Forecast Verification Research - state of the art and future challenges

By: WWRP/WGNE Joint Working Group Forecast Verification Research

JWGFVR

Presenter: Martin Göber, DWD, Hans-Ertel-Centre for Weather Research

THORPEX-HIW Workshop, 18-19 March 2013

“Summary”

State:

“*The state of forecast verification is clearly improving”.* (Ebert et al., 2013)

Challenges:

“All post-THORPEX projects should have a clearly articulated and user-oriented verification plan”. (L. Wilson, personal com.)

Forecast Verification Research - state of the art and future challenges

By: WWRP/WGNE Joint Working Group Forecast Verification Research

Presenter: Martin Göber, DWD, Hans-Ertel-Centre for Weather Research

THORPEX-HIW Workshop, 18-19 March 2013

Working Group

Members

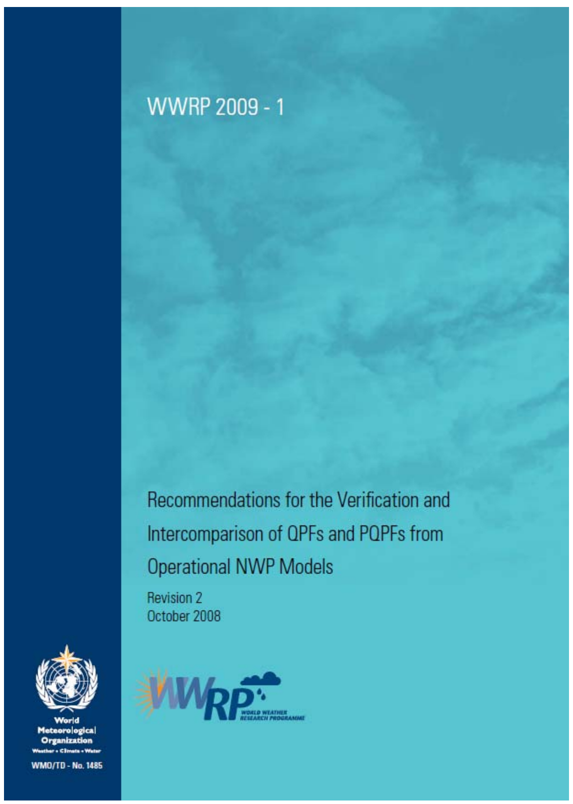
- ❖Beth Ebert (BOM, Australia, co-chair)
- ❖Laurie Wilson (CMC, Canada, co-chair)
- Barb Brown (NCAR, USA)
- Barbara Casati (Ouranos, Canada)
- Caio Coelho (CPTEC, Brazil)
- Anna Ghelli (ECMWF, UK)
- Martin Göber (DWD, Germany)
- Simon Mason (IRI, USA)
- Marion Mittermaier (Met Office, UK)
- Pertti Nurmi (FMI, Finland)
- Joel Stein (Météo-France)
- Yuejian Zhu (NCEP, USA)

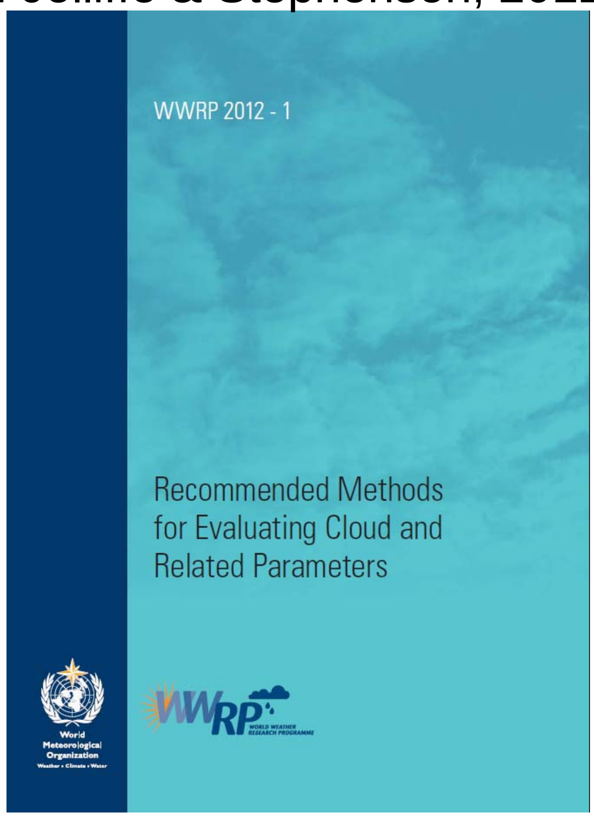
Aims

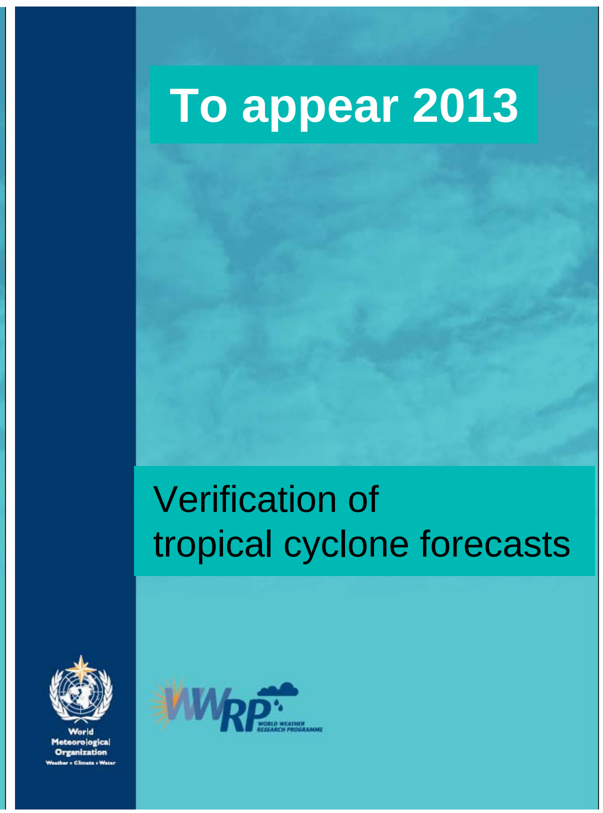
- Develop and promote new verification methods
- Training on verification methodologies
- Ensure forecast verification is relevant to users
- Encourage sharing of observational data
- Promote importance of verification as a vital part of experiments
- Promote collaboration among verification scientists, model developers and forecast providers

Resources

- 2013, special issue of *Meteorological Applications* on 5th verification workshop
    - including: Ebert et al.: **Progress and challenges in forecast verification.**
- „Spatial forecast verification methods **I**nter **C**omparison **P**roject ICP"
    - „AMS-Special collection": http://journals.ametsoc.org/page/ICP
    - Review article: Gilleland et al., 2009:., Intercomparison of spatial forecast verification methods. *Wea. Forecasting*
- google „forecast verification"→ http://www.cawcr.gov.au/projects/verification/
- new edition of THE book: Jolliffe & Stephenson, 2012: *Forecast verification*







JWGFVR

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:

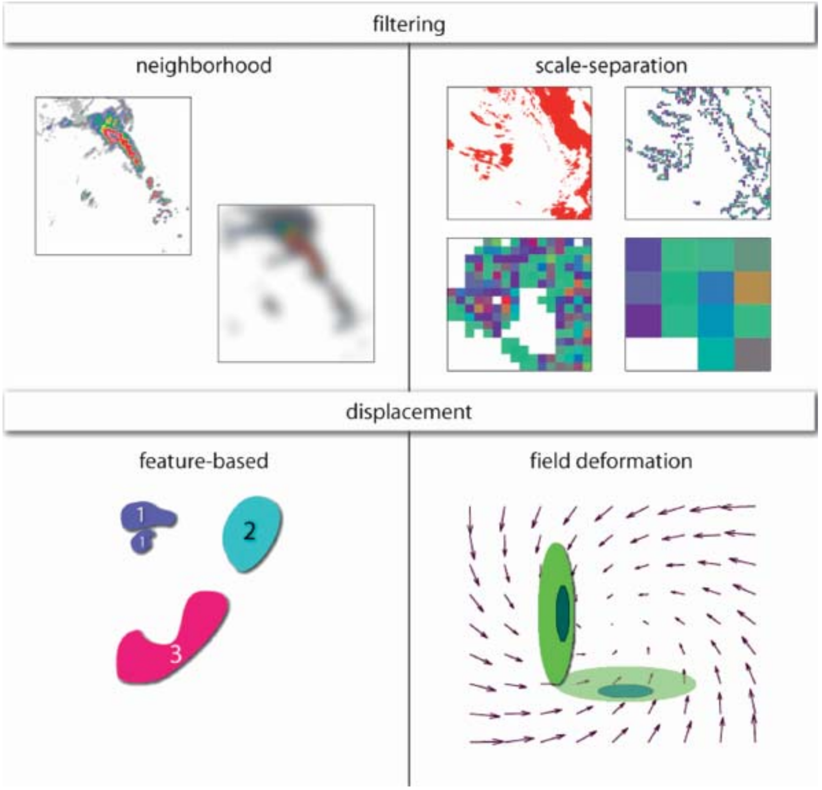
- Verification of high impact weather forecasts and warnings
- Verification of ensembles and probability forecasts
- Spatial forecast verification
- Climate projection evaluation
- Seasonal forecast verification
- Tropical cyclone verification
- Aviation forecast verification
- User issues including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

Next workshop likely in spring 2014 (p = 2*42 %)

- Improved verification practise
  - Confidence intervals and distribution of errors used widely
  - Yet large uncertainty for HIW because of small samples → Bayesian approach might alleviate problem ?
- New diagnostic methods becoming mainstream
  - Should detect nature of forecast errors and give clues about sources of errors
  - Use of spatial verification methods has become far more frequent and widespread, especially for high resolution NWP, where traditional verification methods are strongly affected by „double penalty“ when weather features are predicted, but at the wrong place (time) → proliferation of new spatial ver methods
  - Well suited for HIW, since often well defined features

Spatial Verification Method Intercomparison Project ICP

- International comparison of many new spatial verification methods
- Phase 1 (precipitation) completed
  - Methods applied by researchers to same datasets (precipitation; perturbed cases; idealized cases)
  - Subjective forecast evaluations
  - *Weather and Forecasting* special collection 2009-2010
- ICP 2 in planning stage
  - Complex terrain
  - MAP D-PHASE / COPS dataset
  - Wind and precipitation, timing errors



- Further challenges for diagnostic verification
  - Timing error needs greater attention
    - Neighbourhood in time (Weusthoff , 2011)
    - Spatiotemporal errors of coherent weather features (Bullock, 2011)
    - SAL+Time (Zimmer and Wernli, 2011)
    - Lead time for warnings (Wilson, 2011)
  - Determining ensemble spread of object attributes (Gallus, 2009)
  - Quantifying location based uncertainty (“radius of relibability”, Ebert 2011)

- New scores for difficult issues:
  - Scores for rare (often HIW) event forecasts
  - Scores for consistent performance information across forecasts of different types (continuous vs categorical vs probabilistic)
  - Aggregated scores for precipitation, which enable consistent forecast system monitoring (“SEEPS”,

Extreme events

- ***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

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

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

Base rate independence ⇔ Functions of H and F

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

Extremal Dependency Index - EDI

Symmetric Extremal Dependency Index - SEDI

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

Score for different types of forecasts

- 2AFC – "General Discrimination Score" (Mason and Weigel, 2009)
- *"The goal of (ensemble) forecasting is to maximise resolution (sharpness, discrimination) subject to reliability".* Gneiting (2008)
- **Discrimination:** the ability of a forecast system to distinguish those situations leading to the occurrence of an event from those which don't
- Works for categorical, continuous and probabilistic forecasts and observations

- Make greater use of exciting new observations
  - Remotely sensed data (A-train satellite constellation, many countries polarise their network, 3D-water vapour from GPS)
  - „smart devices“ allow road vehicles, pedestrians, … to measure and transmit atmospheric observations (good for data sparse areas, BUT: quality control, how to use ?)
  - In multi-sensor analyses strengths of one instrument can compensate for weakness of other (blended radar-rain gauge)
    - E.g. OPERA program (Huuskonen et al. ,2012) → what are errors in these analysis‘ and how to account for those in verification ?
  - Contined efforts needed for harmonisation of obs across national borders and disciplines, make them accessible for use in verification

- Synergy of verification with data assimilation
  - Advantages of analysis from DA: efficient and convenient qc, data on a regular and model grid
  - Shortcomings: obs with large deviation from model background wrongly rejected, because of model errors or observed local effects
  - Resulting in: Park (2008) showed (with THORPEX data), that the advantage from verifying against one‘s own analysis can be as large as the real differences between accuracy among different models (see also Hamill 2011)
    - → use ensemble of analysis or determine uncertainty of analysis (Gorgas and Dorninger 2012)
    - Verify only those grid boxes with sufficient obs

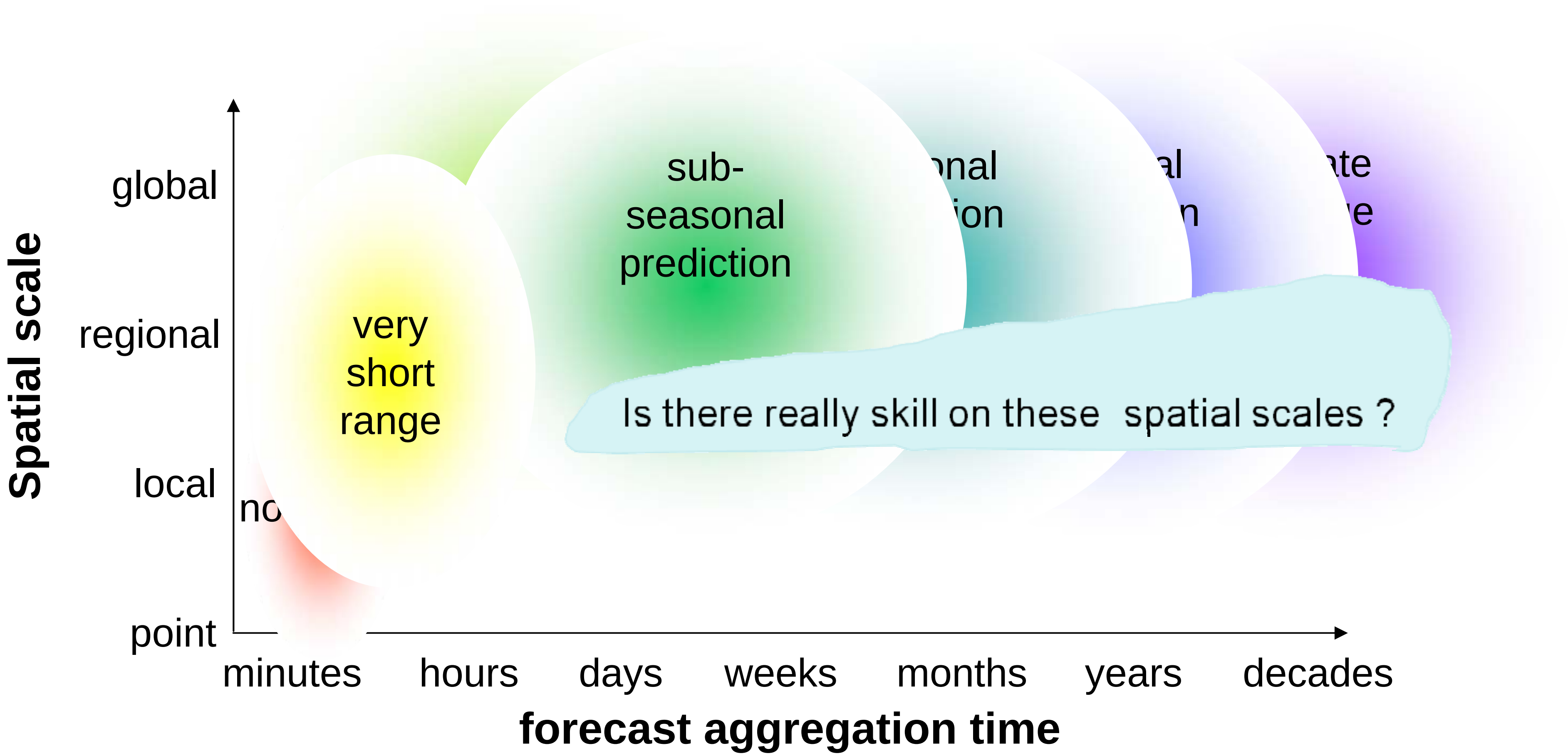
- Verification of **seamless** predictions
  - „seamless“: running same model over all time scales →optimisation of model for one application must be checked across the seamless system for other applications
  - Needs to be applicable to both deterministic and probabilistic forecasts
  - Same scores, but different scales ? →

JWGFVR

Goal size= spatial aggregation scale

Shooting distance = lead and aggregation time

Seamless verification



Adapted from Ebert, 2009

- Seamless verification (cont.)
  - Modelling perspective – is my model doing the right thing?
    - Process approaches
      - LES-style verification of NWP runs (first few hours)
      - T-AMIP style verification of coupled / climate runs (first few days)
      - Single column model
    - Statistical approaches
      - Spatial and temporal spectra
      - Spread-skill
      - Marginal distributions (histograms, etc.)

- Seamless verification (cont.)
  - User perspective – can I use this forecast to help me make a better decision?
    - Neighborhood approaches - spatial and temporal scales with useful skill
    - Generalized discrimination score (Mason & Weigel, *MWR* 2009)
      - consistent treatment of binary, multi-category, continuous, probabilistic forecasts
    - Calibration - accounting for space-time dependence of bias and accuracy?
    - Conditional verification based on larger scale regime
    - Extreme Forecast Index (EFI) approach for extremes

- Multidimensional verification – how to treat the joint distribution of multiple variables
  - Multivariate rank histogram (Gneiting et al 2008)
  - Minimum spanning tree histogram for ensembles (Wilks 2004)
  - Bounding boxes (Weisheimer et al. 2005)
  - Growing need for verification of physically processes (e.g. heat budget terms instead of temperature only)
  - Joint evaluation, e.g. low soil moisture and hot temperatures (Seneviratne, 2011)

- Verification of “downstream products”
  - predicted power generation (Madsen et al., 2005)
  - “Ramping events” (Pocernich, 2010)
  - Value of diagnostic verification for river stage forecasts (Welles and Sorooshian, 2009)
  - comprehensive hydrological and meteorological verification for river forecasts (Demargne et al., 2009)
  - Aviation turbulence (Gill, 2012)
- user oriented verification
  - Research on way of understandable presentation of complex results (Joslyn et al., 2012)
  - Optimisation of user dependent decision thresholds (Ambühl 2010)
  - HIW forecast are often low probability forecasts → how do you react on that? Joint projects with SERA?

“Summary”

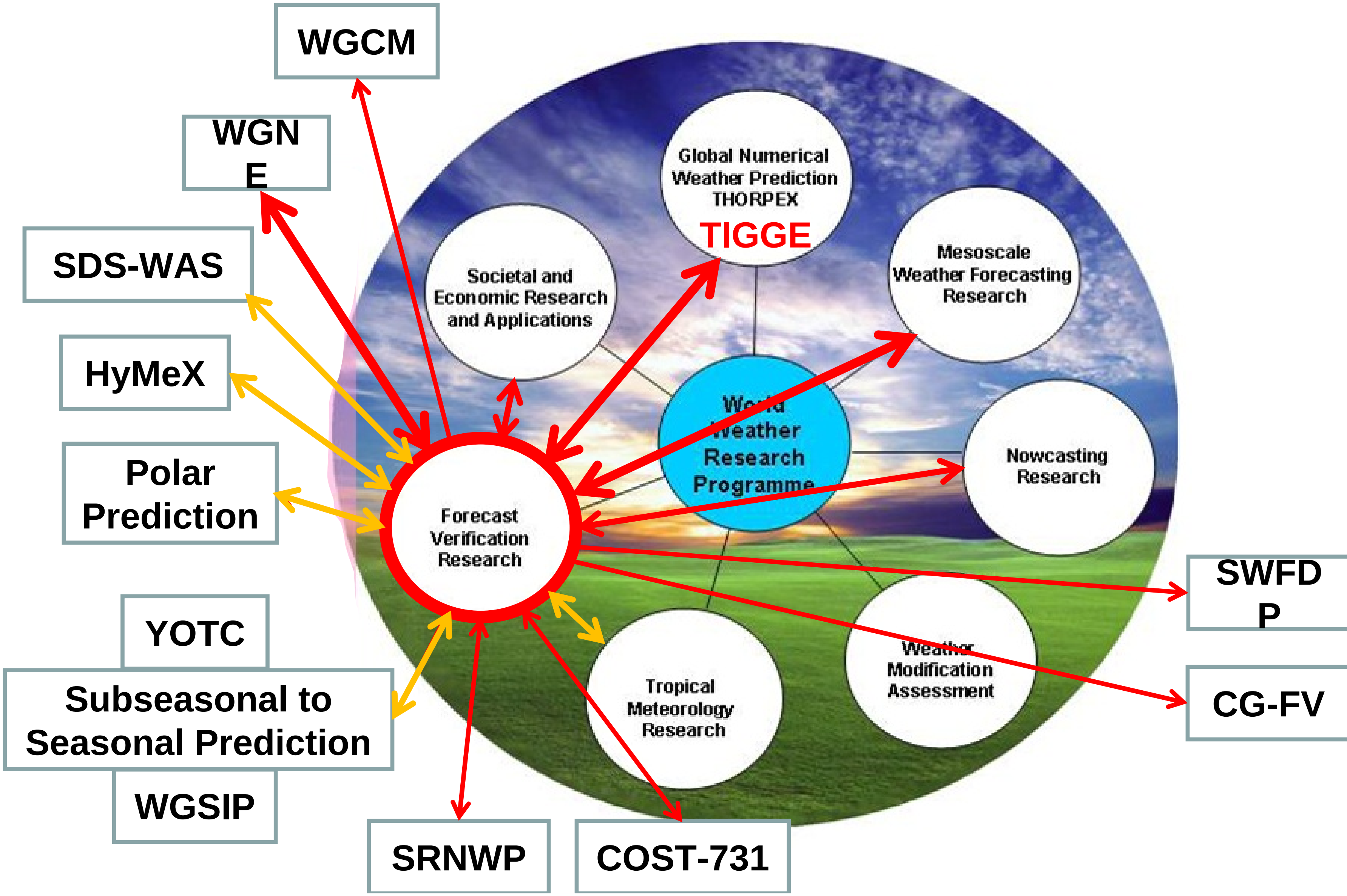
State:

“*The state of forecast verification is clearly improving”.* (Ebert et al., 2013)

Challenges:

“All post-THORPEX projects should have a clearly articulated and user-oriented verification plan”. (L. Wilson, personal com.)

Relationships / collaboration

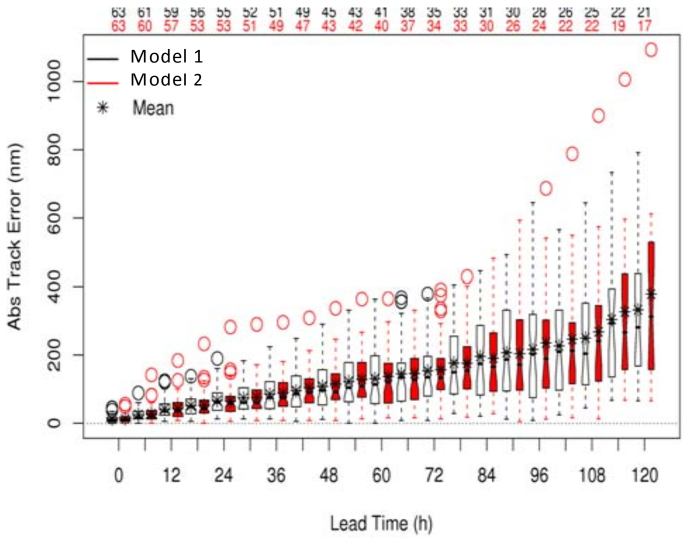


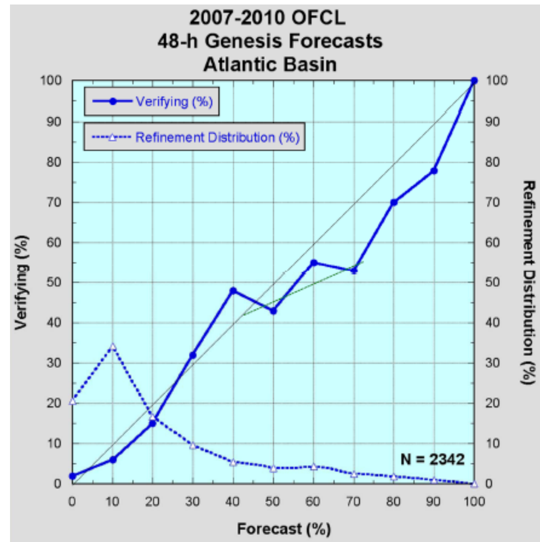
Issues in TC verification

- Observations contain large uncertainties
- Some additional important variables:
  - Storm structure and size
  - Rapid intensification
  - Landfall time, position, and intensity
  - Precipitation
  - Storm surge
  - Consistency
  - Uncertainty
  - Info to help forecasters (e.g., steering flow)
- Tailoring verification to help forecasters with their high-pressure job and multiple sources of guidance information
- ***False alarms*** (incl. forecast storms outliving actual storm) and ***misses*** (unforecasted storms) currently ignored
- How best to evaluate ensemble TC predictions?

Beyond track and intensity…

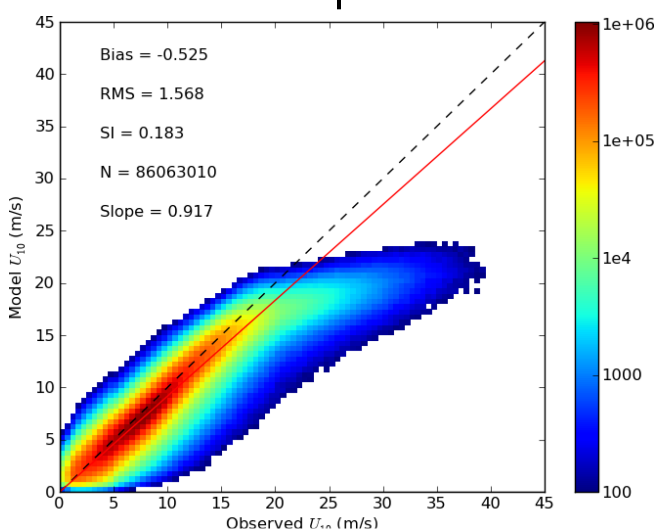
Track error distribution



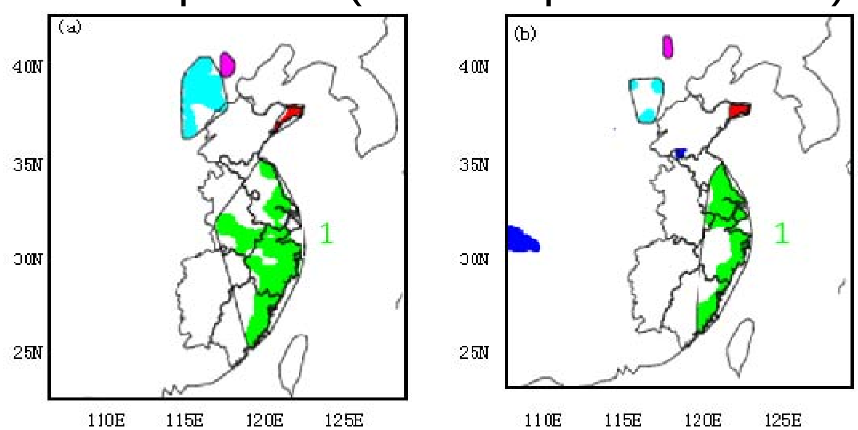


TC genesis

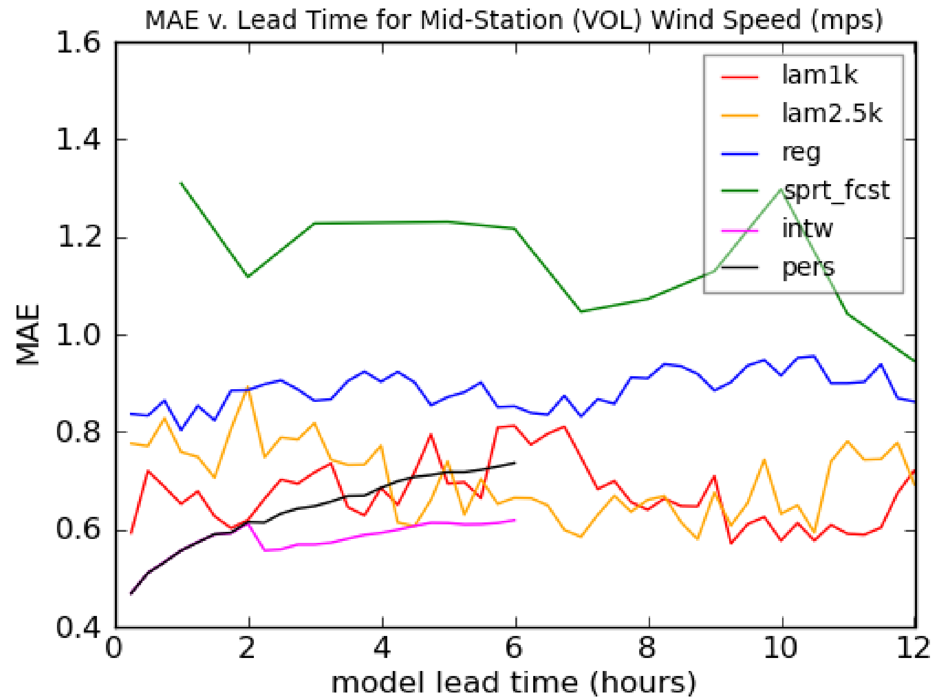
Wind speed



Precipitation (MODE spatial method)



Wind speed verification (model-oriented)



Visibility verification (user-oriented)

lam1k Min. Visibility (m) at VOL HSS=0.095

	Observed						
Forecast	< 30	30 ≤ x < 50	50 ≤ x < 200	200 ≤ x < 300	300 ≤ x < 500	> 500	Total
< 30	0	0	0	0	0	0	0
30 ≤ x < 50	0	0	0	0	0	0	0
50 ≤ x < 200	0	0	52	20	22	43	137
200 ≤ x < 300	0	0	76	18	19	103	216
300 ≤ x < 500	0	1	26	15	12	60	114
> 500	0	9	831	246	170	3743	4999
Total	0	10	985	299	223	3949	5466

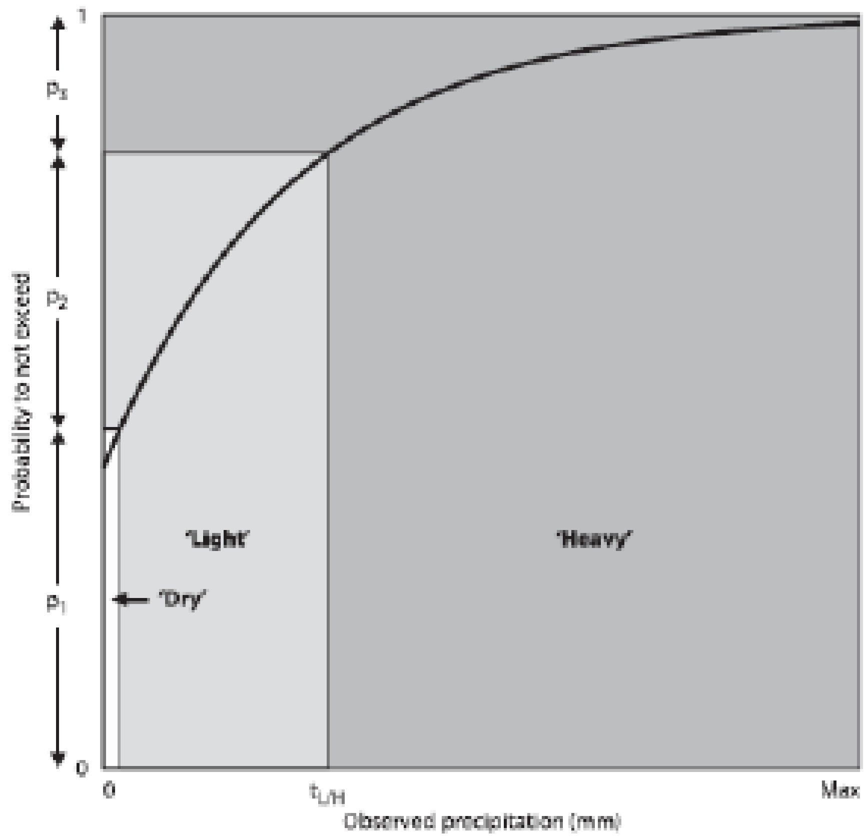
Aggregated scores for monitoring precipitation

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



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