# Verification of numerical models – what are the biggest challenges?

www.cawcr.gov.au

**Beth Ebert (CAWCR / BOM)**

**Barb Brown (NCAR)**

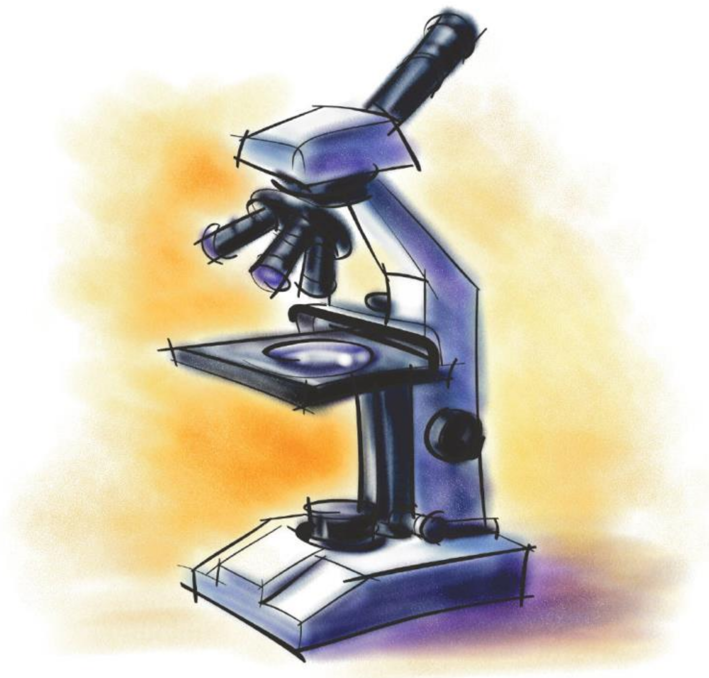
**OzEWEX 2014, 28-29 October 2014, Canberra**

Australian Government
Bureau of Meteorology

**The Centre for Australian Weather and Climate Research**
**A partnership between CSIRO and the Bureau of Meteorology**

CSIRO

# Validation and Verification




***Does my model do the right thing?***

Process studies
Field experiments
Special observations

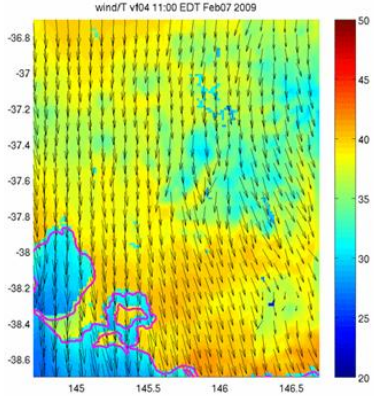
***Did my model get the right answer?***

Systematic verification
Diagnostic verification
Routine observations

# Trends in numerical prediction

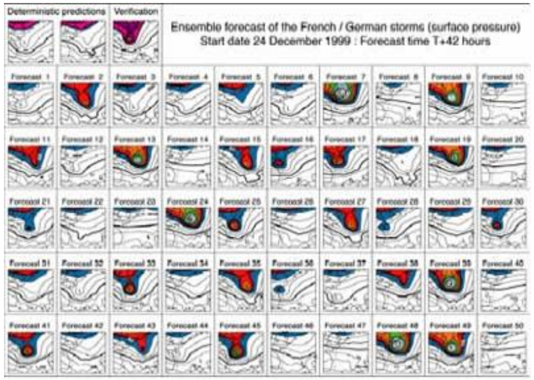
## Higher resolution

- Focus on surface weather



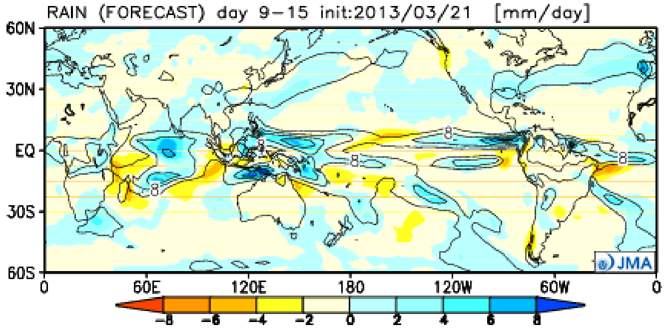

## Ensembles

- Focus on uncertainty



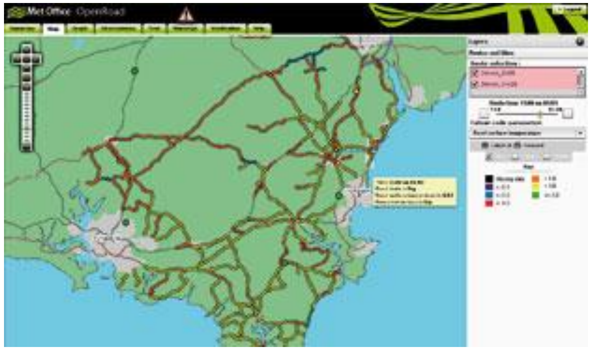

## Coupled extended range

- Focus on longer range




## Impact models

- Focus on user decisions



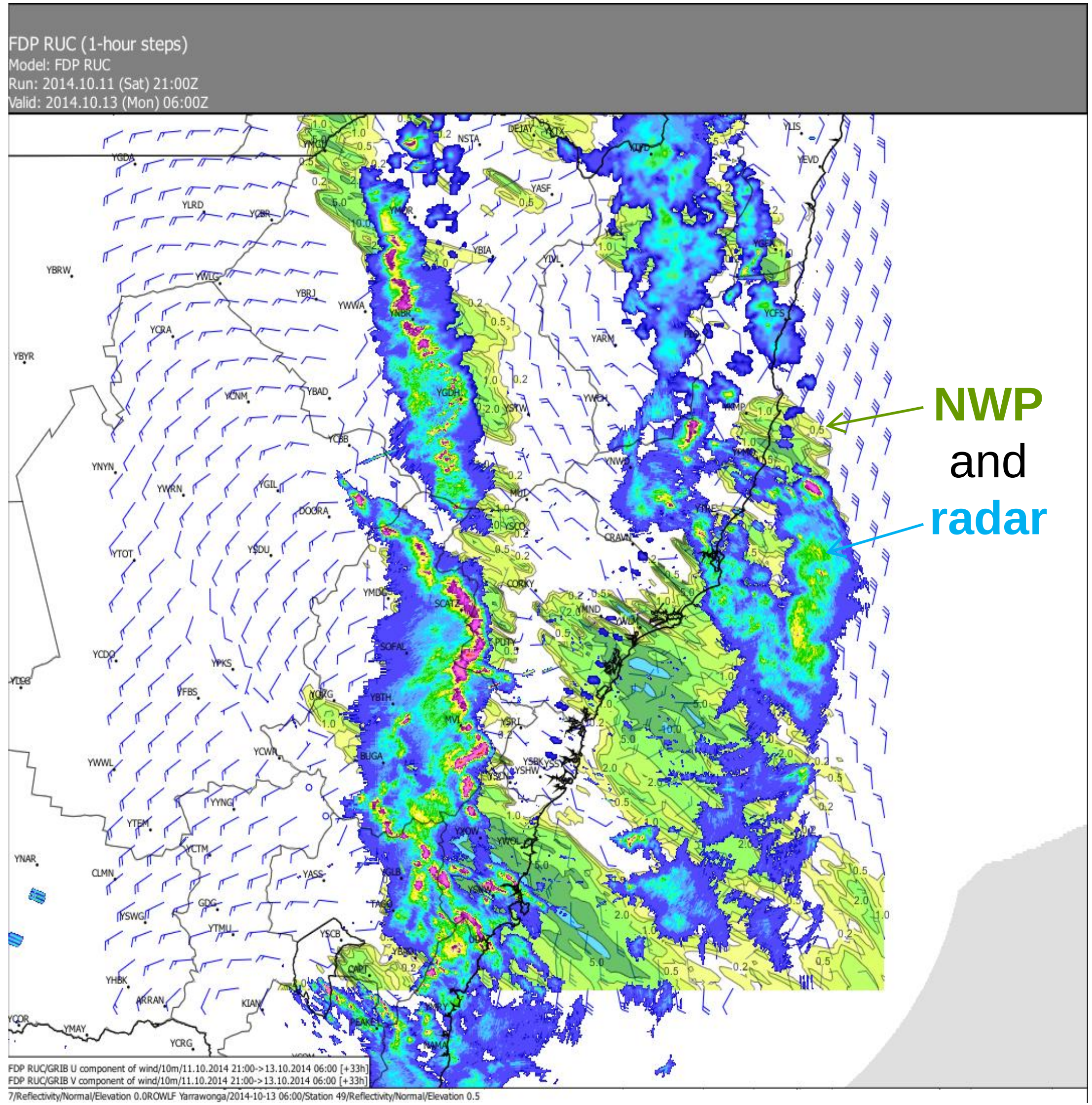
# High resolution NWP

## Benefits

- Surface weather
- Greater realism
- Extreme values

## Verification challenges

- Observations
- Double penalty
- Rare extreme values

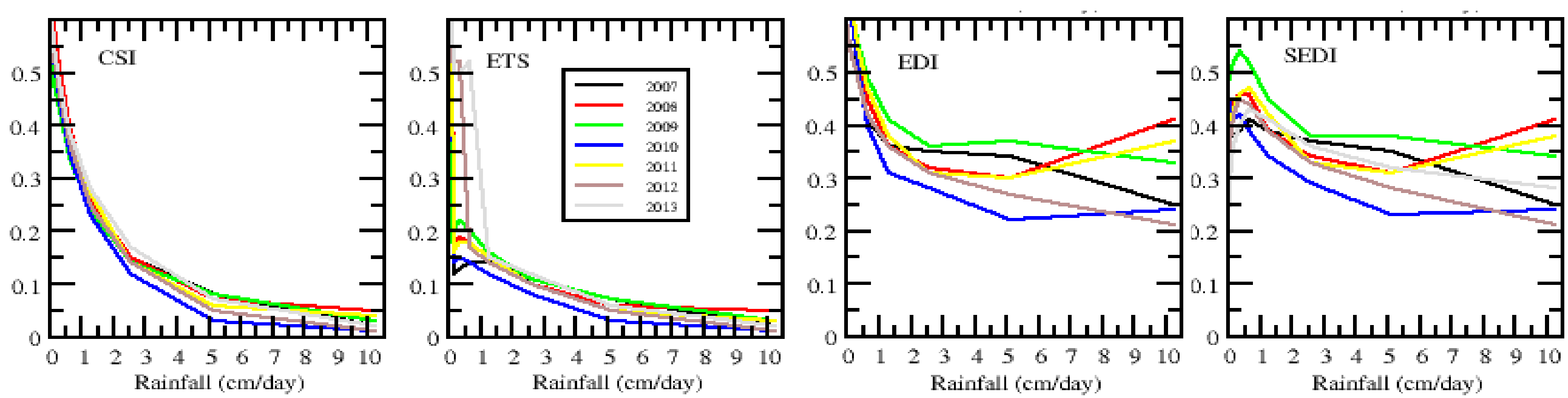


# Verifying rare extreme values

- Hard to observe
- Categorical scores more robust
  - Metrics should reward hits, penalise misses and false alarms
  - For rare events, usual summary scores (e.g., CSI, ETS, HSS, …) → 0
  - New extremal dependence scores:

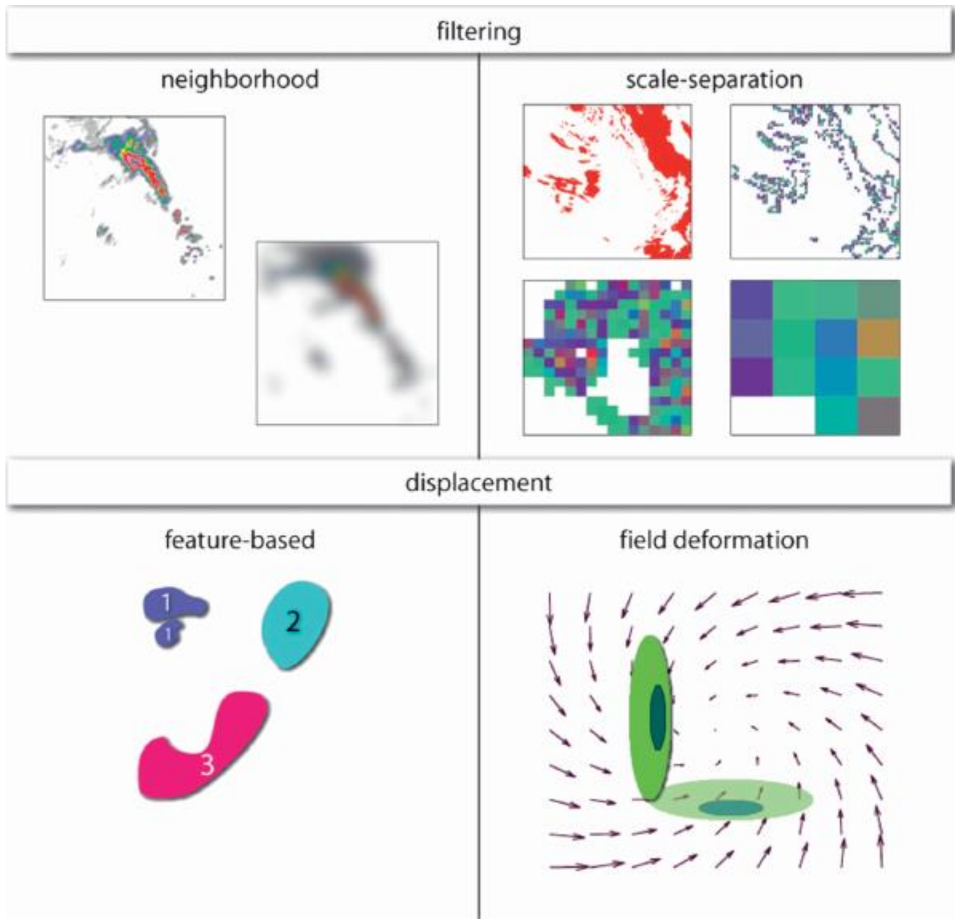
$$EDI = \frac{\log F - \log H}{\log F + \log H} \qquad SEDI = \frac{\log F - \log H - \log(1-F) + \log(1-H)}{\log F + \log H + \log(1-F) + \log(1-H)}$$

Ferro & Stephenson, *Weather and Forecasting*, 2011




Ashrit et al., *Mausam*, 2014

# Spatial verification methods




Credit for "close" forecasts

Scale-dependent error

Attributes of features

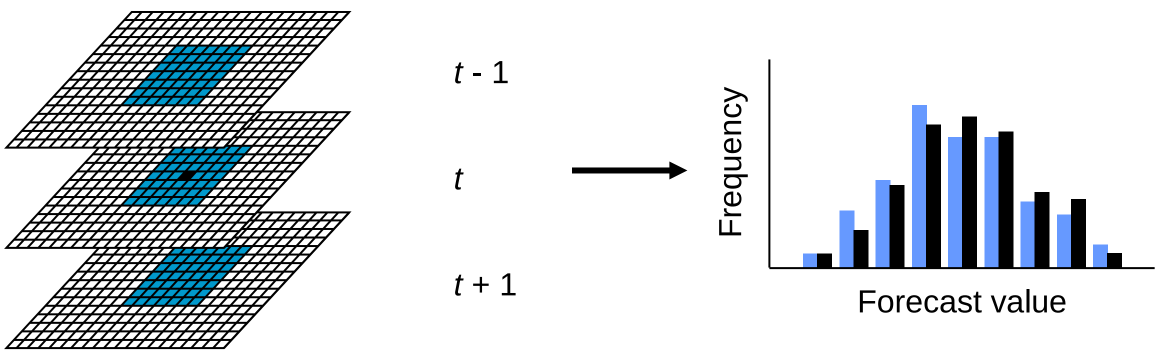
Phase and amplitude errors

Gilleland et al., *Bulletin of the American Meteorological Society,* 2010

# Neighbourhood verification

- Don't require an *exact* match between forecasts and observations
  - Unpredictable scales
  - Uncertainty in observations

Look in a space / time neighborhood around the point of interest




Evaluate using categorical, continuous, probabilistic scores / methods

# Feature-based verification

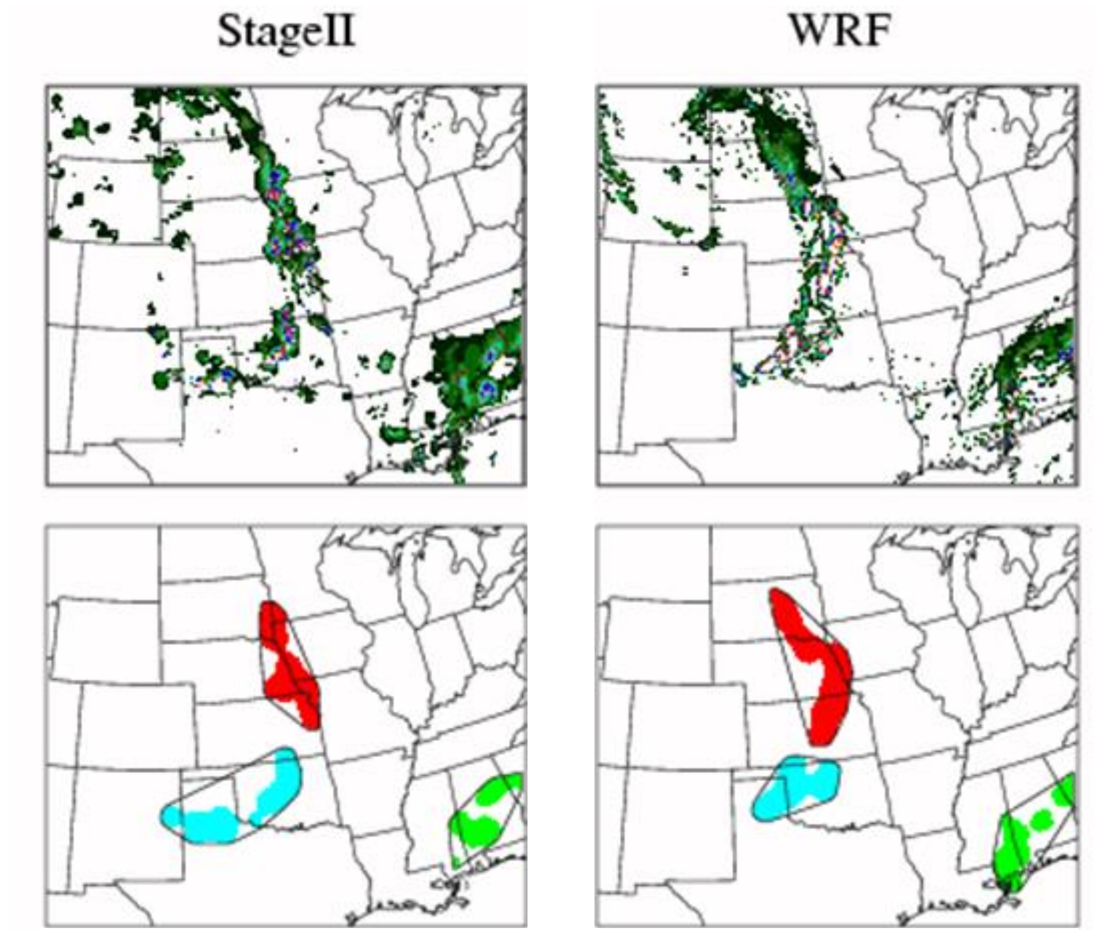
## Compare attributes:

- centroid location
- intensity distribution
- area
- orientation
- etc.

## When objects not matched:

- false alarms
- missed events
- rain volume
- etc.

Method for Object-based Diagnostic Evaluation (MODE)



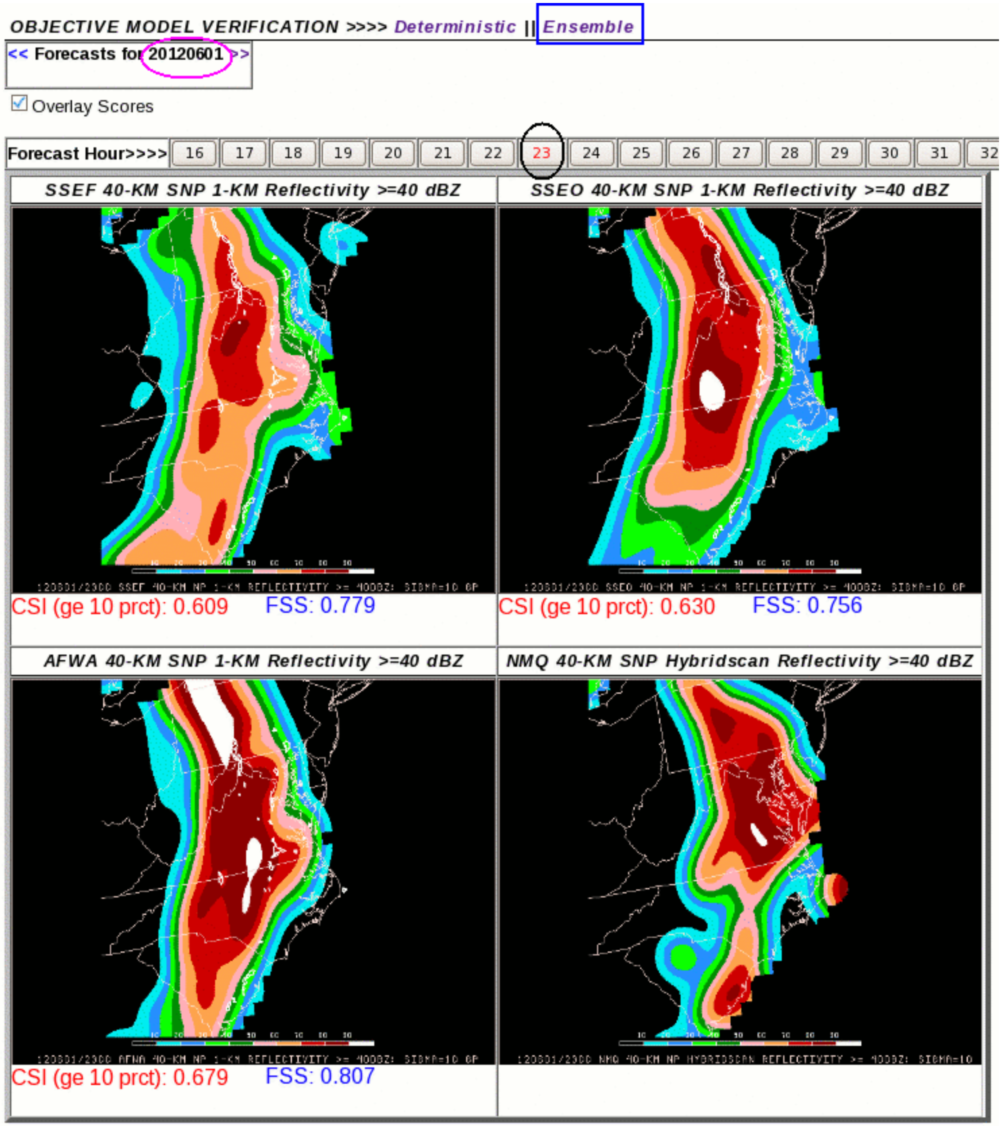

24h forecast of 1h rainfall on 1 June 2005

# Spatial Verification Methods Intercomparison

| Category | Scales with skill | Location errors | Intensity errors | Structure errors | Occurrence (hits, misses, false alarms) |
|---|---|---|---|---|---|
| Traditional (gridpoint) | × | × | ✓ | × | ✓ |
| Neighbourhood | ✓ | × | ✓ | × | ✓ |
| Scale separation | ✓ | × | ✓ | × | ✓ |
| Features based | × | ✓ | ✓ | ✓ | ✓ |
| Deformation | × | ✓ | ✓ | × | × |

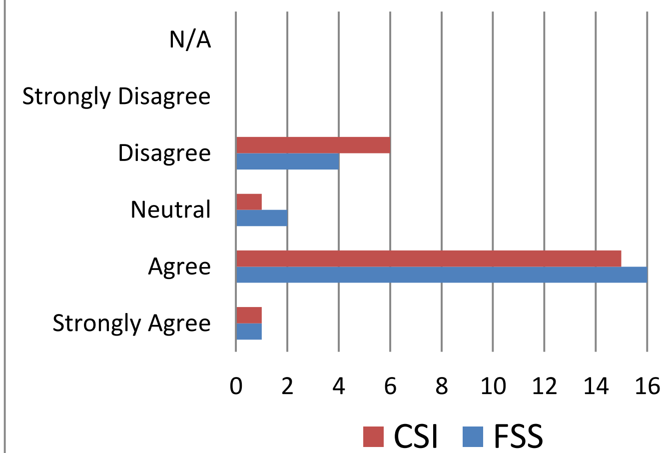
- Conclusions from 1st phase
  - Different methods have different strengths
  - *All address bias*
- 2nd phase
  - Wind and precipitation in complex terrain
  - Ensemble forecasts
  - Point observations, ensemble observations

# Neighbourhood ensemble verification



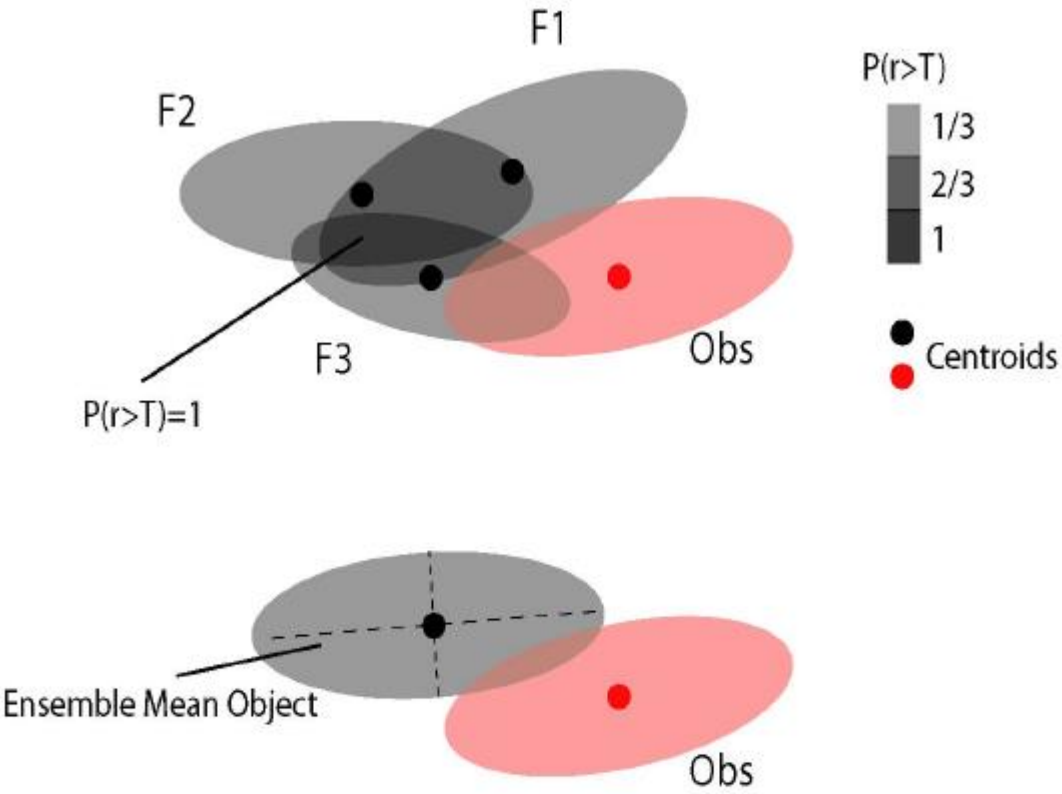

## Survey Question Feedback:

**Ensemble Reflectivity:** Do the objective metrics agree with your subjective impressions of forecast skill?




Melick *et al.*, *NWA*, 2012

# Feature-based ensemble verification

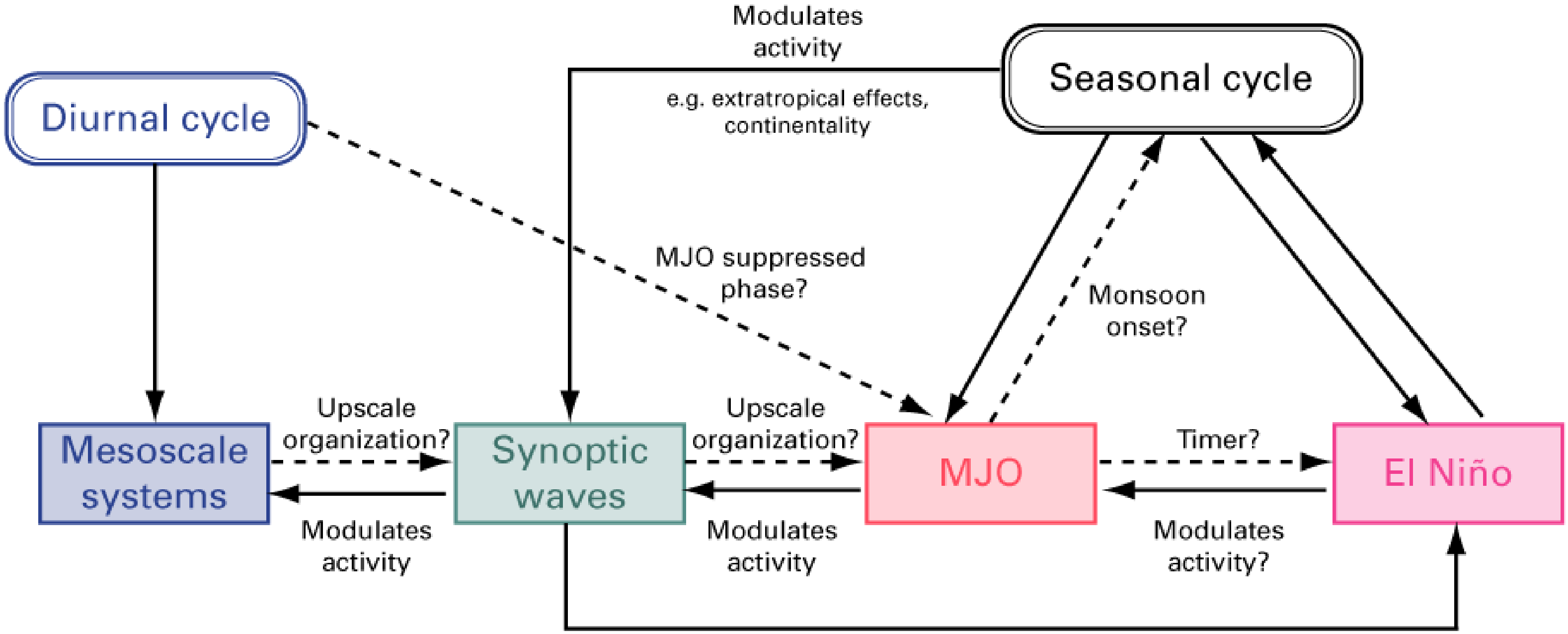



Possible strategies for verifying an ensemble of "objects"

1. Verify objects in probability maps
2. Verify "ensemble mean"
3. Verify distributions of object properties

# Interacting weather & climate processes

Moncrieff et al., *WMO Bulletin*, 2007




1-100 km Hours

100's km Days

1000's km Weeks

10 000's km Months

Forecast *values*

Forecast *anomalies*

Seamless prediction: How to verify across time scales?

# Generalized Discrimination Score (GDS)

Two-alternative forced choice:

Observation 1 | Forecast 1 | Obs correctly discriminated? | YES / NO
Observation 2 | Forecast 2

Observation 1 | Forecast 1 | Obs correctly discriminated? | YES / NO
Observation 3 | Forecast 3

Observation *N*-1 | Forecast *N*-1 | Obs correctly discriminated? | YES / NO
Observation *N* | Forecast *N*

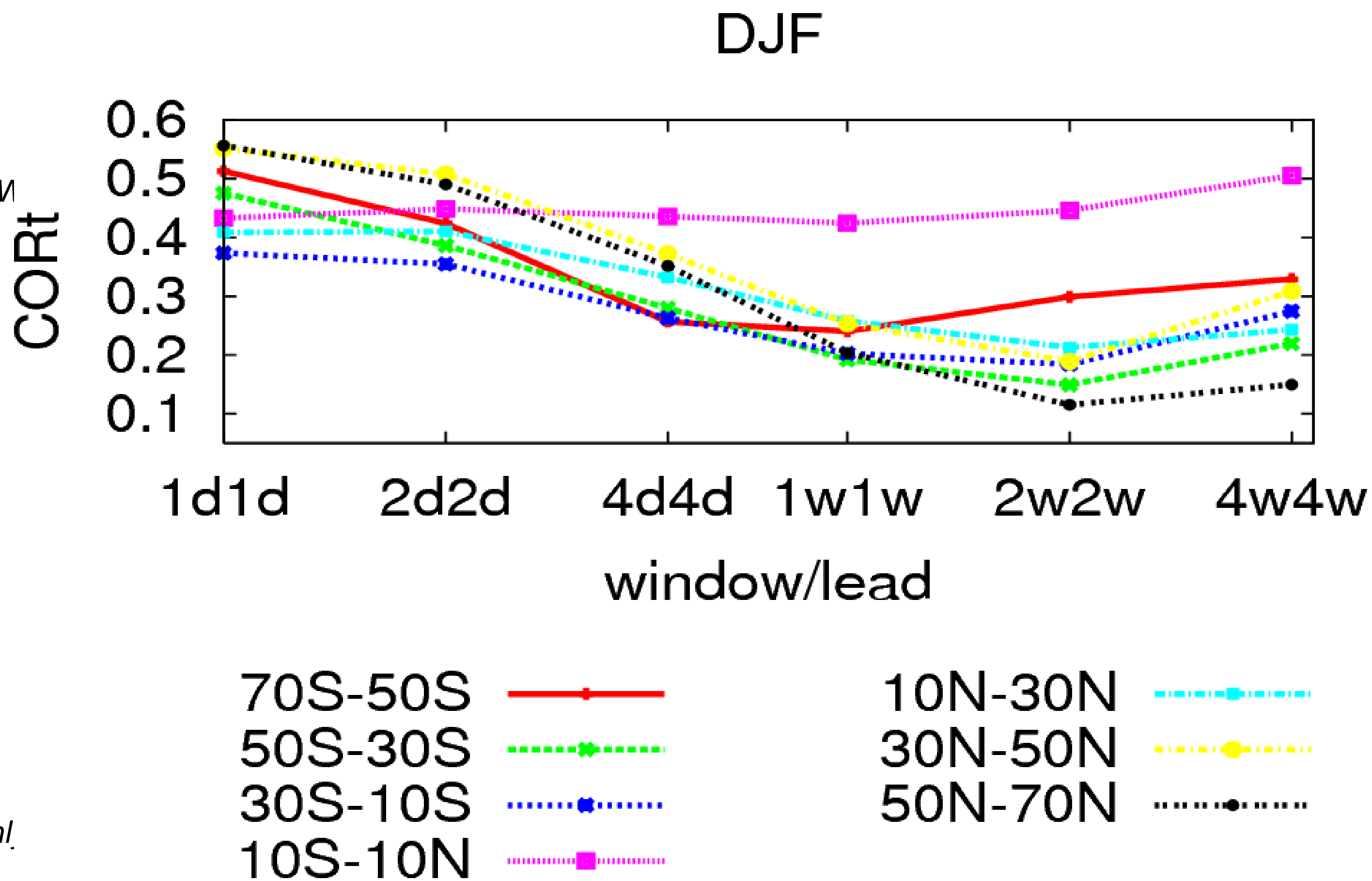
GDS = proportion of successful rankings (no skill = 50%)

Mason & Weigel, *Monthly Weather Review*, 2009

# Multi-temporal verification

- Compute skill for a large range of lead times.
- As lead time is increased, also increase the time-averaging window for a seamless transition from weather to climate.

*Schematic of w*
*definitions*




Zhu et al., *Monthl*

# Transpose AMIP

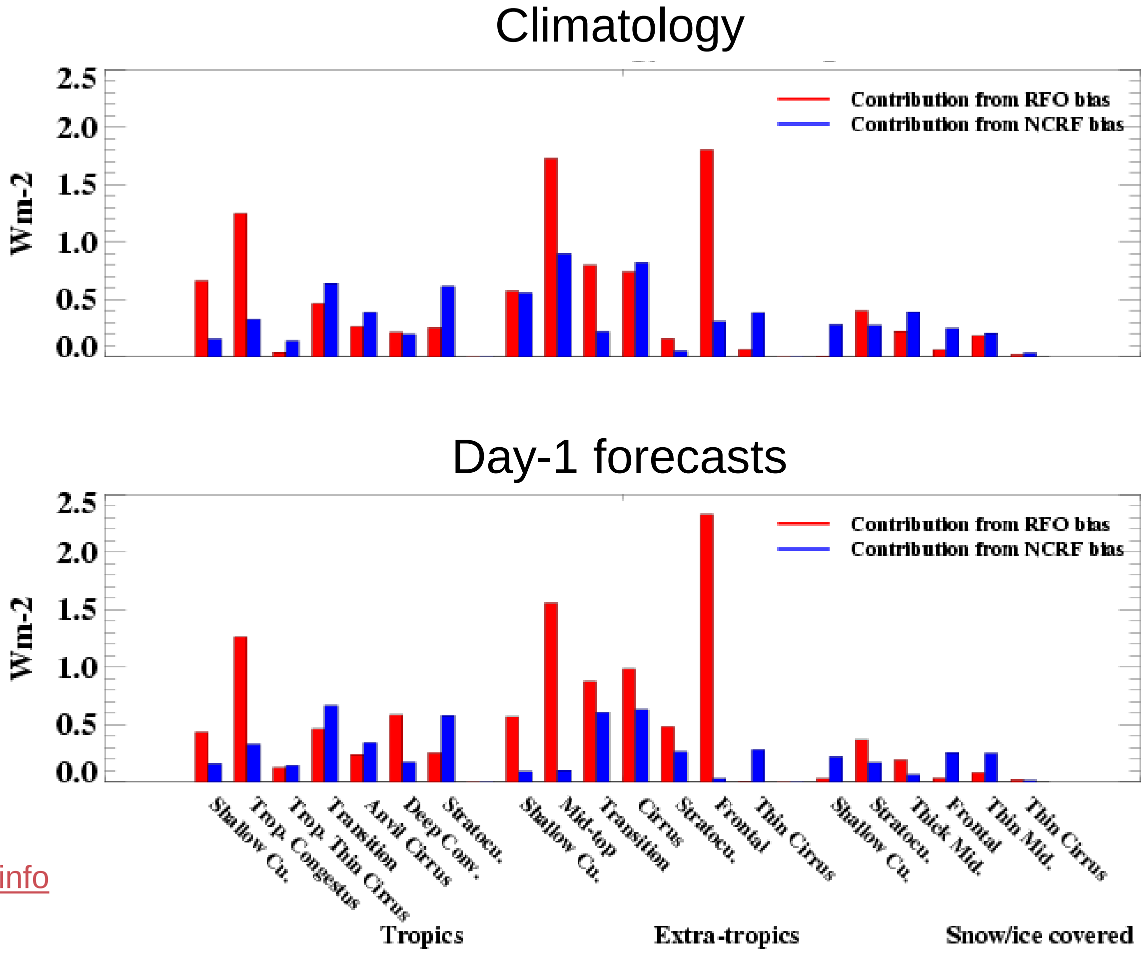
## Run climate models in NWP mode

- Verification against observations → evaluation of processes

SW radiation biases by cloud regime

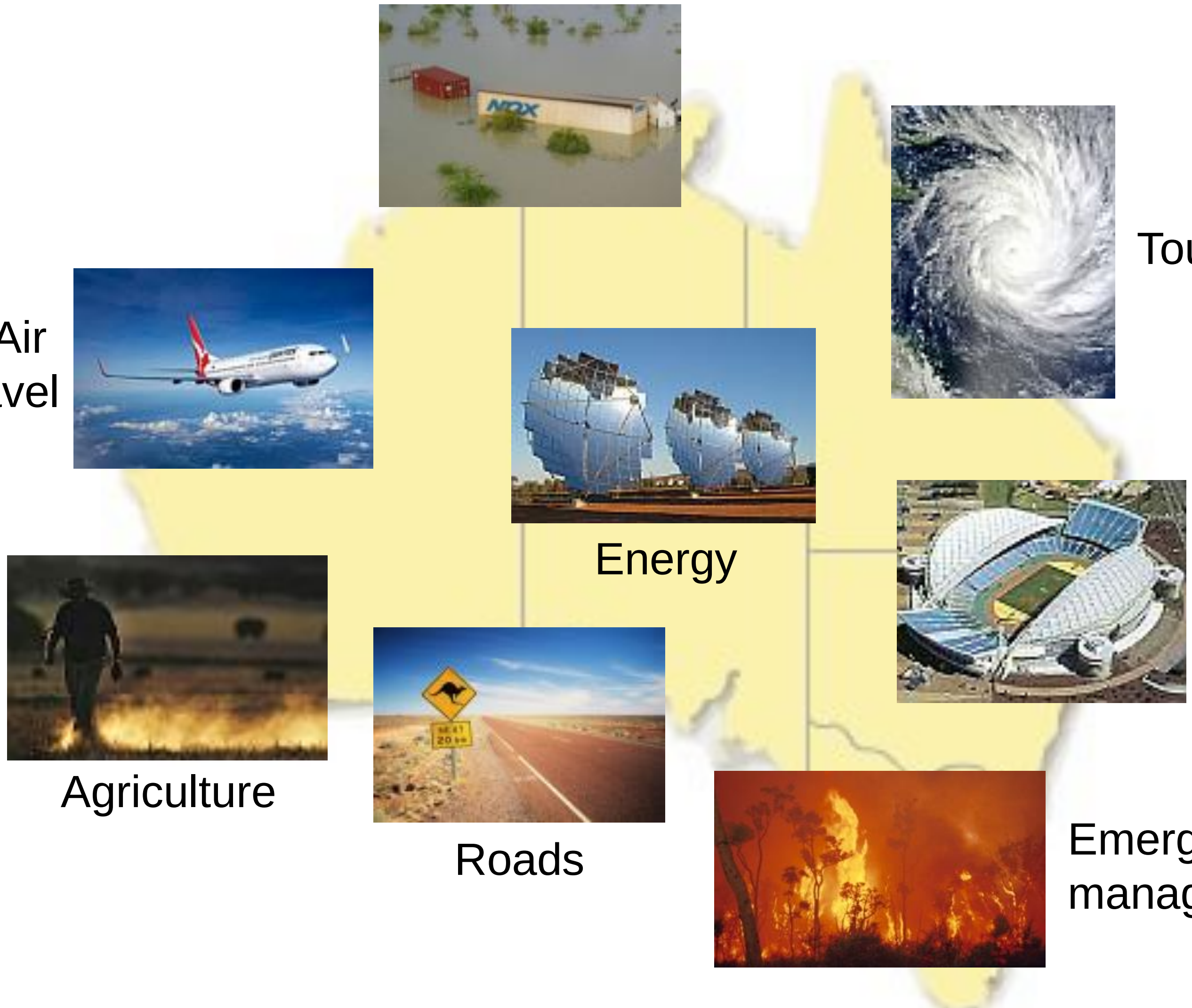
64 5-day hindcasts during Oct 2008-July 2009

(courtesy K. Williams)




http://www.transpose-amip.info

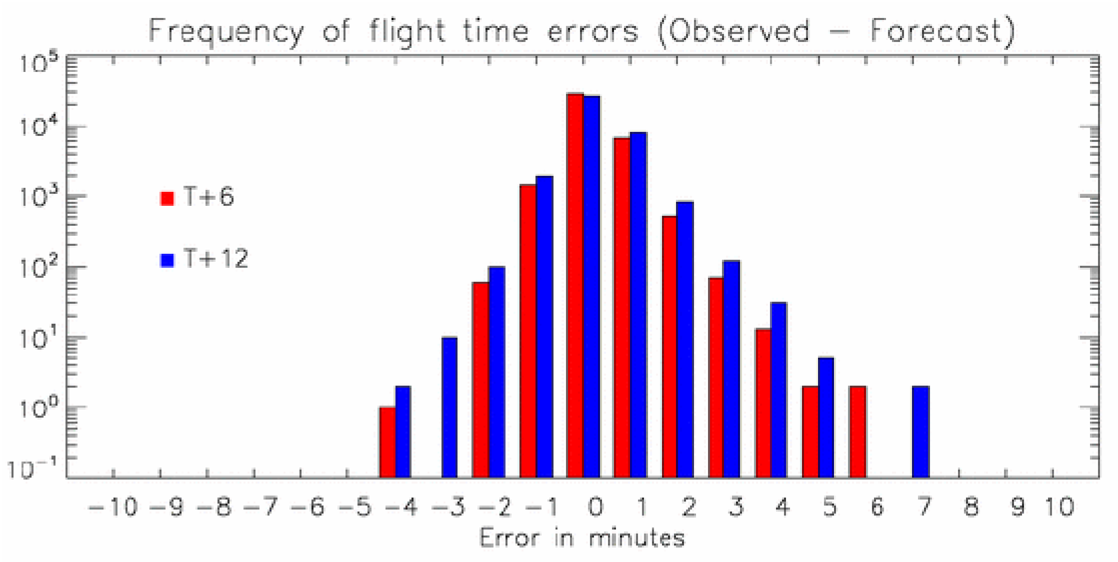
# Weather modelling → impact modelling



# User-relevant verification - Aviation

Flight time error (FTE) = $flight\_time_{obs} - flight\_time_{fcst}$

- Accurate measure of wind forecast accuracy *directly relevant to airlines*
- Calculated using the track that the aircraft actually took
- Uses AMDAR observations from real flights rather than model analyses or radiosondes



# Uncertainty in observations

- As models improve, we can no longer ignore observation error!

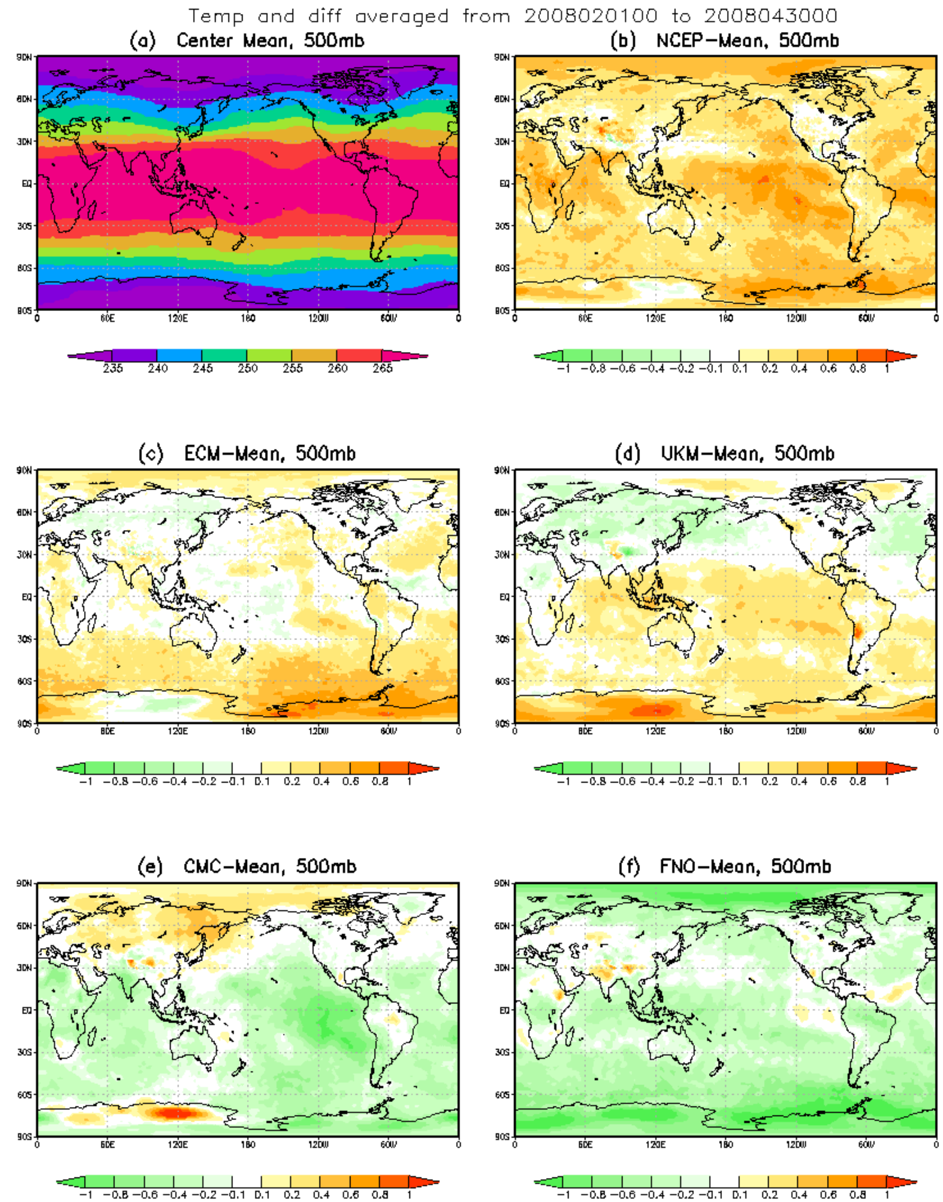
- Remove observation *bias* errors where possible
- Effects of *random* obs error on verification
  - “Noise” leads to poorer scores for deterministic forecasts
  - Ensemble forecasts have poorer reliability & ROC

- What can we do?
  - Error bars in scatter plots
  - Quantitative reference to “gold standard”
    - Correct for systematic error in observations
    - RMSE – Ciach & Krajewski (*Adv. Water Res.,*1999)
    - Categorical scores – Briggs et al. (*MWR*, 2005), Bowler (*MWR*, 2006)
  - Multiple observation sources / analysis methods

# Verification against own model analyses

- Pros
  - Convenient
    - Available in-house
    - Matched grid
  - Spatially complete

- Cons
  - Analysis contains bias
    - Inherited from model first guess
    - Different satellite processing
    - Different observations assimilated
    - Poor models of error covariance
  - → Misleading model skill

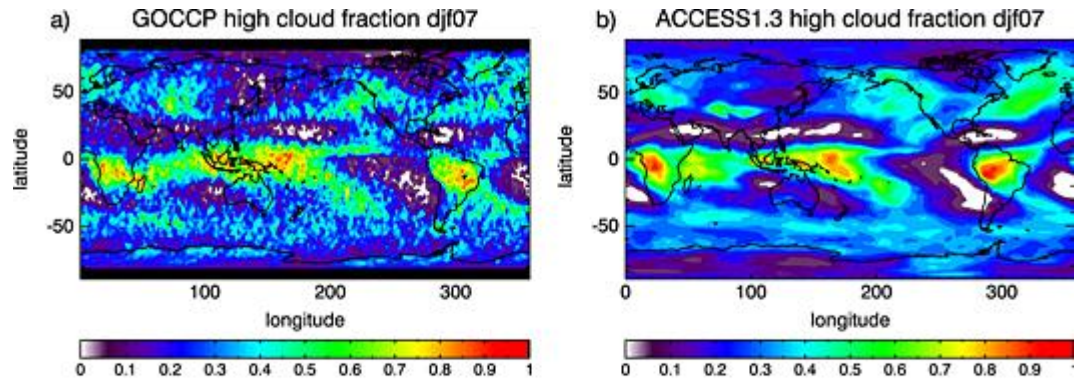



Wei et al., *AMOJ*, 2010

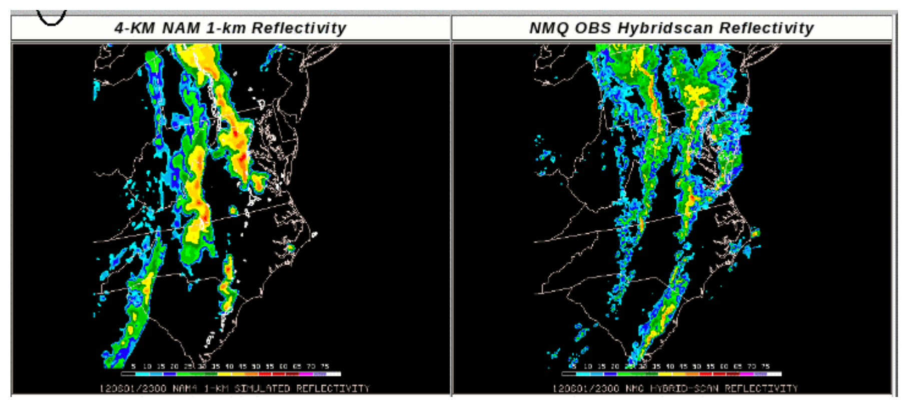
# Observations quantity and quality

## Verification in "obs" space

- Satellite
  - A-Train
  - Himawari-8/9
  - GPM
  - etc.
- Radar
  - National & int'l networks
  - Polarimetric & phased array
- GPS
- 3rd party data
  - Mobile phone technology
- Multi-sensor analyses



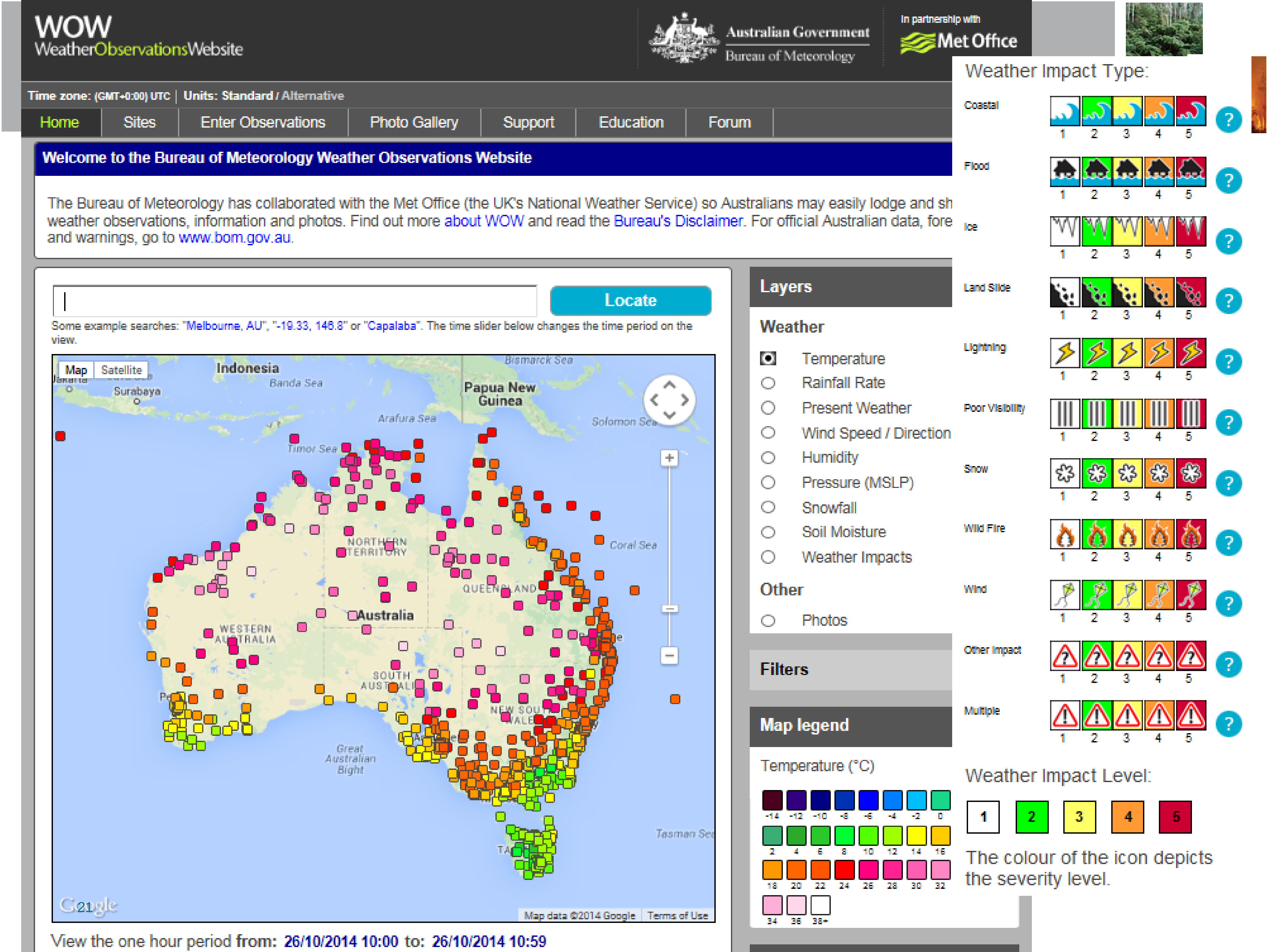

Franklin *et al.*, *JGR*, 2013




Model reflectivity

Obs reflectivity

Melick *et al.*, *NWA*, 2012


WOW
WeatherObservationsWebsite
Australian Government
Bureau of Meteorology
In partnership with
Met Office
Time zone: (GMT+0:00) UTC | Units: Standard / Alternative
Home
Sites
Enter Observations
Photo Gallery
Support
Education
Forum
Welcome to the Bureau of Meteorology Weather Observations Website
The Bureau of Meteorology has collaborated with the Met Office (the UK's National Weather Service) so Australians may easily lodge and sh
weather observations, information and photos. Find out more about WOW and read the Bureau's Disclaimer. For official Australian data, fore
and warnings, go to www.bom.gov.au.
Locate
Some example searches: "Melbourne, AU", "-19.33, 146.8" or "Capalaba". The time slider below changes the time period on the view.
Map
Satellite
Indonesia
Banda Sea
Surabaya
Bismarck Sea
Papua New Guinea
Arafura Sea
Solomon Sea
Timor Sea
NORTHERN TERRITORY
QUEENSLAND
Coral Sea
Australia
WESTERN AUSTRALIA
SOUTH AUSTRALIA
NEW SOUTH WALES
Great Australian Bight
Tasman Sea
Map data ©2014 Google
Terms of Use
View the one hour period from: 26/10/2014 10:00 to: 26/10/2014 10:59
Layers
Weather
Temperature
Rainfall Rate
Present Weather
Wind Speed / Direction
Humidity
Pressure (MSLP)
Snowfall
Soil Moisture
Weather Impacts
Other
Photos
Filters
Map legend
Temperature (°C)
-14 -12 -10 -8 -6 -4 -2 0
2 4 6 8 10 12 14 16
18 20 22 24 26 28 30 32
34 36 38+
Weather Impact Type:
Coastal
Flood
Ice
Land Slide
Lightning
Poor Visibility
Snow
Wild Fire
Wind
Other Impact
Multiple
1 2 3 4 5
Weather Impact Level:
1 2 3 4 5
The colour of the icon depicts the severity level.

# Conclusions

## Progress and challenges in model verification

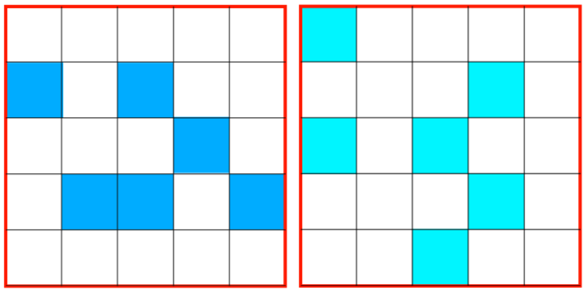
- Spatial verification becoming mainstream
- New scores for extreme events
- Evaluating ensembles
- Verification across time scales
- Relevant metrics for weather impacts
- Observation quality/quantity
- Verification and data assimilation

Thank you!

# More slides

# Fractions skill score

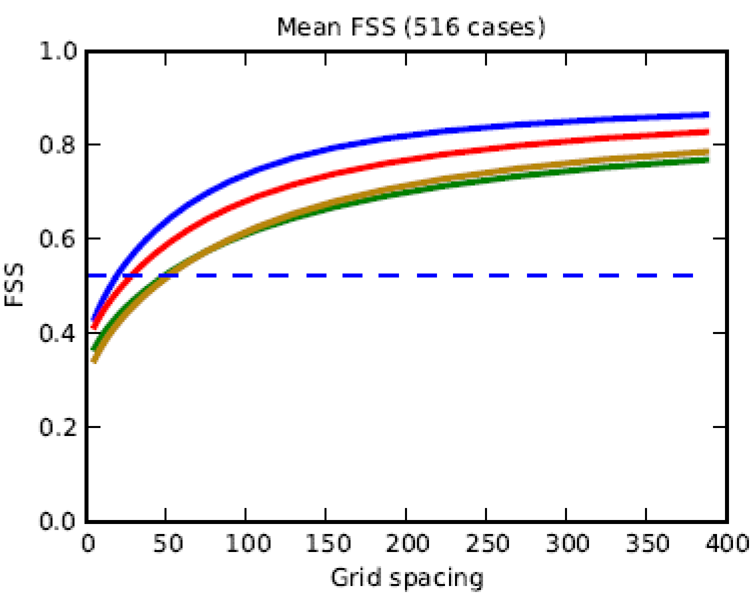
Compare forecast fractions with observed fractions (radar) *probabilistically* over different sized neighbourhoods




Fraction = 6/25 = 0.24    Fraction = 6/25 = 0.24

$$FBS = \frac{1}{N}\sum_{i=1}^{N}(P_{fcst} - P_{obs})^2$$

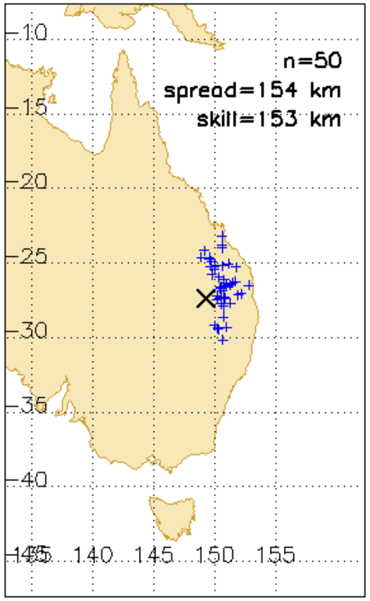
$$FSS = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(P_{fcst} - P_{obs})^2}{\frac{1}{N}\sum_{i=1}^{N}P_{fcst}^{\ 2} + \frac{1}{N}\sum_{i=1}^{N}P_{obs}^{\ 2}}$$





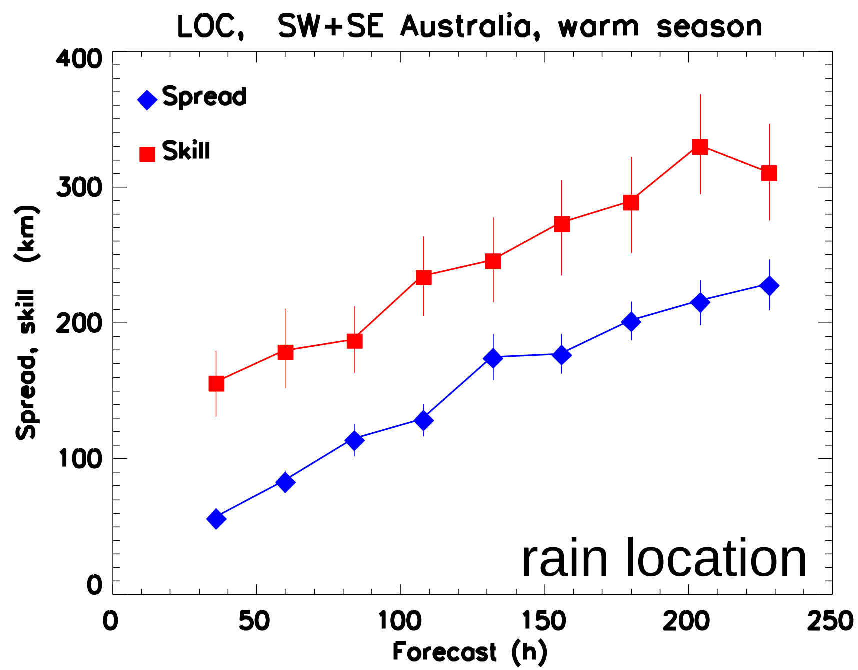

Roberts & Lean, *Monthly Weather Review*, 2008

# Spread and skill for location forecasts




Contiguous rain areas with max rain ≥ 20 mm d$^{-1}$
Warm season, southern Australia

Mean values for 112 events

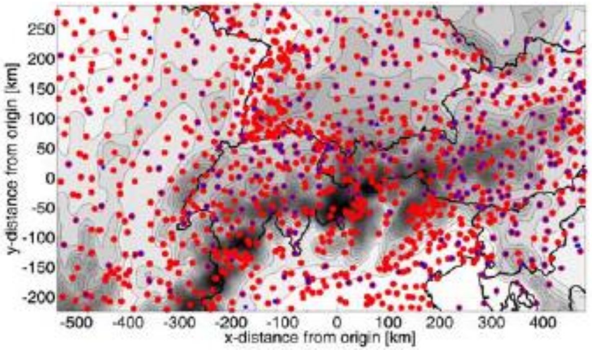



Spread = average distance to ensemble median location

Skill = distance between ensemble median and observed location

Ebert, *IVMW4*, 2009

# Uncertainty in reference data




**6 NWP models:**
COSMO-EU COSMO-DE
COSMO-7 COSMO-2
CMC-GEML CMC-GEMH

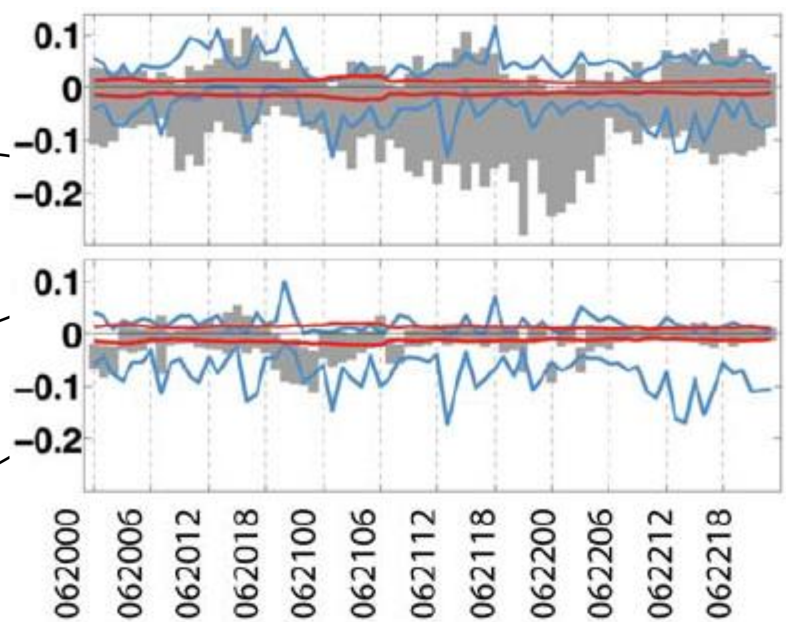
Approaches to model-independent analysis

**3 interpolation tools:**
VERA
Ordinary Kriging
Barnes Interpolation

**2 data networks:**
JDC + GTS

**4 grid spacings:**
4, 8, 16, 32 km

Correlation coeff. for wind speed (difference from standard)




Approaches to verification uncertainty

Poor Man's analysis ensemble

Resampling (bootstrapping)

Sequential Gaussian Simulation

Obs-error based analysis ensemble



Gorgas & Dorninger, *MetZ*, 2012