WORLD METEOROLOGICAL ORGANIZATION

WORLD WEATHER RESEARCH PROGRAMME

WWRP 2010 - 1

4th WMO INTERNATIONAL VERIFICATION METHODS WORKSHOP

Helsinki, Finland, 8 - 10 June 2009

WMO/TD-No.1540

Fourth International Verification Methods Workshop

Monday – 8 June 2009

8:10 **Registration** (FMI, Dynamicum Building Entrance Hall)

9:00 **Opening and Welcome** (Exactum Building Auditorium)

Session 1: User-oriented Verification
Chair: Pertti Nurmi

9:10	1.1	Juhani Damski: Customer-oriented services at FMI Laurence Wilson: What Is A Good Forecast: The importance of user-orientation in verification
9:50	1.2	Clive Wilson: Do key performance targets work?
10:10	1.3	Tressa Fowler: Wind forecast verification
10:30	1.4	Robert Maisha: UM model and Kalman Filter forecast verification at SAWS

10:50 – 11:20 Tea/Coffee

Session 2: Verification Tools and Systems
Chair: Matt Pocernich

11:20	2.1	Tressa Fowler: The Model Evaluation (MET 2.0), Overview and Recent Enhancements
11:40	2.2	Barbara Brown: MODE-3D: Incorporation of the time dimension
12:00	2.3	James Brown: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at point locations.
12:20	2.4	Yin Kong Leung: Objective verification of weather forecast in Hong Kong
12:40	2.5	Marcus Paulat: COSMO-DE-EPS – construction and verification of a limited-area ensemble prediction system on the convective scale

---

13:00 – 14:00 Lunch

Session 3: **Coping with Uncertainty in Verification Data**
Chair: Chiara Marsigli

14:00 3.1 Carlos Santos: Introducing observational uncertainty in the scoring of a multi-model limited area model over the European area
14:20 3.2 Otto Hyvärinen: Visualising verification results when no true reference is available
14:40 3.3 Theresa Gorgas: The Challenge of Finding "Good" Reference Data for Verification
15:00 3.4 Benedikt Bica: High resolution precipitation analysis and forecast validation over complex terrain using an inverse VERA approach
15:20 3.5 Laurence Wilson: Verification of multi-model ensemble forecasts using the TIGGE data
15:40 3.6 Matthew Pocernich: Feature-oriented verification of wind speed forecasts

---

16:00 – 16:30 Tea/Coffee

Session 4: **Properties of Verification Methods**
Chair: Chris Ferro

16:30 4.1 Ian Jolliffe: Probability forecasts with observation error: Is the Brier score proper?
16:50 4.2 Lovro Kalin: Is ETSS Really Equitable?

Session P: **Poster Introductions (one-minute each)**
Chair: Laurie Wilson
17:15 -

P01. Julie Demargne and James Brown: Communicating hydrologic verification information for operational forecasting and real-time decision making in the U.S. National Weather Service

P02. James McGregor: The value of GFS data for precipitation forecasting for the Waikato River catchment, New Zealand

P03. Guenther Mahringer: TAF Verification in the MET Alliance

P04. Sultan Al-Yahyai: Development of a portable verification Package

P05. Adriano Raspanti: VERSUS: Unified Verification Package in COSMO

P06. Yasutaka Ikuta: Evaluation of hydrometeors of a high-resolution model using a radar simulator

P07. Kristian Pagh Nielsen: Verification of cloud physical properties

P08. Robert Fawcett and Elizabeth Ebert: Base rates and skill scores

P09. Marek Jerczynski: Some robust scale separation methods at work

P10. Marion Mittermaier: Time-series analysis of scale-selective verification:
Can we use it for operational forecast monitoring?

P11. Pertti Nurmi: SAL Verification in Hydrological Catchments

P12. Juan Bazo: Verification of Seasonal Rainfall Prediction in the Rimac River Basin

P13. Jonathan Eden: Assessing the skill of GCM-simulated precipitation

P14. David Mendes: Improving Meteorological Downscaling Methods with Neural Network Models: South America Rainfall

P15. David Mendes: Artificial Neural Network (ANN) Application for South America Rainfall

P16. Gezu Mengistu: Long range Ethiopian Weather forecast

P17. Marcel Vallee: A block bootstrapping method for the Canadian NWP model

P18. Ewan J. O'Connor: Cloud verification using radar:
What is the half-life of a cloud-fraction forecast?

P19. Pascal Mailier: Assessing the operational skill of predictions of forecast error

P20. Petra Roiha: Analysis of Marine Seasonal Ensemble Forecasts for the Baltic Sea

P21. Petr Zacharov: An estimation of QPF uncertainty by forecasting the radar-based ensemble skill

P22. Matias Armanini : Extreme temperatures verifications on Argentina forecast by NWP GFS NCEP

P23. Monika Bailey: Verification of nowcasting methods in the context of high-impact weather events for the Canadian Airport Nowcasting (CAN-Now) project

P24. Kadarsah Binsukandar Riadi: Reliability Evaluation of HyBMG by Using the ROC Curve

P25. Girmaw Bogale: Rainfall prediction performance of WRF model over complex terrain of Ethiopia

P26. Kalle Eerola: Verification of the Hirlam NWP forecasts and the connection between the scores and improvements in the model.

P27. Ata Hussain: High resolution Regional Model (HRM) performance as NWP tool in Pakistan

P28. Oluseun Idowu : Verification of numerical weather predictions for the western Sahel by the United Kingdom Met Office Limited Area Model over Africa

P29. Anna Lindenberg: Applicability of Common Verification Methods for Comparisons between Measured Wind Data and Simulated Wind Field

P30. Jordi More: The Impact of Applying Different Verification Techniques and Precipitation Analyses in QPF Verification

P31. Maria Stefania Tesini: Verification and Statistical properties of COSMO-17 QPF

P32. Marco Turco: The forecaster's added value in QPF

P33. Daan Vogelezang: Verification of statistical forecasts of low visibility at the Amsterdam Airport

P34. Ji-Won Yoon and Sei-Young Park: Verification of ensemble forecast using the physical parameterization schemes of WRF model during the Changma period over Korea

P35. Xiaoli Li: Comparisons of global and regional ensemble prediction systems at NMC

P36. Joel Stein and Marielle Amodei: Another look at the contingency tables: Scores based on Manhattan distances in the error space

---

18:00 – 19:30 Ice-breaker and Poster Session (FMI, Dynamicum Building)

Fourth International Verification Methods Workshop

Tuesday – 9 June 2009

Session 5:		**Verification of Weather Warnings** Chair: Tressa Fowler
9:00	5.1	David Stephenson: The Verification of Weather Warnings: Did the Boy Cry Wolf or Was It Just a Sheep? (invited)
9:40	5.2	Martin Göber: Approaches to process and event oriented verification of warnings
10:00	5.3	Michael Sharpe: The challenge of verifying severe weather warnings
10:20	5.4	Clive Wilson: A critical look at the verification of Met Office "Flash" Warnings
10:40	5.5	Marion Mittermaier: Verifying extreme rainfall alerts for surface water flooding

---

11:00 – 11:20 Tea/Coffee

Session 6:		**Spatial and Scale-sensitive Methods** Chair: Francis Schubiger
11:20	6.1	Barbara Brown: Spatial Verification Methods
12:00	6.2	Eric Gilleland: Spatial Forecast Verification: The Image Warp
12:20	6.3	Chermelle Engel: A scale-based distortion metric for mesoscale weather verification
12:40	6.4	Stefano Mariani: On evaluating the applicability of CRA over small verification domains

---

13:00 – 14:00 Lunch

Session 6:		**Spatial and Scale-sensitive Methods (cont'd)** Chair: Johannes Jenkner
14:00	6.5	Marion Mittermaier: Identifying skillful spatial scales using the Fraction skill Score
14:20	6.6	Barbara Casati: New Developments of the Intensity Scale Verification Technique within the Special Verification Methods Intercomparison
14:40	6.7	Elizabeth Ebert: Feature-specific verification of ensemble forecasts

Session 7: **Spatial and Scale-sensitive Methods: High-resolution Models**
Chair: Marcus Paulat

15:00 7.1 Ulrich Damrath: Some experiences during verification of precipitation forecasts using fuzzy techniques
15:20 7.2 Marielle Amodei: Deterministic and fuzzy verification of the cloudiness of High Resolution operational models
15:40 7.3 Kees Kok: Valuing information from high resolution forecasts

---

16:00 – 16:20 Tea/Coffee

Session 7: **Spatial and Scale-sensitive Methods: High-resolution Models (cont'd)**
Chair: Stefano Mariani

16:20 7.4 Chiara Marsigli: QPF Verification of Limited Area Ensemble Systems during the MAP D-PHASE OP
16:40 7.5 Francis Schubiger: Verification of precipitation forecasts of the MAP D-PHASE data set with fuzzy methods
17:00 7.6 Sami Niemelä: Verification of High resolution Precipitation forecasts by Using the SAL Method
17:20 7.7 Mathias Zimmer: Towards Evaluating Timing Errors of Quantitative Precipitation Forecasts with the Feature-Based Technique SAL

17:40 – 18:30 Poster Session (FMI, Dynamicum Building)

20:00 WORKSHOP DINNER (Restaurant "Töölönranta"; self transportation)

Fourth International Verification Methods Workshop

Wednesday – 10 June 2009

Session 8: **Seasonal and Climate Forecast Verification**
Chair: Marcel Vallée

9:00	8.1	Simon Mason: Towards Standardized verification of seasonal climate forecasts (invited)
9:40	8.2	Barbara Casati: Extreme Value theory to Analyze, Validate and Improve Extreme Climate Projections
10:00	8.3	Pascal Mailier: Can you really trust long-range weather predictions? Confessions of a rogue forecaster.

10:20 – 11:00 Tea/Coffee

Session 9: **Tutorial Presentations**
Chair: Anna Ghelli

11:00 – 13:00
13:00 – 14:00 Lunch

Session 10: **New Ideas in Verification**
Chair: Theresa Gorgas

14:00	10.1	Christopher Ferro: Verification measures for rare-event forecasts
14:20	10.2	Deborah Glueck: Bias in trials comparing paired continuous tests can cause researchers to choose the wrong screening modality
14:40	10.3	Zoran Pasaric: Polychoric correlation coefficient in forecast verification
15:00	10.4	Johannes Jenkner: Verification of Probabilistic Calibrations for Deterministic GFS Precipitation Forecasts

15:20 **General Discussion**

16:00 **END OF WORKSHOP**

Session 1: User-Oriented Verification

1.1

What Is A Good Forecast: The importance of user-orientation in verification

Laurence J. Wilson

Meteorological Service of Canada

Allan Murphy has said that forecast verification is a useful activity only if the results lead to some decision about the forecast product being verified. This implies there must be a user for the verification output. And, it also implies that the verification methodology must be designed to tell the user what he or she wants to know about the quality of the forecast. User-relevant verification is a necessary but often ignored or under-emphasized component of current prediction systems. Verification summaries which use traditional measures, for example those routinely prepared by operational centers, may be useful for some, but are typically widely disseminated as if they should satisfy all users of forecasts. Generally, the more care that is taken to design a verification system to meet the specific needs of a particular user group, the more likely it is that the verification system will produce output that is useful in decision-making.

The meaning of “user-relevant” verification will be discussed, along with examples of the impact of user orientation on the design of verification systems.

Email: lawrence.wilson@ec.gc.ca

1.2
Do key performance targets work?
Clive Wilson

Met Office, Exeter, UK

Administrative verification is generally concerned with how well forecasting systems or National Meteorological Services perform overall. There is a need to justify the cost and investment decisions and to show, over time that these are delivering improved forecasts. By necessity a small number of overall measures are used and these are often composited into either a single or a few summary scores. Much of the standard verification literature has little to say on administrative verification beyond, quite rightly, warning against excessive summarising and the impossibility of reducing the multi-dimensional nature of verification to a single score. However governments and other interested stakeholders increasingly demand an objective overall measurement of forecast performance by which to judge and assess forecast providers. In recent years public sector agencies have seen the introduction of key performance indicators and the setting of targets to encourage changes aimed at improving their service. For the last decade the Meteorological Office has set targets for the quality of its NWP forecasts and reported on these annually. A composite summary index formed from skill scores of a number of forecast parameters for the UK and the globe has been used, with some (generally) minor changes in formulation over time. Annual targets are set for the index and meeting or exceeding these is judged a success. To encourage staff to work towards the desired outcome, successful achievement counts towards a corporate bonus scheme. A critical review of the experience in setting and seeking to achieve the targets reveals several difficulties: simple and easily understood scores are preferred by customers and fund-holders but may not be sufficient to show the underlying performance; timescales of major model developments and investments are generally longer than a year; predictability and natural variability have to be minimised or accounted for in setting targets ; a dichotomous success/fail assessment can influence decisions and behaviour, sometimes in a malign way. Recognizing that key performance targets are unlikely to be abandoned some suggestions for improving their formulation and use will be made.

Email: clive.wilson@metoffice.gov.uk

1.3
Wind forecast verification
Tressa L Fowler, Matt Pocernich and Barbara G. Brown

National Centre of Atmospheric Research, Boulder, CO, USA

Wind forecasts pose a unique verification problem for numerical modelers and energy producers. The energy producers' true interest is not in the wind forecast at all, but in the total power generation, of which the time-series of local wind speed is a primary component. Wind speed errors are thus of the greatest interest, but only in a range of wind speeds. In this type of situation, categorical statistics might be preferred to continuous. Since modelers have more information in space and less in time, the spatial features of wind may be useful in evaluation. Many variables other than wind speed may affect power output and the quality of wind forecasts including: wind direction, timing and location of changes, variability, threshold exceedances and event identification. Some preliminary methods of user-relevant wind verification are presented.

Email: tressa@ucar.edu

1.4

UM Model and Kalman Filter Forecasts Verification at SAWS

Robert Maisha Thizwilondi

South African Weather Service, Pretoria, South Africa

The South African Weather Service (SAWS) uses the United Kingdom Meteorological (UKMET) Office's Unified Model to produce forecast guidance. This model has been operational at SAWS from January 2007 and it is also used for research purposes. It produces the forecasts for South Africa and the Southern African region respectively. This model has a horizontal resolution of 15, 12 and also a 4 kilometers resolution that only covers South Africa. This model produces forecast up to 48 hours ahead and it is run once a day, but the 4 km model run forecasts up to 30 hours ahead. Numerical weather prediction models have been found to have systematic and non-systematic bias, especially when forecasting near-surface variables. The non-systematic errors in models were found to be due to physical parameterizations. The model systematic errors vary with geographical locations, times of the day and seasons. Like-wise, the systematic errors are difficult to quantify due to the complexity in separating model inaccuracies and initial conditions. Model systematic errors were found to be due to model's resolutions. Model post processing techniques were found to be very effective in solving model bias. The Kalman filtering system was found to be one of the most effective techniques of removing model bias.

The UM model forecasts are constantly monitored to evaluate its performance by computing monthly verification scores, in order to gain an understanding of the bias and also to come up with strategies to correct bias. Variables such as mean sea level pressure, geo-potential heights, temperature on different pressure levels, relative humidity, 10 meter level winds, dew point temperatures, specific humidity and wet bulb temperatures over the South African domain are evaluated. Verification scores such as bias, correlation coefficient and root mean square error were computed. This was done for the period February 2007 up to April 2009. Also a Kalman filtering system is applied to South African stations daily maximum and minimum temperature to correct the UM model forecasts bias. Also the forecaster's forecasts at SA major stations are also compared with UM model and Kalman filter forecasts.

In this talk the UM model forecasts error scores (bias and rmse) will be presented. Also the model bias removal technique i.e. Kalman filter forecasts results will be presented.

Email: robert.maisha@weathersa.co.za

Session 2: Verification Tools and Systems

2.1
The Model Evaluation Tools (MET 2.0), Overview and Recent Enhancements
Tressa Fowler

National Centre of Atmospheric Research, Boulder, CO, USA

Model Evaluation Tools (MET) is a freely-available software package for forecast verification. It is distributed through the Developmental Testbed Center (DTC) for testing and evaluation of the Weather Research and Forecasting (WRF) model. Development has been led by the community: including WRF users, the DTC, and verification experts through workshops and user meetings. MET allows users to verify forecasts via traditional, neighborhood, and object-based methods. To account for the uncertainty associated with these measures, methods for estimating confidence intervals for the verification statistics are an integral part of MET.

The latest release includes many new features. Verification of probabilistic forecasts is supported and includes appropriate statistics like the Brier score. Wavelet decompositions can be used to examine forecast performance at different spatial scales. User-defined polyline verification regions can be more efficiently defined and applied. The MET tools have been modified to enable more general comparisons between any two variable types. File preprocessing has been simplified, and the output statistics file format has been enhanced. The MET website has been updated to include a graphical user interface for producing configuration files used by the MET tools. Additional scripts to produce graphics have also been added to the MET website. Examples of the existing and new verification capabilities will be shown.

Email: tressa@ucar.edu

2.2
MODE-3D: Incorporation of the time dimension
Barbara Brown

National Centre of Atmospheric Research, Boulder , CO, USA

The Method for Object-based Diagnostic Evaluation (MODE) has been extended to the evaluation three-dimensional rainfall systems, with time as the third dimension. Precipitation objects are identified in three-dimensions (x, y, and t), where x and y are the two horizontal spatial dimensions and t is time, using a convolution/thresholding process. Relevant temporal and spatial attributes – including object centroid location, volume, velocity, and lifetime – are identified and assigned interest values as a component of a fuzzy logic approach for matching objects in space and time. New interest maps are applied in this extension of MODE.

The advantage of considering 3-D objects is that each forecast or observed dataset can be viewed in terms of a relatively small number of rain systems. Furthermore, the evolution of the predicted systems can be verified using basic geometric properties of the objects. For instance, with time pointing vertically, the angle with respect to the horizontal plane indicates the speed and direction of propagation of the rain system. The temporal centroid, beginning and ending times all pertain to the timing of the system. Using 3-D object attributes, we devise a metric that accounts for spatial, temporal and propagation errors and results in a relative measure of forecast quality that can be used to compare forecasts from different models, or forecasts from a given model on different days.

The 3-D objects identified in numerical forecasts and observations from the IHOP period and the 2005 NSSL/SPC (National Severe Storms Laboratory/ Storm Prediction Center) Spring Experiment are used to demonstrate MODE-3D.

Email: bgb@ucar.edu

2.3

The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at point locations

James Brown

NOAA/NWS/OHD, Maryland, USA

Ensemble forecasting is widely used in meteorology and, increasingly, in hydrology to quantify and propagate modeling uncertainty. In practice, ensemble forecasts cannot account for every source of uncertainty, and some sources are difficult to quantify accurately. Ensemble forecasts are, therefore, subject to errors. Ensemble verification is necessary to quantify these errors, and to better understand the sources of predictive error and skill in particular modeling situations. The Ensemble Verification System (EVS) is a flexible, user-friendly, software tool that is designed to verify ensemble forecasts of continuous numeric variables, such as temperature, precipitation and streamflow. It can be applied to forecasts from any number of point locations and issued with any frequency and lead time. The EVS can also produce and verify aggregated forecasts, such as daily precipitation totals based on hourly forecasts or basin-averaged precipitation derived from point precipitation. This paper is separated into three parts. It begins with an overview of the EVS and the structure of the Graphical User Interface. The verification metrics available in the EVS are then described. These include metrics that verify the forecast probabilities (probabilistic metrics) and metrics that verify the ensemble mean forecast (deterministic metrics). Several new verification metrics are also presented. Finally, the EVS is illustrated with two examples from the National Weather Service, one focusing on ensemble forecasts of precipitation from the Ensemble Pre-Processor and one focusing on ensemble forecasts of streamflow from the Ensemble Streamflow Prediction system. The conclusions address future enhancements and applications of the EVS.

Email: james.d.brown@noaa.gov

2.4
Objective verification of weather forecast in Hong Kong

Yin Kong Leung

Hong Kong Observatory, China

One of the main responsibilities of the Hong Kong Observatory (HKO) is to provide weather forecasts for the public of Hong Kong in order to reduce loss of life and damage to property. The accuracy of weather forecasts is one of the key indicators of the performance of HKO.

In order to assess the accuracy of weather forecasts provided by HKO, an objective weather forecast verification scheme for Hong Kong was developed and computerized. The scheme was devised with a view to reflecting as closely as possible how the public would evaluate the accuracy of weather forecasts from their point of views. In the scheme, a score based on the accuracy of various kinds of weather elements: wind (speed and direction), state of sky, precipitation, temperature and visibility is given for each forecast issued. To take into account the significance of individual weather elements at different times of the year, different weightings are assigned to different weather elements for each month according to climatology and the relative importance of the elements for that month in the eyes of the public. This paper describes details of the methodology employed such as categorical forecasts, spatial and user-oriented verification. It also discusses the relative merit of the scheme.

Using this verification scheme, both the accuracy of weather forecasts issued by weather forecasters of HKO and the persistence forecast are assessed so as to measure the skill of forecasters in identifying weather changes. A persistence forecast is a forecast coded up by the computer based on the actual weather conditions, assuming that the weather for tomorrow will be the same as today.

HKO has commissioned since 1989 an independent consultant to conduct two public opinion surveys every year (April and October) to find out the subjective perception of the public on accuracy of weather forecasts issued by HKO. In this paper, a comparison is made on these subjective ratings of the accuracy of weather forecasts given by the opinion surveys with the HKO objective verification scores by using time-lagged stepwise regression.

Email: jykleung@hko.gov.hk

2.5
COSMO-DE-EPS: construction and verification of a limited-area ensemble prediction system on the convective scale
Marcus Paulat

German Weather Service (DWD), Offenbach, Germany

Aiming to improve the very short-range forecast of severe weather triggered by deep moist convection and interaction with small-scale topography, DWD has developed the convection-permitting limited-area model COSMO-DE. This model has a horizontal grid-spacing of 2.8 km, covers the area of Germany and is in operational mode since April 2007.

To properly take into account the limited predictability of processes on this small spatial scale, the DWD project COSMO-DE-EPS is developing an ensemble prediction system based on COSMO-DE. The project aims to quantify forecast uncertainty and to support the beneficial use of COSMO-DE forecasts in warning and decision-making processes.

Project activities comprise the generation, verification, statistical postprocessing, and visualization of ensemble forecasts. A pre-operational mode is foreseen to start in the beginning of 2010 with an approximate number of 20 members differing in physics parameters, lateral boundary conditions, and initial conditions. Operational implementation with about 40 members is envisaged to start in 2011.

The ensemble perturbation strategy focuses on model physics, lateral boundary conditions, and initial conditions. Model physics is perturbed by altering distinct parameters of the physical parameterization schemes either individually or in combination. Lateral boundary conditions are perturbed by nesting the COSMO-DE-EPS members into members of the COSMO-SREPS (ARPA-SIM, Bologna) which itself is a nested EPS with a grid-spacing of 10 km. Thus, the COSMO-DE-EPS represents the small-scale end of an ensemble chain which transfers uncertainty from large scales down to the convective scale. The development of initial condition perturbations is in its early stages. First tests perturb relevant parameters in the data assimilation scheme and also perturb initial fields based on differences between COSMO-SREPS members.

The quality of the current and further developed versions of COSMO-DE-EPS is assessed by PACprove, a probabilistic verification tool developed as part of the project. The tool is able to calculate numerous probabilistic verification scores like the Brier Skill Score, the Reliability Diagram, ROC Curve and Area and the Rank Histogram. Also a part of PACprove is the verification of single members with standard error scores like the FBI and others. This is done to check the quality of single members within the developing process because one can not be sure of realistic results while changing physics parameters.

Some results of 2m-temperature and accumulated precipitation of the current version of COSMO-DE-EPS will be presented. Therefore, two forecast experiments were performed for a few days with 20 members based on the combination of perturbed physics parameters and boundary conditions. Very typical results are that the EPS has too little spread and a skill getting worse for higher thresholds. It is clear, that a longer time period is needed and that we can expect a larger spread by involving the perturbation of initial conditions. But also it is clear that a useful calibration technique is needed.

Email: marcus.paulat@dwd.de

Session 3: Coping with Uncertainty in Verification Data

3.1

Introducing observational uncertainty in the scoring of a multi-model limited area model over the European area

Carlos Santos

Spanish Met Agency (AEMET), Madrid, Spain

Observations are usually treated as the true representation of the status of the atmosphere with no uncertainty associated to their values. This assumption is in general not true, since observations are affected by errors/ uncertainties. Different methodologies have been discussed in the literature to estimate these uncertainties. In the present talk, the uncertainty associated to the 24-hourly accumulated precipitation is inferred from the observed rainfall distribution within a model grid-box. The precipitation forecast is an areal quantity; therefore it is appropriate to use all the available observed information within a grid-box to construct a PDF of the observed rainfall. The observation uncertainty has been estimated from the European high-density network of rain-gauges.

A one-year period of short-range forecasts from a multi-model limited area system has been used to investigate the changes in model skill when observation uncertainty is introduced in the verification. The Spanish Meteorological Agency (AEMET) Short-Range Ensemble Prediction System (AEMET-SREPS) has been selected. In this system, five different limited area models (LAMs) are run twice daily (HIRLAM, HRM, MM5, LM and UM), using four global models as initial and boundary conditions (GFS, ECMWF, GME, UM global). This combination gives a 20 member multi-model multi-boundaries ensemble. The short-range EPS forecasts are assessed over the European area. The Brier Score and its component are calculated including the information provided by the observation PDF.

The scores obtained are expected to better describe the actual performance of the ensemble prediction system forecasts, which is usually degraded when the observations are considered without uncertainty. Comparisons of ensemble and observation uncertainties are beyond the scope of the talk.

Email: csantos@inm.es

3.2
Visualising verification results when no true reference is available
Otto Hyvärinen

Finnish Meteorological Institute, Helsinki, Finland

In verification, one or more data sets (usually forecasts) are compared with one data set (usually observations) that is deemed to be true or at least approximately true. But sometimes no such truth is available. Then one way to proceed is that all data sets are defined, one by one, as the truth and other data sets are compared with the selected data set using some verification measure. From these comparisons, a matrix of comparisons is easily constructed, where results from different truths can be compared and assessed. This matrix includes all information from the comparisons, but can be cumbersome to interpret. We will show how this comparison matrix can be visualized using Multidimensional Scaling (MDS). The way forward is to consider the verification measure as the distance, or more generally as the dissimilarity, between different data sets. In this study only verification measures and metrics for categorical variables were considered. Using the comparison matrix as the distance matrix a simpler two-dimensional mapping of results can be constructed. At the simplest, MDS is similar to Principal Component Analysis where only the first two components are plotted, but better results can be achieved using iterative methods. We concentrate on Sammon mapping which tries to retain the distances between variables in the mapping, so that distances between variables in a map should be proportional to the actual distances.

As a case study a comparison of different snow analyses is presented. Satellite-based snow cover analyses from NOAA, NASA and EUMETSAT and numerical weather prediction (NWP) model snow analyses from High Resolution Limited Area Model (HIRLAM) and ECMWF were compared using data from January to June 2006. Because no analyses were independent and available in situ measurements were already used in the NWP analyses, no independent ground truth was available but MDS enabled us to assess the consistence of models.

This is still a work in progress. Open questions include how to assess the uncertainty in mapping (e.g., by constructing the confidence intervals) and what verification measures that are useful as dissimilarities. In addition, only categorical verification measures have been considered in this study, but continuous measures could also be used.

Email: otto.hyvarinen@fmi.fi

3.3
The challenge of finding "good" reference data for verification

Theresa Gorgas

Department of Meteorology and Geophysics, University of Vienna, Austria

Whenever facing verification problems the question arises about what data should be defined as "truth" related to a specific problem. In the case of NWP model verification common approaches are to verify the model forecasts by using the model's analysis or, alternatively, some irregularly distributed or gridded observation data. Both types of reference data have their advantages as well as drawbacks, like model analyses are easily available but are not independent from the model's structure and observation data are probably more ''true'' at specific station locations but otherwise partially lack spatial representativeness. Regardless of which type of reference data is chosen for verification it has an impact on the results of the procedure.

At the workshop cases of basic NWP-model verification using different types of reference data shall be compared and discussed. A high-resolution GTS and Non-GTS dataset of surface observations for Central Europe which has been set up in the frame work of the programmes COPS and MAP D-PHASE for 2007 will be introduced as a possible source of data. Also gridded VERA (Vienna Enhanced Resolution Analysis) analysis fields based on different networks of observation data will be used for the comparison.

Email: theresa.gorgas@univie.ac.at

3.4

High resolution precipitation analysis and forecast validation over complex terrain using an inverse VERA approach

Benedikt Bica

Dept. of Meteorology, University of Vienna, Austria

Precipitation diagnosis and prognostics over mountainous terrain still poses a challenge due to the complex influence of topography. Both stratiform and convective precipitation usually show patterns that can hardly be resolved by an observation network. Similarly, the data quality control, analysis and modelling of precipitation fields is subject to limitations that arise from the steep gradients which might occur in precipitation fields over mountain regions.

In the proposed presentation, the VERA (Vienna Enhanced Resolution Analysis) method is applied to precipitation fields in order to assess its suitability for analysis, nowcasting and model validation purposes. VERA is based on the variational principle and further allows the inclusion of supplemental knowledge of typical patterns of meteorological parameters (fingerprint) in the analysis process.

In an inverse approach, the fingerprint weighting factors that are gained in the course of the analysis process can be used to evaluate local agreement of forecast models and observations in an innovative way. This is shown for a MM5 field of the August 2005 flooding event in western Austria, Switzerland and Bavaria.

The results prove that the VERA fingerprint technique may lead to a significant improvement of analysis quality and that it further facilitates an innovative approach for local model validation.

Email: Benedikt.Bica@univie.ac.at

3.5

Verification of multi-model ensemble forecasts using the TIGGE dataset

Wilson Laurence

Meteorological Research Division, Environment Canada

The Thorpex Interactive Grand Global Ensemble (TIGGE) is a project designed to facilitate studies designed to determine the benefits of combined multi-model ensemble forecasts compared to individual ensemble forecasts. The TIGGE database began in October 2006, and by Oct. 2007, 10 centers were contributing their global ensemble forecasts to the archive on a daily basis. In addition to standard upper air variables, the archive contains surface variables and information on tropical cyclones. This archive is an excellent source of data for the verification of ensembles and of combined ensembles.

Verification methodology for ensemble forecasts has always had to respond to some interesting issues, mostly relating to the interpretation of the forecasts as probabilities, and the need to compare a probability forecast with a deterministic observation. For example, what is a perfect ensemble forecast? How should a probability distribution function be estimated from the ensemble? The verification of multi-model ensembles adds a few other issues which must be considered, such as whether or not to debias the ensembles for verification, and, given the possibly irresistible temptation to do comparative verification of the ensembles which make up the TIGGE archive, how to ensure that comparative verification is fair. Some of the verification issues relating to ensemble forecasts will be illustrated by means of early studies based on the TIGGE archive.

The presentation will also include a description of a project the authors are undertaking to verify precipitation forecasts from the TIGGE archive, for different areas of the world, using a consistent methodology. Early results from that project will be shown.

Email: lawrence.wilson@ec.gc.ca
Email: laurence.wilson@sympatico.ca

3.6

Feature-oriented verification of wind speed forecasts

Matthew Pocernich

National Center for Atmospheric Research, Boulder (CO), USA

Incorporating wind-generated energy into a portfolio of energy sources is a challenge due largely to the difficulty in predicting wind speeds. Since utilities are responsible for continuously supplying electricity to consumers, any deficiency in energy production must be covered by power bought on the open market. Depending upon the circumstances, this can be costly. Ramping events occur when the wind speed increases or decreases sharply. As they unfold in real time, ramping events cause utilities a great deal of stress. Just as the human eye can readily detect spatial clusters, ramping events are easy to identify visually in a time series, but somewhat more difficult to define algorithmically.

In many respects, verifying ramping events is a one-dimensional example of many problems faced in verifying spatial forecasts. As temporal resolution increases, the chances of a forecast exactly matching the observed values decreases to zero. Just as there are features in spatial forecasts, ramping features must be identified in time series of predicted and observed values.

This talk briefly describes methods for identifying ramps in series of wind speed forecasts and observations. These events are described in terms of maximum change, total magnitude and duration. To describe the accuracy of the forecasts two approaches are taken. For an observed ramp, the magnitude of the error associated with the forecast for the same period is described. Secondly, to quantify errors due to timing, within a window, optimal shifts in forecast times are calculated. Finally, to make the results meaningful for users, errors can be expressed in terms of power generated instead of expressing them in terms of wind speeds,. Since costs differ for over or under forecasted events, errors are partitioned separately to differentiate both types of costs.

Email: pocernic@ucar.edu

Session 4: Properties of Verification Methods

4.1
Probability forecasts with observation error: is the Brier score proper?
Ian Jolliffe

University of Exeter, UK

For probability forecasts of binary events the Brier score is a well-known and widely reported verification measure. One of its virtues is that it is strictly proper and hence cannot be hedged. If observations of the forecast event are prone to error, does this property still hold? It is shown that the answer can be Yes or No depending on definitions. Simple illustrative examples are given.

Email: ian@sandloch.fsnet.co.uk

4.2
Is ETSS really equitable?
Lovro Kalin

Meteorological and Hydrological Service of Croatia, Zagreb, Croatia

This paper investigates the defectiveness of equitable skill scores for multicategorical table. History of methods to evaluate forecasts through multicategory tables is not so long (Vernon, 1953) as history of common verification scores (Murphy, 1996). Methods can be divided into three groups: application of scores designed for standard 2x2 tables, LEPS score and Gandin-Murphy ETSS score. Gerrity (1992) showed that Gandin-Murphy score for multicategorical table can be calculated as mean of N-1 values of Pierce score. Since Pierce score has certain deficiencies they are inherited in the ETSS also. On some examples (Juras and Pasaric, 2006) it can be showed that overforecasting of the extreme categories increases the value of the score compared to the same unbiased forecasts.

Email: kalin@cirus.dhz.hr

Session 5: Verification of Weather Warnings

5.1

The Verification of Weather Warnings: Did the boy cry wolf or was it just a sheep?

D.B. Stephenson and I.T. Jolliffe

School of Engineering, Computer Science and Mathematics, University of Exeter, UK

This talk will review the types and formats of deterministic warnings currently issued by weather services such as the Met Office and how such warnings are typically evaluated. The talk will discuss some of the main issues in the verification of warnings such as how observed events can be defined, the timing of events, the triviality of commonly used scores for rare events, and the missing-d problem (number of correct rejections unavailable). Recommendations and ideas for further research will be presented.

Email: d.b.stephenson@exeter.ac.uk

5.2
Approaches to process and event oriented verification of warnings

Martin Goeber

Deutscher Wetterdienst (DWD), Offenbach, Germany

In most countries the continuous and detailed verification of warnings is a relatively new task compared to model verification. This is also reflected in the state of available information on operational practises. This presentation will present some of the particular issues involved based on information collected internationally and on specific experiences from the German national weather service DWD.
There is a surprisingly great variety of approaches, which can be partly explained by the user driven nature of warning verification, in contrast to the more scientifically oriented model verification. Compared to model verification, there are 2 additional free parameters (lead time and duration), which have to be decided upon by the forecaster or be fixed by process management and which also have a big influence on the quality of the warnings.
Warnings are mostly for areas which leads to observational undersampling for most variables. Thus some ‘’soft touch’’ is required because of the overestimate of false alarms and the underestimate of misses. Data quality is specifically important for warning verification, since a false report of an event would wrongly lead to a miss, i.e. the most serious warning error. Ultimately, a probabilistic, multivariate analysis of events is needed.
The largest difference to model verification occurs in the matching of warnings and observations. On a temporal scale the matching is done in various ways: either on an hourly scale, or from the moment an observation exceeds a threshold or on an even broader scale as an ‘’extreme event’’. Thus the number of ‘’extreme events’’ per year can vary between a few dozen to hundreds of thousands. Furthermore the lead time of the warning has an user dependent influence on the definition of a ‘’hit’’ or ‘’miss’’. On a spatial scale the verification is sometimes down ‘’vaguely by hand’’, or as the ‘’worst thing in an area’’, thus leading to a strong influence of the area size on the warning verification.
Standard contingency table based measures are frequently used to summarise verification results, yet also user defined metrics occur. The detailed analysis of case studies remains popular. The setting of performance targets for warning verification remains an unsolved problem because of the high interannual variability of extreme events. There is a strong influence of the change of the observational networks on warning verification since ‘’if you detect more, it’s easier to forecast’’.
User based assessments play an important role, especially during the process of setting up warning procedures and during the subsequent fine tuning of the process.

A ‘’process oriented’’ approach based on hourly verification and an event based approach has been used for the verification of county warnings at DWD for the last 6 years. Those show a high probability to detect extreme events, yet accompanied by a high number of false alarms. The high false alarm ratios have been reduced substantially through a higher temporal and spatial precision of the warnings.

Email: martin.goeber@dwd.de

5.3
The Challenge of Verifying Severe Weather Warnings
Michael Sharpe

Met Office, Exeter, UK

The UK Met Office issues severe weather warnings for every county and unitary authority in the United Kingdom. The verification of these warnings is a conceptually simple process. However in reality the way operational weather warnings are issued ensures that the verification process is fraught with complex challenges. These complexities include; disparities in the size of each warning area, differences in the length of each warning, the complete freedom to issue warnings at any time and the non-persistence of above event-threshold conditions during the warning period. Each of these issues is discussed and solutions are explored together with the use of fuzzy verification ideas. These ideas have been introduced into a new verification system from which some preliminary results are given. A new measure of the quality of each warning is proposed.

Email: michael.sharpe@metoffice.gov.uk

5.4

A critical look at the verification of Met Office ''Flash'' Warnings

Clive Wilson

Met Office, Exeter, UK

Met Office operational forecasters aim to issue ''flash'' warnings at least 2 hours ahead of a warning event for the National Severe Weather Warning Service (NSWWS). The warnings may be for a number of parameters including severe gales over land, heavy rain, heavy snow, blizzards or drifting, freezing rain or widespread icy roads and fog. The warnings are issued for local authority areas (county, district, borough or unitary - hereafter ''counties'') which vary greatly in extent. Formal warnings are issued when it is expected that the threshold criteria is exceeded at, at least one of, the specified reference sites, which are chosen to represent 80% of the population in the area. To issue a warning the guidelines state that confidence should be 80% or higher. A number of different approaches to verify these warnings have been used and a target for accuracy has recently been set. Concentrating on warnings for severe gales over land and heavy rain, I show how different truth data influences the results, as well as the variation in county size. The widely-used threat score is dependent on the base rate. This dependence is shown to be much stronger for relatively rare events such as heavy rain and severe gales which makes it unsuitable as a summary measure of how good the forecasting process is or how it may, or may not, be improving. The threat score is also influenced by the bias so that it may be hedged. A performance plot of hit rates against false alarm ratio (or 1-success) helps to compare scores using different truth types and their interpretation. Model forecasts for the same warning criteria have also been verified against radar estimates and are contrasted and compared with the human forecasters' performance. Forecasts from both the 12km and 4km models are verified and interesting differences in verification are found.

Email: clive.wilson@metoffice.gov.uk

5.5
Verifying extreme rainfall alerts for surface water flooding
Marion Mittermaier and Nigel Roberts

Met Office, Exeter, UK

During the summer of 2008 a new pluvial (surface water) flood forecasting service was launched at the Met Office in conjunction with England's Environment Agency. This is in recognition that urban areas are particularly vulnerable to heavy rainfall due to the altered land surface characteristics of catchments. The service issues three levels of warnings: advisories (a day ahead), early and imminent alerts. The forecasts are based on first guess probabilities from a variety of model and radar nowcasting sources. Three thresholds are considered: 30 mm/h, 40 mm/3h and 50 mm/6h. This paper will present the results from the analysis of the warnings and events that occurred during the pilot stage of this service. Two contingency table-based methods will be contrasted and compared: (1) which takes the macroscopic ''event'' approach and (2) an hour-by-hour time series approach. The results show that the choice of method preconditions the outcome of the analysis, and that potentially contradictory results may come to light, if not interpreted correctly.

Email: marion.mittermaier@metoffice.gov.uk

Session 6: Spatial and Scale-sensitive Methods

6.1
MODE-3D: Incorporation of the time dimension
Barbara Brown

National Centre of Atmospheric Research, Boulder, CO, USA

The Method for Object-based Diagnostic Evaluation (MODE) has been extended to the evaluation three-dimensional rainfall systems, with time as the third dimension. Precipitation objects are identified in three-dimensions (x, y, and t), where x and y are the two horizontal spatial dimensions and t is time, using a convolution/thresholding process. Relevant temporal and spatial attributes – including object centroid location, volume, velocity, and lifetime – are identified and assigned interest values as a component of a fuzzy logic approach for matching objects in space and time. New interest maps are applied in this extension of MODE.

The advantage of considering 3-D objects is that each forecast or observed dataset can be viewed in terms of a relatively small number of rain systems. Furthermore, the evolution of the predicted systems can be verified using basic geometric properties of the objects. For instance, with time pointing vertically, the angle with respect to the horizontal plane indicates the speed and direction of propagation of the rain system. The temporal centroid, beginning and ending times all pertain to the timing of the system. Using 3-D object attributes, we devise a metric that accounts for spatial, temporal and propagation errors and results in a relative measure of forecast quality that can be used to compare forecasts from different models, or forecasts from a given model on different days.

The 3-D objects identified in numerical forecasts and observations from the IHOP period and the 2005 NSSL/SPC (National Severe Storms Laboratory/ Storm Prediction Center) Spring Experiment are used to demonstrate MODE-3D.

Email: bgb@ucar.edu

6.2
Spatial Forecast Veri f cation\: The Image Warp
Eric Gilleland

National Centre of Atmospheric Research, Boulder, CO, USA

The image warp is a spatial statistical technique for deforming one image to better fit another image. In the context of forecast verification, this technique can be used to deform a forecast field to better match an observed field. Subsequent metrics such as average amount and direction of movement, bending energy, and percent reduction in traditional verification scores can be computed. Therefore, the technique can give more useful information about forecast performance than traditional verification techniques alone. For example, information about spatial displacement and spatial extent errors can be directly obtained. More detailed information, such as false alarms and misses can also potentially be obtained, but to do so automatically would be challenging.

The technique is performed on the test cases from the Spatial Forecast Verification Inter-Comparison Project (ICP), and a method for ranking multiple forecasts based on the results is demonstrated on the perturbed real cases.

Email: ericg@ucar.edu

6.3
A scale-based distortion metric for mesoscale weather verification
Chermelle Engel

University of Melbourne, Australia

Verification of high-resolution mesoscale weather forecasts has become increasingly important in recent years due to increases in model resolution and modelling of mesoscale physical processes. Traditional verification scores such as mean square error or variance have limited use in terms of assessing the value of these types of forecasts, and can actually produce misleading results. In order to get around this problem, new verification measures are currently being developed.

One avenue of verification development has been the use of algorithms based on distortion- or optical-flow. While these methodologies show promise when applied to test cases such as those from the Spatial Forecast Verification Intercomparison Project, they may encounter less favourable results when applied to meteorological fields with field motion/placement error dependent upon scale.

This talk will address a new type of verification measure combining scale-decomposition and a distortion- or optical-flow based technique to characterize scale-dependent distortion error. The capability of this technique to perform will be assessed using both simple and more complex idealized examples. Comparisons with existing methodologies will be discussed along with links to data assimilation.

Email: c.engel@bom.gov.au

6.4

On evaluating the applicability of CRA over small verification domains

Stefano Mariani

Institute for Environmental Protection and Research (ISPRA), Rome, Italy

When verifying numerical weather prediction models, object-oriented techniques provide a useful way to quantify and qualify (in terms of error sources) the forecast error. They give quantitative support to the standard “eyeball” verification, since they measure the spatial displacements perceived in the numerical forecasts. Moreover, results obtained by applying these methods are not affected by the double penalty effect, as it can happen when using traditional categorical skill scores.

This is the case of the contiguous rain area (CRA) analysis applied to the verification of quantitative precipitation forecasts (QPFs). The method is based on pattern matching of two contiguous areas defined as the observed and forecast precipitation areas delimited by a pre-selected isohyet. The pattern matching is then obtained by translating in the x and y directions the forecast features over the observed ones, until a best-fit criterion (e.g., the correlation maximization or the mean square error minimization) is satisfied.

Recent studies have evidenced that, especially over small areas, CRA may lead to unphysical and unreliable results, because the pattern matching may be obtained by shifting the forecast field out of the domain. In small verification areas, there is indeed the possibility of not completely enclosing the observed feature of interest to be compared with the modelled feature. However, by performing quality check tests or by imposing complex matching procedure, for instance based on the correlation maximization conditioned to the mean square error minimization, suspicious displacements may be detected and reliable results may be localized.

Where two or more observed rainfall peaks are present within the small verification area, CRA results might be sensitive to the displacement of the main peak with respect to the displacement of the secondary peaks. Thus, to evaluate the impact of the localization errors associated to the secondary peaks, it is necessary to apply the CRA method over both the entire verification domain and the sub-domains centred over each observed peak. By comparing and contrasting such results is then possible to correctly detect the shift associated to the peaks presents within the verification area and to assess the impact on the overall forecast evaluation of each of them.

By analysing intense precipitation events occurred over the Alpine area and the Cyprus Island, the present study investigates the sensitivity and the applicability of the CRA methodology for QPF verification over small domains. On the purpose, precipitation fields forecast by limited area models and observed by ground-based instruments are employed in the verification study.

Email: stefano.mariani@isprambiente.it

6.5

Identifying skillful spatial scales using the Fraction skill Score

Marion Mittermaier

Met Office, Exeter, UK

The Fractions Skill Score (FSS) was one of the measures which formed part of the ‘’Intercomparison of Spatial Forecast Verification Methods’’ project. The FSS was used to assess a common data set which consisted of real and perturbed WRF forecasts, as well as fake geometric cases, all based on the NCEP g240 grid (which translates to approximately 4 km resolution) over the contiguous USA.

The fake geometric cases showed that the FSS can provide a truthful assessment of displacement errors and forecast skill. The study of the perturbed cases showed that the FSS and the spatial scale of a skillful forecast are relatively insensitive to small shifts or timing errors. Changes in perturbed rainfall totals have greater impact, where subtle differences introduced through near-threshold misses seem to lead to large changes in FSS magnitude.

The outcome of the study also shows that domain size does matter, especially when the proportion of the domain that is ‘’wet’’ is small (often referred to as the wet-area ratio).

Whilst frequency thresholds are potentially useful (by removing the bias) the benefits are greatly reduced when the domain is large, since the cumulative distribution is dominated by zeros. Overall the absolute value of the FSS is perhaps less useful than the scale where an acceptable level of skill is reached.

Email: marion.mittermaier@metoffice.gov.uk

6.6

New developments of the Intensity-Scale verification technique within the Spatial Verification Methods Inter-Comparison

Barbara Casati

Ouranos – Consortium of Research in Regional Climate and Adaptation to Climate Change, Montreal, Canada

The Intensity-Scale technique introduced by Casati et al (2004) is revisited and improved. Recalibration is no longer performed, and the Intensity-Scale skill score for biased forecasts is evaluated. The energy and its percentages are introduced in order to assess the bias on different scales and to characterize the overall scale structure of the precipitation fields. Aggregation of the Intensity-Scale statistics for multiple cases is performed, and confidence intervals are provided by bootstrapping. Four different approaches for addressing the dyadic domain constraints are illustrated and critically compared.

The Intensity-Scale verification is applied to the Spatial Verification Methods Inter-Comparison Project case study data set. The geometric and synthetically perturbed cases show that the Intensity-Scale verification statistics are sensitive to displacement and bias errors. The Intensity-Scale skill score assesses the skill for different precipitation intensities and on different spatial scales, separately. The spatial scales of the error are attributed to both the size of the features and their displacement. The energy percentages allow one to objectively analyze the scale structure of the fields and to understand the intensity-scale relationship. Aggregated statistics for the Spring 2005 case studies show no significant differences among model skills, however WRF4 NCEP over-forecasts to a greater extent than WRF2 and WRF4 NCAR. Tiling provides the most robust approach to address the dyadic domain constraints, since it smooths the effects of the discrete wavelet support and does not alter the original precipitation fields.

Email: b.casati@gmail.com
Email: casati.barbara@ouranos.ca

6.7
Feature-specif c verif cation of ensemble forecasts
Elizabeth Ebert

Bureau of Meteorology, Melbourne, Australia

For high impact events, forecasters want to see evidence that ensemble prediction systems (EPSs) provide information that is as useful as deterministic forecasts and poor man's ensembles. In a recent paper by Novak et al. (Wea. Forecasting, Dec. 2008) forecasters in the United States indicated a desire to see feature-specific verification approaches such as CRA or MODE applied to ensemble forecasts. Several strategies could be considered, including verifying features in the individual ensemble members, verifying features in the ensemble mean forecast, verifying consensus features derived from the ensemble of attributes (location, size, shape, etc.), and verifying features in the probability map.

This presentation will show results from a study that used the Contiguous Rain Area (CRA) method to verify forecasts of 24 h rainfall accumulation in individual members of the ECMWF EPS over the period April 2008-March 2009. The focus was on heavy rain events, with maxima of at least 20 mm d^{-1} and 50 mm d^{-1} in the forecast and/or observations. The ensemble distribution of attributes for each feature was evaluated to explore relationships between the occurrence and properties of forecast and observed events. A strong relationship was found between the number of members predicting an event and the frequency of the event occurring, as well the success of the ensemble to envelope the observed properties of the feature. Examination of the relative operating characteristic (ROC) suggests greater ensemble skill for heavier rain events. The ensemble spread was less than the error of the ensemble mean for all attributes (location, rain area, mean rainfall, maximum rainfall) at all forecast projections, which may mean that the heavy rain features were too similar in the EPS forecasts. In spite of this apparent under-dispersion there was a strong relationship between the spread of the attributes and the skill of the ensemble mean, suggesting that the spread can be used to predict the uncertainty. Based on these early results, and others of Gallus (Wea. Forecasting, in press), object-based verification looks promising for evaluation of heavy rain events and other weather features.

Email: e.ebert@bom.gov.au

Session 7: Spatial and Scale-sensitive methods: High Resolution Models

7.1

Some experiences during verification of precipitation forecasts using ‘’Fuzzy” techniques

Ulrich Damrath

DWD, Germany

Some results concerning verification of precipitation forecasts over Germany with different models (GME, COSMO-EU and COSMO-DE) using the “Fuzzy”-technique (Ebert: Meteorological Applications, vol. 15(2008), issue 1, pp. 51-64) are presented. Gridded observations (in the grid of COSMO-DE) are based on the radar network over Germany with hourly sums of precipitation. Favourite scores are those that can be got from the upscaling method, the fractions skill score and a modification of Casatis energy squared score. Especially during summer time advantages of modells with high horizontal and vertical resolution can be seen clearly. This is mainly true for large scales and low precipitation values. But COSMO-DE was also able to produce more realistic precipitation forecast than GME and COSMO-EU during summer for precipitation values up to 5mm/12h. During winter time the advantages of high resolution models are also visible and the level of scores is higher than during summer.

Email: Ulrich.Damrath@dwd.de

7.2

Deterministic and Fuzzy Verification of the Cloudiness of High Resolution Operational Models

Marielle Amodei

Meteo France, Toulouse, France

The brightness temperature (BT) observed by the Infrared channel of SEVIRI, present in Meteosat 9 is used to verify the forecast qualities of 2 high resolution models Aladin (horizontal mesh 10 km) and Arome (2.5 km) operational in Meteo-France. The observed temperatures are directly related to the atmosphere cloudiness and their forecasted counterparts are obtained through the radiative transfer model RTTOV. The temporal period used for the comparison is summer and autumn 2008.

2x2 tables of contingences are built for different thresholds (defining the events) covering the range data and used to compute deterministic scores. A fuzzy approach is performed by transforming the deterministic forecast in frequencies of events in a neighbourhood around the observation point. Brier skill scores against the persistence forecast are obtained by comparing theses frequencies either to the local yes or no observation or to the observed frequency in the neighbourhood (Amodei and Stein 2009).

The stratification of the results in function of BT allows to document the relative merits of the forecasts all along the troposphere. Thus, it is shown that both models under-estimate the real BT by lack of cloudiness and especially the Arome model for large scale perturbations. Moreover, the high-tropospheric clouds are quasi-absent in the Aladin forecasts and Arome bias is better for this category in convective situations but its clouds are often displaced leading to poor deterministic scores. This drawback is corrected by the fuzzy approach and its probabilistic scores beats the Aladin counterparts. This conclusion is in complete accordance with the verification of the precipitation performed over the same period.

Email: marielle.amodei@meteo.fr

7.3
Valuing information from high resolution forecasts
Kees Kok

KNMI, Netherlands

High resolution forecasting is particularly aiming at predicting the smaller scale atmospheric phenomena. However, traditional verification scores fail to recognize the added value of high resolution forecasts, mainly due to the double penalty. Nevertheless, it is generally accepted from subjective verification that these models have predictive potential for small scale weather phenomena. In this presentation a probabilistic approach is suggested which offers the possibility not only to objectively assess the skill of small scale information but also to quantify the additional value of high resolution models over lower resolution ones. A Model Output Statistics (MOS) technique is used incorporating concepts from fuzzy verification. The MOS approach objectively weighs different forecast quality measures and as such can be regarded an essential extension of fuzzy methods. A few preliminary experiments on real atmospheric data will be discussed.

Email: kokc@knmi.nl

7.4
QPF verification of limited-area ensemble systems during the the MAP D-PHASE OP

Chiara Marsigli

ARPA-SIMC, Bologna, Italy

Ensemble predictions with mesocale limited-area models are nowadays performed in many meteorological centres. The use of these ensembles permit to provide Probabilistic QPF with high spatial detail. Hence, spatial verification methodologies should be adopted in the computation of the standard probabilistic indices used for ensemble verification.

During the MAP D-PHASE Operations Period (OP), six different limited-area ensemble systems were providing data to a common database: INM- (now AEMET-) SREPS, COSMO-LEPS, COSMO-SREPS, LAM-EPS Austria, PEPS and Micro-PEPS. These ensembles are characterised by different perturbation methodologies, different horizontal resolutions and by the use of different mesoscale models. Though they are not exhaustive of the full variety of mesoscale ensemble systems, it is extremely interesting to evaluate and compare their performances, in order to provide a first assessment of the state-of-the-art of ensemble forecast with limited-area models.

In this work, an objective verification of the PQPF issued by these systems is performed over the D-PHASE area, using high resolution observations covering an Alpine area, for the two seasons included in the OP, (summer and autumn 2007).

Due to the focus on precipitation and to the different resolution of the ensembles, a spatial verification method is adopted. DIST is a very simple method, consisting in a comparison of the predicted and observed precipitation distributions in terms of some distribution parameters (mean, median, maximum and percentiles), within boxes of pre-defined size. The box size is chosen so that enough grid points and station points are included in each box. Furthermore, since several forecast systems are compared, the size of the box should be large enough to contain a sufficient number of points of the lower resolution model.

The analysis addresses how the different ensemble systems are able to predict precipitation at high resolution over the Alps, enlightening the different characteristics of each system in forecasting, e.g., average or extreme values. Furthermore, the different amount and quality of the spread of the ensembles is evaluated.

Email: cmarsigli@arpa.emr.it

7.5

Verification of precipitation forecasts of the MAP D-PHASE data set with fuzzy methods

Tanja Weusthoff (1), Francis Schubiger (1), and Felix Ament (2)
(1) MeteoSwiss, Switzerland
(2) University of Hamburg, Germany

In the scope of the forecast demonstration project D-PHASE a real-time end-to-end forecasting system for heavy precipitation and subsequent flood events in the Alpine regions was set up. Based on probabilistic and deterministic atmospheric and hydrological models as well as on nowcasting tools, warnings were issued for specified river catchments. The forecast data for the full D-PHASE observation period (DOP, June - November 2007) was stored providing a huge amount of data which is well suited to investigate the use of atmospheric/hydrological models for flood forecasting in mountainous regions. Various methods are applied to get an objective verification of the model forecasts.

The precipitation forecasts of the Swiss COSMO models are evaluated with reference to Swiss radar measurements by means of Fuzzy verification methods (Ebert, 2008). Based on previously performed idealized case studies, two different methods were chosen for the identification of an event: (i) the mean value over the respective window exceeding a given threshold (Upscaling) and (ii) the fraction of grid points exceeding the threshold (Fractions Skill Score), respectively. The verification is done on different spatial scales by varying the window sizes and for various thresholds defining an event. The scores are determined for 3h and 24h rain accumulations and are then aggregated on different time scales like a month or the whole DOP of 6 months. Several sensitivity studies are performed: The robustness of the results is determined by means of a bootstrapping procedure. The sensitivity of the scores is evaluated by considering only cases where the rain amount exceeds predefined thresholds at a minimum number of grid points. In addition, the dependency of the models performance on the weather type, classified in a subjective way mainly based on the 500 hPa wind field over the Alpine area, is investigated.

The main focus of the here presented verification is on the comparison of the two operational COSMO models at MeteoSwiss with a resolution of 2.2 and 7 km. By varying the spatial scales the scores are calculated for comparable scales allowing for a direct comparison of the model's skill. In the near future, the Fuzzy verification technique will be applied operationally at MeteoSwiss. Both operational COSMO models will be evaluated on a seasonal basis and with respect to the weather type.

Reference:
Ebert, E. 2008: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. Meteorol. Appl. 15, 51-64.
Email: francis.schubiger@meteoswiss.ch

7.6

Verification of high resolution precipitation forecast by using SAL method

Sami Niemelä

Finnish Meteorological Institute, Helsinki, Finland

Operational limited area Numerical Weather Prediction (NWP) has typically been carried out at the so-called synoptic scale, with the model grid size in the 10-20 km range. At these scales, the hydrostatic approximation is still valid and convective phenomena are parameterized instead of being explicitly simulated. Due to the need to forecast smaller-scale phenomena (e.g. strong precipitation, flash floods) there is a push to develop NWP models for these spatial and temporal scales. Since 2006, Finnish Meteorological Institute (FMI) has addressed this need by running a non-hydrostatic, convective permitting (2.5 km horizontal resolution) NWP model AROME. The model is run twice a day (00 and 12 UTC) in a dynamical adaptation mode, where initial and boundary conditions are taken from the coarser resolution synoptic scale model (HIRLAM, 7.5km). The obvious question is, do we gain any added value by using km-scale NWP models.

The purpose of this study evaluate the precipitation forecast of AROME in different regimes from June 2008 to November 2008. The main emphasize is to evaluate AROME's ability to predict the structure, intensity and location of precipitation events by using SAL-methodolgy (Structure-Amplitude-Location). Since SAL-verification is so-called grid-to-grid method, it requires high resolution gridded observation data. In this case, the only available data set is from Finnish radar network. However, radars are measuring reflectivity (i.e. precipitation intensity) several hundreds of meters above surface, so there is a clear geometrical mismatch between the modelled and radar-observed precipitation. This mismatch is addressed by using Radar Simulation Model (RSM), which simulates an individual radar measurement in model atmosphere. This way we avoid the difficult problem of linking the measured reflectivity to the precipitation at the surface. Consequently, the model evaluation by SAL-method is based on model-produced and observed radar reflectivities [dBZ].

In addition to SAL-method, the traditional verification methods are exploited to asses the possible added value of the km-scale AROME model compared to FMI's synoptic scale HIRLAM model (7.5 km). Here the focus is to use high resolution Helsinki Testbed (http://testbed.fmi.fi) precipitation observations measured by WXT weather transmitters (warm rain only).

Email: sami.niemela@fmi.fi

7.7

Towards evaluating timing errors of quantitative precipitation forecasts with the feature based technique SAL

Matthias Zimmer

Institute for Atmospheric Physics, University Mainz, Germany

The assessment of the quality of quantitative precipitation forecasts (QPF) with short accumulation times (~1 hour or less) is one of the challenging topics in the field of forecast verification. Precipitation fields are often characterized by a high variability and complexity in time and space. On hourly time scales, QPFs are typically characterized by errors both in the spatial representation as well as in the temporal behaviour (e.g. propagation speed of fronts, onset of convection, etc). In this study, we present an extension of the three component feature-based verification technique SAL for the additional evaluation of timing errors. Hourly QPFs from COSMO-DE, the operational numerical weather prediction model from the German weather service which is operated without parameterization of deep convection, have been analyzed to reveal errors in time and space of QPFs in Germany during summer 2007.

The feature-based verification technique SAL, which contains three independent spatial components that measure the quality of the structure (S), amplitude (A) and location (L) of a QPF is complemented by a forth component that determines the timing error of QPFs with short accumulation times. This will be done by a fuzzy assessment of temporal forecast errors. The timing error is determined as the time shift for which L attains its minimum value. This extended version of SAL, with a temporal component (T) for the timing error is referred to as SALT.

The observational data set used for a first application of SALT has an hourly time resolution and is based upon a disaggregating technique, which combines the high temporal resolution of radar data with the fairly high accuracy of the amount of precipitation obtained from rain gauge measurements. All precipitation fields have been transformed onto the same grid with a horizontal grid spacing of 7 km.

Evaluations of hourly QPFs with SALT show that the extension of SAL leads to valuable additional information about the performance of high-resolution NWP models for short QPF accumulation times. Additionally, weather-type based investigations will be performed to reveal differences in the timing error behaviour for different synoptic situations, e.g. the transit of frontal rain showers or the onset of air mass convection.

Email: zimmernm@uni-mainz.de

Session 8: Seasonal and Climate Forecast Verification

8.1

Towards Standardized Verification of Seasonal Climate Forecasts

Simon Mason

The Earth Institute of Columbia University, New York, USA

Under the auspices of the WMO Commission for Climatology (CCl), a set of recommended verification procedures for operational probabilistic seasonal forecasts, including those from the Regional Climate Outlook Forums (RCOFs), National Meteorological and Hydrological Services (NMHSs) and other forecasting centres, has been drafted. The recommendations are meant to complement the WMO's Commission for Basic Systems (CBS) Standardized Verification System for Long-Range Forecasts (SVSLRF), which provides guidlines for the verification of Global Producing Centre (GPC) products. The procedures defined under the CCl are targeted partly at end-users of the forecasts, and partly at the forecasters themselves. Two related underlying principles in making the CCl recommendations were: (a) to focus on verification procedures that measure specific forecast attributes, and to avoid those procedures (even if they are widely used) that measure multiple attributes; and (b) to recommend procedures that have a reasonably simple interpretation. In this presentation an overview of the CCl recommendations will be provided.

Email: simon@iri.columbia.edu

8.2

Extreme Value Theory to analyze, validate and improve extreme climate projections

Barbara Casati (1), Louis Lefaivre (2)

(1) Ouranos – Consortium of Research in Regional Climate and Adaptation to Climate Change, Montreal, Canada
(2) Canadian Meteorological Service, Dorval, Canada

Extreme weather events can cause large damages and losses, and have high societal and economic impacts. Climate model integrations predict increases in both frequency and intensity of extreme events under enhanced greenhouse conditions. Better understanding of the capabilities of climate models in representing the present climate extremes, together with the analysis of the future climate projections, can help to forewarn society of future high-impact events, and possibly develop better adaptation strategies.

Extreme Value Theory (EVT) provides a well established and robust framework to analyze the behaviour of extreme events for the present climate and future projections. In this study, stationary Generalized Extreme Value (GEV) distributions are used to analyze and validate observed and modelled extremes in the present climate, while a non-stationary GEV fit is used to analyze the trends of the extreme distributions in the context of a changing climate. GEV distributions of annual extremes for 24-hours precipitation accumulations and daily minimum and maximum temperatures are analyzed for 12 climatologically homogeneous regions over North America. The analysis is performed on an ensemble of Canadian Regional Climate Model (CRCM) simulations, under a SRES A2 emission scenario. Significant positive trends for the location of the CRCM distributions are found in most regions, indicating an expected increase in extreme value intensities, whereas the scale (variability) and shape (tail values) of the extreme distributions seem not to vary significantly.

Climate model projections can be affected by biases and representativeness errors, due to the scale mismatch between model resolution versus localized extreme phenomena. The representativeness issue is particularly relevant for extreme precipitation events, which are often associated with small-scale features (e.g. convective precipitation). In this work, biases and representativeness mismatch are quantified by comparing the GEV distributions of the CRCM extremes to those obtained from station measurements, in the present climate (1961-2000). As expected, the greatest discrepancies are associated with the location and scale of precipitation distributions (extreme precipitation intensities and their variability are underestimated). These occur in a systematic fashion and can lead to the definition of a downscaling relation. Temperature extremes, on the other hand, exhibit some warm or cold biases, which vary by region. More realistic future extreme projections are obtained by applying the trends of the CRCM distributions to the observed GEV distributions.

Email: b.casati@gmail.com

8.3

Can you really trust long-range weather predictions? Confessions of a rogue forecaster

Pascal Mailier

Royal Meteorological Institute of Belgium, Bruxelles, Belgium

In November 2005, the onset of a cold spell in Europe triggered a considerable rise in UK wholesale gas prices. The main factor which had made energy markets particularly sensitive was the expectation by the Met Office that a negative phase of the North Atlantic Oscillation (NAO) would favour colder-than-usual conditions in NW Europe over the winter.

First, the NAO and its relationship with the European winter weather will be explained in simple terms. Then, the Met Office statistical forecast model of the winter NAO will be introduced. The Met Office rates the skill of its winter NAO forecasts as “reasonable”, with the sign of the winter NAO (negative=cold and dry, positive=mild and wet) being predicted correctly two times out of three. However, it will be shown that this model is in fact not more reliable than a simple ‘toy’ model with no useful predictive skill. This case demonstrates how tricky it can be to gauge the quality of weather forecasts, more especially when it comes to assessing their true usefulness.

Email: pascal.mailier@oma.be

Session 10: New Ideas in Verification

10.1
Verification measures for rare-event forecasts
Christopher Ferro

School of Engineering, Computing and Mathematics, University of Exeter, UK

Verifying forecasts of rare events is challenging, in part because traditional performance measures degenerate to trivial values as event rarity increases. The extreme dependency score was proposed recently as a non-vanishing measure for the quality of deterministic forecasts of rare, binary events. This measure, however, has some undesirable properties, including dependence on the base rate. Two modified versions of the extreme dependency score are proposed here, which overcome all of its shortcomings. The new measures are non-vanishing, base-rate independent, equitable, and have regular isopleths that correspond to symmetric and asymmetric relative operating characteristic curves.

Email: c.a.t.ferro@exeter.ac.uk

10.2

Bias in trials comparing paired continuous tests can cause researchers to choose the wrong screening modality

Deborah Glueck

Colorado School of Public Health, University of Colorado, Denver, USA

Background: To compare the diagnostic accuracy of two continuous screening tests, a common approach is to test the difference between the areas under the receiver operating characteristic (ROC) curves. After study participants are screened with both screening tests, the disease status is determined as accurately as possible, either by an invasive, sensitive and specific secondary test, or by a less invasive, but less sensitive approach. For most participants, disease status is approximated through the less sensitive approach. The invasive test must be limited to the fraction of the participants whose results on either or both screening tests exceed a threshold of suspicion, or who develop signs and symptoms of the disease after the initial screening tests. The limitations of this study design lead to a bias in the ROC curves we call paired screening trial bias. This bias reflects the synergistic effects of inappropriate reference standard bias, differential verification bias, and partial verification bias. The absence of a gold reference standard leads to inappropriate reference standard bias. When different reference standards are used to ascertain disease status, it creates differential verification bias. When only suspicious screening test scores trigger a sensitive and specific secondary test, the result is a form of partial verification bias.

Methods: For paired screening tests with bivariate normally distributed scores, we give formulae and programs to quantify the effect of paired screening trial bias on a paired comparison of area under the curves. We fix the prevalence of disease, and the chance a diseased subject manifests signs and symptoms. We derive the formulas for true sensitivity and specificity, and those for the sensitivity and specificity observed by the study investigator.

Results: The observed area under the ROC curves is quite different from the true area under the ROC curves. The typical direction of the bias is a strong inflation in sensitivity, paired with a concomitant slight deflation of specificity.

Conclusion: In paired trials of screening tests, when area under the ROC curve is used as the metric, bias may lead researchers to make the wrong decision as to which screening test is better.

10.3

Polychoric correlation coeff cient in forecast veri f cation

Zoran Pasaric

Geophysical Institute, Faculty of Science, University of Zagreb, Croatia

Forecast verification based on KxK contingency tables is not yet standardized. In the measure-oriented approach various scores are calculated and used to condense some aspects of the forecast quality, each score resulting in a single number. The final goal is to assess particular forecasting system or to compare various such systems. Beside classical measures like the Heidke or the Pierce ones, the Gandin-Murphy family of scores are used. The latter includes the Gerrity and the LEPS sub-families. On the other side, verification problem is multifaceted and no score can comprehend all information that is contained in the contingency table. For this reason the distribution-oriented approach has been proposed by Murphy and Winkler. Here, the joint empirical distribution of forecasts and observations as given by the KxK table is analyzed as a whole. In the present work a measure of association in the KxK table, known in social sciences as polychoric correlation coefficient (PCC) is applied. A standardized bivariate normal distribution is related to the table in a natural way. This normal distribution is fully specified by its correlation coefficient which in turn is the PCC of the table. The PCC possesses several desirable properties including the weak sensitivity on the number of categories. Moreover, from the bivariate normal distribution, which is determined by the PCC, and from the marginal frequencies, it is possible to reconstruct fairly well the original table. In this way the dimensionality of the problem is reduced from KxK to 2K, while the differences between the original and the reconstructed table could be further analyzed from the distributional point of view. The method is systematically applied to a large set of 6x6 contingency tables on verification of quantitative precipitation forecasts.

Email: pasaric@irb.hr

10.4

Verifications of probabilistic calibrations for deterministic GFS precipitation forecasts

Johannes Jenkner

International Research Institute for Climate and Society, Columbia University, New York, USA

Probabilistic predictions of heavy precipitation events provide useful guidance for risk management and action planning. In the present study, deterministic precipitation forecasts from the GFS model are subjected to a probabilistic calibration based on re-forecasts with a consistent model version. Daily rainfall estimates are taken from a gridded observational analysis based on multi-satellite data. Altogether, 11.5 years of global forecast and observational data are available to be subdivided in a training and test period. Logistic regression is applied to model the observed estimates from the spatial configuration of the forecast rainfall. To this end, principal component scores are derived in the neighborhood of individual target grid points. Then, these scores are taken as empirically independent predictors. Multiple model setups are used for the logistic regression which encompass varying sample sizes, neighborhood extents and predictor subsets.

The predictive power of individual forecast calibrations is investigated with respect to discrimination and reliability. The verification methodology is based on 2afc tests which are complemented by reliability diagrams. To provide an overview of the performance, the verification measures are aggregated over different regions and lead times. The results highlight different degrees of forecast quality, reflect specific forecast enhancements for individual predictors and pinpoint an optimal neighborhood size.

Email: jenkner@iri.columbia.edu

Posters

PO1

Communicating hydrologic verification information for operational forecasting and real-time decision making in the U.S. National Weather Service

Julie Demargne (1,2), James Brown (1,2), Yuqiong Liu (1,3), Dong-Jun Seo (1,2), Kevin Werner (4) and Lisa Holts (4)

(1) Office of Hydrologic Development, NOAA/National Weather Service, United States of America
(2) University Corporation for Atmospheric Research, United States of America
(3) Riverside Technology, Inc., United States of America
(4) Colorado Basin River Forecast Center, United States of America

In this presentation, we report recent progress in the NOAA's National Weather Service Hydrologic Services Program towards systematically verifying hydrologic forecasts and effectively communicating verification information to all users. The NWS produces hydrologic forecasts across time scales from hours to months to support a wide variety of applications, such as public safety during flooding and economic well-being for large-scale water management. Forecast verification is essential to monitor forecast quality over time, analyze the different sources of uncertainty and skill across the entire river forecasting process, and compare the quality of forecasts from different methodologies in order to evaluate forecast skill improvement from new science and technology. Such verification activity is fruitful only if the information generated leads to decisions about the forecast or forecasting system being verified. Therefore NWS developers and forecasters are working together and with users to develop meaningful verification products and capabilities to effectively help forecasters and external users in their decision making. Such verification requires a combination of diagnostic verification information, with summary verification results on the quality of past forecasts given similar conditions, and real-time verification information, with techniques to identify historical analogs to real-time forecasts and bias-correct the forecasts before the corresponding observations occur. This includes the selection of verification measures that can be easily integrated into the user decision making. Recent developments to provide such verification information in real-time, including a web application for water supply forecast verification, will be presented.

Email: julie.demargne@noaa.gov

PO2

The value of GFS data for precipitation forecasting for the Waikato River catchment, New Zealand

Stacey Dravitzki and James McGregor

Victoria University of Wellington, Wellington, New Zealand

The hydroelectric system on the Waikato River provides 13% of New Zealand's electricity. Hydroelectric operations can be optimised and floods mitigated if predictions of precipitation inputs can be improved. The Global Forecast System (GFS) model provides valuable information, but is limited in model resolution: strictly containing only one grid point within the 12,000 km2 river catchment. To verify these precipitation forecasts, we have compared the six-hourly precipitation forecast runs to rain gauge data from 22 stations near the Waikato.

Forecast verification statistics were calculated over a two-year period of forecasts out to 180 hours lag time. Discrete categorical analysis calculated hit rates of 0.8 for a 6-hour lag and 0.72 for a 180-hour lag forecast. In moderate to high precipitation categories, the false alarm rate is high and probability of detection is low. When comparing lagged time series, the mean error was between -0.3 and 0.7 mm per six-hour period for different stations near the Waikato, while the root mean squared errors range between 2.8 and 4.5 mm despite negative skill scores when compared to assuming the climatological mean. These results are dominated by the model's ability to correctly identify dry periods, which account for 85% of all periods in this study.

The timing and consistency of predicted precipitation were investigated as a lag ensemble. Precipitation predictions were simplified to either: a binary wet or dry, or with precipitation falling into one of six precipitation categories. Once again, this showed that the model is more skilful in predicting dry periods.

Email: j.mcgregor@vuw.ac.nz

PO3

TAF Verification in the MET Alliance

Guenter Mahringer

Austro Control MET Linz, Austria

The MET Alliance is a group of national aeronautical meteorological service providers from Belgium, Germany, Ireland, The Netherlands, Switzerland and Austria. They cooperate in the improvement and rationalisation of the meteorological services for aviation.
One of the Met Alliance projects deals with Forecast Verification. The development and implementation of internationally accredited and applicable forecast verification systems has long been a great source of difficulty and complexity for aviation MET service providers. In May 2008, the Met Alliance agreed on a common TAF verification system. By this cooperation, comparisons and the determination of best practice in forecast production can more easily be achieved, and the duplication of efforts can be avoided.
The TAF verification method is described in Mahringer and Frey (2007) and Mahringer (2008). The common and individual requirements of all Met Alliance members were included in the operational verification system, which is operational since November 2008.
This presentation focuses firstly on the ways to display verification results to forecasters, management and customers, secondly on verification measures, their advantages and limitations in comparing the quality of forecasts for different airports.
In a TAF, the forecaster gives a range of possible conditions of wind, visibility, present weather and clouds, by using different types of change groups. These conditions are valid for time intervals, the shortest being 1 hour. A TAF thus contains a range of forecast conditions. For each hour of the TAF, the highest (most favourable) observed value is used to score the highest forecast value, and the lowest (most adverse) observed value is used to score the lowest forecast value. All available observations within the hour are used.
For forecasters, a website is available where they can check individual TAFs. Weather conditions are highlighted which have been observed and forecast, observed but not forecast, and forecast but not observed. In this way, special attention is not only drawn to misses but also to over-forecasting, which is a great challenge in forecasting for short time intervals. When longer periods are evaluated, contingency tables give valuable insight in strengths and weaknesses of the forecasts.
For quality management and comparisons, the Gerrity Score (GS) and the Heidke Skill Score (HSS) are used. To address ICAO requirements, the contingency table diagonal and similar numbers are used as measures for the fraction of correct forecasts. The possibilities and limitations for comparisons between different airports and between TAFs and AUTOTAFs are investigated.
For customer information, results are shown for events of individual importance. A way to show the forecast quality for single events is the combination of event frequency p(E), p(E) when the event is forecast, and p(E) when the event is not forecast.

References:
Mahringer, G and Frey, H (2007). Austro Control TAF and TREND Verification. Third International Workshop on Verification Methods, ECMWF, Reading, UK.
Mahringer, G (2008). Terminal aerodrome forecast verification in Austro Control using time windows and ranges of forecast conditions. Meteorological Applications, Volume 15, Issue 1, Pages 113-123.

Email: guenter.mahringer@austrocontrol.at

PO4

Development of portable verification Package

Sultan Al-Yahyai

DGMAN, Oman

Directorate General of Civil Aviation and meteorology DGMAN has developed a portable verification package to be used to verify the NWP models against the observations. DMO, MOS and OBS are stored in a relational database to be used in the verifications. The package has a user friendly GUI to help the user selecting the input parameters such as (verification scheme, model type, variable name, statistic to be used and the OBS time and the forecast range). The system will plot the statistical scores (Bias, RMSE, Hit Rate … ect) to be used by the researchers. The package can also plot the diurnal cycle of different parameters from OBS, DMO and MOS. Conditional verification of different parameters can be generated using the relational database feature of the package.

Email: s.alyahyai@gmail.com

PO5

VERSUS - Unified verification package in COSMO

Adriano Raspanti

Italian Met Service - COSMO consortium, Pomezia (RM), Italy

The development of a complete Conditional and Standard Verification Tool has been the first priority and outcome of the VERSUS project.

From a more general point of view the main purpose of VERSUS is the systematic evaluation of model performances in order to reveal, in a way different from the usual classical verification tools, model weaknesses.

The typical approach to Conditional Verification consist of the selection of one or several forecast products and one or several mask variables or conditions, which would be used to define for example thresholds for the product verification (e.g. verification of T2M only for grid points with zero cloud cover in model and observations).

Through the development of VERSUS software a unified tool able to perform operational standard verifications, operational conditional verifications along with experimental standard and conditional verifications, in batch and interactive mode has been achievd.

The modularity of VERSUS easily allows updates and use of new verifications methods through the use and implementation of ‘’R’’ language software package or even ‘’ad hoc’’ algorithms (Fortran, C++, PHP).

The verification software has been developed with an User friendly graphic user interface, that makes easier of all the verification activities. The GUI is based on standard Web interface. The architecture is DB based.

Email: raspanti@meteam.it

PO6

Evaluation of hydrometeors of a high-resolution model using a radar simulator

Yasutaka Ikuta

Numerical Prediction Division, Japan Meteorological Agency, Tokyo, Japan

Japan Meteorological Agency plans to operate a high-resolution local forecast model (LFM) in the future. The main purpose of operating the LFM is the improvement of the disaster prevention information. For this purpose the improvement of short-range precipitation forecasts is an important subject. And the development of a sophisticated microphysics parameterization scheme is underway as one of our efforts to address this issue. This paper presents a new verification approach developed to evaluate the performance of the microphysics parameterization scheme.

As the verification approach, simulated equivalent reflectivity by a new radar simulator is compared with observed reflectivity using the fractions skill score. The new radar simulator is developed in consideration of refractivity distribution, which diagnoses the reflectivity from hydrometeors using the bulk microphysics parameterization scheme. The characteristic of this verification approach is that the three-dimensional distribution of hydrometeors is estimated through the verification of reflectivity.

This verification technique is applied to the LFM. The results show the verification is sensitive to different configurations of the reflectivity diagnostic method. Moreover, in high-resolution forecasts of the LFM the necessity for the spatial verification technique is shown by the problem of time-space matching. It is demonstrated from the result that verification approach in this study makes it possible to quantitatively discuss the reproducibility of the structure of precipitation system using the explicitly evaluated three-dimensional distribution of hydrometeors.

Email: ikutamet.kishou.go.jp

PO7

Verification of cloud physical properties

Kristian Pagh Nielsen

DMI, Copenhagen, Denmark

For the first time, field verification of NWP cloud physical properties against measurements from the MSG satellite are presented. In classical observations clouds are described in terms of 1/8 fractions of cloud cover. The cloud fraction, however, is only one of the factors that determine the reflectance and transmittance of a cloud field. The other factors are the inherent cloud physical properties. Therefore, satellite measurements of cloud physical properties, much improves the capability of cloud verification. With these measurements, processes such as aerosol indirect effects can also be assessed in detail. Results will be presented and discussed.

Email: kpn@dmi.dk

PO8

Base rates and skill scores

Robert Fawcett and Elizabeth Ebert

Bureau of Meteorology, Melbourne, Australia

In assessing skill in the forecasting of relatively rare events (e.g., the tails of a predictand distribution), several factors are important. Among these may be included (i) the climatological probability of the event (also known as the base rate), (ii) the strength of the relationship between the predictor and the predictand, (iii) the choice of skill metric, and (iv) the calibration of the forecast system in relation to the verifying observations. In this presentation, we explore connections between the first three of these factors in the context of a simple theoretical forecast model for which it is possible to derive exact results (without resorting to large Monte Carlo simulations). The model assumes a simple linear relationship between the predictor and predictand, together with a noise component whose magnitude is tied to the correlation between the predictor and predictand.

Skill measures assessed include the LEPS2 and half-Brier skill scores, and an extensive list of scores associated with 2×2 contingency tables (such as the proportion correct, hit rate, false alarm rate, false alarm ratio, Peirce's skill score, critical success index, Gilbert skill score, Yule's Q, extreme dependency score, likelihood ratio and odds ratio). [For the purposes of this study, issues of calibration (e.g., those arising from departures from reliability) are ignored, as the simple theoretical model is constructed to be correctly calibrated.]

For a given level of the relationship between predictor and predictand, some skill scores (e.g., LEPS2 skill score, false alarm ratio, critical success index, Gilbert skill score) show a decreasing level of skill as the climatological probability of the tail category decreases, while others (e.g., hit rate, Peirce's skill score, Yule's Q, odds ratio, extreme dependency score) show the opposite behaviour. The results obtained here confirm previous findings that the relationships between skill scores and the underlying strength of the connection between predictor and predictand vary markedly from skill score to skill score, leading to the conclusion that a particular magnitude of skill score is a highly relative thing.

Email: r.fawcett@bom.gov.au

PO9

Some robust scale separation methods at work

Marek Jerczynski

Institute of Meteorology and Water Management, Krakow, Poland

Various meteorological forecasting and observing systems work in their characteristic spectra of space scales and for this reason there is a real need to adequately separate scales to properly compare and merge data of various origin. There are many separation methods applied in contemporary operational verification/validation systems but usually they suffer because of sensitivity to outliers, dependence on assumptions on error distribution, also because of deficiencies of other kinds. Due to huge amount of data analyzed by automatic operational systems unsupervised scale separation methods seems to be of prime importance in mentioned above tasks. Some examples of such attitude are analyzed in this presentation. Robustness is a key feature of examined methods.

Email: zijerczy@cyf-kr.edu.pl

PO10

Time-series analysis of scale-selective verification: can we use it for operational forecast monitoring?

Marion Mittermaier

Met Office, Exeter, UK

A scale-selective verification package was introduced into the Met Office operational verification suite early 2006. Whilst the need for scale-selective verification has been recognised (e.g. to overcome issues such as the double-penalty problem), especially at finer horizontal resolution, the pragmatic usefulness of such verification techniques for operational monitoring of forecasts remains unknown. In this paper we have the luxury of exploring a three-year (and growing) time series to answer some of the questions that remain. We can compare different resolution models and different lead times to consider what such scale-selective metrics can tell us about trends in model performance over time. As ever, the answers are not all clear cut, and we conclude with a number of areas that still need improving.

Email: marion.mittermaier@metoffice.gov.uk

PO11

SAL Verification in Hydrological Catchments

Pertti Nurmi and Sigbritt Näsman

Finnish Meteorological Institute

SAL (Structure-Amplitude-Location) is an object-based verification method which is convenient for the verification of QPF forecasts within specified domains like hydrological catchments. SAL has been applied for chosen river catchments of various sizes in Finland by verifying deterministic forecasts originating from the global ECMWF and the regional HIRLAM_RCR and HIRLAM_MB71 models. Human forecaster generated QPF fields, by utilizing in-house grid editing production tools, are analyzed to estimate the potential added value of human intervention. Radar-derived QPE fields are used as main source of observed “truth” information. However, also and rain gauge data may be and have been utilized. The smaller river basin covers 3000 and the larger one 30000 square kilometers. The results show that the higher resolution models perform better than the coarser ones which would indicate that SAL is a useful tool to take into account the effects of the notorious “double penalty” issue.

Email: pertti.nurmi@fmi.fi

PO12

Veri f cation of the Seasonal Rainfall Prediction in the Rimac River Basin

Juan Bazo and Carmen Reyes

Peruvian Meteorological and Hydrological Service, Lima , Perù

This research is based on the verification of the seasonal statistics forecasts, by using probabilistic techniques, considering the scope of the forecast is common to use many variables of very different characteristics, as well as the binary variables (rainfall, non rainfall), categorical (below normal, normal and above normal) and continuous variables (temperature, humidity and others). In addition, a same variable can be considered as binary, categorical or continuous according to the particular application. It is often used the range of a variable, considering a finite number of intervals (below, normal and above); In general, the percentiles of the climatic series are considered to define the thresholds. We focus on applying this technique to verify the skill of the seasonal rainfall prediction by means of the statistical tool Climate Predictability Tool (CPT), from the International Research Institute for Climate and Society (IRI), these forecasts were done for the Rímac River Basin from 2006 to 2009; as a result we are going to show the forecasts and their verifications corresponding for the rainfall season (DJF), in which we obtained high scores (up to 80% of hit). The results are interpreted based on the discussion of advantages and limitations of the used methodology.

Email: jbazo@senamhi.gob.pe

PO13

Reassessing the skill of GCM-simulated precipitation

Jonathan Eden

School of Geography, Earth and Environmental Sciences, University of Birmingham, UK

GCMs are the most important tool in estimating future climate change. Although projections of large-scale circulation made by GCMs are relatively skilful, precipitation is still considered to be poorly represented. In order to accurately reproduce precipitation fields, models must be able to represent a number of processes, such as condensation, evapotranspiration and the orographic influence on the movement of air (Randall et al., 2007). These processes occur at finer scales than those used in climate change analysis and are thus parameterised within the model. In general, GCMs are able to reproduce the main features of global precipitation. However, it is difficult to compare monthly means of simulated precipitation with observations as GCMs are not able to reproduce interannual variability.

Several studies have evaluated the ability of atmospheric reanalyses to produce reliable precipitation estimates. A reanalysis is able to assimilate a range of observed atmospheric variables within a background forecast model (Bosilovich et al., 2008). As such, it can be considered an 'ideal' GCM in which the large-scale circulation is in good agreement with reality (Widmann and Bretherton, 2000). Precipitation observations are not included in the reanalysis and the model's precipitation field is subsequently parameterised in the same way that it would be in a GCM. These studies have shown reanalysed precipitation to have good skill in reproducing observations in many parts of the world.

It is possible extend this validation of precipitation to GCMs by ensuring that the large-scale circulation is in good agreement with observations. Here we use a nudging technique, based on Newtonian relaxation, to force the key circulation and temperature fields within the ECHAM5 GCM to corresponding values from the ERA-40 reanalysis. The nudging procedure allows for divergence, vorticity and surface temperature to be independently calculated at each model timestep before being subjected to a pre-defined relaxation coefficient. Thus the GCM is able to perform its own physics and dynamics freely whilst being guided towards the observed large-scale circulation over an historical period.

By comparison with global GPCP data and land-only GPCC data, we show that in many areas the precipitation from the nudged ECHAM5 simulation is in excellent agreement with both the quantity and interannual variability of observations. Correlation of simulated and observed monthly precipitation means is greater than 0.9 across large parts of the northern hemispheric land mass. Oceanic correlations exhibit greater variation, with the exception

of the equatorial trough, and correlations are notably weaker over tropical land masses. For much of Africa (especially during the boreal summer) and South America, this is perhaps attributable to a sparse observational network compared with North America and Eurasia. We further demonstrate that simulated precipitation can thus be considered an excellent predictor for regional precipitation as part of a statistical downscaling methodology. It is hoped that downscaling relationships between simulated and observed precipitation fields can be applied to future simulations.

Email: jme184@bham.ac.uk

PO14

Improving Meteorological downscaling Methods with Neural Network Models: South America Rainfall

David Mendes

Earth System Science Center - CCST / National Institute for Space Research – INPE, São Paulo, Brazil

As history embraces the beginning of a new millennium, old problem still constitute enormous challenges to the population in general, and to the academic world in particular. The atmospheric sciences modelling community spends a large share of its research activity to improve three different aspect of atmospheric modelling, namely: a) Short-term weather forecasting (e.g. Bengtsson, 1999), b) Development of climate change scenarios for several decades or centuries ahead (e.g. Tett et al., 1997, and others), and c) Reconstruction of historical past climate (e.g. Valdes and Crowley, 1998).

In recent years, several different methods have been applied to bridge the scale gap between coarse resolution GCMs and the finer spatial resolution required for climate impact studies (Frei and Gavin, 2005; Raphael and Holland, 2006). These methods have become known in the literature as downscaling methods. Two main types of downscaling methods have emerged in the last decade. First, techniques that are based on regional dynamical models with finer resolution that GCMs. The second group of downscaling methods is based on the establishment of empirical statistical transfer functions between the large-scale circulation and local climate variables (Palutikof et al., 1997).

In recent years, an increasing number of paper within the meteorological community have adopted Artificial Neural Networks (ANN), to model many different variables, including seasonal forecast of the Indian Monsoon (Navone and Ceccatto, 1994). Others authors have used ANNs to perform short-term forecast of precipitation (e.g. Kuligowsky and Barros, 1998a; Luck et al., 2000; Boulanger et al., 2006).

The main aim of this project is to develop and test a novel type of statistical downscaling technique based on the use of Artificial Networks, applied of the climate change. The models will be constructed using observed data (South America sector) and then applied to CGM output in order to evaluate their ability to produce higher resolution climate change scenarios and improved short-term weather forecast over South America.

These projects will asses, as objectively as possible, the potential advantages of using ANN Model to solve three different types of meteorological/climatological downscaling problem.

The ANNs models can be trained to find the best mathematical relationship between the atmospheric circulation and local climate, without predefined restrictions. Thus the method is able to capture some of the non-linear relationships between local climate and the large-scale circulation (Sahai et al., 2000; Knutti et al., 2003).

The result preliminary introduces a methodology of downscaling applied to GCMs model output using an Artificial Neural Network (ANN) approach and linear regression. The method is proposed for downscaling daily precipitation series for South America Continent. The performance of the temporal neural network downscaling model is compared to a regression-based statistical downscaling model with emphasis on their ability in reproducing the observed climate variability and tendency for the period 1970-2000. Furthermore, the different model test results indicate that the Neural Network Model significantly outperforms the statistical models for the downscaling of daily precipitation variability.

Email: david.mendes@cptec.inpe.br

PO15

Artificial Neural Network (ANN) Application for South America Rainfall

David Mendes

Earth System Science Center - CCST / National Institute for Space Research – INPE, São Paulo, Brazil

The atmospheric sciences modeling community spends a large share of its research activity to improve three different aspect of atmospheric modeling, namely; Short-term weather forecasting (e.g. Bengtsson, 1999), development of climate change scenarios for several decades or centuries ahead (e.g. Tett et al., 1997), and reconstruction of historical past climate (e.g. Valdes and Crowley, 1998).
It is now widely accepted that General Circulation Models (GCMs) represent the most satisfactory technique to answer these challenges (IPCC, 1996). Furthermore the present low spatial resolution of most GCMs (~ 150 km - forecast and ~ 350 km climate model) makes it problematic to use their output in local and even regional studies.

These are a variety of downscaling techniques in the literature, but in practice two major approaches can be identified at the moment.
1) Dynamic downscaling approach is a method of extracting local scale information by developing and using limited-area (LAMs) or regional climate model (RCMs) with the GCMs data used as boundary conditions.
2) Empirical (statistical) downscaling starts with the premise that the regional climate is the result of interplay of the overall atmospheric, or oceanic, circulation as well as of regional topography. In recent years, a number of papers within the climatology community have adopted artificial neural network (ANNs) as a tool to downscaling from the large scale atmospheric circulation (e.g. Hewitson and Crane, 1992; Cavazos, 1997).

The Artificial Neural Network (ANN) is a system based on the operation of biological neural network, in other works is an emulation of biological neural system. The advantages of ANN are: 1) can perform tasks that a linear program can not; 2) when elements of the ANN fails, it can continue without any problem by their parallel nature.

In this work the Neural Network models are developed using MatLab. Input to the ANN are the 5 predictor variables derived from the IPCC AR4 models (CGCM3.1, CSIRO-MK3.5, ECHAM5, GFDL 2.1, MIROC-MEDRES) and are predictor variable derived from the daily observed data while the output are daily precipitation amounts. The ANN structure adopted is a Multi-Layer Perceptron (MLP) with a feed-forward configuration.

Specifically an ANN-MLP structure with 3 layer and 5 nodes per layer was selected (Figure 1). A non-linear transfer function was selected for all nodes and layer and a back-propagation algorithm (Rumelhart et al., 1996) was used for training the ANN. The number of hidden nodes (optimal) (over a ranger of 4 and 20 with a 5 nodes step) and the proper learning rate and momentum were determined through sensitivity analyses.

Sensitivity analysis provides a measure of the relative importance among the predictors (input) by calculation how the model output varies in response to variation of an input. The results provide a measure of the relative importance of each input (predictor) in the particular input-output transformation. From the 30 years of observed data daily representing of climate, first (1971-1990) are considered for calibrating the downscaling models while the remaining 10 years of data (1991-2000) are used to validate those models.

Email: david.mendes@cptec.inpe.br

PO16

How Different Rainfall Seasons Respond Differently to the same ENSO event in Ethiopia

Gezu Mengistu

National Meteorological Agency, Ethiopia

It is well recognized that Ethiopia is one of the most adversely affected countries to climate change .Low economic strength, inadequate infrastructure, low level of social development, lack of institutional capacity, and higher dependency on natural resources base make the country more vulnerable to climate stimuli(including both variability and extreme events). The Ethiopian economy is based on rain-fed agriculture. Regardless, rains seriously affects the country's food production prospects and in the human and animals lives.

The relation of ENSO and climate cause Ethiopia rainfall distribution by making it "above normal" and "below normal" depending on the type of episodic event.
Season has different meaning but meteorologically it is a period when an air mass characterized by homogeneity in meteorological parameters, which influences a region. Seasons are generally classified into four namely, winter, spring, summer and autumn, while in low latitude they categorized as wet and dry only. In the case of Ethiopia, seasons are unique and are classified mainly based on rainfall and its distribution (Werkneh Degefu 1987).

By using rainfall data from different station and SST I found that, ElNino tends to decrease kiremt (June-September) rainfall while it tends to enhance Belg (February-May) and Bega(October-January) rainfall. Similarly, LaNina tends to decrease Belg (February-May) rainfall, enhances the dryness of the Bega (October-January) season and tends to increase the kiremt (June-September) rainfall.

Traditionally, many people assume that ENSO episodic events (EliNino and LaNina) significantly affect the annual total rainfall amount in Ethiopia and they may tend to build total rainfall. However, my results clearly indicate that the impact of extended ENSO events affect individual rainfall seasons in an opposite manner. Excess rainfall is one season is compensated by rainfall deficiency in the other season resulting in the near normal annual rainfall. Hence, in an opposite manner excess rainfall in one season is compensated by rainfall deficiency in the other season resulting in near normal annual rainfall. Hence irrespective of extended ENSO events over two or more seasons the annual total rainfall amount tends to be near normal over most of our area of interest.

Hence most of the sever drought years of the past are due to failures of individual rainfall seasons, while the annual total rainfall amount of most of these years are near normal .Hence, ENSO related impacts needs to be based on individual seasons. Climate monitoring based on annual rainfall could be misleading during extended ENSO events.

Email: Gezumengistu@yahoo.com

PO17

A block bootstrapping method for verification of the Canadian NWP model

Marcel Vallee

Meteorological Research Division, Environment Canada

When the verification period is a valid representation of your verification population and each event is independent and part of the same distribution then bootstrapping your verification sampling period is equivalent to estimating your verification population. Unfortunately, meteorological data is spatially and temporally correlated which forces the use of bootstrapping in blocks. The width of the confidence interval provides an estimate of the uncertainty inherent in the process of population sampling.

This poster will present results based on the bootstrapping technique with confidence intervals performed in the verification package for surface weather variables at the Meteorological Research Branch. The goal is to compare the operational numerical model against a new proposed model currently being implemented at the Canadian Meteorological Center in Canada.

Email: marcel.vallee@ec.gc.ca

PO18

Cloud verification using radar: what is the half-life of a cloud-fraction forecast?

Robin J. Hogan (1), Ewan J. O'Connor (1,2) and Anthony J. Illingworth (1)

(1) Dept. of Meteorology, University of Reading, Reading, UK
(2) FMI, Helsinki, Finland

Cloud radar and lidar can be used to evaluate the representation of clouds in models, but usually only the model climatology is tested, for example by comparing the cloud-fraction PDF to observations. This says nothing about the skill of the model at simulating clouds at the right time. Skill scores have been used before to evaluate clouds in models, but usually only simple scores have been used, which often have deficiencies such as being easy to hedge and tending to a meaningless value (usually zero) for rare events.

In this poster we assess the suitability of a number of scores for verifying cloud-fraction forecasts based on five desirable properties, and recommend the use of the new "Symmetric Extreme Dependency Score" (SEDS), which is equitable (for large samples), relatively resistant to hedging and independent of the frequency of cloud occurrence (or base rate), even for very rare events. We then use data from five European ground-based sites and seven operational models, processed using the Cloudnet analysis system, to investigate the dependence of forecast skill on (1) cloud fraction threshold, (2) height, (3) horizontal scale, and (4) forecast lead time. The models are found to be most skillful at predicting the occurrence of mid-level clouds and least skillful at predicting boundary-layer clouds. It is found that skill decreases approximately inverse-exponentially with forecast lead time, enabling a forecast "half-life", to be estimated. When considering the skill of instantaneous model snapshots, we find that typical values range between 2.5 and 4.5 days.

This work suggests that the cloud radar sites could be used for the routine verification (and improvement) of clouds in numerical weather prediction (NWP) models; currently the clouds in these models are only tested against human cloud observations, which can be very misleading. Moreover, the SEDS measure is well suited for use in other areas of forecast verification.

Email: e.j.oconnor@reading.ac.uk

PO19

Assessing the operational skill of predictions of forecast error

Pascal Mailier

Royal Meteorological Institute of Belgium, Bruxelles, Belgium

For energy companies, the ability to guess the most likely direction and magnitude of errors made in deterministic forecasts of surface temperature would provide a substantial competitive advantage. An automated, but expensive proprietary system (black box) designed for this purpose has been tested on a set of daily minimum/maximum forecasts out to 9 days at one single location in the UK (London Heathrow). A simple and cheap alternative system based on the ECMWF Ensemble Prediction System was devised to serve as a benchmark. The two competing systems had to predict one of three categories: blue (observed temperature will be more than 2 C lower than predicted), red (observed temperature will be more than 2 C higher than predicted), and white (observed temperature will be within 2 C of the predicted value). Performance was quantitatively assessed using several metrics in order to examine various attributes: association between observed and forecast categories, proportion of correct forecasts, hit rate vs. false-alarm rate (ROC score), and finally frequency bias. Results from this test did not provide evidence of the superiority of the automated system and therefore its purchase was not recommended.

Email: pascal.mailier@oma.be

PO20

Analysis of marine seasonal ensemble forecasts for the Baltic Sea

Petra Roiha

Finnish Meteorological Institute, Helsinki, Finland

Ensemble forecasts have for long been an essential tool in meteorology. Nowadays seasonal ensemble forecasts are becoming prevalent also in oceanography. Due to the longer marine timescales, the useful time span of marine forecasts may be longer than in atmospheric applications.

In this work we used the ensemble approach to forecast physical and chemical changes in Baltic Sea. We present results of ensemble forecasting in the Baltic, and discuss the usefulness of this method. FMI's operational 3-dimensional biogeochemical model was used to produce monthly ensemble forecasts for different physical, chemical and biological variables. The modelled variables were temperature, salinity, velocity, silicate, phosphate, nitrate, diatoms, flagellates and two species of potentially toxic filamentous cyanobacteria.

Ensembles were produced by running several 30 day runs of the biogeochemical model. The model was forced every run with a different member of the ECMWF's orthogonally perturbed monthly ensemble prediction forecasts. A deterministic control run was also included for comparison purposes. The initial conditions for the marine ensemble were provided by the deterministic short range operational Baltic Sea forecast which is based on the same ocean model.

The ensembles were then analysed by statistical methods and the median, quartiles, minimum and maximum values were calculated for model output variables to gain insight into the applicability of the results. Verification for the forecast method was made by comparing the results against in-situ temperature data gathered with fixed wave buoys. Different verification methods were evaluated
from the perspective of marine science where observations are often fewer than in meteorology. The results of the model demonstrated that ensemble ocean forecasting is a viable tool and it indeed is possible to forecast with useful accuracy the Baltic Sea with these time spans.

Email: petra.roiha@fmi.fi

PO21

An estimation of QPF uncertainty by forecasting the radar-based ensemble skill

Petr Zacharov

Institute of Atmospheric Physics, Academy of Sciences of the Czech Republic, Prague, Czech Republic

We summarize the results dealing with estimating PMP uncertainty by using the relationship between ensemble skill and ensemble spread. The results refer to five convective storms, which occurred over the Czech territory and caused local flash flooding. The regional ensembles were formed by using COSMO model in an experimental mode. A driving COSMO (LLM) was run with the horizontal resolution of 11 km and with initial and boundary conditions derived from ECMWF analyses. The driven COSMO (SLM) used the horizontal resolution of 2.8 km and the initial and lateral data from LLM. The SLM integration started at 06 UTC and finished at 24 UTC of the same day. The regional ensemble of 13 members was formed by a simple modification of LLM results.

We focused on the assessment of the relationship between ensemble spread and ensemble skill. The ensemble spread and skill values were calculated by using traditional (Root Mean Square Error) and fuzzy (Fractions Skill Score) measures. The ensemble skill was evaluated by comparing the ensemble member forecasts with radar-based rainfalls and the spread was estimated comparing the ensemble member forecasts with the undisturbed control forecast. The effect of scale was assessed by considering square elementary areas of various sizes that were centered in grid points of the verification domain.

The relationship between ensemble spread and skill used was expressed by regression curves that were constructed for 4 events. The predicted skill values for the fifth event were evaluated. The FSS-based skill forecast appears to be useful in uncertainty forecasting. It represents quite well the real skill and could be used as a combined input to the hydrological models together with deterministic or ensemble precipitation forecast.

The work is supported by the project OC112 (COST731) and by the grant GACR 205/07/0905.

Email: petas@ufa.cas.cz

PO22

Extreme temperatures verifications on Argentina forecast by NWP GFS/NCEP

Matias Armanini

National Weather Service, Buenos Aires, Argentina

Due to both the impressive topographical variation which the Argentine Republic presents on its western side and the sea temperature influence over its eastern part, it is of highest importance to acknowledge how different models of numerical forecast represent these issues.

In this work, verification on extreme temperature forecasts up to 24 hours derived from the numerical forecast model GFS/NCEP has been carried out in order to quantify the average seasonal error, which implies its subsequent use as maximum and minimum temperature forecast over the Argentine territory. Therefore, the most frequent statistical parameters used for model assessment have been analysed: the root mean squared error (RMSE), the bias and the success percentage.

The model reliability along the Argentinean littoral region all over the year concerning the minimum temperatures forecast is one of the most relevant results obtained in this work. Likewise, the country south-western area reflects the worst performance of the model, with marked errors mainly dealing with underestimation. Indeed, these errors are product of two major factors: the low density of meteorological seasons and an extremely variable topography. Finally, the seasonally average temperatures display a minor amplitude in relation to the observed ones, due to overestimation errors regarding minimum temperatures broadcast and underestimation errors in the maximum, showing the models conservative role in the region.

Email: matiasarmanini@yahoo.com.ar

PO23

Verification of nowcasting methods in the context of high-impact weather events for the Canadian Airport Nowcasting (CAN-Now) project

Monika Bailey (1), Janti Reid (1), George Isaac (1), Faisal Boudala1(1), Norbert Driedger (1), Marc Fournier (2), and Laurence Wilson (3)

(1) Cloud Physics and Severe Weather Research Section, Environment Canada
(2) Canadian Meteorological Aviation Centre-East, Environment Canada
(3) Meteorological Research Division, Environment Canada

A prototype nowcasting system for forecasting high impact weather is under development at Pearson International airport (CYYZ) as part of the Canadian Airport Nowcasting Project (CAN-Now). The CAN-Now system inputs data from multiple sources: model data, hourly meteorological observations, radar, satellite, and high time-resolution measurements from a large suite of on-site instruments measuring parameters such as visibility, precipitation, ceiling and winds. The study presented here compares forecasts from the Canadian GEM Regional and GEM-LAM models with forecasts obtained using measurements from on-site instrumentation at CYYZ. Earlier verification studies were performed on the entire data set or on data stratified by season or on individual case studies, here we focus specifically on high impact weather events observed at CYYZ.

Data collection for CAN-Now began in February 2007 and a two year archive is now available for such studies. The study is organized as follows. (1) High- impact aviation related weather scenarios are defined in terms of observables such as winds, temperature, relative humidity, occurrence of precipitation, ceiling and visibility. It is necessary to keep the definitions broad to ensure a sufficient number of events for statistical analysis. (2) The start and end times of significant events are identified by searching the hourly surface observations obtained from the Environment Canada archives. (3) High temporal-resolution data is extracted from the archive of on-site instrument data for these times. (4) Verification is performed, separately for each scenario type, of forecasts of temperature, relative humidity, winds and precipitation occurrence. The forecasting performances of the GEM models, of extrapolated observations and of persistence are compared statistically in terms of the start time, duration and intensity of events. Our goal is to use these results to develop a methodology for generating verifications in real time.

Email: monika.bailey@ec.gc.ca

PO24

Reliability Evaluation of HyBMG by Using ROC Curve

Kadarsah Binsukandar Riadi

Meteorological and Geophysical Agency, Jakarta, Indonesia

Reliability evaluation of HyBMG model has been done by using Relative Operating Characteristics (ROC) which is created by plotting the hit and false-alarm rate. The Evaluation model is use rainfall data from only 34 cities over 10 years from 1998 to 2007. The result is ROC's curve that describes the reliability HyBMG to predict rainfall. HyBMG has a reliability to predict the rainfall in a particular region.

Keywords: HyBMG, ROC, False Alarm Rate, Hit Rate.

Email: kadarsah@yahoo.com

PO25

Rainfall prediction performance of WRF model over complex terrain of Ethiopia

Girmaw Bogale

National Meteorological Agency, Addis Ababa, Ethiopia

This paper presents WRF model performance to predict rainfall over complex terrain of Ethiopia .The study conducted during main rainy season of the country (June-September, 2008). For this period, we took 33 days of WRF forecast with forecast length of 3 days. The observed rainfall is compared with the forecast of WRF for 89 days. Time series and spatial analysis as well as the dichotomous (contingency table) methods are used in this study.

Results from time series analysis showed that WRF model can capture the natural variability of the rainfall. Spatial analysis indicated that rainfall reproduced by WRF model has the same spatial coverage to that of the observed rainfall for most of the days. Nevertheless, the model did not reproduced the exact amount rainfall like the other models do. In addition, dichotomous method is used to test how the model can capture the rainfall event. The accuracy of the model is greater than 80% over central, western and northwestern parts of the country for the first 24 hours and its accuracy decreased relatively for 48 hours and 72 hours forecast. In this part of the country, the model over forecast the rainfall events and do not have missing rainfall events.

WRF model performance is relatively weak over northeastern, eastern and southern parts of the country. The accuracy of the model lies between 50 and 70%. Likewise, the model missed some of rainfall events in this parts of the country.

Email: girmaw.bogale@gmail.com

PO26

Verification of the Hirlam NWP forecasts and the connection between the scores and improvements in the model

Kalle Eerola

Finnish Meteorological Institute, Helsinki, Finland

The forecasts of the FMI operational Hirlam forecasts have been verified for the period 1990-2008. The aim is on one hand to see, how much the forecasts have been improved during the years, when measured with monthly verification scores, like rms-error and bias. The other purpose is to connect the improvements in scores to the changes and improvements in the Hirlam data-assimilation/forecasting system.

The forecasts are compared to the corresponding numerical analysis (so-called field verification concept) on two areas, one consisting of Europe and northern Atlantic and the other of Scandinavia. The monthly scores contain both the spatial and temporal variation and the verified parameters are mean sea level pressure and geopotential, temperature and wind components on constant pressure levels.

The Hirlam forecast system has undergone many changes. For instance, the horizontal resolution has increased from 0.5 deg. to 0.15 deg. and the vertical resolution from 16 to 60 levels. The data assimilation system has gone from optimal interpolation via three-dimensional variational assimilation to four-dimensional variational assimilation. The model has undergone many changes both in the dynamics and physics.

The verification scores show a substantial improvement during the years. For instance, the rms-error of the mean sea level pressure is nowadays about half of the value it used to be in the early 1990's, in other words, two days' forecasts are now as good as one day's forecasts 20 years ago. The same feature is even more pronounced for the upper and middle troposphere fields. Especially, the reduction in rms-error has been very prominent during the latest years. It seems that this reduction is connected to the introduction of the re-run concept and related mixing of the large-scale features from the high-quality ECMWF analysis to the first-guess in the Hirlan data assimilation.

On the other hand, the similar reduction in error cannot be seen in the lower troposphere. Improvements in the 850 hPa and 925 hPa temperature can be seen only in the first years. Then reduction in error was related to the too moist boundary layer and was corrected when introducing the new radiation scheme.

Also the reduction in the bias of mean sea level pressure can be connected to the change in the model: an artificial turning of the surface stress vector, introduced in 2004, seemed to cure the problem: the bias was almost totally removed.

Above we showed some examples, where the improvements in scores could be connected to the changes in the Hirlam system. However, statistical verification scores do not tell the reason to the quality of the forecasts or tell what should be corrected in the model. However, these examples demonstrate that they can be useful in assessing the effect of changes and corrections afterwards. On the other hand, the long-term time-series of the verifications scores also tell, if the efforts to improving the models have been successful or not.

Email: kalle.eerola@fmi.fi

PO27

High-resolution Regional Model (HRM) performance as NWP tool in Pakistan

Ata Hussain

Pakistan Meteorological Department

Owing to the complexity of the dynamics of the South Asian Summer Monsoon (June-September), the prediction of monsoon weather systems and associated extreme weather events have always been very challenging to the meteorologists of South Asian countries. Although, Pakistan receives about 150 mm during the summer monsoon season (which is 50% of the annual rainfall), seasonal weather prediction as well as the day-to-day weather forecasting during this season have always been exigent. Pakistan Meteorological Department (PMD) has been making use of various tools and techniques for weather forecasting in Pakistan. In the recent years, PMD has implemented the High resolution Regional Model (HRM) of DWD (the national meteorological service of Germany) as an operational model for numerical weather prediction in Pakistan. The initial and later boundary conditions for HRM are taken from DWD's global model GME with the multilayer soil model. The model is run with the resolution of 22 Km. In this study, the performance of HRM has been examined for various extreme weather events in Pakistan. The model has been found to be a very useful tool and a valuable addition in the forecasting practices of PMD. Different extreme weather events at various locations in Pakistan as predicted by the HRM have been looked at and compared with the observed weather conditions. The weather events in addition to others include the rainfall events at Multan (a major city about 500 km south of Islamabad) and Islamabad on 4th July, 2005 and 5th August 2006 respectively, wide spread rainfall on 10th February, 2007, a hailstorm at Islamabad on 10th January, 2008, and tropical cyclones of Gonu and Yemyin which formed over the North Arabian Sea during June, 2007. It has been found that spatial and temporal distributions of predicted weather conditions may vary from the observed events. Some times, the variations are quite significant. In some cases, however, the model has captured the events very well especially, the hailstorm event at Islamabad and the tropical cyclone Yemyin.

Email: atahussaingill@yahoo.co.in

PO28

Verification of numerical weather predictions for the western Sahel by the United Kingdom Met Office Limited Area Model over Africa

Oluseun Idowu

University of Missouri, Kansas City, USA

Numerical weather predictions (NWPs) are subject to systematic errors and biases. Hence, the continuous verification of NWP model outputs in order to contribute to model improvement became very important over recent years. This study investigates the potential of the 20km x 20km resolution Limited Area Model over Africa (Africa LAM) developed by the United Kingdom Meteorological Office (UK Met Office) to be used as a supplementary tool to improve weather forecast output to end-users over the Western Sahel (WS) and Nigeria. The Africa LAM T+24h forecasts dataset was verified against daily observed rainfall, maximum and minimum temperature data, of 36 selected meteorological point stations over the WS from January 2005 to December 2006. The verification algorithms and measures used in this study are in accordance with the WMO NWP verification standards. Results indicate that the Africa LAM model temperature forecasts show skill, more so during the raining seasons (AMJ and JAS) than during the dry seasons (JFM and OND) over the WS. The model rainfall forecasts, however, show more skill during the dry seasons (JFM and OND) than during the raining seasons (AMJ and JAS). The results further indicate that, on a regional basis, the model temperature forecasts show more spatial skill over the Wet Equatorial (WE) climate zone than over the Wet and Dry Tropics (WDT) and Semi Arid (SA) climate zones of the WS, while rainfall forecasts show more skill over the SA climate zone than over the WDT and WE climate zones of the WS. These results give a better understanding of the model forecast errors and also provide the feedback necessary for a possible improvement of Africa LAM forecasts by the UK Met Office.

Email: osif38@umkc.edu

PO29

Applicability of common verification methods for comparisons between measured wind data and simulated wind fields

Anna Lindenberg and Anne Paetzold

Institute for Coastal Research, GKSS Research Centre, Geesthacht, Germany

Within the verification of wind data sets one has to distinguish between two major points:
The first step is to obtain useful measuring data, which should be representative for the simulated wind fields that shall be investigated. This task turns out to be more difficult than usually expected, because each kind of data set holds certain deficiencies. Near surface wind speed measurements at national weather stations show disturbances due to the environment. Therefore they are rarely representative for the wind conditions of the size of one model grid box. Tower data from high measurement masts show a better representativity but are less available. Additionally, the influence of the tower itself must be considered and corrected. Another possible data source are wind turbine production data from wind parks. They possess more spatial coverage, but no information about low and high wind speeds and they are affected by wind park effects. All these issues must be considered before applying any verification method.

After selecting and preparing a useful data set the applicability of verification methods must be checked depending on the kind of information that shall be extracted. In this presentation common statistical methods regarding the comparison of observed and simulated data are presented. They are applied to compare wind speed measurements with simulated wind fields from the regional climate model COSMO-CLM. Their advantages and disadvantages are discussed especially from a mathematical point of view. Additionally, common statistical visualization techniques are investigated and their limitations for such wind field comparisons are demonstrated.

Email: janna.linderberg@gkss.de

PO30

The Impact of Applying Different Verification Techniques and Precipitation Analyses in QPF Verification

J. Moré (1), A. Sairouni(1), T. Rigo(1), M. Bravo(1), and J. Mercader(2)

(1) Meteorological Service of Catalonia (SMC), Barcelona, Spain
(2) Department of Astronomy and Meteorology. University of Barcelona, Spain

Verification of QPF in NWP models has been always challenging not only for knowing what scores are better to quantify a particular skill of a model but also for choosing the more appropriate methodology when comparing forecasts with observations. On the one hand, an objective verification technique can provide conclusions that are not in agreement with those ones obtained by the ''eyeball'' method. Consequently, QPF can provide valuable information to forecasters in spite of having poor scores. On the other hand, there are difficulties in knowing the ''truth'' so different results can be achieved depending on the procedures used to obtain the precipitation analysis.

The aim of this study is to show the importance of combining different precipitation analyses and verification methodologies to obtain a better knowledge of the skills of a forecasting system. In particular, a short range precipitation forecasting system based on MM5 at 12 km coupled with LAPS is studied in a convective precipitation event that took place in NE Iberian Peninsula on October 3rd 2008. For this purpose, a variety of verification methods (dichotomous, recalibration, scale decomposition and object oriented methods) are used to verify this case study. At the same time, different precipitation analyses are used in the verification process: some obtained by using rain gauges and others obtained by interpolating radar data using different techniques (nearest neighbour, maximum value and mean box).

Email: jmoremeteo.cat

PO31

Veri f cation and statistical properties of COSMO-I7 QPF

Maria Stefania Tesini

ARPA-SIMC /CIMA, Bologna, Italy

Limited Area Models are able to produce Quantitative Precipitation Forecast (QPF) with high spatial and temporal resolution, showing a large variability also in the amount of rain falling in a restricted area. However frequent errors in time and space positioning make difficult a grid-point based employment of models QPF. In order to appreciate the properties and the additional information provided by LAMs respect to coarser resolution models we devised a strategy based on the aggregation of observations and forecasts that fall within a predefined geographical area, several times wider than the model grid-box.

The first aim of this work is the assessment of observed and forecast precipitation climatology over the well-defined areas by the study of the distribution function (pdf) and the evaluation of summarizing quantities such as mean, maximum values and quantiles, in order to analyse the capability of the models in reproducing precipitation statistical properties. Moreover we evaluate the “day-by-day” quality of the QPFs, making use both of descriptive methods and of quantitative measure (such as POD,FAR,TS and BIAS) in each defined area, throughout the selected period and separating according thresholds deduced by the climatological results obtained in the previous point.

We will present how COSMO-I7 (7 km horizontal resolution) performs on the Italian territory, also in comparison with ECMWF model, focusing on relevant rain events useful for the Italian Civil Protection monitoring alert service.

Email: mstesini@arpa.emr.it

PO32

The forecaster's added value in QPF

Marco Turco

ARPA Piemonte, Torino, Italy

To the authors' knowledge there are relatively few studies that try to answer this topic: "Are humans able to add value to computer-generated forecasts and warnings ?". Moreover, the answers are not always positive. In particular some postprocessing method is competitive or superior to human forecast (see for instance Baars et al., 2005, Charba et al., 2002, Doswell C., 2003, Roebber et al., 1996, Sanders F., 1986).

Within the alert system of ARPA Piemonte it is possible to study in an objective manner if the human forecaster is able to add value with respect to computer-generated forecasts. Every day the meteorology group of the Centro Funzionale of Regione Piemonte produces the HQPF (Human QPF) in terms of an areal average for each of the 13 regional warning areas, which have been created according to meteo-hydrological criteria. This allows the decision makers to produce an evaluation of the expected effects by comparing these HQPFs with predefined rainfall thresholds. Another important ingredient in this study is the very dense non-GTS network of rain gauges available that makes possible a high resolution verification.

In this context the most useful verification approach is the measure of the QPF and HQPF skills by first converting precipitation expressed as continuous amounts into ''exceedance'' categories (yes/no statements indicating whether precipitation equals or exceeds selected thresholds) and then computing the performances for each threshold. In particular in this work we compare the performances of the latest three years of QPF derived from two meteorological models COSMO-I7 (the Italian version of the COSMO Model, a mesoscale model developed in the framework of the COSMO Consortium) and IFS (the ECMWF global model) with the HQPF. In this analysis it is possible to introduce the hypothesis test developed by Hamill (1999), in which a confidence interval is calculated with the bootstrap method in order to establish the real difference between the skill scores of two competitive forecast.

It is important to underline that the conclusions refer to the analysis of the Piemonte operational alert system, so they cannot be directly taken as universally true. But we think that some of the main lessons that can be derived from this study could be useful for the meteorological community. In details, the main conclusions are the following:
- despite the overall improvement in global scale and the fact that the resolution of the limited area models has increased considerably over recent years, the QPF produced by the meteorological models involved in this study has not improved enough to allow its direct use: the subjective HQPF continues to offer the best performance;
- in the forecast process, the step where humans have the largest added value with respect to mathematical models, is the communication. In fact the human characterisation and communication of the forecast uncertainty to end users cannot be replaced by any computer code;
- eventually, although there is no novelty in this study, we would like to show that the correct application of appropriated statistical techniques permits a better definition and quantification of the errors and, mostly important, allows a correct (unbiased) communication between forecasters and decision makers.

Email: marco.turco@apra.piemonte.it

PO33

Verification of statistical forecasts of low visibility at Amsterdam Airport

Daan Vogelezang, Nico Maat, and Janet Wijngaard

KNMI, Netherlands

Accurate, reliable and unambiguous information concerning the actual and expected (low) visibility conditions at Amsterdam Airport Schiphol is very important for the available operational flow capacity. Visibility forecast errors have therefore a negative impact on safety and operational expenses. KNMI has performed an update of the visibility forecast system in close collaboration with the main users of the forecasts (Air Traffic Control, the airport authorities and KLM airlines). This automatic forecasting system consists of a Numerical Weather Prediction Model (Hirlam) with a statistical post processing module on top of it. Output of both components is supplied to a human forecaster who adds resolution and accuracy to it and tailors it to the user's requirements by issuing a special probabilistic forecast bulletin.

Probabilities for Runway Visual Range (RVR) are calculated whereas formerly only the Meteorological Optical Range (MOR) values were forecasted. Since RVR depends on both MOR and the local Background Luminance, a (deterministic) statistical forecast for the latter had to be developed.

A second improvement was achieved by calculating joint probabilities for specific combinations of visibility and cloud base height for thresholds which have direct impact on the flow capacity at the airport. Formerly separate forecasts for visibility and cloudbase were combined afterwards, assuming full dependence between both.

Verification of the modified system shows strongly increased reliability (on dependent data) and enhanced resolution for several visibility thresholds and lead times. Verification with independent data is underway, but the first results are promising.

Finally a simple guideline model is developed that shows how to optimize a threshold percentage, in case the users (i.e. Air Traffic Managers) have to make a categorical choice / decision from the full probabilistic visibility forecast. It is shown that the performance of the forecast system combined with a user-specific sensitivity to false alarms and misses can result in a cost-optimal decision threshold percentage.

Email: daan.vogelezang@knmi.nl

PO34

Veri f cation of ensemble forecast using the physical parametrization schemes of WRF model during the Changma period over Korea

Ji-Won Yoon, Yong Hee Lee, Jong-Chul Ha, Hee choon Lee, Dong-Eon Chang

National Institute of Meteorological Research, Seoul, Korea

In this study, we composed the Ensemble Prediction System (EPS) with 120 ensemble members by using the combinations of different physical parameterization schemes. We conducted the numerical simulation with this EPS during the Changma period (from 25 June to 10 July 2006), four times a day ie, 00, 06, 12, 18UTC for each day. The simulated 6-h accumulated precipitation amounts were verified against with 610 Automatic Weather Station (AWS) rain gauge data over Korea.

In terms of the equitable threat score (ETS), we found that the combination of NCEP 3 class microphysics, BMJ cumulus parameterization and MRF PBL scheme revealed the best forecast skill for both light rainfall event (>1mm/6hr) and moderate rainfall event (>25mm/6hr) have. The EPS using the ensembles of microphysics showed more sensitivity for light rainfall events, while the experiment using the ensembles of cumulus parameterization scheme showed more sensitivity for moderate rainfall events.

We assessed the ensemble forecast of the cumulus parameterization schemes by using the Receiver Operating Characteristic (ROC), reliability diagram, and odds ratio. For light rainfall events, ensemble forecast of NOC (No Cumulus), KF2 (Kain-Fritsch II) and BMJ (Betts-Miller-Janjic) scheme combination showed similar pattern in ROC area. On the other hand, the ensemble forecast using the KF2 scheme showed different pattern compared with the NOC and BMJ scheme for the moderate rainfall events. Overall, the ensemble forecast of KF2 scheme combination showed low reliability, especially, it showed poor skill for the range of high probability forecast. Ensemble forecast of BMJ scheme combination showed high odd ratio. Consequently, we found that ensemble forecast of BMJ scheme combination showed better forecast skill than other ensemble of physical schemes.

- Odds ratio is forecast skill that can be judged by comparing the odds of making a good forecast (a hit) to the odds of making a bad forecast (a false alarm)

Email: jwyoon@metri.re.kr

PO35

Comparisons of global and regional ensemble prediction systems at NMC

Xiaoli Li, Hua Tian, Guo Deng

National Meteorological Center (NMC), China Meteorological Administration (CMA)

The performance comparison of global ensemble prediction system (GEPS) and regional ensemble prediction system (REPS) has recently been given great attention at several operational numerical forecast centers. The quasi-operational GEPS at NMC has been upgraded since December 2006 by the use of the breeding of growing mode (BGM) as initial perturbation strategy. Since 2005 World Weather Research Programme (WWRP) has launched Beijing 2008 Olympic Games Meso-scale Ensemble Prediction Research and Development Project (B08RDP), based on B08RDP and practical needs of short-range weather forecasting service for Beijing 2008 Olympics Games, a regional ensemble prediction system (REPS) based on Weather Forecasting and Research (WRF) Model has been developed at NMC. The initial perturbation strategy of the REPS is BGM, and lateral boundary conditions are from the GEPS. The multi-physics technique is used to represent the model perturbation in the REPS. Although both GEPS and REPS have provided the useful information for weather forecasting service in the summer of 2008 from the viewpoint of forecasters, the objective comparison of GEPS and REPS is still of interest to investigate.

The 36-day (July 21-August 24 2008) forecast results from both systems are selected to be validated. The verification is performed at stations to avoid the impacts of systematic error of analysis. The following scores are used to compare the general skill, and reliability and resolution attribute of two systems as well: 1) continuous ranked probability score (CRPS) and its decomposition (reliability and resolution); 2) reduced centered random variables (RCRV); 3) Brier skill score (BSS) and its decomposition; 4) relative operating characteristic (ROC) curve and area under ROC (AROC). The bootstrap resampling technique has been applied into the above-mentioned scores to determine the significance of skill difference of GEPS and REPS.

Results indicate that compared to GEPS, the REPS generally performs significantly better than GEPS for the short-range precipitation forecast. The decomposition of CRPS and BSS show that the advantages of REPS over GEPS come from its better reliability and resolution attributes both. Also, AROC score shows the better discrimination ability in REPS. The further investigation of reliability difference between two systems can be found in the bias and dispersion terms of RCRV, showing that REPS has significantly less bias and better dispersion attribute. The above results are based on the preliminary comparisons, the comprehensive validation still need to be done in the future.

Email: lixl@cma.gov.en

P36

Another look at the contingency tables: Scores based on Manhattan distances in the error space

Joel Stein and Marielle Amodei

Meteo-France, Toulouse, France

Alternative presentation of scores is based on the Manhattan distance in the phase space of the forecasts. The key factor is represented by the ratio of the weights assigned to misses and false alarms. This ratio is 1 for the Heidke skill score, is equal to the ratio of the number of non-events to the number of events for the true skill statistics. A score based on the deterministic limit leads to assign this ratio to 2. Some other values can be found by taking into account the economic value.

Applications to the Finley tornadoes and the comparison of two quantitative precipitation forecasts performed by operational models in Meteo-France show the interest of the graphical representation in this error space to collect the qualities and the drawbacks of the forecasts. The choice of the metrics used to measure both types of errors strongly influences the vision of the quality of the forecast. The skill scores deduced from these distances summarize the quality of a forecast. Thus, the basic assumptions made to choose one score or another are easy to understand in this framework. This presentation covers at the same time dichotomous and polichotomous cases. Moreover, the Finley tornadoes allows the discussion of the limit for rare events.

Email: joel.stein@meteo.fr

Workshop Participants

Matias Armanini	Argentina	P22, tutorial
Elizabeth Ebert	Australia	O6.7, P8
Chermelle Engel	Australia	O6.3
Benedikt Bica	Austria	O3.4
Theresa Gorgas	Austria	O3.3
Alexander Kann	Austria	-
Guenther Mahringer	Austria	P3
Stefan Schneider	Austria	-
Christoph Zingerle	Austria	-
Pascal Mailier	Belgium	O8.3, P19
Nilza Barros da Silva	Brazil	tutorial
David Mendes	Brazil	P14, P15
Monika Bailey	Canada	P23, tutorial
Timothy Bullock	Canada	-
Barbara Casati	Canada	O6.6, O8.2
Darlene Langlois	Canada	-
Francois Lemay	Canada	tutorial
Michel Moreau	Canada	-
Tom Robinson	Canada	-
Marcel Vallee	Canada	P17
Laurence Wilson	Canada	O1.1, O3.5
Yin Kong Leung	China	O2.4, tutorial
Xiaoli Li	China	P35
Lovro Kalin	Croatia	O4.2
Zoran Pasaric	Croatia	O10.3
Daniela Rezacova	Czech Republic	-
Petr Zacharov	Czech Republic	P21
Kristian Pagh Nielsen	Denmark	P7
Xiaohua Yang	Denmark	-
Anna Ghelli	ECMWF	-
Girmaw Bogale	Ethiopia	P25, tutorial
Gezu Mengistu	Ethiopia	P16
Alberto Blanco	Finland	tutorial
Juhani Damski	Finland	O1.1
Kalle Eerola	Finland	P26
Otto Hyvärinen	Finland	O3.2
Sami Niemelä	Finland	O7.6
Pertti Nurmi	Finland	P11
Heikki Pohjola	Finland	-
Petra Roiha	Finland	P20, tutorial
Marielle Amodei	France	O7.2, P36
Christine Le Bot	France	tutorial
Ulrich Damrath	Germany	O7.1
Martin Goeber	Germany	O5.2

Anna Lindenberg	Germany	P29
Marcus Paulat	Germany	O2.5
Matthias Zimmer	Germany	O7.7
Flora Gofa	Greece	tutorial
Kadarsah Binsukandar Riadi	Indonesia	P24, tutorial
Stefano Mariani	Italy	O6.4
Chiara Marsigli	Italy	O7.4
Arturo Pucillo	Italy	tutorial
Adriano Raspanti	Italy	P5
Maria Stefania Tesini	Italy	P31
Marco Turco	Italy	P32
Yasutaka Ikuta	Japan	P6
Luc Yannick Andreas Randriamarolaza	Madagascar	tutorial
Bun-Liong Saw	Malaysia	tutorial
Premchand Goolaup	Mauritius	tutorial
Sergio Buque	Mozamique	tutorial
Kees Kok	Netherlands	O7.3
Daan Vogelezang	Netherlands	P33
James McGregor	New Zealand	P2
Dag Johan Steinskog	Norway	-
Bente Marie Wahl	Norway	tutorial
Sultan Al-Yahyai	Oman	P4, tutorial
Ata Hussain	Pakistan	P27
Juan Bazo	Peru	P12, tutorial
Marek Jerczynski	Poland	P9
Joanna Linkowska	Poland	tutorial
Joao Rio	Portugal	tutorial
Sei-Young Park	Rep. Korea	P34, tutorial
Thizwilondi Robert Maisha	South Africa	O1.4, tutorial
Jordi Mercarder-Carbo	Spain	tutorial
Jordi More	Spain	P30
Carlos Santos	Spain	O3.1
Francis Schubiger	Switzerland	O7.5
Jonathan Eden	UK	P13
Christopher Ferro	UK	O10.1
Ian Jolliffe	UK	O4.1
Marion Mittermaier	UK	O5.5, O6.5, P10
Ewan J.O'Connor	UK	P18
Michael Sharpe	UK	O5.3
David Stephenson	UK	O5.1
Simon Thompson	UK	-
Clive Wilson	UK	O1.2, O5.4
Vitalii Shpyg	Ukraine	tutorial
Barbara Brown	USA	O2.2, O6.1

James Brown	USA	O2.3, P1
Tressa Fowler	USA	O1.3, O2.1
Eric Gilleland	USA	O6.2
Deborah Glueck	USA	O10.2
Oluseun Idowu	USA	P28
Johannes Jenkner	USA	O10.4
Simon Mason	USA	O8.1
Matthew Pocernich	USA	O3.6
Nanette Lomarda	WMO	-
Tirivanhu Muhwati	Zimbabwe	tutorial

World Weather Research Programme (WWRP)
Report Series

Sixth WMO International Workshop on Tropical Cyclones (IWTC-VI), San Jose, Costa Rica, 21-30 November 2006 (WMO TD No. 1383) (WWRP 2007 - 1).

Third WMO International Verification Workshop Emphasizing Training Aspects, ECMWF, Reading, UK, 29 January - 2 February 2007) (WMO TD No. 1391) (WWRP 2007 - 2).

WMO International Training Workshop on Tropical Cyclone Disaster Reduction (Guangzhou, China, 26 - 31 March 2007) (WMO TD No. 1392) (WWRP 2007 - 3).

Report of the WMO/CAS Working Group on Tropical Meteorology Research (Guangzhou, China, 22-24 March 2007) (WMO TD No. 1393) (WWRP 2007 - 4).

Report of the First Session of the Joint Scientific Committee (JSC) for the World Weather Research Programme (WWRP), (Geneva, Switzerland, 23-25 April 2007) (WMO TD No. 1412) (WWRP 2007 – 5).

Report of the CAS Working Group on Tropical Meteorology Research (Shenzhen, China, 12-16 December 2005) (WMO TD No. 1414) (WWRP 2007 – 6).

Preprints of Abstracts of Papers for the Fourth WMO International Workshop on Monsoons (IWM-IV) (Beijing, China, 20-25 October 2008) (WMO TD No. 1446) (WWRP 2008 – 1).

Proceedings of the Fourth WMO International Workshop on Monsoons (IWM-IV) (Beijing, China, 20-25 October 2008) (WMO TD No. 1447) (WWRP 2008 – 2).

WMO Training Workshop on Operational Monsoon Research and Forecast Issues – Lecture Notes, Beijing, China, 24-25 October 2008 (WMO TD No. 1453) (WWRP 2008 – 3).

Expert Meeting to Evaluate Skill of Tropical Cyclone Seasonal Forecasts (Boulder, Colorado, USA, 24-25 April 2008) (WMO TD No. 1455) (WWRP 2008 – 4).

Recommendations for the Verification and Intercomparison of QPFS and PQPFS from Operational NWP Models – Revision 2 - October 2008 (WMO TD No. 1485) (WWRP 2009 - 1).

Strategic Plan for the Implementation of WMO's World Weather Research Programme (WWRP): 2009-2017 (WMO TD No. 1505) (WWRP 2009 – 2).