Joint Working Group on Forecast Verification Research

Report to WGNE-28, Nov 8, 2012

Laurence Wilson

With thanks to the members of the working group

Membership

- Unchanged since last year

Beth Ebert (BOM) (co-chair)	Laurie Wilson (Env. Canada) (co-chair)
Barbara Brown (NCAR)	Barbara Casati (Ouranos)
Pertti Nurmi (FMI)	Martin Göber (DWD)
Caio Coehlo (CPTEC)	Simon Mason (IRI)
Yuejian Zhu (NCEP)	Anna Ghelli (ECMWF)
Joel Stein (Meteo France)	Marion Mittermaier (UKMO)

Outline

- Update on activities (Current projects)
  - Melbourne workshop and followup
  - TC and Cloud documents
- Planned activities
  - Sochi Olympics
  - IPC2
  - Other WMO: Polar prediction project and subseasonal to seasonal prediction project
- Training and Outreach
  - Verification tutorial Melbourne
  - SWFDP related activities
- Progress and challenges in Verification methodology
  - Progress:
    - Improved verification practices
    - Spatial methods increasing
    - New scores for specific or general purposes
    - Verification of “downstream products”
  - Challenges:
    - Observations and use of remote-sensed data
    - Use of model-tainted data as truth (use of the analysis)
    - Verification of “seamless” forecasts
    - Multi-dimensional verification
    - Spatial verification of ensembles
    - User-oriented verification

5th International Verification Methods Workshop

December 1–7, 2011

Bureau of Meteorology, Melbourne , Australia

Australian Government

Bureau of Meteorology

-107 papers, about 40% oral, 60% poster
-3 days
-approximately 150 attendees
TOPICS:

- Verification of high impact weather forecasts and warnings
- Verification of ensembles and probability forecasts
- Spatial forecast verification
- Climate projection evaluation
- Seasonal forecast verification
- Tropical cyclone verification
- Aviation forecast verification
- User issues including communicating verification to decision makers
- Verification tools

Many papers to appear in special issue of MetApps
Information and presentations at: http://cawcr.gov.au/events/verif2011/

JWGFVR

WMO Documents

1. Document on methods for verification of quantitative precipitation forecasts

Available on WWRP web site
http://www.wmo.int/pages/prog/arep/wwrp/

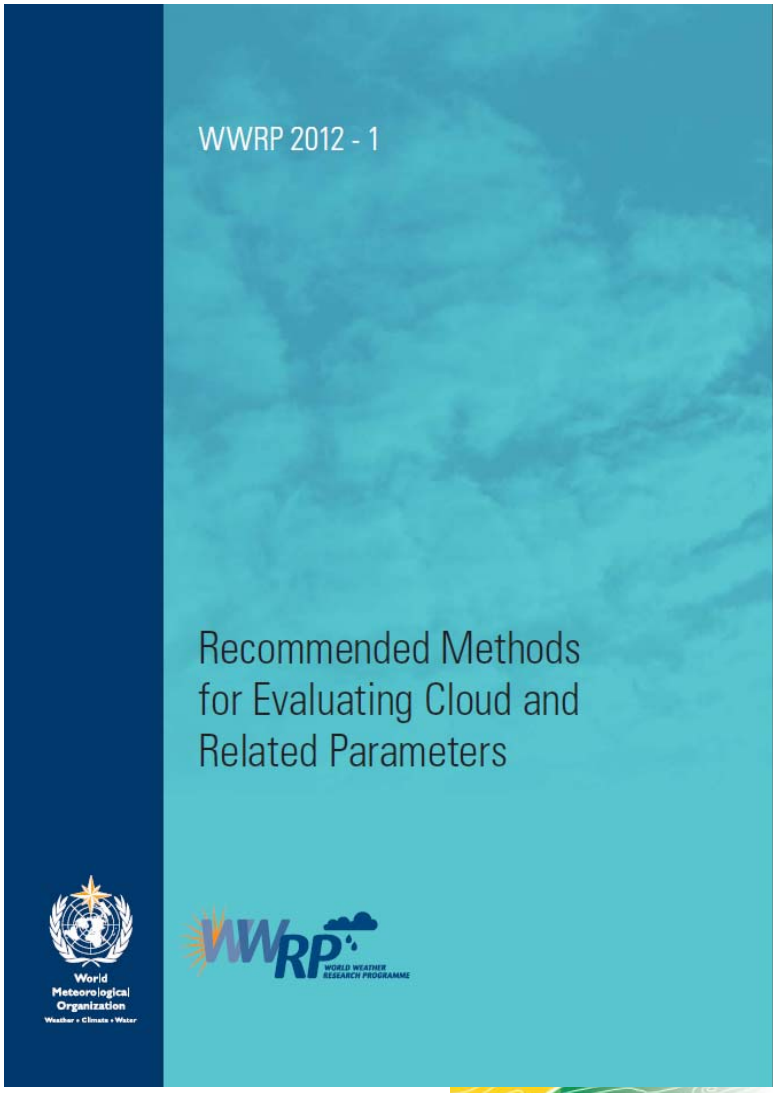
2. **Cloud verification document**:

-NOW available at:
http://www.wmo.int/pages/prog/arep/wwrp/new/documents/WWRP_2012_1_web.pdf

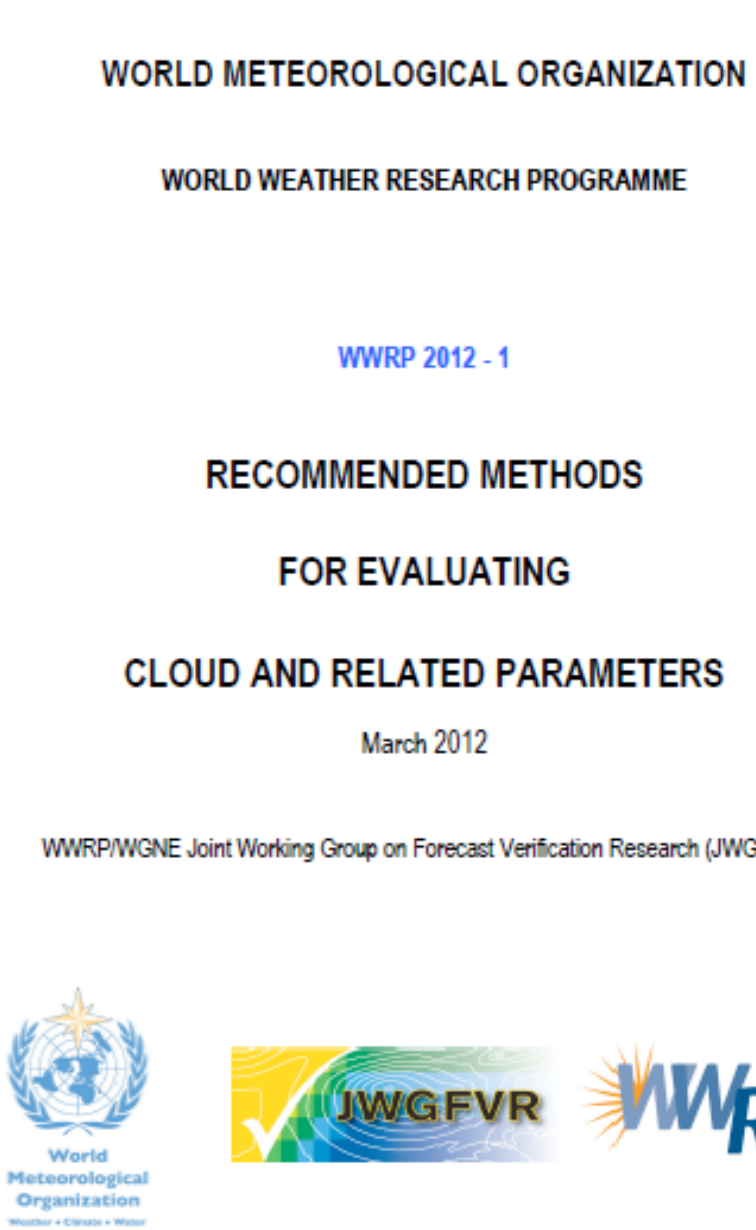
3. Tropical cyclone verification

-Nearly finished. Has been used in training in China

Publish as review papers?



JWGFVR



Cloud verification document

- Some Issues:
  - Systematic sfc obs differences manual vs auto
  - Model tainting of satellite data
  - Use of satellite data and related management
  - To thin or not to thin

JWGFVR

Tropical Cyclone Document

- Outline
  - Current practices, parameter by parameter survey of what centers now do routinely
    - Description of methods of estimating parameters such as intensity
  - Experimental methods (published or about to be), not in general use
    - Mostly for ensemble forecasts
  - Reporting guidelines
  - Summary of recommendations
  - Appendices
- More of a literature survey, with identification of promising methods

Spatial method intercomparison project 2 (ICP2)

- IPC(1) – over US, pcpn only, great success, ~15 papers 09-10
- IPC2 in planning, commitment to action is imminent (Barb Brown's group)
  - Over Europe, pcpn and other variables (wind?)
  - Identify set of interesting cases with necessary high resolution observation datasets (MAP D-Phase?)
  - Test spatial and scale-resolving methods (FSS, SAL, MODE, upscaling, CRA, wavelet method, image warping….
  - Link with SRNWP
- A joint project of JWGFVR and Mesoscale WG

Training and outreach: 5th International Verification Methods Tutorial and Workshop Melbourne, Australia, Dec 1-7 2011

TUTORIALS:

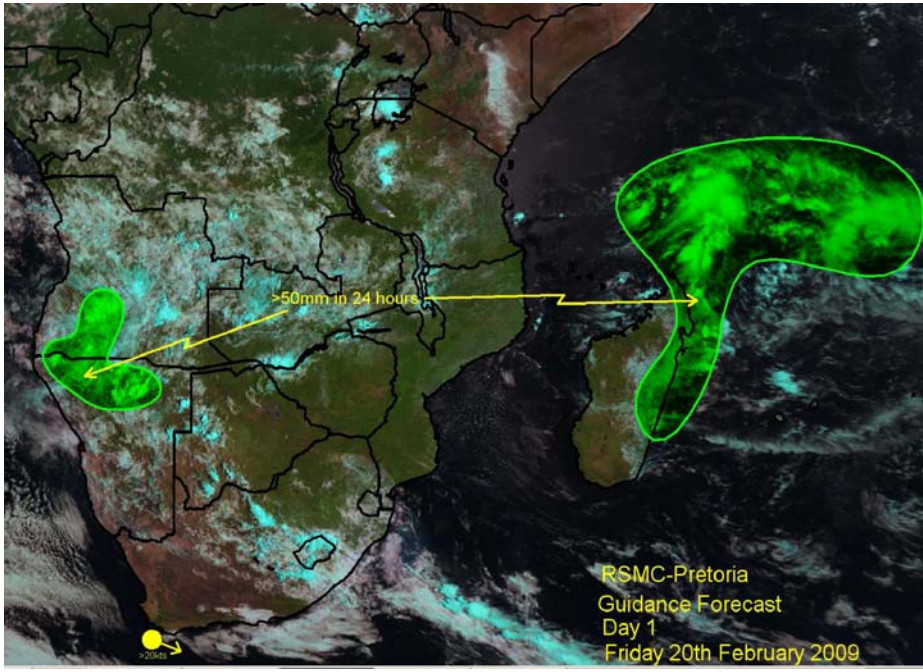
3 days

-basic verification; hands on exercises, students can bring their own data; includes “R” instruction – 34 participants selected from 89 applicants, 44 countries

View from break-out area

Training activities (capacity-building)

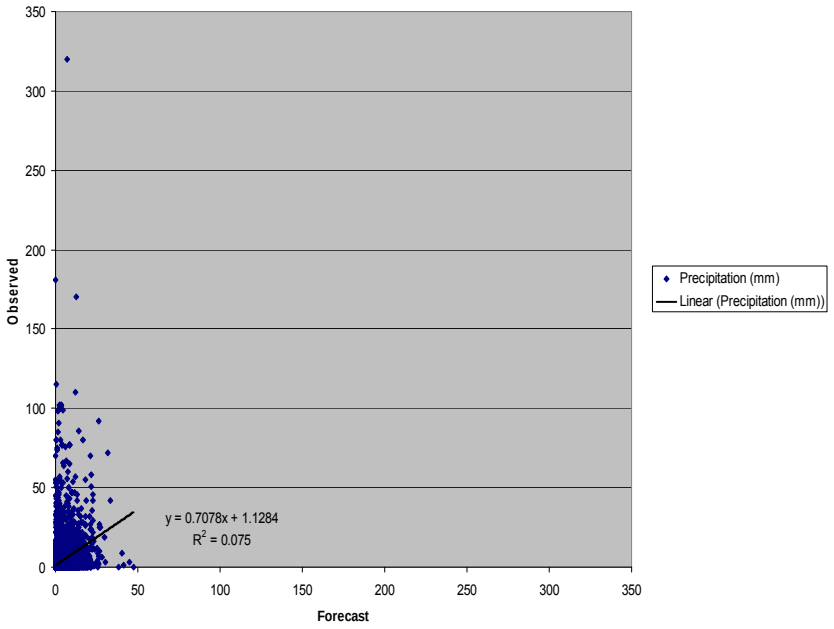
- Training new scientists; training activities in connection with RDPs and FDPs; training the trainers
  - Workshops every 2 years or so
  - Southern Africa SWFDP, 2009-2011
    - Difficulty getting global centers to help with verification
  - Eastern Africa SWFDP
  - South Pacific SWFDDP
  - Planned travelling seminars



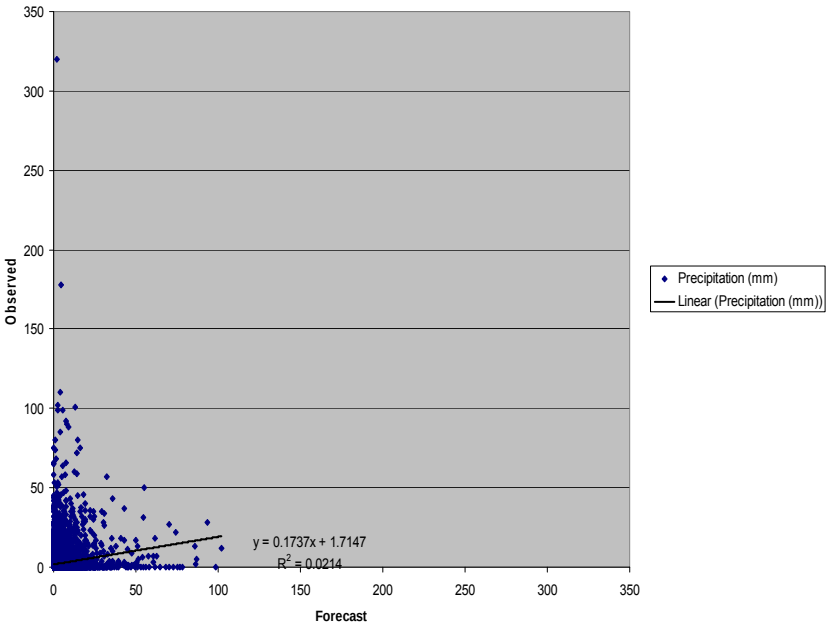
JWGFVR

24 h Precipitation fcsts – E. Africa Day 1

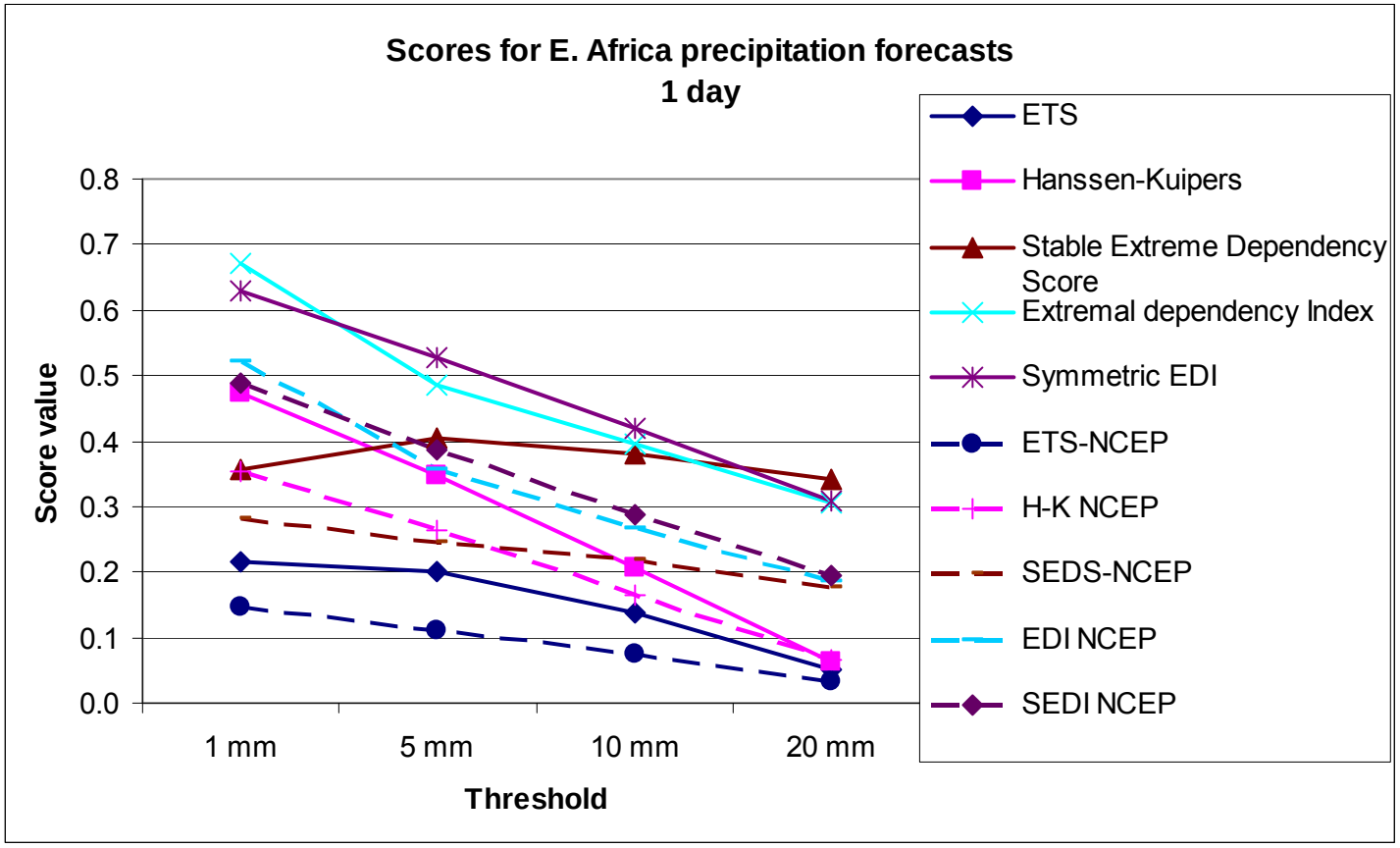
ECMWF



NCEP



Verification of pcpn – E. Africa



Outreach: Web activities

- *Strong* focus of the WG
- EUMETCAL training modules completed and available

  ww.eumetcal.org.uk/eumetcal/verification/www/english/courses/msgcrs/index.htm
- Verification web page

  Google search:
  "forecast verification"
- Vx-discuss: a forum for discussing verification issues
  - Hosted at NCAR
- Sharing of tools: making it easier to do verification
  - MET – MODE
  - R
  - Climate predictability tool software (CPT) – from IRI
  - Spatial methods as available
  - Others as available and agreed

JWGFVR

Progress in Verification

- 1. Lots of evidence of use of CI in verification results (bootstrapping)
- 2. Increased use of diagnostic methods – both spatial and pointwise.
  - FSS, CRA, MODE especially are used outside their development community
- 3. New Scores:-
  - EDS-EDI family
  - SEEPS
  - 2AFC

Verification of extreme, high-impact weather

- **EDS – EDI – SEDS - SEDI** ⇔ **Novelty categorical measures!**

Standard scores tend to zero for rare events

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

H = a / (a+c), hit rate

F = b / (b+d), false alarm rate

p = (a+c) / n, base rate

q = (a+b) / n, relative frequency of forecasted events

$$\mathrm{EDS} = \frac{\log p - \log H}{\log p + \log H} \qquad \mathrm{SEDS} = \frac{\log q - \log H}{\log p + \log H}$$

Ferro & Stephenson, 2010: Improved verification measures for deterministic forecasts of rare, binary events. *Wea. and Forecasting*

Base rate independence ⇔ Functions of H and F

$$\mathrm{EDI} = \frac{\log F - \log H}{\log F + \log H}$$

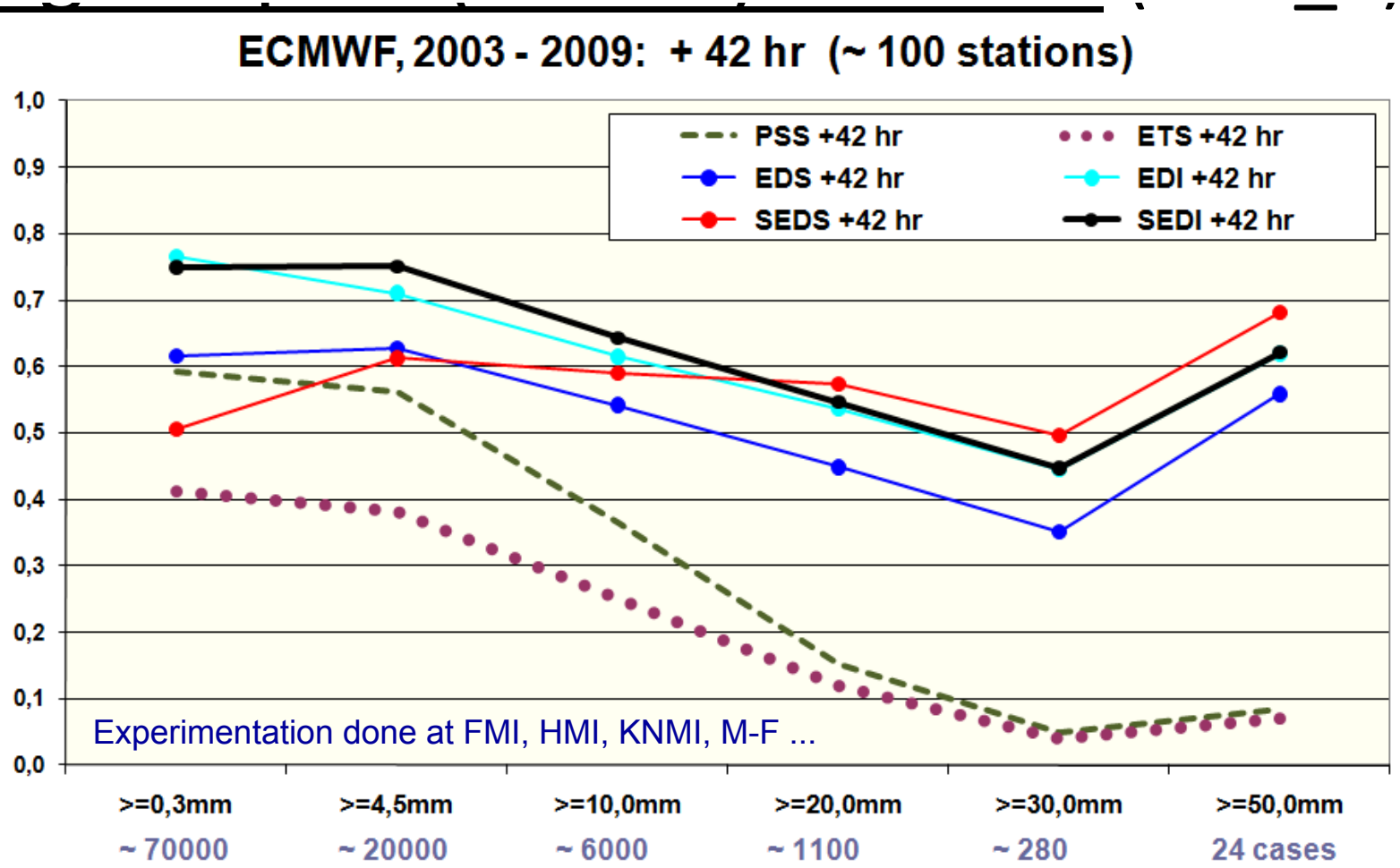
Extremal Dependency Index - EDI

Symmetric Extremal Dependency Index - SEDI

$$\mathrm{SEDI} = \frac{\log F - \log H - \log(1-F) + \log(1-H)}{\log F + \log H + \log(1-F) + \log(1-H)}$$

JWGFVR

Verification of extreme, high-impact weather



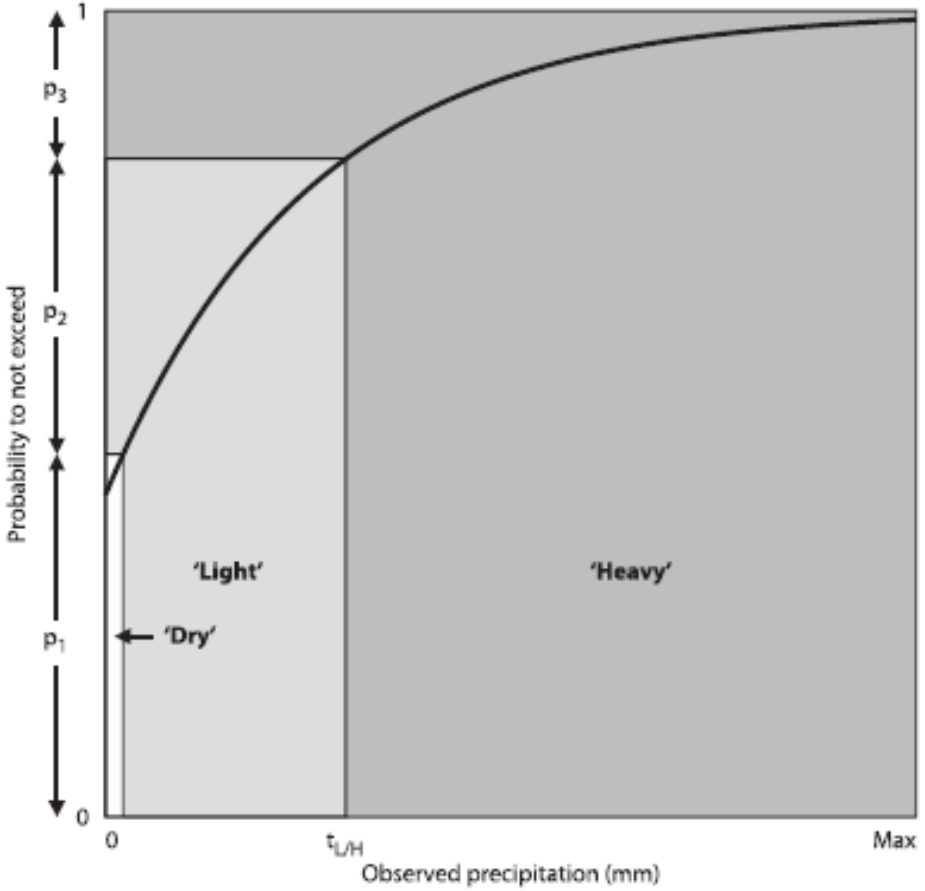
- *More work is needed to assess their potential as scores for severe weather events* **(ECMWF Verif. Sub-group**

JWGFVR

1 - SEEPS = ½ (TSS $_{dry/wet}$ + TSS $_{light/heavy}$) is a 3-category score

SEEPS ⇔ Stable Equitable Error in Probability Space

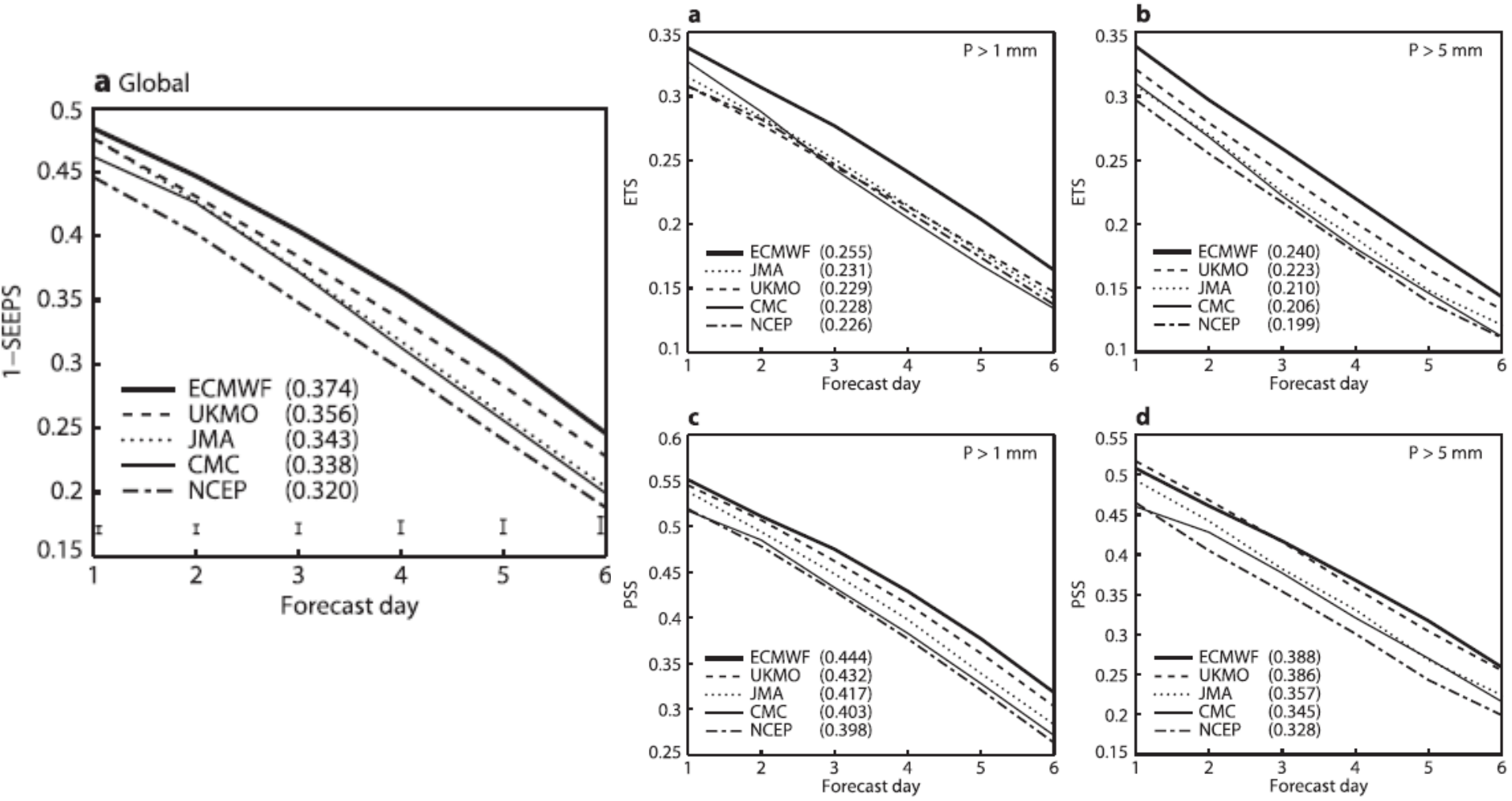
- **M.J. Rodwell et al., 2010: QJRMS, 136,** 1344-1363**.**
- **Derived from LEPS score ⇔ Linear Error in Probability Space**
  - **uses the climatological cumulative distribution function**
- **3 categories: (i) "dry" (ii) "light precipitation" (iii) "heavy precipitation"**
  - **Needs long-term climatological precipitation categories at given SYNOP stations ⇔ Accounts for climate differences between stations**
- **Negatively oriented error measure ⇔ Perfect score =0 => *1 – SEEPS***



- *Status –*
  - **Further testing (Haiden et al, 2012)**
  - **Likely to be proposed for the CBS standard model verification for precip**

$$\mathbf{S} = \frac{1}{2}\begin{Bmatrix} 0 & \frac{1}{1-p_1} & \frac{4}{1-p_1} \\ \frac{1}{p_1} & 0 & \frac{3}{1-p_1} \\ \frac{1}{p_1} + \frac{3}{2+p_1} & \frac{3}{2+p_1} & 0 \end{Bmatrix}.$$

Seeps, ETS, PSS 24 h precipitation test results



Haiden et al 2012 conclusions

- **D+1 to D+6 differences among models are ~ 1 forecast day for Seeps, ETS and PSS**
- **Differences in accuracy between tropics and mid-latitudes are very large: Mid- lat D+6 ~ tropics D+1**
- **Seeps results suggest both overprediction of light rain and underprediction of heavy precipitation**
- **Nearly half of the forecast error at D+1 due to comparing grid box averages with point values.**
- **Upscaling to a common grid by averaging did not affect the results**
- **Need to drop stations with very dry climates from SEEPS for stability reasons – threshold chosen for this didn't affect the model ranking**

2AFC – "Discrimination score" (Mason and Weigel, 2009)

- Consider two cases at a time, one with occurrence of the event, one without:
  - Score is 0 if incorrectly discriminated (wrong one forecast)
  - Score is 1 if correctly selected
  - Score is 0.5 if a toss up (forecasts are equal, both yes or both no)
- General – works for categorical, continuous forecasts and observations, also probabilistic
- Relates directly to trapezoidal area under ROC curve.
- Paper mentions common error of ROC – the binning of data – for discrimination all forecastable values should be considered.
- **DISCRIMINATION: The ability of a forecast system to distinguish those situations leading to the occurrence of an event from those which don't.**
- "The goal of (ensemble) forecasting is to maximize resolution (sharpness) (discrimination) subject to reliability" --Tilman Gneiting

JWGFVR

Verification reported at WGNE 28 – some comments

- Good stuff!
  - Increased discussion and attention to verification
  - Use of fractions skill score to diagnose scales for which there is skill (Investigate some of the other spatial methods)
  - Session on precipitation verification:
    - Only MF included confidence intervals on results.
  - Progress towards surface verification
- Not so good stuff
  - Too little use of (bootstrapped) confidence intervals
  - Love affair with 500 mb verification continues
    - Despite encouraging comments about the importance of surface verification at WGNE 26
- Some suggestions:
  - For verification results that purport to state the accuracy or skill of the model for general user community…..
    - Need to evaluate surface variables
    - Verify with respect to observations
  - For precipitation – use the new SEDS or EDI score especially for higher thresholds

JWGFVR

Challenges in Verification Research

- Observations and use of remotely-sensed data
  - Remote sensing errors
  - Data management issues
- Verification of seamless forecasts
  - Focus on "attributes"? E.g. 2afc and other general scores
- Multi-dimensional verification
  - How to treat the joint distribution of multiple variables?
  - Stratification according to criteria of interest such as flow patterns
  - Examples: MST histogram for ensembles (Wilks, 2004)
    - Multivariate rank histograms (Gneiting et al 2008)
    - Bounding box (Weisheimer et al 2005)
- Spatial verification of Ensembles
  - Some work in this area: e.g Gallus (2009), Zacharov and Rezacova (2009) – FSS; Duc et al (2011) – multi dimension FSS
- User oriented verification
  - Example "balanced scorecard" – may include non-meteorological performance factors
- Verification of high resolution models
  - Data sources, measurements?
  - Standard vs non-standard

JWGFVR

Use of model-tainted data in verification (verification against own analysis)

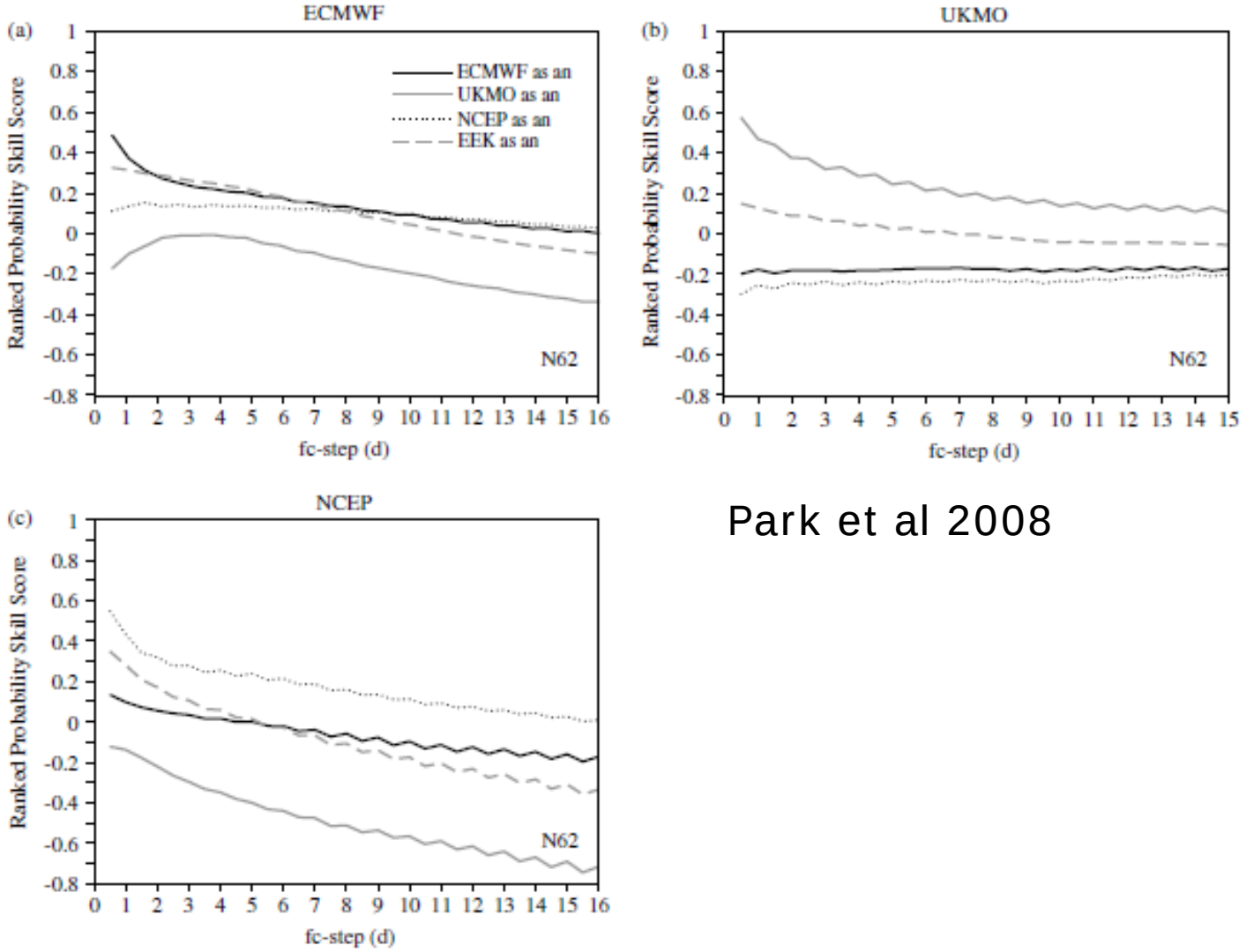
- Discussion motivated in part by suggestions we could use DA system for verif too…
- Recall last year: Park et al (2008) showed for ECMWF, UKMet and NCEP, that best comparative verification found if use own analysis
  - Advantage of own analysis >> real differences in accuracy between eps forecasts
  - Need more evaluation of this effect --
- Effective "pure" "clean" model verification means:
  - Truth data is not influenced by any model preferably, especially not one's own
  - Qc of obs used in verification is independent of models
  - Applies also to reanalyses, used as truth or as "climatology" in skill scores.
  - "representativeness" error
- BUT….
  - Verification against analysis is relatively easy and practical
  - And, the methodology of DA (Variational, KF) has something to offer…
- Alternatives (depends on whether comparative or not):
  - Verify only where grid points supported by sufficient obs data
  - Use model-independent analysis methods and qc. E.g. US 4 km precip analysis
  - If comparing models – use ensemble of analyses, random selection of verifying analysis.

JWGFVR

Any questions?

- What else does WGNE want help with from JWGFVR?
  - Aside from finishing the long-awaited TC document

Verification results depend on analysis used



Park et al 2008

Why is there a problem?

- Most likely due to differences between model climatologies
  - Includes scaling effects, smoothing of small scales
  - Differences between model climatologies, which are carried by the background field.

Alternatives

- If one must verify against an analysis….
  - May be OK for diagnostic studies when only one model is involved.
- For comparison, selecting a single analysis is unfair. To make it fairer:
  - Each own analysis (WMO method) (but still will overstate accuracy)
  - Use analysis which is independent of all models in the comparison
  - Use analysis which doesn't depend on a model background – verify only where there is data
  - Use an ensemble of analyses from all models in the comparison
  - Randomly select the verifying analysis from among the analyses
- But better still, verification against observations (not qc'd with respect to model)
  - E.g. precipitation data from special networks, radar-enhanced precip analyses
  - Remotely sensed data – do the retrieval algorithms depend on a model?
- Model tainted data issue also affects reanalysis data used as climatology – may have a negative effect on results
  - Long term climatology should be from observations

JWGFVR

Thanks!

JWGFVR

Aerosol Verification