WORLD METEOROLOGICAL ORGANIZATION
COMMISSION FOR ATMOSPHERIC SCIENCES

Original: ENGLISH

WWRP/WGNE
Joint Working Group on Forecast Verification Research (JWGFVR)

Freie Universität Berlin, Berlin
8-10 September 2011

REPORT

Attendees: Beth Ebert (co-chair), Laurie Wilson (co-chair), Barb Brown, Caio Coelho, Martin Göber, Simon Mason, Marion Mittermaier, Pertti Nurmi, Joel Stein (by phone), Yuejian Zhu (by phone), Nanette Lomarda (WMO)

Apologies: Anna Ghelli, Barbara Casati

1. Welcome and introductions

The meeting was opened by JWGFVR co-chair Beth Ebert and the agenda was approved. Note-takers were recruited from among the attendees.

2. Minutes and actions from 2010 Toulouse meeting

The minutes from the 2010 JWGFVR meeting in Toulouse were approved. The actions arising from that meeting were reviewed. An updated action list arising from the 2011 meeting is provided as an appendix to this report.

3. Verification activities in ongoing projects

3.1 Forecast and research demonstration projects

SWFDP

Laurie reported that the project has 3 tiers – (a) global modeling (ECMWF, NCEP, UKMO), (b) regional centers, and (c) national meteorological services. There are three existing projects, one in development, and another proposed:

- Southern Africa – completed as demonstration, now in operation, with 16 countries, including Indian Ocean island nations. The final meeting was held in Mauritius last July.
- Eastern Africa – Nairobi is the regional center, with Tanzania very involved (also in the Southern African one).
- Southwestern Pacific – regional center is Met Service New Zealand in Wellington.
- SE Asia – newest SWFDP being developed, with the regional center in Hanoi, Vietnam.
- Bay of Bengal – proposed new SWFDP.

Peter Chen is very supportive of the verification component of the SWFDPs. Laurie has been involved in many of these providing verification management, planning, training and assistance particularly for the African SWFDPs. Rick Jones has conducted some training activities as well. Some national centers are now doing their own verification, mainly of heavy rain but also strong wind. Important thresholds have been identified.

All forecasts and variables should be verified, started at NMS level. So far there has been little success getting regional or global centers to do verification. In S. Africa there has

been no verification of regional forecasts (spatial verification of hand-drawn graphical chart). A graphical analogy of the contingency table proposed for verification of spatial forecasts but has not been used – still hoping it will happen. The Met Office will soon be doing some verification using SEEPS approach. Anna is doing some verification of ECMWF forecasts but the details are not clear. Apparently ECMWF cannot redistribute GTS data. The US has indicated it will verify its NWP forecasts, subject to resources.

SNOW-V10

Laurie reported that verification of SNOW-V10 (Vancouver Olympics RDP forecasts) was slow to get going, but some preliminary results are starting to come in now. There continues to be some effort required to get the observations in good shape to use for verification. The verification in being done in two parts;

- User focus – verify forecasts for user-supplied thresholds for Olympics period (three weeks starting in Feb 2010)
- Model verification – more general verification for more thresholds; longer time period (Feb-Aug 2010)

The Canadian nowcasting and mesoscale modelling groups are doing some of the verification work.

London 2012 Olympics

Marion contacted key people in the Met Office but there was little interest in having a FDP/RDP. There is some interest in creating boundaries for finer resolution modelling. There will be no special observations for the Olympics period, although some extra instruments have been installed in the venue area. Since the Met Office isn't doing anything special, our group does not need to be involved or to consider this in the future.

Lake Victoria project

Laurie reported that the focus of this project is on impacts for fishing and agriculture. There are 5,000 deaths annually due to sudden storms. The World Bank is funding this project, which is also tightly linked to Eastern Africa SWFDP. Finland is also involved (cell phone technology), and the Met Office is running a 4-km model.

There is little or no quantitative verification so far, although Marion has experimented in the past using TRMM data for qualitative precipitation verification. This project is still in development – not sure who is working on this.

FROST-14

Pertti will represent JWGFVR in this project, and is currently helping in planning the verification, which is likely to be similar to SNOW-V10. It is still being discussed whether it will be a FDP or RDP. The weather and associated forecasting issues sound similar to Whistler.

New observations will include sensors along the new road to Sochi, radar located along coast – the latter will not be very useful for mountainous regions but probably will be for Sochi and coastal regions. There may be a vertically pointing radar. Multiple modelling groups are involved (high resolution runs from COSMO, etc.), also TIGGE, NOAA, ECMWF. They will have two winters to set up and run the models to allow tuning. The VERSUS package is one option for verification.

Action: Barb, Laurie and Pertti agreed to meet during the EMS/ECAM to further discuss the FROST-14 verification plan **(done)**

3.2 Spatial Verification Methods Intercomparison Project

There was significant interest in conducting a second phase of the ICP in complex terrain, focusing not only on precipitation but also on wind and possibly clouds. Forecasts and

analyses from COPS / MAP D-PHASE could be used. INCA and VERA both provide Austrian analysis over Alps. The SRNWP could have a role in setting up datasets and possibly coordinating. Barb has some project money to help support ICP-2. (see also 12.4 in this report)

3.3 UK STORMS DEMON project

Marion described a NERC project called STORMS DEMON that focuses on better prediction of floods and storm impacts. It will consider inputs, datasets to better quantify rainfall amounts, and quantify biases in forecast models (short and long-term). It will consider uncertainty in measurement as well as prediction. The JWGFVR precipitation verification document might provide some useful guidance for this project.

Actions: Marion will attend a 1-day workshop on STORMS DEMON in November.
Pertti to provide some COST731 input for the workshop.

4. Welcome and talk by Prof. Uwe Ulbrich

Professor Uwe Ulbrich welcomed the JWGFVR to the Freie Universität Berlin, and gave a short time describing the history and activity of the Meteorological Institute.

5. Joint activities with other working groups

5.1 WWRP Joint Scientific Committee

Beth represents the JWGFVR at the WWRP JSC. The committee now has ten independent scientists to keep WG chairs honest – Celeste Saulo (U. Buenos Aires) is our advisor. Nice things were said about JWGFVR. Some of the topics discussed at the JWC meeting included new methods for mesoscale verification (including the desire to use standard metrics in a similar way as is done for the WMO CBS Standard Verification), ways to handle observational uncertainty, verifying the onset of cloud / timing errors, use of the COPS and D-PHASE datasets. Celeste requested some South American verification training, which could be part of Nowcasting symposium in Brazil later 2011. Gilbert Brunet is keen for us to develop methods for seamless verification; HYMEX data might be appropriate for exploring this.

Dr Duan, chair of the Tropical WG, has indicated that they would like help on verification in the Shanghai Landfall Typhoon project. In accordance, Anna provided some verification training at a TC workshop in Shanghai in May 2011, and Yuejian will provide some ensemble TC verification training at a workshop in Nanjing in December 2011. We will need to work with this WG to strengthen connections.

FROST-14 will be an area for collaboration with the Nowcasting WG. In discussions with Paul Joe on the SOCHI project, they are interested in verifying additional quantities including the 0° isotherm and forecast timing errors. There is a by-invitation workshop in October 2011 in Boulder on the use of NWP in nowcasting.

Action: Barb to attend the October 2011 workshop in Boulder on NWP in nowcasting and give a talk on verification.

5.2 WWRP Mesoscale Forecasting Research WG

As this is our second joint meeting in a few years, it is clear that we have many interests in common. Further discussion on this topic was deferred until the joint meeting on Sept 10.

5.3 WWRP SERA WG

JWGFVR met with SERA last year in Toulouse. It appears that we would need to drive any joint project since they are a very diverse group. A project on TCs would be ideal but interaction with SERA has been difficult to maintain. Simon mentioned that we might want to join with SERA to become involved in a project with the International Red Cross on how disaster

occurrences are linked to weather and seasonal forecasts, especially precipitation and flooding. This would require some verification information regarding impacts of timing and spatial errors. The group agreed that this is possibly a better fit with SERA.

Action: A subgroup (Barb, Beth, Martin and Simon) will discuss a potential joint (with SERA) project on disaster occurrences.

5.4 GIFS/TIGGE

Laurie attended the 9th GIFS/TIGGE meeting in Geneva in September 2011. Verification featured strongly on the agenda, especially regarding TCs and heavy precipitation. Highlights included ECMWF ensemble verification of TC tracks and intensities in their annual report using a modified form of the ROC with the false alarm ratio instead of false alarm rate. NCDC keeps historical information on tropical cyclones so in principle all can be verified historically. Strike probabilities are also verified. Tom Hamill recently submitted a paper verifying TIGGE forecasts over the US using Stage 4 precipitation analyses showing that a multi-model ensemble outperforms single model ensembles.

The GIFS demonstration project will build on the SWFDP framework. Mio Matsueda showed an exciting new precipitation product based on four ensembles from the TIGGE archive showing areas where each are forecasting more than 90% of the model climatology (based on ~5 years of TIGGE data). This is not considered operational given the 48h delay in data reception plus 1 day delay for processing. Matsueda is moving on to Oxford to do something else. The La Plata basin project was mentioned again, and it was suggested that they apply for official status as a RDP/FDP.

GEO-WOW is a EU funded project to improve accessibility and attractiveness of data sets, including site-specific information from TIGGE. ECMWF will begin this work starting in approximately 12 months.

5.5 WWRP/THORPEX Polar Initiative

Laurie attended a WCRP/WWRP meeting on this initiative in late 2010. It was suggested at that meeting that CBS provide standard NWP verification results for regions poleward of 60N and S. The initiative will include data assimilation experiments that exploit satellite data in polar regions, field campaigns, and coupled modelling in the polar regions. Verification is considered to be an important component. A scientific steering committee is being formed with a chair to be confirmed (likely to be Thomas Jung from Alfred-Wegener-Institute).

Action: Pertti to represent JWGFVR in the scientific steering committee for the WWRP/THORPEX Polar Initiative.

5.6 Sub-seasonal to seasonal prediction

Laurie represented JWGFVR at a WWRP/WCRP/THORPEX meeting on sub-seasonal to seasonal prediction held at the Met Office in December 2010. A project is being planned, with David Anderson wrote an interesting document on the research issues and will write the implementation plan. They have requested Barb to be a member of the scientific steering group, particularly since she also has an interest in SERA.

Action: Barb to represent JWGFVR in the scientific steering committee for the WWRP/WCRP/THORPEX sub-seasonal to seasonal prediction project being developed.

5.7 CLIVAR

CLIVAR also participated in the workshop on sub-seasonal to seasonal prediction. A question of interest is how to conduct time-scale independent verification of forecast uncertainty

(e.g., spread-skill, included as part of "seamless verification"). At longer time scales (e.g., decadal) the sample size is quite small so this may be an impossible task – usually a proxy that can be measured/observed is used to infer information about scales that we can't. Simon noted that there is a guidance document for verification in CLIVAR.

> **Action:** Simon to distribute a copy of the CLIVAR verification document to JWGFVR members.

5.8 WGNE/WGCM Climate Metrics Panel

Beth represents JWGFVR on this panel, which is highly dependent on the efforts of its leader (Peter Gleckler, PCMDI). The panel is specifying and using standard metrics to verify CMIP climate model runs. A Wiki page is being developed to describe the panel's work and evaluation of the CMIP results.

5.9 CBS CG-FV

Marion reported on the changes to the CBS Standard Verification for NWP, including finer resolution, use of a standard climatology (ERA Interim), and addition of surface verification. Daily scores will be collected in order to compute confidence intervals on the monthly scores. JMA is considering how to verify surface variables, where quality control provides a challenge.

5.10 CBS Commission for Aeronautical Meteorology

No real news.

5.11 CBS Public Weather Services Programme

No news.

5.12 SRNWP

In the SRNWP Verification project the SEEPS score will be implemented in Europe, and scores for verifying extremes are also being explored. Verification against European radar data probably will not happen due to time and quality control constraints. The verification project will end soon. Suggest freer access of results.

5.13 Sand and Dust Storm project

There will be a meeting 25-26 November in Turkey for the North African and European Regional Steering Group. The Met Office has some experience verifying dust forecasts.

> **Action:** Marion or Anna to represent JWGFVR at the Sand and Dust Storm steering group meeting in November in Turkey, pending approval from the Met Office.

6. JWGFVR documents

6.1 Recommendations for verification of cloud forecasts

No further progress on this document, but Marion has submitted a paper on cloud verification to *QJRMS*, and some of this may be useful for the document. Probably not too much more needs to be done. An October 2011 finish date would permit publication as a Technical Document this year, but that seems like an unrealistic goal. A December 2011 target date would still permit a publication date of 2011.

> **Actions:** Marion and Anna to agree on the latest draft of cloud verification recommendations.

Marion to update the document and send it to JWGVFR members for feedback.

6.2 Recommendations for verification of tropical cyclone forecasts

This is being (slowly) written by Laurie, Barb, Beth, and Eric Gilleland (NCAR). It is required at least in draft form to be used at an ensemble TC training course in Nanjing in December 2011. The document therefore needs to include ensemble verification.

Actions: Beth, Laurie and Barb to meet Sunday morning to discuss next steps for TC paper (done).
Barb to coordinate follow-up phone discussion in September.

Simon noted that to get maximum impact it is necessary to publish JWGFVR recommendations documents as review papers in established journals.

Action: Beth to take the lead on converting the QPF/PQPF verification document to a journal article.

7. Outreach

7.1 Verification web page (http://www.cawcr.gov.au/projects/verification)

Beth thanked all people who provided new references on forecast verification to be included in the webpage.

Actions: Beth to add new references to the web page.
Simon will write a response for FAQ #1: "How many samples are needed to get reliable verification results?"
Marion will look for broken links in the web page.

7.2 Community verification packages (R, MET, etc.)

Simon mentioned that IRI is working on a new version of Climate Predictability Tool software, http://portal.iri.columbia.edu/portal/server.pt?open=512&objID=697&PageID=7264&mode=2. CPT is designed specifically for producing and verifying seasonal forecasts of tercile probabilities. Simon explained that in order to run CPT one needs the climatology of observational data and a set of retrospective climate model forecasts (hindcasts).

Caio pointed out KNMI's online Climate Explorer (http://climexp.knmi.nl/) which assists in the display and verification of several types of climate products.

Barb mentioned the need for someone to continue the development and maintenance of the R verification package developed at NCAR because Matt Pocernich has moved field of work. Simon suggested contacting Exeter University and/or MeteoSwiss (Andreas Weigel) to ask whether they could continue work on development and maintenance.

Barb noted that it is now possible to read TRMM data in MET. Martin mentioned DWD is currently handing satellite data for verification using R. Marion is testing MET functionality as an external (to the US) user.

Barb mentioned that NCAR has the option to host visitors for couple of weeks to work on proposals.

Actions: Simon to write a paragraph about IRI/CPT software capabilities on seasonal forecast verification to be added to the web page.
Caio to write a paragraph about KNMI Climate Explorer forecast verification tool to be added to the web page.

Beth to put information about available forecast verification packages on the web page.

Marion will ask Chris Ferro and David Stephenson if there is any interest at Exeter University in maintaining the R verification package.

7.3 Verification tutorials

(a) Kuala Lumpur, Malaysia

Nanette mentioned that the current proposed dates are March or May 2012 (May preferred). The tutorial will be for one week and given in English with Malaysian funding for students and WMO funding for lecturers. Beth suggested tutorials should not only target Malaysian students but also students from other countries of the region. The idea is to use material similar to previous tutorials already given by JWGFVR, with lectures and exercises. At least 2 lecturers would be needed. Laurie, Pertti, Barb and Caio volunteered to be the lectures.

Action: Nanette to follow up with Malaysian contacts to start planning for 2012 tutorial.

(b) With EMS/ECAM, Reading, 2013

Pertti suggested that this training could focus on subject areas of interest to forecasters based on the experiences of the Melbourne verification workshop. In particular tutorials could be focused on warning verification. Martin suggested focusing on basics followed by more advanced verification topics later. There was an overall agreement that this will depend on the target audience. We should have a better idea after the Melbourne workshop in December 2011.

Action: JWGFVR members make a decision on tutorial topics for EMS/ECAM-related verification workshop when we meet in Melbourne in December 2011.

(c) South America (Peru, Colombia, Ecuador, Brazil)

Pertti reported on training he conducted in Peru, which had an aim not only on verification, but capacity building for various countries including Peru, with funding provided by the Finnish Ministry. In his two visits to Peru Pertti had the chance to talk to management people in Peru and explain why forecast verification and its training are important. Peru training was based on EUMETCAL material. In his first visit Pertti used four full days to teach the following topics: basics of forecast verification, categorical verification, continuous verification and probabilistic verification (1 topic per day), with exercises from the Helsinki tutorial on each of the topics but no student projects. In his second visit he used the first day to review what he had taught in the first visit and the second day to teach probabilistic verification with some examples. Pertti still needs to provide a document with recommendations on verification scores for SENAMHI. The general feeling was that Peru students were very enthusiastic and satisfied with the training, rating it at 4.25 out of 5 on a post-course survey.

Pertti mentioned that Ecuador and Colombia are also interested in having similar verification tutorials funded by the Finnish Ministry.

Simon mentioned he recently ran similar training in Colombia (five days spent in verification and hands on exercises and a couple of other days on other things). The focus was on tercile based probabilistic seasonal forecasts. This training was funded by the Colombian Ministry of Health. Lectures were given in English with simultaneous translation to Spanish.

Nanette mentioned the next Nowcasting Symposium will be held August 15-24 2012 in Brazil (Rio de Janeiro). JWGFVR has been requested to run some verification training in this workshop. She also mentioned that if this training is to be combined with a JWGFVR meeting and the nowcasting workshop it is possible to be funded by WMO.

Action: Beth to follow up with Nowcasting Symposium organizing committee and Celeste Saulo (U. Buenos Aires) regarding verification training needs.

7.4 Conferences

The following conferences/workshops/meetings were mentioned as events that might interest JWGFVR members or where they could provide some help/support:

- International Road Weather Conference, SIRWEC, Helsinki, May 2012 (www.sirwec2012.fi)
- CMOS meeting, joint with AMS meeting early June 2012 in Montreal
- Nowcasting workshop, 15-24 August 2012, Rio de Janeiro (Brazil)

8. Planning for 5th International Verification Methods Workshop and Tutorial

Half a day was spent planning the tutorial and program for the 5th International Verification Methods Workshop to be held 1-7 December 2011. For details see Appendix 2.

Actions:

1. Beth and local organizing committee (LOC) to check R and Excel installations on BMTC computers
2. Beth and LOC to organize a welcome event for tutorial students and a Sunday outing
3. Laurie to design and prepare data for the group projects
4. Group (Pertti, Anna, Marion, Laurie) to select students from pool of applicants (**done**)
5. Pertti to send follow-up email to Helsinki students to find out how they are using what they learned
6. Marion to take the lead in assembling the program, with help from Barb, Laurie, Martin, Pertti, Beth, Simon
7. Caio to contact Paco Doblas-Reyes about a WGSIP invited talk by Oscar Alves (**done**)
8. Lecturers to send their talks out ahead of time to the JWGFVR
9. Pertti to investigate bulk purchase of data sticks for tutorial participants (for R and data)
10. Anna to investigate whether Wiley could provide bags

9. Ideas for work in new research areas

9.1 Ensemble verification

Simon mentioned a few approaches that we should consider:

(a) Conditional exceedance probabilities – a way to measure reliability of continuous forecasts without using bins and binary variables. Checks whether there are conditional biases in the forecast. Better than rank histogram, and can be applied to individual forecasts as well as a collection of forecasts. See Mason et al. (MWR, 2007) and comments to come out in MWR in a few months time.

(b) Recent work with Andreas Weigel to extend the concept of the two-alternative forced choice (2AFC), which measures forecast discrimination for categorical, probabilistic, and deterministic forecasts of continuous variables, to verify ensemble forecasts. A general discrimination score could be the best to use for subseasonal forecasts, i.e., it is a score that one could use for all time scales.

Yuejian gave a remote talk in which he described the verification of forecast 10% and 90% quantiles derived from ensembles. The simplest method was:

Count 0 if: obs > f(10%) or obs < f(90%)

Count 1 if: obs < f(10%) or obs > f(90%)
Average over many cases. Could be converted to a skill score by applying the above method to both the forecast and to climatology. A useful diagnostic diagram is a plot of the forecast and observed quantile as a function of lead time.

Marion found that the double penalty applies when verifying high resolution ensemble forecasts, and that the Brier score penalizes a mis-located forecast similar to RMSE for deterministic forecast verification

Simon recommended that JWGFVR publish a paper that describes how to verify several attributes of ensemble forecasts, and users could then choose methods which evaluate attributes that are important to them. Illustration of the Murphy-Winkler framework for ensemble verification.

9.2 Seasonal forecasts

Caio discussed the need to verify mapped seasonal forecasts. Charts show where warm / cold (or wet / dry) regions were forecast and observed. The Wilson score may be a good approach to use with spatial forecasts.

9.3 Seamless verification

Simon noted that the 2AFC approach could be used in a seamless verification context, as it has similar interpretation even when applied to different kinds of forecasts.

Beth noted that the context for "seamless verification" is coming from the need to verify sub-seasonal forecasts, which connect NWP and seasonal prediction. The seamless verification approach used would depend on how the prediction is made, and whether one wanted to know whether the model producing a realistic forecast (rather than a correct forecast). Approaches that could be used to evaluate multi-week dynamical processes like blocking, Rossby waves, etc. include phase verification, EOFs, and weather regime prediction (including transition from one type to another). Rather than simply assessing overall skill, it may be more important to focus on particular attributes. For example, verify forecast intensities and frequencies and whether these are predicted better than climatology.

Simon pointed out that the predictand is often not well defined – does the forecast refer to point scale or grid scale? Focus on what one could reasonably hope to predict.

9.4 Warnings / extreme events (also discussed at joint meeting)

Martin led a discussion on verifying warnings. There is growing interest in fuzzy or near-miss definition of events – note that a forecast can verify as both a near miss and a near hit at the same time (e.g., Michael Sharpe's work at UKMO). Warnings also frequently don't have a clear meaning – is an event the "worst" thing that happens in an area, or does it need to cover some reasonable fraction of the area? The Met Office is starting to issue probabilistic warnings based on their high resolution ensemble. Different verification approaches may be needed for different users.

9.5 Data assimilation

This discussion considered to what extent the data assimilation system should provide input for the verification system (e.g., quality control of the observations). Verification against model analyses may not be the best approach, particularly for surface / low level variables, as they're not a good substitute for the truth (no matter how convenient) and may contain significant model bias. Some analyses are better than others (e.g., ECMWF); verification against an ensemble of analyses may get at observation uncertainty.

The Met Office uses QCed observations from its DA system as verification data. The group recommended that a sensitivity experiment be conducted to investigate how the tolerances used in data quality control affect the verification of forecasts against QCed observations.

(These points were discussed further with the Mesoscale WG the following day.)

Action: JWGFVR to work with an appropriate modeling partner to conduct a sensitivity experiment on the effect of data assimilation QC tolerances on forecast verification results.

Joint meeting with the Mesoscale Weather Forecasting Research Working Group

10. Adoption of the agenda; working arrangements

This is the second time that the JWGFVR has met with the Mesoscale WG; the first was in Shanghai in 2008. Each topic in the joint meeting was co-led by a member from each working group.

11. General overview of both WG's activities

Jeanette provided an overview of the Mesoscale WG's activities, focusing on issues that they felt JWGFVR would be interested in. They are involved in two upcoming workshops, one on the use of NWP in nowcasting (Boulder, Oct 2011), and the other on modeling in complex terrain (spring 2012). She indicated that an important issue for their group is the limits of predictability, especially as we move to higher resolution models. They are interested in getting some guidance from JWGFVR on standard metrics to use for mesoscale model verification.

Laurie provided an overview on the Verification WG's activities including outreach and capacity building, advisory documents, participation in ECMWF Advisory Committee, and web activities (FAQ, sharing of tools).

12. Project/activities in which both groups are involved

12.1 Vancouver 2010

Stephane Belair and Laurie reported on SNOW V10. LAM models were run at 2.5 and 1 km resolution, producing mapped forecasts and meteograms of a number of variables. The data from this experiment are available to be used for further research. The observations are still being worked up; multiple observations at some sites allows for the estimation of observation error. (See also 3.1)

12.2 Sochi FROST-14

Stephane Belair and Pertti reported on FROST-14. The radar (to be installed) will look up the valley toward Krasnaya Poliana (ski venue) but unfortunately won't have a clear view, there is lots of blockage. A vertical profiler will also be installed.

The focus of this experiment is (a) mesoscale forecasts in complex terrain with models run down to 250 m resolution, (b) regional EPS, (c) nowcasts of high impact weather including fog. TIGGE forecasts will be made available in real time. Want to examine the value chain: global NWP → high-res models → statistical post-processing → official forecasts. The verification is likely to be similar to that being done in SNOW V10, i.e., user-oriented point-based verification of forecasts and research-oriented verification of high resolution model output.

Standard formats for model outputs will be very important in order to facilitate their use and their verification. Tiziana Pacagnella will look for resources within the COSMO consortium that might be able to help. There are four working groups in FROST-14, each headed by a Russian scientist. Ideally Roshydromet would perform the verification with guidance from JWGFVR.

See http://frost2014.meteoinfo.ru for more information. (See also 3.1)

Actions: Pertti, Laurie, Barb to prepare a verification plan for FROST-14.

Define a strong recommendation for standard model format and (ideally) centralized verification by Roshydromet to push forward to the JSC.

12.3 HyMeX

Volker Wulfmeyer gave an overview of HyMeX, which is a WWRP project that focuses on the hydrometeorology of the Mediterranean at a variety of scales. The Mesoscale WG is particularly interested in heavy rain and flash floods. The Special Observation Periods will be Sept-Oct 2012, Mar-Apr 2013, Sept-Oct 2013, Mar-Apr 2014. There will be three supersites making detailed measurements and strong use of radar data from Spain, France, and Italy (via OPERA).

A newTask Team on Modeling 5 (TTM5) has been proposed by Manfred Dorninger and Volker Wulfmeyer to plan the verification during HyMeX. They are currently leading this effort but they would very much like input from JWGFVR. There is a need to coordinate the verification efforts of the project participants, and to make sure that the surface data are harmonized so that they can be used for verification. Tiziana noted the need for better data sharing, as it is currently so difficult to get permissions for data related to each project. Multiple analyses over Europe could provide different verification options.

Action: JWGFVR to suggest a representative to work with Manfred Dorninger and Volker Wulfmeyer to plan the verification of HyMeX QPFs and hydrological forecasts

12.4 NWP in nowcasting – what does application of NWP for nowcasting imply in terms of verification needs "beyond the usual for NWP"?

Martin noted that model forecasts and manually produced warnings can be quite different - warnings have two additional free parameters to be decided by the forecaster – lead time and duration, which are not usually an issue for longer range NWP. Severity is frequently another factor. Humans and models have different natures, in that humans deliberately over-warn. Nevertheless, it may be possible to approximate a warning service using model output only.

To help people take more notice of small probabilities associated with rare events, probabilistic warnings should be presented relative to the climatological probabilities. The odds ratio can be used to assess warnings.

There is an interplay between hits, misses and false alarms such that in order to achieve a certain POD or threat score one would need to make a decision to warn at a fairly low probability, which introduces a high over-forecasting bias. As forecasting systems become more accurate this problem lessens.

Dale Barker noted that NCAR's VDRAS is a nowcasting application of NWP, with storms already spun up. 3-hourly data assimilation is not really useful for nowcasting since frequent cycling means spin-up hasn't completed and the forecast actually degrades. Clouds take a long time to spin up, so better assimilation of cloud data is a high priority.

13. Scores to be promoted for routine mesoscale model long-term quality assessment and intercomparison (proposal reprinted as Appendix 3)

Laurie presented a draft paper discussing verification of forecasts from mesoscale models. Rather than specifying how it should be done, it delves into a number of issues including the purpose of the verification; choices to be made regarding surface or upper air verification and point-wise vs. spatial verification; and issues to be considered when verifying different meteorological variables like precipitation, wind, etc. The paper proposes that the JWGFVR and SRNWP (which involves some in the Mesoscale WG) collaborate on a second verification methods intercomparison that would include point-wise verification of extremes using new scores like SEEPS, in addition to standard and spatial verification methods.

The two groups discussed the interaction of data assimilation with verification. The idea would be to take advantage of verification information in the DA forward model and OPS (Obs Preprocessing System). One could split the observations so that only some are assimilated while the rest are available for verification. The main advantage is that it provides information on uncertainty (based on difference between observations and forecasts plus representativeness error, measurement errors). However, representativeness error is different from model to model – this might have an impact on comparing models, but as we are not taking into account observation error at this point the loss might not be so great. One problem is that some of the observations are thrown out by the DA system and would not be available for verification. This topic needs further consideration, and a test of the impact of "lost" observations. (see 9.5)

Barb related the JWGFVR's thoughts regarding the next phase of the Spatial Verification Methods Intercomparison Project, i.e., complex terrain, wind as well as precipitation (see 3.2). She recommended that the experiment also include a protocol for ranking cases. The Mesoscale WG was quite keen for COPS and MAP D-PHASE data to be used for ICP-2. The INCA and VERA analyses could be used for verification data (see paper by Kann et al. WAF 2011 on sensitivity of analysis to data density). The Mesoscale WG are also interested in verifying humidity and cloud forecasts.

Actions: Barb, Marion, Laurie, Beth, Volker, Jeanette, Mathias, and Eric Gilleland to better define the ICP-2, perhaps having a special meeting at the Melbourne verification workshop.

This group to also work on the mesoscale verification methods document.

14. "Non-standard" verification

14.1 Cloud and cloud properties verification and methods / suitable data sources

Marion noted that modelers want verification of regional forecasts, diurnal variation, vertical distribution, bias, etc. It is important to know whether the model provides instantaneous or time-averaged cloud forecasts. Traditional and spatial metrics can be used to look at spatial and temporal behaviour of model clouds. It is desirable that the cloud analysis be model-independent, though most cloud analyses make use of at least some model information (e.g., for cloud base). For 3D structure PDFs of model clouds can be compared to CloudSat to evaluate bulk properties. CloudSat data is more difficult to use for direct model evaluation.

User needs for aviation and model verification can be different – for aviation use surface stations, while bulk statistics may be fine for model evaluations. But even surface observations have biases (e.g., manual vs. automated). Many cloud products are available (e.g., for climate), but they need some validation. Volker noted that the Climate Monitoring SAF's SEVERI and AVHRR cloud products seem to be very good quality.

14.2 Measures suitable for extreme weather

Timing and duration are not verified well though they are important aspects of extreme weather. Pertti noted the lack of availability of extreme observations for real-time verification, and those that are may not be reliable (e.g., attenuation of radar signal in extreme heavy rain). Real-time verification – even just a picture or a map – is valuable for forecasters. Possible new sources of data include aircraft MODE-S transponders (with 2s frequency) and GSM telephone masts.

Although no score has all of the desirable properties, the new scores for verifying deterministic forecasts of extremes (EDS, SEDS, EDI, SEDI) appear to be useful, with symmetric scores preferred. Information on location and timing errors is important to forecasters. For rare events, quantifying the significance of the score is problematic. Use quantiles of the distribution to be able to combine results from many locations to build sample size.

14.3 Verification of convection-permitting LAM EPS

Laurie noted that when verifying EPS forecasts against surface observations, representativeness (scale) error should not be taken into account, except to ensure that there is an appropriate match to the forecast. Sometimes remotely sensed data are closer to what the model is producing than basic variables. An example is brightness temperature.

Martin suggested that upscaling is important, and that it is necessary to use multiple measures to evaluate the forecasts. Kazuo Saito showed that to handle the double penalty that occurs for high resolution EPS forecasts, the neighborhood approach (e.g., fractions skill score) is useful for showing the value of a high resolution forecast over a low resolution one. Dale Barker pointed out that one can't assume that regional model is better for everything (e.g., for upper air variables, global models are still better) but we hope the higher-resolution models are better for weather variables.

14.4 Verification over "difficult" terrain (not representative for typical observation conditions) (e.g. urban, steep orography)

Mathias Rotach discussed the model handling of difficult terrain as smoother terrain but with a change in roughness length, rather than something more realistic. Verification can be done at levels above the surface to get a more valid picture of model skill, or else physical downscaling can be used to predict the variable at the actual height of the observations. There was some debate as to whether physical downscaling removes the benefit of high resolution. Marion stated that statistically post-processed global model will generally do better than post-processed LAM. Stephane said he would prefer not to alter the model values.

Showing forecast value is difficult using every-day forecasts; it may be necessary to consider forecasts of extremes.

14.5 Seamless verification across space/time scales

Beth considered the needs for seamless verification across scales as (a) is my model doing the right thing? and (b) can I use this forecast to make better decisions? Process studies and examination of marginal distributions are appropriate verification approaches for the first question, while neighborhood, 2AFC, and conditional verification approaches might be more appropriate for user-focused verification. Dale noted that evaluating the early time steps of a forecast gives valuable performance information.

Seamless verification implies changing the resolution of observations as you move from longer ranges to shorter ranges. A seamless verification approach may be to "stretch" forecasts to do more than what they were intended to do originally (e.g., making a 2-week forecast look more like a 1-week forecast rather than like a 1-month forecast), and evaluate how well they can do that. Peter Steinle noted that most users simply want verification information at scales that are relevant for them, and aren't concerned about the other scales.

Jeanette raised additional issues including the need to distinguish normal and extreme weather when evaluating models, and need to place greater focus on timing at short ranges.

14.6 Typhoon verification

Barb and Yu Hui raised a number of important issues. Track and intensity verification have been done for many years but are not enough, particularly to help forecasters use TC forecasts. Additional important information is needed for storm structure, precipitation, storm surge, landfall time/position/intensity, consistency, uncertainty, and additional information to assist forecasters (e.g., steering flow). What about false alarms and missed events? Genesis – how far in advance can TCs be predicted?

Ensemble TC forecasts pose a particular challenge. The current methods are inadequate and not often applied. This is an area for new research.

What are the observational errors in such an extreme environment? Does evaluating maximum wind speed make sense when observations are unreliable? Perhaps the 90th percentile would be more robust. Satellite data should also be used more often to verify TC forecasts.

Appendix 1. JWGFVR Action Items – Updated at Berlin meeting (8 Sept 11); updates are in blue

	Action	Responsible	Status
	Berlin – September 2011		
1	Attend a 1-day workshop on STORMS DEMON in November	Marion	
1.1	Pertti to provide some COST731 input for the workshop.	Pertti	
2	Attend the October 2011 workshop in Boulder on NWP in nowcasting and give a talk on verification.	Barb	
3	A subgroup will discuss with SERA a potential joint project on disaster occurrences.	Barb, Beth, Martin, Simon	
4	Represent JWGFVR in the scientific steering committee for the WWRP/THORPEX Polar Initiative.	Pertti	
5	Represent JWGFVR in the scientific steering committee for the WWRP/WCRP/THORPEX sub-seasonal to seasonal prediction project being developed.	Barb	
6	Distribute a copy of the CLIVAR verification document to JWGFVR members.	Simon	
7	Represent JWGFVR at the Sand and Dust Storm steering group meeting in November in Turkey, pending approval from the Met Office.	Marion	Marion will not be able to attend. Perhaps Anna could attend, or else Ric from the Met Office verification team.
8	Agree on the latest draft of cloud verification recommendations.	Marion, Anna	
9	Update the cloud verification document and send it to JWGVFR members for feedback.	Marion	
10	Discuss next steps for TC verification paper	Beth, Laurie, Barb	done
11	Coordinate follow-up phone discussion on TC verification paper in September.	Barb	done
12	Take the lead on converting the QPF/PQPF verification document to a journal article.	Beth	
13	Add new references to the FAQ web page.	Beth	
14	Write a response for FAQ #1: "How many samples are needed to get reliable verification results?"	Simon	
15	Look for broken links in the web page.	Marion	

16	Write a paragraph about IRI/CPT software capabilities on seasonal forecast verification to be added to the web page.	Simon	
117	Write a paragraph about KNMI Climate Explorer forecast verification tool to be added to the web page.	Caio	
18	Put information about available forecast verification packages on the web page.	Beth	
19	Ask Chris Ferro and David Stephenson if there is any interest at Exeter University in maintaining the R verification package.	Marion	
20	Follow up with Malaysian contacts to start planning for 2012 tutorial, to be led by Laurie, Pertti, Barb and Caio.	Nanette	
21	Make a decision on tutorial topics for EMS/ECAM-related verification workshop when we meet in Melbourne in December 2011.	all	
22	Follow up with Nowcasting Symposium organizing committee and Celeste Saulo (U. Buenos Aires) regarding verification training needs.	Beth	begun
23	Workshop - With local organizing committee (LOC), check R and Excel installations on BMTC computers	Beth	
24	Workshop - With LOC, organize a welcome event for tutorial students and a Sunday outing	Beth	
25	Workshop - Design and prepare data for the group projects	Laurie	
26	Workshop - Select students from pool of applicants	Pertti, Anna, Marion, Laurie	done
27	Workshop - Send follow-up email to Helsinki students to find out how they are using what they learned	Pertti	
28	Workshop - Take the lead in assembling the program, with help from many others	Marion, with Barb, Laurie, Martin, Pertti, Beth, Simon, Anna	
29	Workshop - Contact Paco Doblas-Reyes about a WGSIP invited talk by Oscar Alves	Caio	done
30	Workshop - Lecturers to send their talks out ahead of time to the JWGFVR	Marion, Barb, Laurie, Martin, Pertti, Beth, Simon, Caio, Ian	

31	Workshop - Investigate bulk purchase of data sticks for tutorial participants (for R and data)	Pertti	
32	Workshop - Investigate whether Wiley could provide bags	Anna	done
33	JWGFVR to work with an appropriate modeling partner to conduct a sensitivity experiment on the effect of data assimilation QC tolerances on forecast verification results.	who?	
34	Prepare a verification plan for FROST-14.	Pertti, Laurie, Barb	
35	Define a strong recommendation for standard model format and (ideally) centralized verification by Roshydromet to push forward to the JSC.	Pertti, Laurie, Barb, Beth	
36	Choose a JWGFVR representative to work with Manfred Dorninger and Volker Wulfmeyer to plan the verification of HyMeX QPFs and hydrological forecasts	Barb, Beth	
37	Barb, Marion, Laurie, Beth, Volker, Jeanette, Mathias, and Eric Gilleland to better define the ICP-2, perhaps having a special meeting at the Melbourne verification workshop.	Barb, Marion, Laurie, Beth	
38	Above group work on the mesoscale verification methods document.	Barb, Marion, Laurie, Beth	
Toulouse – September 2010			
2	Talk to telecommunication officer at WMO to help get African observation data for verification of SWFDP products	Nanette Lomarda	Nanette – no extra verification data other than GTS. Laurie – not all data getting into GTS, need to get more in, especially a problem in E Africa; Peter Chen is aware – maybe he needs to do something. Hamse Kabela (Tanzania) will be working on getting more GTS data into system.
8	Spatial Verification Intercomparison project potential work on timing verification - ask physics colleagues about their problems with timing (diurnal cycle and/or others)	Anna Ghelli	
10	Provide example of verification seasonal TC predictions to TC Panel by November 2010. Convey to TC Panel that it would be easier to simply do the verification on an occasional basis rather than prepare/support a community code.	Barb Brown	Not done; Barb will talk with Nanette

13	Inform JWGFVR when COST 731 final report is ready (~March 2011), share with WG	Pertti	Final seminar 1 year ago. Report not done.
16	Follow up with Herb Peumpel as to whether advice on verification of aviation products is desired.	Nanette Lomarda	Nothing new – no action Mentioned the importance of AMDAR – it is proprietary?
18	Write a section on verification of extended/seasonal range TC forecasts for TC verification document, after talking with Frederick Vitart and Fernando Prates (ECMWF) on verification of monthly TC forecasts	Anna Ghelli	
26	*JWGFVR web page FAQs:* 3. Review FAQs and see whether some editing is needed 4. Go through the COMET site on verification and see whether new FAQ topics are suggested	All Beth Ebert	(3) Not done (4) Not done
27	JWGFVR reference list – members to update Beth about articles from literature by end of November 2010 *JAS, Atmospheric Research* *QJRMS, JGR* *WAF, Tellus* *MWR, Atmosphere and Ocean* *Met Applications, Journal of Hydrology* *Journal of Climate, International Journal of Climate* *BAMS, JAMC* *J. Hydrometeorology, Australian Met Society Journal* *Met Z*, *Natural Hazard & System Science* National reports from USA and Canada	Marion Mittermaier Joël Stein Pertti Nurmi Laurie Wilson Anna Ghelli Barbara Casati, Caio Coelho Barb Brown Beth Ebert Martin Göber Yuejian Zhu	Many done, others still need to send info
29	Provide Nanette with a description of topics that JWGFVR could offer as part of a travelling tutorial	Anna Ghelli	Not done
30	Send information on climate model verification methods to put on JWGFVR web site	Barbara Casati	Not done
32	Work with Brian Mills (SERA chair) to draft ideas for joint JWGFVR/SERA project around TIGGE TC tracks (CXML data)	Anna Ghelli, Beth Ebert, Barb Brown, Laurie Wilson, Joel Stein	Written, but no reply. May propose a different project.

Helsinki – June 2009			
8	Verify some of the ECMWF EPS precipitation forecasts for South Africa (SWFDP) if data from rain gauges are provided to her.	Anna Ghelli	In progress; difficult to get replies; Simon will contact Anna to see if she needs help
20	Add a short description of the various verification packages on the JWGFVR web page and provide a link to the package web page itself for more detailed information.	Beth Ebert	Not done
Shanghai – December 2008			
28	Contact European Weather services to find out what it is done on severe weather warnings verification.	Martin Göber	Martin: Workshop has not happened yet; keep this item on the list

Appendix 2 – Workshop planning

8. Planning for 5th International Verification Methods Workshop and Tutorial

8.1 Budget

- WMO will cover the tutorial costs (total of $86K)
- Scientific workshop registration fee: $350 for delegates, $200 for students

8.2 Tutorial

Venue and logistics

- Tutorial will be held at Bureau of Meteorology Training Centre, which is ~10 minutes walk from the workshop at the Bureau Head Office.
- Rooms
    - Two home rooms with 17 computers in each, separate lecture room that seats 48 people
    - Not sure home rooms will be available during scientific workshop (probably yes)
- Computing
    - Check that computers in training room have correct version of R loaded on them
    - Allow people to use their own laptops for projects
    - Ask people to install a particular version of R; will also want Excel
    - Give people data and correct version of R on a stick
    - Encourage participants to do some EUMETCAL tutorials in advance
- Recording presentations
    - Will record lecture (video of person plus slides) using Bureau's software
    - Not sure how it will be distributed (DVD? Web?)
- Food / refreshments
    - Morning and afternoon tea to be provided
    - Lunches – people get their own lunch at small café in building or one of several local restaurants in vicinity (this is standard practice for other WMO courses held in BMTC)
    - Need to organize some kind of social event
        - Barbecue in park nearby on the first day?
        - Perhaps also an outing for the Sunday between the tutorial and workshop

Schedule

- Most sessions to be 90 minutes: 45 minutes lecture and 45 minutes exercises
- Include exercises in sessions
    - Need to allow for people who don't bring a laptop – share?
        - Find out ahead of time who will bring a laptop (Simon says recent experience is that almost everyone brings one)
        - Some exercises may only require pen and paper
- Need for ensemble verification?
    - Many students will not be up to this, but this also is a frequently requested topic
    - Most people are interested in probabilities
    - Need to include tercile probability information for seasonal prediction
- Also need to mention multi-category scores in contingency table talk

- Include parallel session(s) at end on special topics – Warning verification, aviation, tropical cyclones, decadal and climate prediction

Lecturers

- Each topic includes lead plus a couple of people who are "helpers"
- Most of us want to come to tutorial
    - 7 people - ~4K each => $28K total, which is in the budget (Barb still needs to find out about funding from NCAR)
        - Martin needs only accommodation / per diem for the tutorial
        - Joel will come for workshop only. Also Yuejian.
        - Anna does not plan to come.
    - Ian Jolliffe needs support for accommodation / per diem – we should be able to fund this

Projects

- Laurie will take the lead on this
- With 34 students there are likely to be 8 or 9 group projects
- Need to get student list ASAP to start getting data so it can be prepared
- Each teacher should be assigned to one or two projects to focus on
- Assign students to groups ahead of time
    - Give them a chance to select top 4 projects from the list – assign to one
        - Request information about R experience and knowledge to help balance the level of experience in each group

Student selection process

- Group to make selection: Pertti, Marion, Laurie, Anna
- Process is underway, each member of group has made their choices (→ **done 14/9/2011**)
- Approximately 15 can be funded by WMO

Follow-up

- Send email to students from Helsinki – how did they use what they learned? What should we do in next tutorial? Pertti will send follow-up email to Helsinki students
- Need to have a certificate for the students. WMO doesn't have a standard; will use Helsinki certificate as a model

8.3 Scientific Workshop

Venue and logistics

- Conference room on 6th floor of Bureau Head Office holds 160 people
- Offer generic wireless internet plus computers set up for people who don't have laptop
- Poster session to be held in 9th floor Winter Garden (atrium); has security doors
- TravelLodge hotel is next door (AUD$135 per night). Should be enough room for everyone.
- Train station is next door – bus from airport stops there
- Icebreaker 1st night in ground floor restaurant (Summit Café)
    - Public lecture (Neville Nicholls) to be given at the end of the first afternoon before icebreaker

Program topics and presenters

- Topics as advertised on poster and web site, possibly adjusted by abstracts that are submitted
- Marion will work on putting program together
    - Other people will help –Barb, Laurie, Martin, Pertti, Beth, Simon (for some sessions)
        - Can ask for advice from other staff on certain talks
- Local organizing committee will put abstracts into one document to send out for review
- Target date 15 October to have selections made, program assembled, notify people
- Local organizing committee will notify people of whether they will give talk or poster; Barb and Carol will be a back-up

Keynote speakers

- Caio suggested to include Oscar Alves to give a WGSIP talk; Beth will send invitation letter once Paco Doblas-Reyes has mentioned this to him (**done** – 14/9/2011)
- 5 keynote speakers – need to provide funding for three
    - Susan Joslyn – communicating uncertainty
    - Andrew Watkins – verifying seasonal forecasts (does not require funding)
    - Andreas Weigel – verifying ensemble forecasts
    - Phil Gill – verifying aviation forecasts
    - Barb Brown – verifying TC forecasts

Format of workshop sessions

- Determine length of talks when we see how many talks we have
- Each session to end with a discussion
    - Keynote speaker responsible for seeding discussion in those sessions
    - Perhaps also ask chairs of other sessions to lead a similar discussion
    - Need rapporteur for each session
- Posters: Can accommodate as many as needed. Perhaps portrait size would be best… Can also put posters on windows and walls, not just boards

Special issue of Meteorological Applications

- Results of discussion should feed into article for Met Apps (workshop is being supported by Wiley)
- Need a couple of volunteers to be special editor for this
    - Caio volunteers for this; still need at least one more person
    - Marion also? Maybe even one more person
    - Also need someone to put together the lead article

JWGFVR meeting afterward

- JWGFVR to meet on the day after the workshop Dec 8 to debrief

8.4 Miscellaneous

- Everyone should try to arrive no later than Tuesday night
- Tutorial talks should be sent out ahead of time
- Buttons, bags, sticks
    - Memory sticks- just for tutorial; Pertti will look into this
    - Bags – we need to check with Wiley – Anna?

Appendix 3. DRAFT - Verification of forecasts from mesoscale models

(prepared by Laurie Wilson, September 2011)

This discussion paper describes some general principles and methods of verification as applied to forecasts from mesoscale models (often called Local Area Models – LAMs). In some respects, established methods for global models can be used directly for LAMs, though the priorities and emphasis might be different. This document does not present recommendations for a specific set of verification scores to be used, but rather describes many of the issues that must be considered in order to arrive at a well-reasoned verification strategy, with application to high resolution models.

I. Purposes of verification

In general, different users of verification results will have quite different needs, which means that the target user or users must be known before the verification system is designed, and also that the verification system design may need to be varied or broadened to ensure that the needs of all the users can be met. To summarize briefly, the first principle of verification is: Verification activity has value only if the information generated leads to a decision about the forecast or system being verified. Thus, the user and the purpose of the verification must be known in advance.

Purposes of verification can be classified as either administrative or scientific, or rarely a combination of both. Administrative verification includes such goals as justifying the cost of a weather service or the cost of new equipment, or monitoring the quality of forecasts over long periods of time. Administrative verification usually means summarizing the verification information into as few numbers as possible, using scoring rules. Scientific verification, on the other hand, means identifying the strengths and weaknesses of a forecast in enough detail to be able to make decisions about how to improve the product, that is, to direct research and development activity. Scientific verification therefore means more detail in the verification methodology, and less summarizing of the verification information. The term “diagnostic verification” is often applied to verification with specific scientific goals.

To focus the verification activity, it is often useful to articulate the question(s) that are to be answered. Some specific questions that would be of interest in mesoscale model verification are:
1. Are the mesoscale model forecasts more accurate than forecasts from global models at points of interest?
2. In situations where global model and mesoscale model forecasts are simultaneously available, are finer spatial scales more accurately represented in the mesoscale model than they are in the global models?
3. Which configuration of the mesoscale model is most accurate? (e.g. in the SRNWP project)
4. What is the smallest scale at which the model shows positive skill?
5. How well does the model predict spatially defined features such as fronts and contiguous areas of precipitation?
Etc.

The main users of verification information aimed at answering the above questions would be mesoscale modelers and forecasters who may use the mesoscale model as guidance. Of the above, questions 3 and 4 might be most interesting to the modeling community while questions 2 and 5 might be more of interest to forecasters. Question 1 is a basic question which would be of interest to both communities. And, it should be said that the administrators of a model research and development program would be interested in question 1 and 2 especially.

II. Choices to be made

2.1 Surface or upper air verification?

It is safe to say that an important reason for developing LAMs in the first place is to be able to apply higher resolution lower boundary conditions and higher horizontal and temporal resolution to better resolve local processes, most of which either take place in the boundary layer, or are determined by the lower boundary conditions. Thus it would follow that surface-based verification is relatively more appropriate to the assessment of LAMs. "Surface-based" means all parameters defined by observation from the surface, including precipitation rate and accumulation, temperature, relative humidity (or other moisture variable), wind direction and speed, cloud amount and base height, visibility, and mean sea level pressure.

While upper air verification might also be of use for LAMs, the general lack of high resolution observation data sources would limit its applicability to answer questions such as those posed above. One possible exception is the use of satellite data for the verification of cloud distribution. Relatively little work has been done in this area; it remains a promising area for development of verification methodology.

2.2 Point-wise and/or spatial verification?

Most verification methodology has been developed and used for pointwise verification, where model forecasts are matched to observations at points, verification scores computed and summarized over all available forecast-observation pairs. Observations are usually from in situ data networks, and can be considered to accurately measure meteorological variables only in the immediate vicinity of the instrument. These data networks often seriously undersample the spatial variability of the variables measured, an important issue for variables with high small scale variability such as wind and precipitation. Data are sometimes available from high resolution special networks, and these should be used in pointwise verification whenever possible, especially for high resolution models. But it remains true that pointwise verification gives useful verification information only at the points where data is available, and all points are usually treated independently. A simple phase error, for example, will result in an overforecasting error at one location and an underforecasting error at a nearby location. For mesoscale models this becomes a more critical issue when the smaller scales represented in these models are subject to phase errors, whether variable or systematic. Such spatially defined errors are not seen by pointwise verification, which limits the utility of these methods for some users. Nevertheless, for the purposes of comparing the general accuracy or skill of models (e.g. questions 1 and 3), and for users interested in forecast quality at specific locations, it makes sense to apply verification measures on a point-wise basis.

Spatial verification methods, on the other hand, are specifically designed to diagnose the forecast accuracy of spatially-defined features such as fronts or precipitation areas or to evaluate different scales separately. These are particularly useful for mesoscale models and should form part of any diagnostic evaluation of mesoscale models. Their use typically relies on the availability of spatially continuous or at least very high resolution observations, for example, radar or satellite data. Compared to pointwise methods, spatial methods are much newer and their properties are not yet well-understood. Given their potential importance for the verification of high resolution forecasts, it is recommended that a project be undertaken in conjunction with the SRNWP project to comparatively evaluate these methods, to support recommendations on which of the methods are best for general use in the verification of mesoscale models.

2.3 Thoughts on choosing best methods for specific variables

2.3.1 Point-wise verification

In addition to considering the verification question to be answered, and the user of the information, the selection of specific verification methods depends on the nature of the variable

being evaluated. Listed below are some factors to consider in the verification of specific surface variables.

1. Temperature (2m). Since temperature is relatively smoothly-varying and is an approximately normally distributed unbounded variable, verification methods for continuous variables can work well, for example mean error (linear bias), mean absolute error and (root) mean square error. For skill, one could use a skill score based on the MAE or the RMSE. It is recommended that the elevation at which the temperature is represented be matched between observation and model forecast, by processing the model forecast (not the observation) using a standard assumed lapse rate. To be consistent with reduction of pressure to sea level, the same lapse rate could be used as for that purpose, the WMO standard, applied to the temperature at the model's lowest computational level. If the observation height is above the lowest model level, then the forecast value can be obtained by vertical interpolation between model levels. Temperature can certainly be verified as a categorical variable, especially when performance with respect to critical values such as 0 C is important.

2. Wind. 10m winds normally exhibit considerable variability in time and space. It is important that the definition of the wind from the model (averaging period – 1 minute or 10 minute etc) be matched to the observed averaging period. It is also acceptable to upscale the observed wind to the model's temporal scale (one or more timesteps) if the information is available to do so. Wind can be verified as a vector quantity, the average vector error, and this is recommended. More commonly, wind is verified by direction and speed components. When this is done, results are more meaningful if light wind cases are separated out (e.g. < 3 m/s) for both the speed and direction verification. To keep the sample balanced, light winds are excluded if either forecast or observed or both, then the remainder of the sample can be evaluated using continuous scores (bias, MAE, RMSE) or windspeed and direction can be separated into categories and verified by means of a contingency table and associated scores. Category thresholds can be chosen to suit the user(s) of the verification results.

3. Humidity. Here the selection of the form of the variable to be evaluated is dependent on the user and purpose of verification. For general model accuracy assessment, one might verify relative humidity, but for consideration of high impact weather it would be more useful to evaluate dew point temperature, which relates to smog warnings and heat waves (high humidex). Since the moisture variables can all be computed given one of them along with temperature and sometimes surface pressure, it is possible also to evaluate other forms which are not observed directly such as specific humidity and mixing ratio, for diagnostic purposes. As for temperature, dewpoint can be verified using the continuous variable measures (bias, MAE, rmse), but relative humidity and absolute humidity are bounded variables which might be best verified categorically, with the categories chosen according to the use of the verification results.

4. Clouds. Clouds are usually reported as total cloud cover in tenths (or oktas) of the celestial dome. To the extent that this quantity is an output parameter of the model, it can be compared with observations on a pointwise basis and verified. Though it could be verified as a continuous variable (errors in terms of numbers of tenths), it is often most meaningfully verified as a categorical variable in three or four categories covering the range. Another reason for categorical verification of cloud cover is that its climatological distribution is often u-shaped with highest frequencies at the extremes of the distribution. Categorical verification, especially frequency bias, is a convenient way of determining whether the model can reproduce the observed climatological distribution. Cloud base (Ceiling) height is another form of cloud information of interest especially to the aviation community. This quantity is extremely difficult to predict with useable accuracy, and if verified, the verification should be with respect to categories of interest to the user. As a variable, it has a highly skewed and rather non-linear distribution. It might make most sense to verify the existence of a ceiling below, say, 3000 m as a two-category (yes-no) contingency table, and then to verify ceiling height categorically only for those cases where the occurrence of a ceiling is both predicted and observed.

5. Visibility. Surface visibility is highly variable on small scales both spatially and temporally, especially when the more important low visibilities are occurring. Point observations are often inhomogeneous as well, since manned sites may report “prevailing visibility”, estimated in terms of the ability to see specific markers while at automated sites visibility is estimated in many ways, but these do not include any attempt to integrate the information over different directions. This means that any objective verification results for visibility must be interpreted with caution since error levels in both the observations and model are expected to be large. As for ceiling height, visibility should be verified categorically, with the categories set in consideration of the user needs and in terms of the available sample. Also as for ceiling height, visibility can be evaluated as a two category yes-no with respect to a specific threshold, say 5 km, then occurrences and forecasts below that limit can be verified with respect to other thresholds of importance.

6. msl pressure. This is perhaps less important a parameter to verify for mesoscale models, but may be useful to evaluate in comparison with global models for example. Reduction of the model’s pressure field to sea level would be done following the WMO standards, and the temperature lapse used should be consistent with that used to adjust temperature in the vertical. Verification methodology should follow the WMO guidelines for verification of deterministic models, including the scores used. Verification against analysis is possible. This should be done against an independent analysis (perhaps a high resolution global analysis would suffice). If no independent analysis of sufficiently high resolution is available, evaluation against the analysis used to initialize the model is possible. It should be made clear in the verification presentation, however, that verification against the model’s analysis overstates the accuracy. Verification against observations is preferred, and it is best if any quality control of the observations is independent of the model being verified.

7. Precipitation amount and rate. Precipitation accumulation should be verified according to the guidelines described in the WMO document on this subject. For mesoscale models, the interest would be in shorter accumulation periods of 6h or less. Data from high density gauge networks should be used wherever possible. Precipitation rate forecasts can be compared with radar estimates if available. Instantaneous precipitation rate is highly variable also in space and time, and so might not give meaningful results unless the forecast and observed rates are averaged over a suitable period.

2.3.2 New scores for pointwise verification

There are two different new scores which should be computed when possible, for the purpose of gaining experience with these measures.

1. The extreme dependency family of scores. There are four different but similar formulations of these scores: Extreme dependency score (EDS), symmetric extreme dependency score (SEDS), Extremal dependency index (EDI) and symmetric extremal dependency index (SEDI). These are all designed for contingency table evaluation and seek to improve evaluation of contingency tables in situations where the base rate is low (low frequency of occurrence of the event of interest in the verification sample), that is, for extremes. Of the four, the symmetric versions, SEDS and SEDI are preferred.

2. Stable equitable error in probability space (SEEPS). This score was developed for use as a “headline” score at ECMWF, specifically for precipitation forecasts. Its claimed stability properties make it a good candidate for identifying long term trends in forecast accuracy, thus it might be of greatest use for that purpose. It is based on an older score, linear error in probability space (LEPS), which also has useful properties but isn’t widely used. Both LEPS and SEEPS require access to a long term precipitation climatology at the verification locations so that the forecast and observation can be converted to probability space to compute the score. This may be a practical disadvantage, and has probably hindered the use of LEPS over the years, but it is also an advantage because it means that both scores can take account of differences in climatology from one station to another. SEEPS (actually, 1-SEEPS) is being

evaluated at ECMWF for global models; it should also be evaluated for mesoscale models. SEEPS is formulated in terms of a 3-category contingency table; it might also be applicable to the verification of wind speed since wind speed normally has a climatological distribution similar to that of precipitation accumulation in form and shape. LEPS is simpler in concept, is a linear score (not dependent on categorization), and could be applied to any variable for which long term climatology is available.

2.3.3 Comments on matching model forecasts with point observations for verification

In general "model to data" methods are preferred, where the observations are not processed at all. For continuous or slowly varying variables (spatially), simple linear interpolation from the grid to the observation location is recommended. For all other variables which are episodic (precipitation) or are characterized by sharp gradients (e.g. cloud, visibility and wind, it is preferred to match the observation to the nearest grid point forecast. If the modeler believes that the nearest grid point is not representative of the observation location, then a more representative grid point may be chosen instead as long as this is done in advance. For example representative grid points were chosen in advance of the Vancouver winter Olympics for use in verification of the mesoscale model forecasts. This is most often an issue in mountainous areas where there may be large height differences between the observation location and the nearest grid point.

2.3.4 Upscaling

The concept of upscaling observations to the model resolution is mentioned at this point because it effectively lies between point-based and spatial methods. Upscaling is usually done simply by averaging observations over a grid box, which smoothes out the smaller scale information present in a (higher) density observation network, and matches it to the resolution of the model. The aim is to determine the observational counterpart to a "grid-box average" which is what the model often predicts. The size of the grid box can be varied to suit the needs of the users; this would be one way of tracking scale information in the verification. It is advisable to set a lower limit to the number of observations required to estimate the grid box average, to ensure consistency of these estimates over the verification domain. Verification is local to the grid box, and is carried out only where there are enough observations. When there is only one observation in the grid box, this method becomes equivalent to "nearest-grid point matching" as described above.

2.3.5 Spatial verification

Numerous spatial and scale-sensitive verification methods have been proposed and tested in the last decade or so. Through projects such as the Inter-comparison project (ICP), the results of which were published in Weather and Forecasting in 2009 and 2010, experience and knowledge of the characteristics of these new methods is increasing. More testing and experience is required, however, before decisions can be made regarding which of the methods is recommended as standard.

Nearly all the methods require at least a high density network of surface observations to work effectively, but a few of the methods are flexible enough to be useable on standard-density networks of observations. For mesoscale limited domain models, access to high spatial density observations would be pretty much essential to obtain meaningful results. Of course satellite data and radar data should be useful for the spatial verification of the relevant variables.

The methods are:

Method for Object-based Diagnostic Evaluation (MODE)
Contiguous rain area (CRA)
Wavelet-based Intensity-Scale Technique and wavelet-based analysis

Image warp forecast verification method
Optical flow technique
Fractions Skill Score
Structure, amplitude, location (SAL)
Composite method

Descriptions of all of these, and examples of their application to the standard datasets in the ICP are published in the special collection of papers in weather and forecasting and references therein. Two aspects of the comparison of forecasts and observations in the spatial context would seem most important for mesoscale model diagnosis: Scale-tracking and the comparison of meteorological structures of interest ("objects"). Since observation networks and models typically resolve different subsets of the total spectrum of atmospheric variability, methods which implicitly or explicitly provide information on the scale content of forecast and observation would be of diagnostic value. For example, the fractions skill score, a "neighbourhood" method, implicitly limits the scale information content of forecast and observation to those scales larger than the chosen size of the neighbourhood. The score is computed for different sized neighbourhoods and the results plotted on a curve. The wavelet intensity scale technique allows a skill score to be partitioned according to scale, so that the smallest scale at which the forecast is skillful can be identified (question 4 above). Furthermore, the wavelet-based analysis method is a model-independent analysis which keeps track of the scales that can be resolved by an observation network, which varies according to the station density.

Methods which focus on the comparison and tracking of objects include MODE, SAL and the CRA technique. Comparison and verification is usually in the form of errors in specific defined attributes of the objects such as their shape, intensity and/or location. Object-based methods have been mostly applied to the verification of precipitation areas, but could be applied to any spatially defined structure, for example an area of strong winds, bounded by a specific isotach, or perhaps cloud areas, fog areas etc, any region of meteorological interest with well-defined spatial boundaries. One challenge with object based techniques is the matching of forecast and observation ("Which forecast object goes with which observed object"? and "how far apart do they have to be before they are considered not associated"?). These issues are addressed in different ways in the different methods, and are exacerbated by differences in the spatial scale information content of forecasts and observations ("Are all those small objects really related to that big one?"). Probably the simplest conceptually is the SAL method, which does not demand explicit matching of forecast and observed objects, but rather evaluates their attributes collectively over a pre-specified domain. That said, it is not yet clear how well these different methods will work in practice specifically for LAMs, and for variables other than precipitation. This is an important motivation for recommending a second ICP, in collaboration with the SRNWP.

III. A proposed second ICP in collaboration with the SRNWP.

This proposal links to the **Proposed Future of EUMETNET/SRNWP Verification Programme** as drafted by Dr. C. Wilson. The SRNWP proposal includes several components, focused on the intercomparison of different LAM implementations over a common European area by consortium members, for all of the variables listed above plus wind gusts. Verification using standard methods is proposed for the intercomparison (deliverables ND-1 and 2) and using selected spatial and scale-selective methods (ND-3). Severe/high impact weather verification is included (ND-4) , which relates to the new point-wise verification scores described above.

The primary goal of this project is to extend the experience with several spatial and scale-sensitive verification methods to the SRNWP domain and to variables other than precipitation in order to clearly identify their strengths and weaknesses for verification of high resolution models.

Compared to the SRNWP extension, this proposal mostly represents an extension and variant of the focus on spatial and scale-selective methods identified under ND-3. The other components of the proposed SRNWP extension are supported as proposed. The following process is suggested:

1. Following the ICP format, (see http://www.ral.ucar.edu/projects/icp/ for further information) specific meteorologically interesting cases of sufficient duration (a few days) should be identified, for the common domain of the different SRNWP models and for which high resolution spatial data is available, both in situ data and radar data (for precipitation). The case selection should not be limited to high precipitation events, but also include intense winds, and also some “ordinary” cases, meteorologically interesting, but not extremes.

2. The data should be prepared in a standard international format, and made available to all invited participants. Invited participants would include the developers of all the spatial methods listed above and/or their delegates, along with members of the consortium. That is, the evaluation should not be limited to the four methods listed in ND-3. Each group would be responsible for running their method on the standard set of cases, for all the models in the intercomparison, and writing up the results. A distributed approach to the analysis is likely more feasible for the spatial methods than a centralized approach, because the expertise in application of the methods is distributed.

3. Highest priority should be given to evaluation with respect to observations rather than analyses. If evaluation is done against analyses, then these analyses should be independent of all models in the intercomparison. If that is not the case, then the intercomparison would be unfair, unless alternative data matching methods are adopted, for example ensembles of the analyses, or random selection of the verifying analysis at each forecast. It is recommended that the wavelet-based analysis described by B. Casati be included as a candidate analysis in the list shown under ND-3. This method has the advantage that it is local, and sensitive to the variations in scales that can be represented over the observation network. Verification is point-wise, but only those scales represented in both observations and forecasts are verified.

4. The pointwise verification effort (ND-1 and 2) should include the use and testing of the new SEEPS score for precipitation and the SEDS and SEDI for extremes (ND-4). Since the SRNWP comparison includes the ECMWF model as a standard, and the SEEPS score is already being computed for several global models by ECMWF, it shouldn’t take too much effort to test the mesoscale models at the locations within the chosen common domain where data is available at ECMWF. This assumes that the climatology data needed for SEEPS can be released by ECMWF.

It is hoped that a joint SRNWP-ICP project will be of benefit to both the verification community and the SRNWP community, and that, by the end of this project, it will be possible to make clear statements about the applicability of all the verification methods, both point wise and spatial for high resolution models.