Improving the quality of European weather radar composites with the BALTRAD toolbox

Anders Henja

Anders Henja[1] and Daniel Michelson[2]

[1]*Swedish Meteorological and Hydrological Institute, Malmö, Sweden, anders@baltrad.eu*

[2]*Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, dbm@baltrad.eu*

(Dated: 29 May 2012)

1. Introduction

EUMETNET OPERA's operational data centre (ODC, or Odyssey) started generating European weather radar composite products operationally during 2011. The quality of these products has become a major concern within the user communities, especially NWP groups that have fed back their recommendations for improvement. During the second half of 2011, EUMETNET managing bodies and OPERA approved a proposal whereby the BALTRAD toolbox would be trialed for use by Odyssey as a means of introducing quality-control procedures designed to improve the quality of European composites.

An Odyssey development environment (ODE) has been established by SMHI for these purposes. Supporting this ODE is a live feed of European polar (raw) data that have been "bounced" from the UK Met Office in Exeter and Météo France in Toulouse. The concept we have followed is that processing all European data in near-real time gives us the best conditions for trialing improvements to European weather radar data quality in a realistic production setting.

A suggested production chain has been tailored for Odyssey with the BALTRAD toolbox. Individual polar scans are sorted into polar volumes, all volumes are quality controlled, and then composites are generated. Quality-indicator fields are generated by each tool in the quality-controlled volumes. Quality-controlled volumes contain both uncorrected and corrected reflectivities, and so do the composite products. Quality indicators are also composited. European SYNOP observations have been used to evaluate the results using two different methods, and monthly accumulations have also been generated for visual scrutiny.

This paper reports on this trial of the BALTRAD toolbox. The suggested processing chain is described, and the evaluation methods and results are presented. The trial also included an evaluation of system performance and scalability, and these outcomes are also given. We summarize the OPERA working document containing the same information in a little more detail (Michelson and Henja, 2012), with the exception of some of the statistical evaluation methods that have been elaborated since the working document was written.

2. Production chain

In the following description of the production chain, a basic understanding of the ODIM_H5 information model is assumed (Michelson et al. 2011). The radar observables and quality datasets are all referred to using ODIM terminology.

2.1 Polar scans and volumes

Input polar data arrive as individual scans and as polar volumes, both from BALTRAD partners and from Odyssey. In the process of generating the volume from a set of scans, the data are processed using a set of tools from the BALTRAD toolbox designed to improve data quality and add traceability to the production chain. The following is performed:

1. If the TH reflectivity dataset exists previously, it is preserved. Otherwise, the DBZH reflectivity dataset is copied to TH. In subsequent processing steps, TH remains unchanged and the quality of DBZH is improved. (In principle, this violates the definition of TH, but from a processing-chain perspective it is acceptable as it facilitates before vs. after comparisons.)
2. Anomaly detection and removal. Software from FMI (Peura, 2002). Data from scans having elevation angles below 2° are processed using a full set of anomaly detectors. After thresholding (default -24 dBZ), the detectors for biometeors, speckle, emitters, and "ships" (large point targets possibly involving side lobes), are applied. Data at or above 2° are thresholded and processed using the "speck" detector. The BALTRAD toolbox contains an XML configuration file where site-specific settings can be are stored and applied in real time. Data from radars without an entry in this file are processed using default argument values. A quality indicator is added to each scan dataset containing the maximum probability of anomaly for all detectors.
3. Beam-blockage correction. Software from SMHI. This correction is performed in polar space using a 1 km digital elevation model (GTOPO30) and a geometric propagation model that oversamples the radar's geometry. GTOPO30 is the only reliable public domain global DEM. For each scan of data, the percentage of beam blockage is determined, and it is added to this scan as a quality indicator. Each time a scan of data with a new geometry arrives, it is processed and the resulting beam-blockage result is written to a cache file. Such files are read straight into quality indicator fields for scans with previously known geometries. Data in sectors with partial blockage up to 70% are corrected. Data in sectors

with blockage exceeding 70% are assigned the *nodata* value. GTOPO30 tiles are used in unmodified form in the BALTRAD toolbox.

4. "Detection range". Software originally from FMI (Koistinen and Hohti, 2010). This trial is performed during the cold season, and we know that winter precipitation is often shallow. Overshooting is therefore a concern, so knowledge of the radar's real maximum range, according to the given meteorological situation, is extremely valuable because this information helps reduce uncertainty. This quality-control procedure operates in polar space and generates a quality indicator containing the probability of overshooting (or apparent maximum range of detectability) for the input volume. Unlike the beam-blockage correction, data in areas thought to be overshot are not assigned the *nodata* value. This quality tool is purely descriptive and therefore never used to decide whether to correct or reject data.

For data that arrive as complete polar volumes, this same quality-control procedure is applied as a pre-processing step prior to composite generation. Doing this ensures that all input data to the composite generator are processed on equal terms.

Note that this quality-control chain is applied to data in memory; data are read at the beginning of step 1 and they are written to disk after step 4. This is done deliberately to minimize unnecessary file I/O, thereby improving overall system performance.

2.2 Composite generation

The BALTRAD toolbox contains a general framework for generating composites of various kinds in various ways. The concept that the composite generator follows is that all input data are in original polar coordinates, and that these data are navigated directly to the output composite grid with no intermediate steps or storage. Navigation is performed using PROJ.4, and no shortcuts or compromises are made regarding navigational accuracy for the sake of improving execution speed. This concept is sometimes referred to as "slow compositing", because we don't use any lookup tables containing static configuration information that can significantly improve speed. The advantage of not relying on such lookup tables, however, is that the characteristics of the input data can (and will!) change without advance notice, and our algorithm will not experience any negative consequences. Also, such lookup tables require resources to maintain, so we want to avoid this burden. We have devoted a lot of time and effort to optimize performance without compromising quality. Another advantage of our approach is that all the pre-processing steps described above can be chained in memory together with the composite generator, in practice meaning that input data are read just once at the beginning and written once as a composite product.

We have added a maximum-reflectivity criterion to our composite generator, in order to create similar composites to those generated in Toulouse and Exeter. The maximum reflectivity is not the best choice of algorithm, but it is the one decided for use by Odyssey, so we feel compelled to use it despite its problems.

None of the quality indicators generated in polar space (described above) are suitable for use as selection criteria when generating composites. We have experimented using probability of overshooting as the only criterion, and this led to unacceptable artifacts in the composites that cannot justify its use. The beam-blockage results are used implicitly as a criterion wherever beam blockage exceeds 70%. For blockage less than 70%, if the correction raises the reflectivity value so that it becomes the maximum value compared to other data from the same site, then this value is selected. In such cases, this algorithm is indirectly quality-based. Otherwise, the value closest to the Earth's surface value is selected, which is the best choice for minimizing border artifacts in composites.

Our composite generator also allows for generalized multiple selection, which means that several output datasets can be processed simultaneously. The output composite product is written to ODIM_H5, and it contains the structure given in Table 1. The addition of a second quantity, e.g. compositing TH and DBZH at the same time, comes at a performance penalty of around 12% in processing time.

Table 1. Organization of data in composites. It may seem that the quality indicators are duplicated unnecessarily, but they are not because the compositing algorithm will select different input data, and therefore different corresponding quality information, depending on whether the data are quality-controlled or not.

Path	Content
/dataset1/data1	Uncorrected reflectivity, TH
/dataset1/data1/quality1	Maximum probability of anomaly
/dataset1/data1/quality2	Degree of beam blockage
/dataset1/data1/quality3	Probability of overshooting
/dataset1/data1/quality4	Surface distance from the radar. This information is derived dynamically during the composite generation.
/dataset2/data1	Corrected reflectivity, DBZH, based on the corrections applied in the processing chain
/dataset2/data1/quality1	Maximum probability of anomaly
/dataset2/data1/quality2	Degree of beam blockage
/dataset2/data1/quality3	Probability of overshooting
/dataset2/data1/quality4	Surface distance from the radar. This information is derived dynamically during the composite generation.

Note that this composite product contains radar reflectivity factor aloft, ie. the values are not surface estimates. Note also that the surface distance quality indicators will be different for each reflectivity dataset, because of the influence of "nodata" values from the beam-blockage correction (>70% blockage).

3. Evaluation methods

Three types of precipitation information were generated based on the composite products, for evaluation purposes:

1. Radar reflectivity factor, one composite every three hours to be compared with surface observations.
2. Daily precipitation accumulations, 6-6 UTC, to be compared with gauge observations.
3. Monthly accumulations.

Three different evaluation methods were employed, all designed to evaluate accuracy of the composite products in different ways, both qualitatively and quantitatively. Two of these methods involve the use of surface point measurements. These have been co-located with radar data using the radar pixel that the point observation lies inside. No "best fit" method involving neighboring pixels has been used.

3.1 Contingency-table statistics

These statistics are conventional measures based on hits and misses. Input data to this evaluation were the reflectivities from radar and observations pertaining to precipitation from the SYNOP network. It should be noted that three-hourly SYNOP bulletins do not contain accumulated precipitation observations, but rather an indication of whether precipitation of various kinds occurred at observation time. This information is found in the Present Weather (*ww*) observation. These observations are used by SMHI's mesoscale analysis system (Häggmark et al., 2000), where they are quality controlled such that we can reject reports flagged as being suspicious. The radar data were not thresholded. Derived statistical measures were: Probability of Detection (POD), False Alarm Rate (FAR), Percent Correct (PC), Probability of False Detection (POFD), and Hanssen-Kuipers Skill (HKS) all described by Wilson (2001).

Each contingency table was subjected to Barnard's exact test (Barnard, 1945) to test its validity, ie. the null hypothesis. Only times where both tables passed this test were analyzed further. Mean values and standard deviations for the month of February 2012 where derived from three-hour POD, FAR, PC, POFD, and HKS differences. Based on these, 95% confidence intervals were determined along with their margins of error, making it possible to test the statistical significance of each individual difference.

3.2 Radar-gauge relations

It is not possible to compare quantitative precipitation estimates for integration periods less than 12 hours at the continental scale due to of lack of surface observations. We have therefore used 24-hourly SYNOP reports, at 06 UTC, from Europe in order to improve our chances at deriving stable, reliable statistical relations.

Methods used for deriving gauge-radar relations in the Baltic Sea Experiment (Michelson and Koistinen 2000, Koistinen and Michelson 2002) have been adapted for European use. These relations are derived every 24 hours based on radar and gauges that report at least 0.1 mm. The relation between surface distance from the radar and the gauge/radar ratio on a dB scale is derived. Each valid G/R point in the accumulation has a corresponding average distance from the radar in the composites' quality field (quality4). This average distance is used to account for drop-out of a radar in the time series. Because we have saved both uncorrected (TH) and corrected (DBZH) reflectivities in the same product files, we have made it easy to compare statistical relations for each.

The justification for using this evaluation technique is that the gauge-radar relations with distance will (hopefully) have different characteristics with uncorrected and corrected radar data. The uncorrected data would display a larger system bias (bias at 0 km distance), and the corrected data would display a smaller variance (scatter) about the derived statistical relation. Also, owing to the beam-blockage correction, the bias as a function of distance would be lower with quality-controlled data. These are the hypotheses to be tested.

3.3 Visual comparison of monthly accumulations

This is the simplest evaluation technique, but it is also useful because it is intuitive: "seeing is believing". We have simply accumulated all the precipitation for a complete month and generated images of uncorrected and uncorrected results. Both random and systematic errors accumulate over time, so the effects of the applied quality control methods relative to the uncorrected data will illustrate how contaminated the original data are, and how effective the production chain is. This technique is also designed to reveal over-correction.

3. Results

Statistics were derived every three hours during the month of February 2012. Results that are statistically significant using a 95% confidence interval have been plotted and will be shown at the conference. Quality control lowers POD. FAR is also lowered, more than POD. PC is raised, and POFD is lowered. HKS remains about the same. About 8% of the differences between uncorrected and corrected skill scores where not statistically significant. Statistics are rather noisy, but the improvements are generally statistically significant.

Statistics were also derived for all point comparisons together for the complete month. The results are presented in Table 2. These serve to summarize the information in the detailed results from every three hours, and hopefully be more reliable

because the amount of data is much larger. Again, the quality controls lower POD, but not as much as FAR. PC is raised and POFD lowered. HKS is lowered marginally. These results are all statistically significant.

Table 2. Contingency table statistics for all data during the month of February 2012, based on over 230 000 point comparisons.

Skill measure	**Uncorrected**	**Corrected**
Probability of detection	0.356	0.316
False alarm rate	0.478	0.413
Percent correct	77.73	79.19
Probability of false detection	0.097	0.066
Hanssen-Kuipers Skill	0.259	0.250

Statistical relations between daily gauge and radar-based accumulations were derived for the month of February 2012. Monthly statistics are presented in Table 3. Individual daily statistical relations are quite noisy despite the 24-hour integration periods. Daily relations between the gauge-radar relation and surface distance from the radar are not physically meaningful. This can be seen through the so-called "system bias", which is the radar data's bias relative to the gauges at 0 km, in this case a strong over-estimation, which the quality-control procedure reduces a little. The correlation coefficient remains unchanged, but both mean bias (irrespective of range) and standard deviation are lowered. The daily mean biases range between around ±2 dB and are therefore not physically meaningful either.

The improvements from the quality control are small but statistically significant.

Table 3. Statistical measures based on over 12 000 gauge-radar point pairs during the month of February 2012. The "system bias" is the bias at 0 km distance from the radar, taken from the derived gauge-radar ratio (dB) as a function of distance.

Statistic	**Uncorrected**	**Corrected**
Correlation coefficient	0.51	0.51
Mean bias (dB)	0.277	0.124
Standard deviation (dB)	5.5	5.4
"System bias" (dB)	-5.11	-4.78

A detail from uncorrected and quality-controlled monthly accumulations is shown in Fig. 1. Note that border effects between radars are artificial because even radars with very poor data availability have been included. Absolute quantities are irrelevant in this context, and have therefore been omitted deliberately.

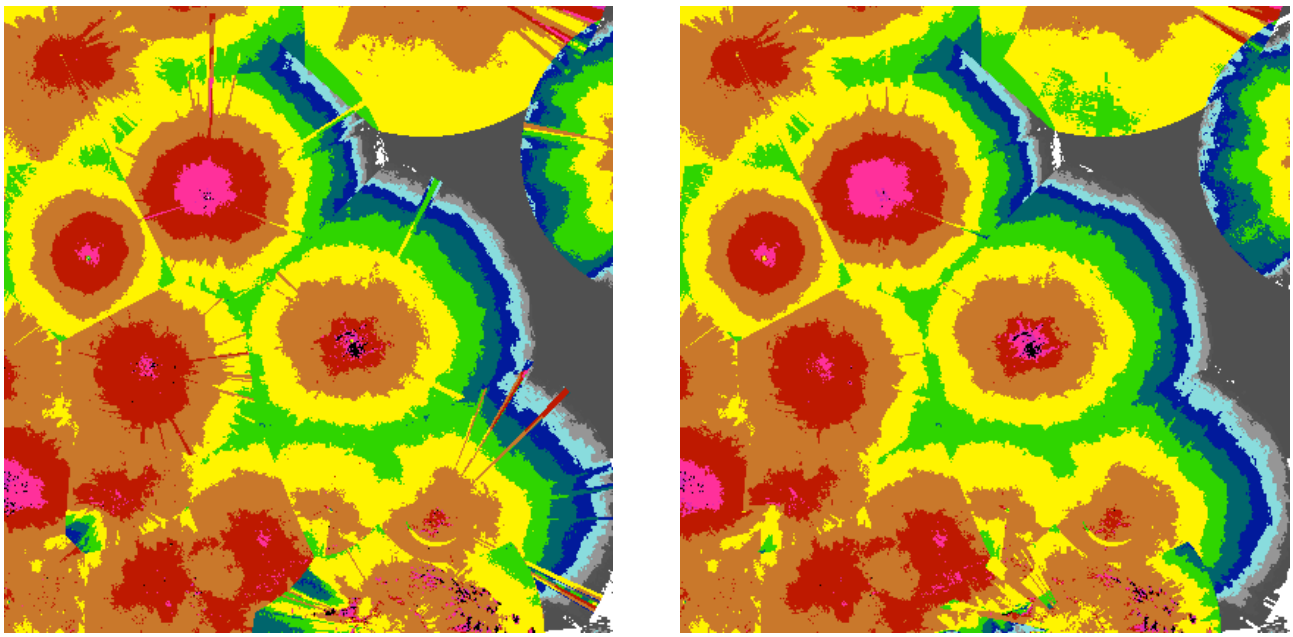

Figure 1. Detail from the European composite domain illustrating the positive impact of using the BALTRAD toolbox on long-term accumulated precipitation. Accumulating data over a long period helps to reveal and emphasize systematic errors, and this example shows how signals from external emitters (left) are significantly reduced (right) using algorithms in the toolbox. Generally, higher-quality data should be more concentric around the radar site, and this is what has been achieved. Border effects between different radars should be ignored in these images, as this illustration is based on accumulated data that has not taken into account different data availabilities from different radars.

Some, but not all all, errors are corrected in quality-controlled data. The “bull’s-eye” effect is clearly visible in data from most radars, and this reveals the radar’s increasing underestimation of precipitation with range which is not yet corrected for by Odyssey. Some very noisy radar data have become much less noisy. External emitter signals have been reduced in many places. The beam-blockage correction has reduced the “notch” effect in some places, although there appear to be many places with blockage that is not based on topography. Some other interesting artifacts are also removed. Generally, a well-sited radar in an area with a homogeneous precipitation climate will have a symmetric “bull’s eye”, so higher-quality data will be more concentric. The quality controls applied in our processing chain have achieved this in many places.

4. Toolbox performance

We are naturally keen on improving the quality of Odyssey’s products in a way that can be realistically achieved in real time. It is one thing to be able to achieve higher quality, but this higher quality must be implementable in real time for it to be meaningful. Using the toolbox data processing functionality, we have tailored a combination of tools that can be executed on the command-line and/or scheduled using an entry in the crontab scheduler table. Every 15 minutes, this processing chain is run, and all data are processed at once from a “cold start”. The combination of tools has been parallelized in order to optimize the processing time on multi-core CPU computers, which is standard in computer hardware today. This parallelization is achieved through rather basic asynchronous multi-processing functionality that is available on any Linux/UNIX platform. This can only work in practice if the code base supports concurrent file I/O, which ours does.

For most operations, ie. combining polar scans into volumes and quality controlling the volumes, the parallelization uses all available threads in all CPU cores. For compositing, the concurrency is limited to six processes, one for each so-called tile that into which we have subdivided the composite domain. What we have done, in principle, is to create a European composite based on concatenated regional composites. In COST 73 (Weather Radar Networking), the need for regional compositing centers in Europe was envisaged as a way of sharing the burden. This burden is manageable today on a single, relatively modest, computer using the same 20-year old concept.

These tiles have not been defined using an objective algorithm. They are defined manually considering two criteria: 1) density of radars, and 2) grid size. High density of radars should be combined with a small grid and vice versa. Tiles located in between other tiles are smaller for this reason, and because of the many radar coverage areas that overlap adjacent tiles.

All radars with coverage areas in any tile are assigned to that tile, no matter how small that coverage area may be. This means that data from several radars are processed in more than one tile. The selection is conducted automatically each time the production chain is run; there are no pre-defined look-ups.

Based on the above, all data available on 31 January 2012 at 12:00 UTC were used for benchmarking tests. The complete data set comprises 695 individual scan files and 53 files already containing polar volumes. The number of polar volumes processed is 102 in total containing 827 scans. Hardware used is a three-year old rack server with one Xeon 2.26 GHz quad-core CPU, two threads per CPU core, 12 Gb of RAM, and 7200 rpm SATA-300 disks.

Table 4. Benchmarking tasks. For all tasks except the first, all 827 scans in 102 volumes have been processed.

Task	**Time (sec) CPU cores**	
	1	**4**
Read and sort 695 scans, write 49 polar volumes	10	1
Anomaly identification and removal only	121	31
Full QC chain, generate all beam-blockage analyses from scratch	1094	436
Full QC chain, look up all beam-blockage analyses	146	42
Generate European composite as single tile	451	-
Generate tiled (parallelized) European composite	-	43

The complete task of organizing input data, quality controlling them, and then building the European composite is achieved within 90 seconds from “cold start” using a single *crontab* entry and modest server hardware. Detailed plots, illustrating performance achieved when processing all data for February 2012, will be shown at the conference.

5. Discussion and conclusions

While the contingency-table skill scores succeed in revealing improved data quality using the toolbox, they must be taken with a grain of salt. The Hanssen-Kuipers Skill is supposed to be an overall score that accounts for information that is both desired and undesired, and it is unbiased in relation to the precipitation frequency, but it is also overly dependent on Probability of Detection (Wilson, 2001). This seems apparent in our experience. Nevertheless, it is encouraging that improvements can be seen despite some data already being quality controlled in unknown ways prior to their being sent to Odyssey. Also, the nature of some problems in the data imply that only a small proportion of the coverage area will be negatively affected, e.g. thin sectors of interference from external emitters, so it is also encouraging that improvements can be seen despite the difficulty in co-locating these areas with surface observations.

The characteristics of gauge-radar comparisons can be very well known within individual countries, but using point observations to validate radar data at the European scale is a different matter. Solutions developed for national or regional use may not be applicable at the continental scale. Our biggest surprise is that the so-called "system biases" were so great, revealing a large over-estimation by radar at close range. We suggest that the main explanation for this is ground clutter that is not properly suppressed yet. Statistical methods that have been developed for gauge adjustment in northern Europe succeed in revealing improvements to data quality achieved when using the BALTRAD toolbox and the suggested processing chain, but the statistical results themselves are too noisy to be used for gauge adjustment, not even a bulk adjustment based on an average bias. Data quality must be improved further before this should be attempted.

Default parameter settings have been used with the anomaly detection and removal tools with data from most radars. Radar-specific settings can be tuned and stored in an XML configuration file, and this has been done for a few radars. This exercise has not been exhaustive, however. One uncertainty is knowing the limitations of the software, so there is definitely scope for future work on this. The local radar owners could install the toolbox and tune the parameters themselves before reporting the results to the Odyssey. The community-based nature of BALTRAD software encouraged this.

Beam-blockage correction appears to have brought a positive impact to data quality. This can be seen through the improved concentricity of accumulated precipitation fields from some radars, and also through the rejection of data in sectors with over 70% blockage where data from overlapping radars are used instead. The results from this trial suggest further that correcting for beam blockage that is unrelated to topography (e.g. buildings, towers, forest) could be a serious concern, and that tools for addressing such errors should be made available.

The data processing chain tailored here for Odyssey is a "serving suggestion" that can be implemented as is, in modified form, or not at all, according to OPERA's wishes. The statistical evaluation methods can also be made available in the form of a quality monitoring tool for measuring data quality continuously, but also off-line when trialing new tools for Odyssey operations.

The issue of quality control of data before they are sent to Odyssey remains a problem. This strategy was never intended when the OPERA data centre was specified. With the suggested production chain's data quality improvements from the BALTRAD toolbox, perhaps OPERA should reconsider whether Odyssey should request unprocessed data from the members so that they can be quality controlled in a harmonized way as originally intended. Doing so would add much needed traceability to Odyssey's production practices, and it would also be easier to isolate improvements to data quality. An example of this is Finnish data that are quality controlled for Odyssey but not for BALTRAD. In this trial, we have used Finnish data received through BALTRAD, which are not quality controlled, and results show clear improvements achieved using the BALTRAD toolbox in a way which is traceable and therefore understandable.

When it comes to performance, additional improvements are surely achievable by optimizing the processing chain further at the code level, but we don't think such optimizations should ever accept a lowering of data quality. Using the parallelized toolbox on modern server hardware with a more than one CPU, each with more than four cores, would certainly speed up processing times. Using more modern hard-disks (than SATA) might help too. But probably the biggest improvement would be if input volumes are quality controlled as soon as they are available instead of waiting and processing everything at once as we have done here. This would surely reduce processing times to sub-minute levels without compromising quality of results.

Using GTOPO30, for use with beam-blockage modeling, is a reasonable starting point at the European scale and for civilian (non-military) applications. It would be interesting to test the beam-blockage correction algorithm using a higher-resolution DEM, e.g. ASTER or similar, but doing so lies outside the scope of this trial.

While the BALTRAD toolbox supports ODIM 2.1, where accurate azimuthal angles may be added which are necessary for accurate beam-blockage analyses, the beam-blockage analysis does not yet make use of the detailed metadata. Consequently, uncertainty in beam-blockage results is obvious with a limited number of radars at present.

Acknowledgments

This trial was funded by EUMETNET OPERA. BALTRAD was, and BALTRAD+ is, part funded by the European Union (European Regional Development Fund and European Neighbourhood and Partnership Instrument).

References

Barnard G.A., 1945: A new test for 2×2 tables. *Nature* 156-177

Häggmark L., Gollvik S., Ivarsson K-I., and Olofsson P-O., 2000: Mesan, an operational mesoscale analysis system. *Tellus* **52**(a), 2-20

Koistinen J. and Hohti H., 2010: Operational diagnosis of precipitation detection range. *Proc.* ERAD 2010 (short abstract).

Koistinen J. and Michelson D.B., 2002: BALTEX Weather Radar-based Precipitation Products and their Accuracies. *Boreal Env. Res.* **7**(3), 253-263

Michelson D.B. and Koistinen J., 2000: Gauge-radar network adjustment for the Baltic Sea Experiment. *Phy. Chem. Earth (B)* **25**, 915-920

Michelson D.B., Lewandowski R., Szewczykowski M., Beekhuis H., 2011: EUMETNET OPERA weather radar information model for implementation with the HDF5 file format. OPERA Working Document WD_2008_03, Version 2.1

Michelson D. and Henja A., 2012: OPERA Work Package 3.6: Odyssey additions. Task 3. Tuning and evaluation of "andre" tool. OPERA WD_2012_02c, 20 pp.

Peura M., 2002: Computer vision methods for anomaly removal. *Proc.* ERAD 2002, 312–317

Wilson C., 2001: Review of current methods and tools for verification of numerical forecasts of precipitation. COST 717 Working Document WDF_02_200109_1. 14 pp.