Report of work on common preprocessing

Budapest, September 14 - 27, 2008
Alena Trojáková
RC LACE Data Manager

1 Introduction

Within the scope of "Project of development of an operational data assimilation system for LACE" G. Bölöni (2008) proposed a centralized observation preprocessing. In order to keep the freedom of using different ALADIN/ODB cycles in the centers the inputs for program BATOR should be prepared in the central way and common preprocessing was proposed to be based on existing system operationally used at Hungarian Meteorological Service (HMS). This short report describes the work performed during DM's two weeks stay in HMS in order to help with analysis and technical realization of improvement of current observation preprocessing system and design of RC LACE centralized one. In the next sections the proposal/plan of the data distribution is described and technical details of some steps/items related to this proposal.

2 Ways and means

Our primary goal is to provide observation for data assimilation suites of LACE members. To cover this need we have found generally two approaches and the second one is going to be adopted for centralized observation preprocessing system.

2.1 First approach

The first approach under consideration was to run centralized observation preprocessing with 6h frequency and roughly at the same cutoff time as ARPEGE (see Table 1) to be able to deliver data at the same time as the first ARPEGE LBC or a bit earlier (see Table 2).

Analysis time	Short cut off	Long cut off
00 UTC	2h 15	8h 10
06 UTC	3h 00	6h 50
12 UTC	1h 50	8h 10
18 UTC	3h 00	6h 50

Table 1: Operational setting of ARPEGE suite in Météo France.

Analysis time	Short cut off	Long cut off
00 UTC	2h 30	8h 10
06 UTC	3h 20	7h 10
12 UTC	2h 10	8h 10
18 UTC	3h 20	7h 10

Table 2: Start of operational suites at CHMI setup based on the first ARPEGE LBC file.

List of necessary modification with respect to current HMS preprocessing system follows:

- prolongation of time-window to +/- 3h for all observation types
- development of the tool to split and merge OBSOUL according to obstype and time slots to give each partner freedom to choose the time-window of given observation type

  (or one can run OULAN for each observation type separately)

Summary of the first approach:

- + relatively simple solution to do and maintain
- + standard/usual amount of observation as some may be delayed by transfer and/or processing

- - large amount of data to download at single moment
- - as such not suitable more frequent cycling (nor nowcasting application Diag/VarPack)
- - ARPEGE time constraints may not be suitable for Members considering use of ECMWF LBCs

2.2 Second approach

The second approach is to run common observation preprocessing hourly with 1hour time-window (+/- 30min) and with cutoff time of 2 hours (for estimation details see Section 3.1.1). There is an idea to have only single cut-off, if possible and for long cutoff download only complementary data for 6h time-window.

List of necessary modification with respect to current HMS preprocessing system follows:

- reduction of time-window to +/- 30min for all observation types
- development of the tool to split and merge OBSOUL according to obstype and time slots.
- investigation of error in observation/amount of observation as some may be delayed by transfer and/or processing)

Summary of the second approach:

- + even more simple solution to do and maintain (if single cutoff is considered)
- + the amount of data should be suitable for both short and long cut-off
- + smaller amount of data to download at the analysis time (some data can be downloaded in advance)
- + as such suitable more frequent cycling
- + possibly extensible for nowcasting application Diag/VarPack, "nowcasting" cut-off time should be defined (roughly 5-15 minutes) and. Maintenance of 3 cut-off seems to be too complicated, so this "nowcasting" cutoff should be considered with lower priority only in case of single cutoff for basic DA suite.

General technical/operational requirements

- hardware configuration (with allowed remote access via FTP)
- naming convention for the data (to avoid mess as some ASCII data will be replace by BUFR ones progressively)
- development of monitoring/alert system for local maintenance stuff and users as well (at least mail in case of problem) !

3 What has been done

This section is summary of work done during the stay. Not all tasks were finished, but the aim was to identify weak points related to proposed observation preprocessing implementation and tackle/touch some further improvements such as change of data formats.

3.1 Feasibility study of hourly observation processing

Based on proposal the feasibility study of hourly run of HMS observation processing was done. There are following modifications essential for setting time-window of 1h (plus/minus 30min around given analysis time):

- decrease of time-window for AMDAR (amdar_cf=0)
- decrease of time-window for ATOVS, HIRS and MHS (PROC_LinkTovs=$d_SCR/link_sat_30.sh) New procedure to link all available data within predefined time-window was developed.
- decrease of time-window for AMV (PROC_GEOWIND=$d_SCR/merge_bufers_save4d_1h) small modification of current procedure was done.
- decrease of time-window for WIDNPROFILER (WP_LH=1 and WP_RH=1) This item is incomplete as time-window was reduced to +/- 1h, not to +/- 30 min, but there is a plan to change preprocessing of WIDNPROFILER soon, thus special development was skipped.

For technical details see atroja@3700a scr/Assim_e13 script and corresponding include_e13 and Monitor_e13 files, which were based on operational versions from mid September.

3.1.1 Overview of observation availability

Following observation types are currently treated in HMS: SYNOP, AMDAR, TEMP, WINDPROFILER, SATOB (AMV) and SATEM (AMSUA, AMSUB, HIRS, MHS from NOAA and SEVIRI from MSG). To investigate an option to have single cut-off for all data assimilation suite, the check of observation availability was done.

- SYNOP are available hourly with delay of 14 to 978 minutes
- AMDAR available irregularly with delay of 17 to 1245 minutes
- SATOB/AMV available from geostationary satellites (currently MSG 2 - ? available every 15min) with several hours delay - should be investigate in obs department
- TEMP available at 00,06,12,18 UTC with delay of 97 to 767 minutes
- WINDPROFILER availability was not investigated as the treatment of data is going to be changed
- SATEM

  - from geostationary satellites (currently SEVIRI from MSG 2 - available every 15min with delay of 20 minutes)

  - from polar satellites (currently AMSUA, AMSUB, HIRS, MHS from NOAA, available irregularly with delay of 20 to 92 minutes

the estimation of data delay was based on difference of data time and time stamp of file which contained the data. Conventional data (SYNOP, TEMP, AMDAR) are decoded to local netCDF database "on a flight" from whole the globe, thus estimation of data availability on the domain of interest was impossible with this approach. From observation department we got estimates of minimum delay for European data 45 minutes for SYNOP and 1h30 or TEMP. Thus we tried to derive the best delay/cutoff by following experiments. We ran preprocessing script with 1, 2 and 7h delay, where 7 hour should delay correspond to long cut off, 2h to short cut off and 1h was the first trial to decrease delay to its minimum. After observation extraction the data were read by BATOR and only data inside domain of interest were selected. Summary of data (corresponding to the number of available parameters of given observation type,e.g. T2m. Rh2m, MSLP, u10, v10 measurement for SYNOP) is in Table 4-6 for each of three experiments. And Table 3. shows the same statistics for operational ALADIN/HU assimilation suite.

- E12 - 1h delay
- E13 - 2h delay
- E14 - 7h delay

	obstype	00 short	00 long	06 short	06 long	12 short	12 long	18 short	18 long
SYNOP	1	3674	3674	4432	4432	4464	4464	4376	4376
AMDAR	2	1101	1101	9864	9864	12153	12174	9711	9711
SATOB	3	2168	2168	2670	2670	4212	4212	2240	2240
TEMP	5	9019	9287	2038	2117	9080	9080	1143	1143
PROFILER	6	826	854	864	864	874	980	940	998
SATEM	7	121825	143386	158628	158628	151910	204447	155290	155290
file size (.gz) [MB]	1,2,3,4,5,6,7	15.0	17.9	18.4	18.4	15.1	20.5	22.2	22.3

Table 3: Number of data within 6h time-window for short and long cutoffs for analysis 20080924 of operation ALADIN/HU assimilation suite.

	21	22	23	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
1	3956	1536	1523	3672	1524	1542	3946	1642	1657	4432	1658	1643	4229	1629	1652	4435	1630	1562	4132	1559	1575	4376	1656	1657	3972
2	3216	1437	939	297	243	771	978	1626	4725	4902	3984	5133	5685	5367	5100	6120	4233	5328	3810	4866	5331	5097	3444	3549	2499
3	2322	2192	2188	2168	1984	2062	2184	2122	2344	2670	3312	3548	3676	3822	3978	4212	4240	4442	4202	3664	2958	2240	2236	2148	2096
5	0	0	0	4837	0	0	0	0	0	1414	0	0	0	0	0	4319	0	0	0	0	0	650	0	0	0
6	726	736	712	748	732	804	802	526	726	754	678	708	770	708	736	774	784	782	744	750	760	794	840	862	862
7	0	0	0	44029	0	0	0	0	0	0	0	0	161076	27650	164357	0	182992	105074	0	209298	0	0	0	0	0

Table 4: Number of data within 1h time-window for 1h cutoff/delay during 2008092321-2008092321.

	21	22	23	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
1	3962	1541	1523	3674	1524	1542	3946	1642	1657	4432	1658	1649	4229	1629	1652	4464	1630	1562	4147	1559	1575	4376	1656	1657	3972
2	3216	1491	963	297	444	822	1026	1674	4776	5013	4152	5241	5685	5382	5487	6306	5295	5328	3813	4920	5451	5223	4242	3642	2502
3	2322	2192	2188	2168	1984	2062	2184	2122	2344	2670	3312	3548	3676	3822	3978	4212	4240	4442	4202	3664	2958	2240	2236	2148	2096
5	0	0	0	8897	0	0	0	0	0	2038	0	0	0	0	0	8992	0	0	0	0	0	1143	0	0	0
6	726	736	712	748	732	804	802	782	726	754	678	736	770	708	736	814	910	782	744	750	760	794	840	862	862
7	0	2045	0	44029	242315	0	58156	0	245309	14929	171	0	161076	27650	164357	90	182992	105074	0	209298	0	16476	0	119918	0
[MB]	7.0	6.7	2.9	5.0	5.8	3.6	12.0	0.1	11.2	11.7	2.3	3.3	7.1	7.5	7.5	5.8	10.5	7.9	9.4	10.8	3.7	13.1	10.0	8.0	4.4

Table 5: Number of data within 1h time-window for 2h cutoff/delay during 2008092321-2008092321.

	21	22	23	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
1	3967	1541	1523	3674	1524	1542	3946	1642	1657	4432	1658	1649	4229	1629	1652	4464	1630	1562	4147	1559	1575	4376	1656	1657	3972
2	3219	1491	972	297	444	828	1035	1689	4776	5013	4161	5241	5694	5388	5487	6327	5298	5328	3813	4920	5451	5226	4242	3642	2505
3	2322	2192	2188	2168	1984	2062	2184	2122	2344	2670	3312	3548	3676	3822	3978	4212	4240	4442	4202	3664	2958	2240	2236	2148	2096
5	0	0	0	9287	0	0	0	0	0	2117	0	0	0	0	0	9080	0	0	0	0	0	1143	0	0	0
6	836	840	782	748	816	834	886	848	726	818	678	736	828	828	858	920	942	822	854	868	878	914	928	964	954
7	0	2045	0	44029	242315	0	58156	0	245309	14929	0	0	161076	27650	164357	90	182992	105074	0	209298	0	16476	0	119918	0

Table 6: Number of data within 1h time-window for 7h cutoff/delay during 2008092321-2008092321.

	obstype	00 short	00 long	06 short	06 long	12 short	12 long	18 short	18 long
SYNOP	1			4460	4460	4509	4509	4110	4110
AMDAR	2			4572	4572	5298	5298	5226	5226
SATOB	3			2820	2820	3864	3864	2016	2016
TEMP	5			1238	1238	8474	8552	861	861
PROFILER	6								
SATEM	7			304145	304145	388420	388420	277124	277124

Table 3: Number of data within 6h time-window for short and long cutoffs for analysis 20081001 of operation ALADIN/HU assimilation suite.

	21	22	23	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
1	3952	1550	1532	4078	1531	1549	4053	1654	1675	4454	1674	1676	4251	1682	1688	4509	1671	1600	4070	1580	1582	4105	1647	1664	3853
2	2967	1140	267	3	336	741	924	2238	4671	4488	3435	4467	4491	4704	5217	5127	4632	5616	5082	5094	4704	5031	5280	3570	2469
3	2256	2246	2200	2252	2270	2314	2286	2286	2504	2820	3410	3660	3666	3654	3668	3864	3924	3852	3886	3204	2498	2016	1974	2048	2046
5	0	0	0	5069	0	0	0	0	0	832	0	0	0	0	0	4614	0	0	0	0	0	836	0	0	0
6	744	728	720	694	698	650	640	682	704	640	606	648	764	632	622	644	562	624	590	650	702	744	790	818	786
7	0	0	0	0	0	224631	0	159148	0	0	5466	35039	0	0	0	200941	0	0	169864	0	23217	0	0	0	152174

Table 4: Number of data within 1h time-window for 1h cutoff/delay during 2008093021-2008100121.

	21	22	23	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
1	3952	1550	1532	4124	1531	1549	4065	1654	1675	4454	1674	1676	4251	1682	1688	4509	1671	1600	4070	1580	1582	4105	1647	1664	3853
2	2961	1161	267	6	381	816	981	2253	4758	4566	3543	4554	4578	4833	5331	5280	4830	5784	5262	5172	4875	5223	5415	3636	2520
3	2256	2246	2200	2252	2270	2314	2286	2286	2504	2820	3410	3660	3666	3654	3668	3864	3924	3852	3886	3204	2498	2016	1974	2048	2046
5	0	0	0	6768	0	0	0	0	0	1238	0	0	0	0	0	6021	0	0	0	0	0	861	0	0	0
6	744	728	720	694	698	650	640	682	704	640	606	648	764	632	622	644	562	624	590	650	702	744	790	818	786
7	109061	0	0	96174	0	224631	79479	159148	78652	23551	5466	35039	160074	5340	172171	200941	75750	0	296084	182886	25345	0	108241	0	152174

Table 6: Number of data within 1h time-window for 1h30 cutoff/delay during 2008093021-2008100121.

	21	22	23	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
1	3952	1550	1532	4124	1531	1549	4065	1654	1675	4454	1674	1676	4251	1682	1688	4509	1671	1600	4070	1580	1582	4110	1647	1664	3853
2	2967	1161	267	9	381	816	1005	2253	4758	4572	3543	4554	4584	4977	5355	5289	4839	5784	5262	5178	4896	5226	5457	3666	2520
3	2256	2246	2200	2252	2270	2314	2286	2286	2504	2820	3410	3660	3666	3654	3668	3864	3924	3852	3886	3204	2498	2016	1974	2048	2046
5	0	0	0	8611	0	0	0	0	0	1238	0	0	0	0	0	8145	0	0	0	0	0	861	0	0	0
6	744	728	720	694	698	650	640	682	704	640	606	648	764	632	622	644	562	624	590	650	702	744	790	818	786
7	109061	0	0	96174	0	224631	79479	159148	78652	23551	5466	35039	160074	5340	172171	200941	75750	0	296084	182886	25345	0	108241	0	152174

Table 7: Number of data within 1h time-window for 2h cutoff/delay during 2008093021-2008100121.

Analysis of results showed that for SATOB and SATEM there are no differences between E13 and E14 and for SYNOP, AMDAR and TEMP there are only very small differences. The questionable are PROFILER data, where prolongation of delay (or even long cutoff) can be considered. Because of planed internal change of PROFILER data treatment, we propose to repeat this statistics when netCDF decoding will be ready. Another unclear issue is related to SATEM observation, where operational suite (short and long cutoff) uses less data (for 00) than available by hourly preprocessing, this difference should be investigated. On the top of that we have realized that for short cutoff (usually 2:30 for 00 and 12UTC or 3:10 for 06 and 18UTC) we will have available less data than required due to 2h delay, e.g. for 00 short cutoff starting at 2:30 we have available data from 21,22,23,00 only, problem is the delay of data measured between 0:30-2:30, which will be delivered only at after 3 and 4 UTC (with 2h delay), thus we don't have enough satellite data for short cutoff. One solution could be to try to shorten the delay, e.g. to 1h30 or we should reconsider the usage of the first approach. This point was found on the last day of the stay, so the conclusion that the second approach will be realized is not final and needs further discussion.

3.2 Fix for date of ATOVS data

During validation of obsoul tools L. Szabó pointed out that the dates of all satellites data have wrong date (for analysis date 2008091712 UTC data had assigned date of previous day 20080916). Current data flow for satellite preprocessing is displayed on following Fig 1.

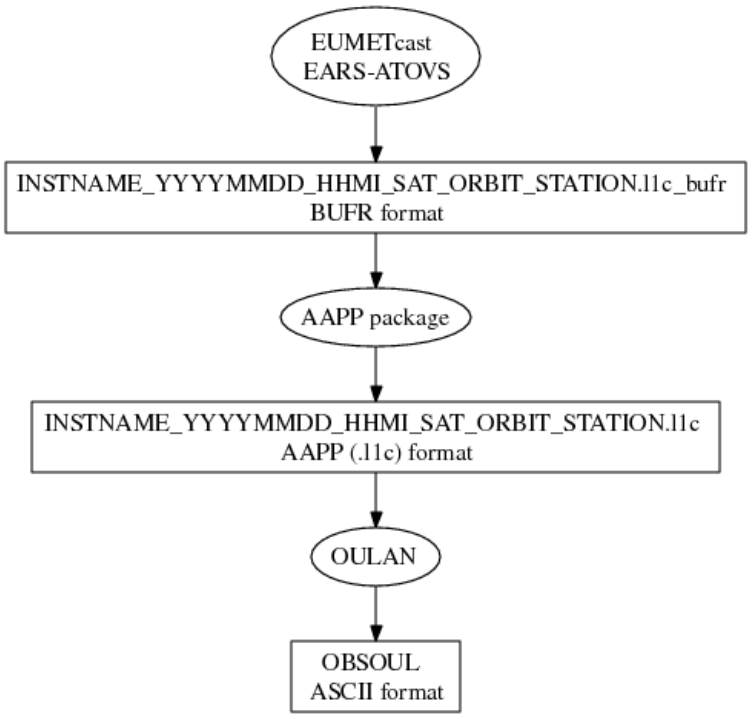

Fig. 1: ATOVS preprocessing used in HMS.

The object of EARS-ATOVS is to provide the sounder data covering data-sparse areas within 30 minutes of the instrument observations. Sounder data is produced by a set of the instruments known as ATOVS. Product (Level 1C BUFR) naming convention of files is as follows: INSTNAME is the instrument name: hirs, amsua, amsub, YYYYMMDD_HHMI is the observation time of the first instrument scan line in the product, SAT is the satellite name: noaa15, noaa16, noaa17, noaa18, ORBIT is the orbit number since launch of the satellite and STATION is the ground station short-name (Athens, Kangerlussuaq, Maspalomas, Lannion, etc.). For more details see http://www.eumetsat.int.

The bug was hidden in the function DAYOFYEAR (aapp_get_1c.F) used in AAPP software package to convert BUFR to AAPP format. The number of days for leap year was wrong (starting from September).

3.3 NBSLOT and ficdate

There is an option to use variable NBSLOT (BATOR_NBSLOT)) and 'ficdate' file to reduce observation into predefined number of time-slots (time intervals) according to dates in the 'ficdate' file, e.g. for 20080917 00UTC analysis and NBSLOT=7, 'ficdate' file looks like:
20080916210000
20080916213000

20080916223000
20080916233000
20080917003000
20080917013000
20080917023000
20080917030000

or for 20080917 00UTC analysis and NBSLOT=3, file looks like:
20080916223000
20080916233000
20080917003000
20080917013000

Unfortunately this option does not work completely for BATOR cy30t1 of HMS model installation. More precisely, for NBSLOT=7 treatment of all data except of satellites looks fine, but satellite data are put to the first time-slot independently of its time. In case of NBSLOT=3 none data are set correctly. This problem was just identified, due to lack of time was not solved and should be investigated in more details later.

3.4 Change of data format

Next issue which was tackled concerns change of data format of some observations (e.g. AMV and ATOVS). From historical reasons all the data are treated by program OULAN and its results is ASCII OBSOUL format in HMS preprocessing. The idea is to change format of some data into BUFR to save the time while reading data and the volume of transfered data and simplify the preprocessing, e.g. Fig 2 shows current ATOVS processing on the left and simplification allowed by direct use of BUFR data.

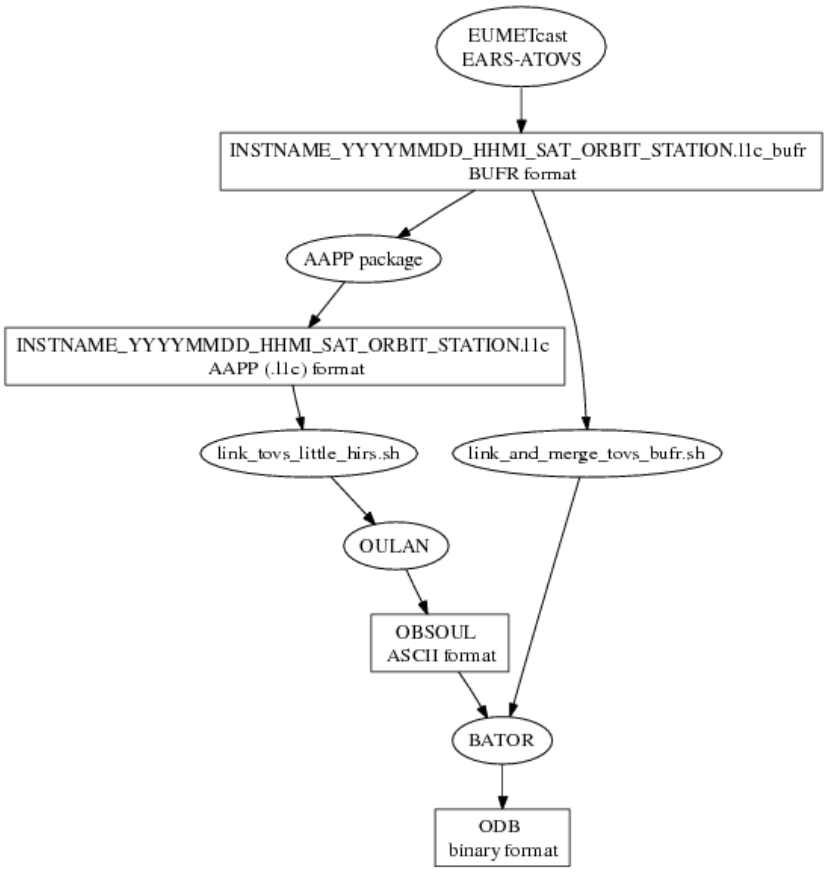

Fig. 2: ATOVS preprocessing used in HMS.

3.4.1 AMV

We have tried to read AMV data in BUFR format with BATOR (CY30T1) with no success, but it seems that BATOR (CY33) is able to read directly data from AMV_1_YYYYMMDD45 file (disseminated by EUMETcast), but there are missing some 10% of data with respect to the treatment with current processing chain - splitting AMV file into small files, decoding to ASCII files which are read by OULAN program to produce OBSOUL file. For the difference see Figures below.

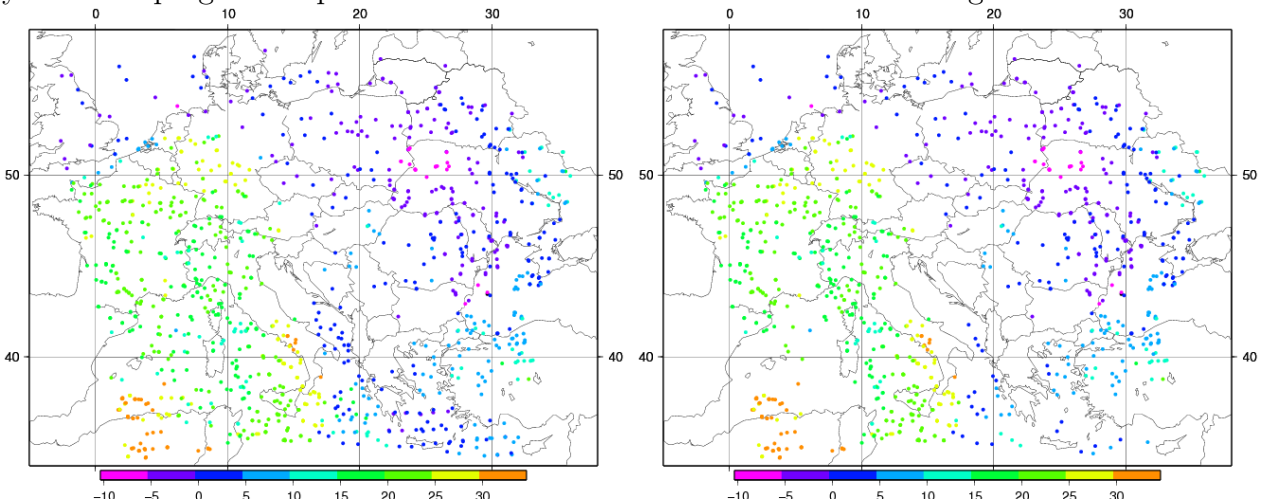

Fig. 3: u-component of AMV generated from OBSOUL (left) and directly from BUFR (right)

For the time being the cause of difference is unknown and should be investigated further.
To read AMV in BUFR format only following changes were needed

- input file name has to be named BUFR.geowind
- special refdata file for BATOR is required with following line
  geowind BUFR geowind n_date n_time

3.4.2 ATOVS

The first tests were done with reading AMSU-A data in BUFR format. Again only two modification were needed:

- input file name has to be named BUFR.tovamsua
- special refdata file for BATOR is required with following line
  tovamsua BUFR amsua n_date n_time

Contrary to AMV the ATOVS data are disseminated in several files, thus further tests with merged data are planned and merging tool should be developed.

4 Summary

During this two weeks stay we have investigated possible approaches to common observation preprocessing implementation, some solutions were proposed with summary of pros and cons. Part of the stay was dedicated to technical improvement of current observation preprocessing system and here follows list of task which are to be done:

- cutoff estimation should be studied further to give final conclusion about observation preprocessing approach
- change of local PROFILER treatment (netCDF)

- development of split/merge tool for OBSOUL file (ongoing work on L. Szabó)
  - should be well documented as should be provided to LACE Members
  - should allow splitting according to time and obstype
- hardware setup of dissemination server

- investigation NBSLOT and 'ficdate' issue
- investigation of progressive change of satellite data format to BUFR
  - investigation of differences in AMV data reading
  - feasibility study for ATOVS data and development of split/merge tool, eventually trial with Metop data can be considered

The first four items should be realized in HMS (by G. Bölöni and L. Szabó) and the last two by A. Trojáková at CHMI.

References

G. Bölöni, 2008: LACE workplan on data assimilation 2008-2010 *www.rclace.eu*