and the possibilities to have a sustained impact of moisture informations.

The operational implementation of 4D-VAR at the ECMWF and the subsequent introduction of a vast amount of satellite data into the operational NWP suite has helped to put ECMWF clearly in the lead of the global NWP competition. It has recently been that ECMWF is approximately 12 h better in forecast verification scores at all forecast time ranges than any other NWP center. Of course this must be taken as a strong argument in favour of 4D-VAR, but it also changes the future scenario for operational NWP in Europe. We have concluded that the HIRLAM community should concentrate on mesoscale modelling including 4D-VAR for nowcasting and very short range forecasting purposes. The successful implementation of a mesoscale 4D-VAR at the Japanese Meteorological Agency also gives strong support to this development strategy.

Main milestones for the development and operational utilization of HIRLAM 4D-VAR may be summarized as follows:

- January 2004: The HIRLAM 4D-VAR based on semi-Lagrangian time integration and including rudimentary moist physics in inner loops is ready for pre-operational testing.
- January 2005: Pre-operational tests of HIRLAM 4D-VAR for 20 km horizontal resolution have been finished and documented. HIRLAM 4D-VAR to be included as an option in the HIRLAM Reference system. Several new components, for example, a weak digital filter constraint and control of lateral boundary conditions have been added to the HIRLAM 4D-VAR.
- 2005: The first HIRLAM group starts to use HIRLAM 4D-VAR operationally.
- 2005: Start development and testing of HIRLAM 4D-VAR for a 3 km mesoscale version of HIRLAM (or the non-hydrostatic model decided for HIRLAM at that time).
- 2007: Operational application of a 3 km model including 4D-VAR.
- 2010: Operational application of a 1 km model including 4D-VAR.

Finally it should be added that the advantage of state of the art data assimilation system in house is not only for operational NWP, but is crucial for attracting and recruiting new research staff and to be able to participate and carry out other international research projects with external funding, in cooperation with other main meteorological services.

Participating in such projects also means that additional development resources is available to the development of 4D-VAR and particularly to its observation usage.

References

- Buizza, R. 1994. *Impact of simple vertical diffusion and of the optimisation time on optimal unstable structures*. ECMWF Tech. Memo. 192, 25pp. Available from the European Centre for Medium Range Weather Forecasting, Shinfield Park, Reading, Berks. RG2 9AX, UK.
- Giering, Ralf. 1999. *Tangent linear and Adjoint Model Compiler, Users manual 1.4*.
- Gustafsson, N. and Huang, X.-Y. 1996. Sensitivity experiments with the spectral HIRLAM and its adjoint. *Tellus*, **48A**, 501–517.
- Gustafsson, N. and McDonald, A. 1996. A comparison of the HIRLAM gridpoint and spectral semi-Lagrangian models. *Mon. Wea. Rev.*, **124**, 1008–2022.
- Gustafsson, N., Lönnberg, P. and Pailleux, J. 1997. Data assimilation for high-resolution limited-area models. *J. Meteorol. Soc. Japan*, **75**, 367–382.
- Gustafsson, N., Källén, E. and Thorsteinsson, S. 1998. Sensitivity of forecast errors to initial and lateral boundary conditions. *Tellus*, **50A**, 167–185.
- Gustafsson, N., Berre, L., Hörnquist, S., Huang, X.-Y., Lindskog, M., Navascuès, B., Mogensen, K.S. and Thorsteinsson, S. 2001. Three-dimensional variational data assimilation for a limited area model. Part I: General formulation and the background error constraint. *Tellus*, **53A**, 425–446.
- Hortal, M. 2002. The development and testing of a new two-time-level semi-Lagrangian scheme (SETTLS) in the ECMWF forecast model. *Quart. J. Roy. Meteor. Soc.*, **128**, 1671–1687.
- Huang, X.-Y., Gustafsson, N. and Källén, E. 1997. Using an adjoint model to improve an optimum interpolation based data assimilation system. *Tellus*, **49A**, 161–176.
- Huang, X.-Y., Yang, X., Gustafsson, N., Mogensen, K.S. and Lindskog, M. 2002. *Four-dimensional variational data assimilation for a limited area model*. Technical Report 57, 44pp. Available from HIRLAM-5, c/o Per Undén, SMHI, S-60176 Norrköping, Sweden.
- Janisková, M., Thépaut, J.-N. and Geleyn, J.-F. 1997. Simplified and regular physical parameterizations for incremental four-dimensional variational assimilation. *Mon. Wea. Rev.*, **127**, 26–45.
- Järvinen, H., Thépaut, J.-N. and Courtier, P. 1996. Quasi-continuous variational data assimilation. *Quart. J. Roy. Meteor. Soc.*, **122**, 515–534.
- Lewis, J. and Derber, J. 1985. The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus*, **37A**, 309–327.
- Park, S.K. and Županski, D. 2003. Four-dimensional variational data assimilation for mesoscale and storm-scale applications. *Meteorol. Atmos. Phys.*, **82**, 173–208.
- Temperton, C., Hortal, M. and Simmons, A. 2001. A two-time-level semi-Lagrangian global spectral model. *Quart. J. Roy. Meteor. Soc.*, **127**, 111–127.

- Yang, X. 1998. Adjoint physics development in the HIRLAM 4DVAR. *Pages 95–102 of: HIRLAM 4 Workshop on Variational Analysis in Limited Area Models, 23-25 February 1998, Meteo-France, Toulouse, France. Available from HIRLAM-5, c/o Per Undén, SMHI, S-60176 Norrköping, Sweden.*
- Yang, X. 2002. Physical adjoint in HIRLAM 4DVAR. *Pages 50–57 of: HIRLAM Workshop on Variational Data Assimilation and Remote Sensing, 21-23 January 2002, Finnish Meteorological Institute, Helsinki, Finland. Available from HIRLAM-5, c/o Per Undén, SMHI, S-60176 Norrköping, Sweden.*