Postprocessing Calibration

The icing on the cake for high-utility stochastic weather prediction

Presentation Overview

Why Calibrate?

- The need
- The goal

Aspects of Calibration

- 1st Moment Calibration– *Model Bias Correction*
- 2nd Moment Calibration– *Ensemble Spread Correction*
- "Reality" Calibration – *Down Scaling*

Ground Truth Issues

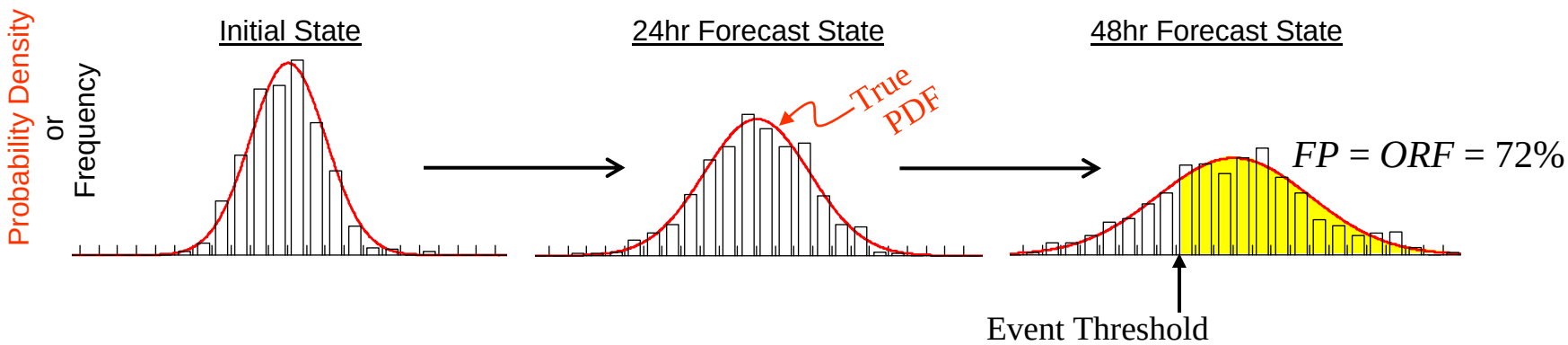
- Choices
- Length of Training

What to Know

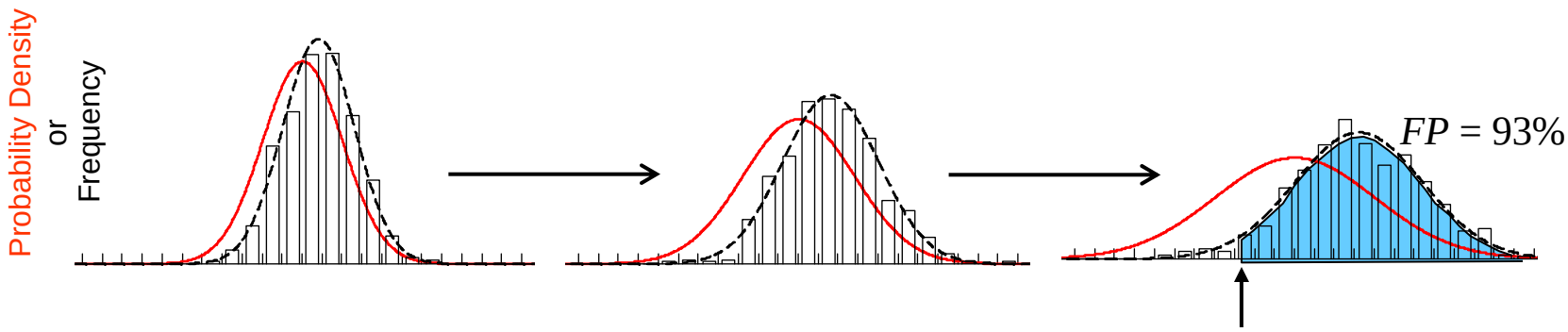
- Checklist
- Sample Ensemble System Evaluations

The Need for Calibration

- Ensemble estimates the true (*long-term verification*) Probability Density Function (PDF)
- In a large, ideal ensemble, Forecast Probability (*FP*) = Observed Relative Frequency (*ORF*)



- In practice, things go awry from…
  - Undersampling (too few ensemble members)
  - Systematic Error
    - -- Incomplete capture of uncertainty (analysis and model) can cause low dispersion
    - -- Model biases can cause shift in mean and inappropriate dispersion
- Ensemble's estimated PDF does not match true PDF, and $FP \neq ORF$



Calibration Goal

Increase ensemble skill and utility by maximizing:

1) **Reliability**

- Forecast Probability = Observed Relative Frequency

- Statistical Consistency
  Mean Square Error of EF Mean = Ensemble Variance
  …*or…*
  Over time, verifying obs. indistinguishable from ensemble members

2) **Sharpness**

- Ability to distinguish between events and non-events

- Narrow (sharp) ensemble PDF is better
  -- Results in probabilities closer to 0% or 100%
  -- BUT, must maximize Sharpness while ensuring reliability

End State: Reliable and sharp PDF of **P(Obs | Fcst)**

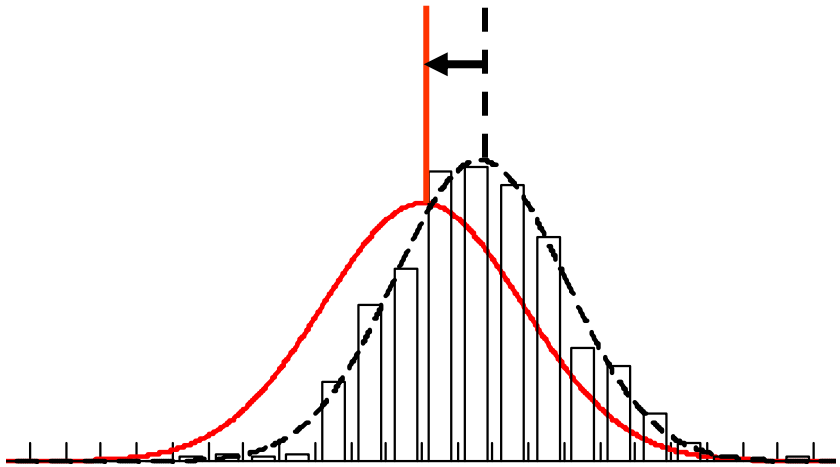
… a "translation" from model output to reality.

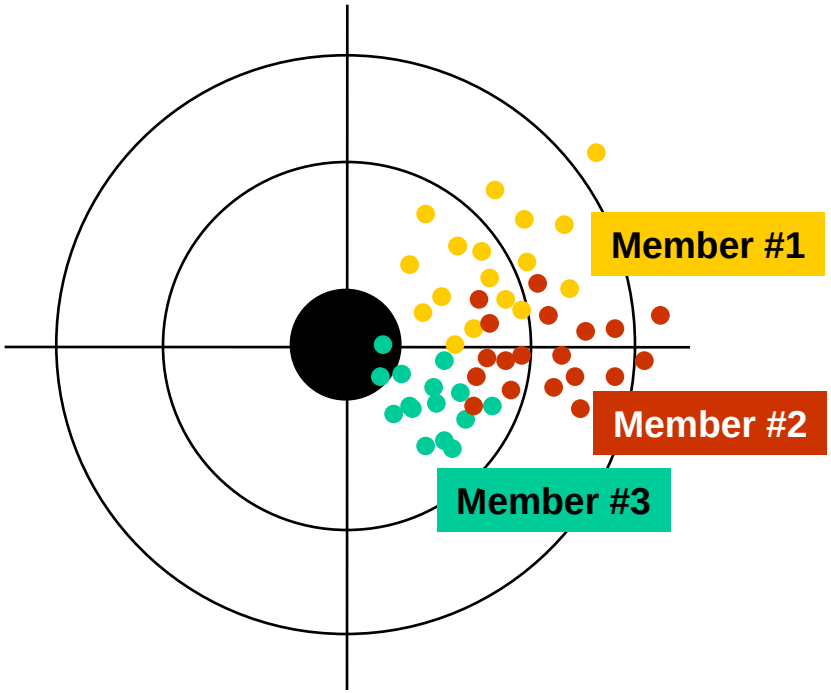
1st Moment Calibration

(Model Bias Correction)

1st Moment Calibration
(Model Bias Correction)

FOCUS: Increase reliability by “recentering” forecast PDF





For multi-model ensemble, this is a deterministic calibration done separately on each ensemble member

-- extremely significant to ensemble skill

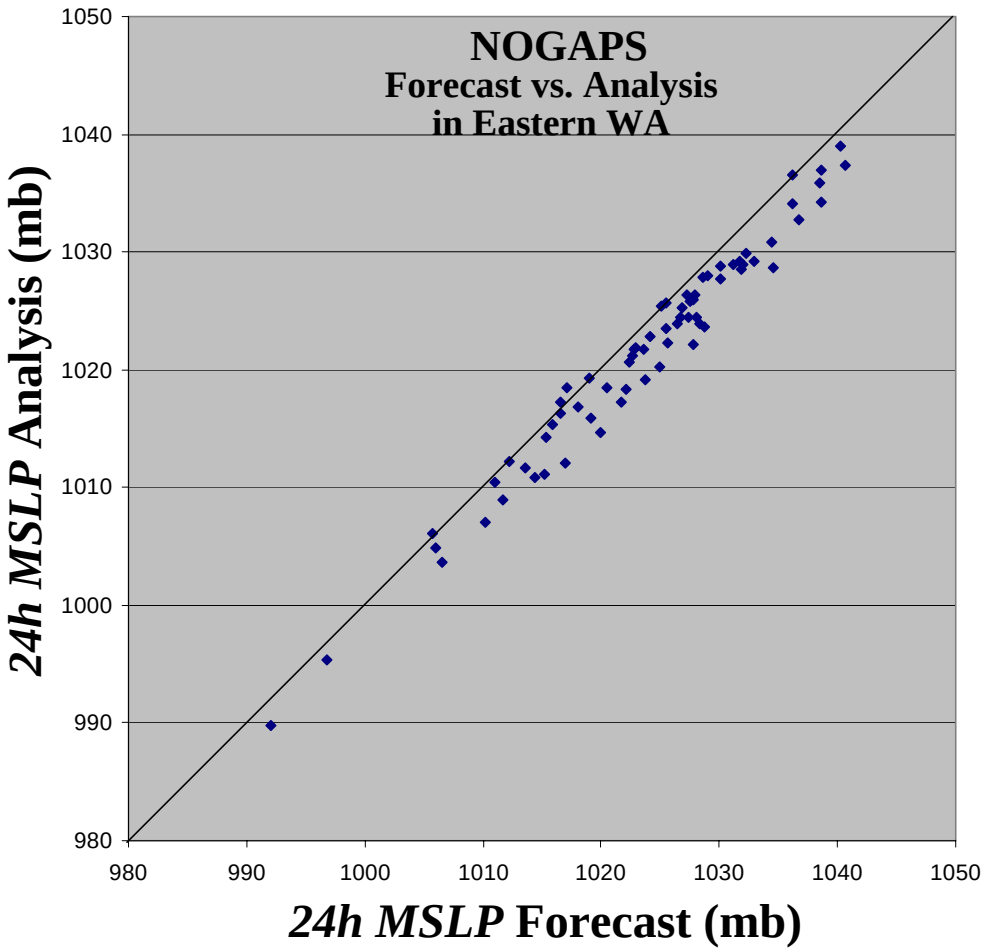
Model Bias

- **Statistically significant value of forecast mean error**

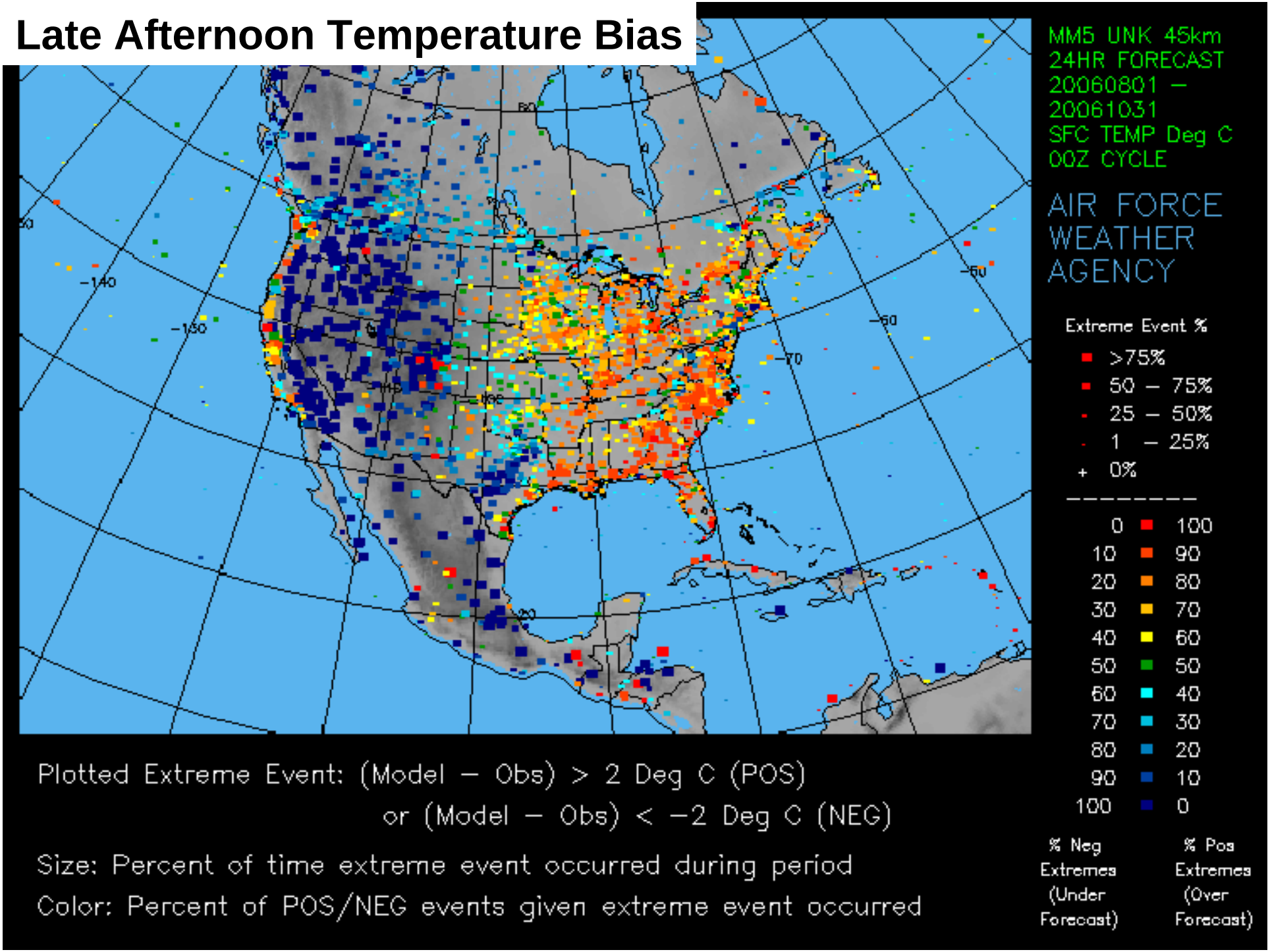
$$Bias = \frac{1}{N}\sum_{n=1}^{N}\left(fcst_n - obs_n\right)$$

  - **But relationship may be linear or nonlinear**

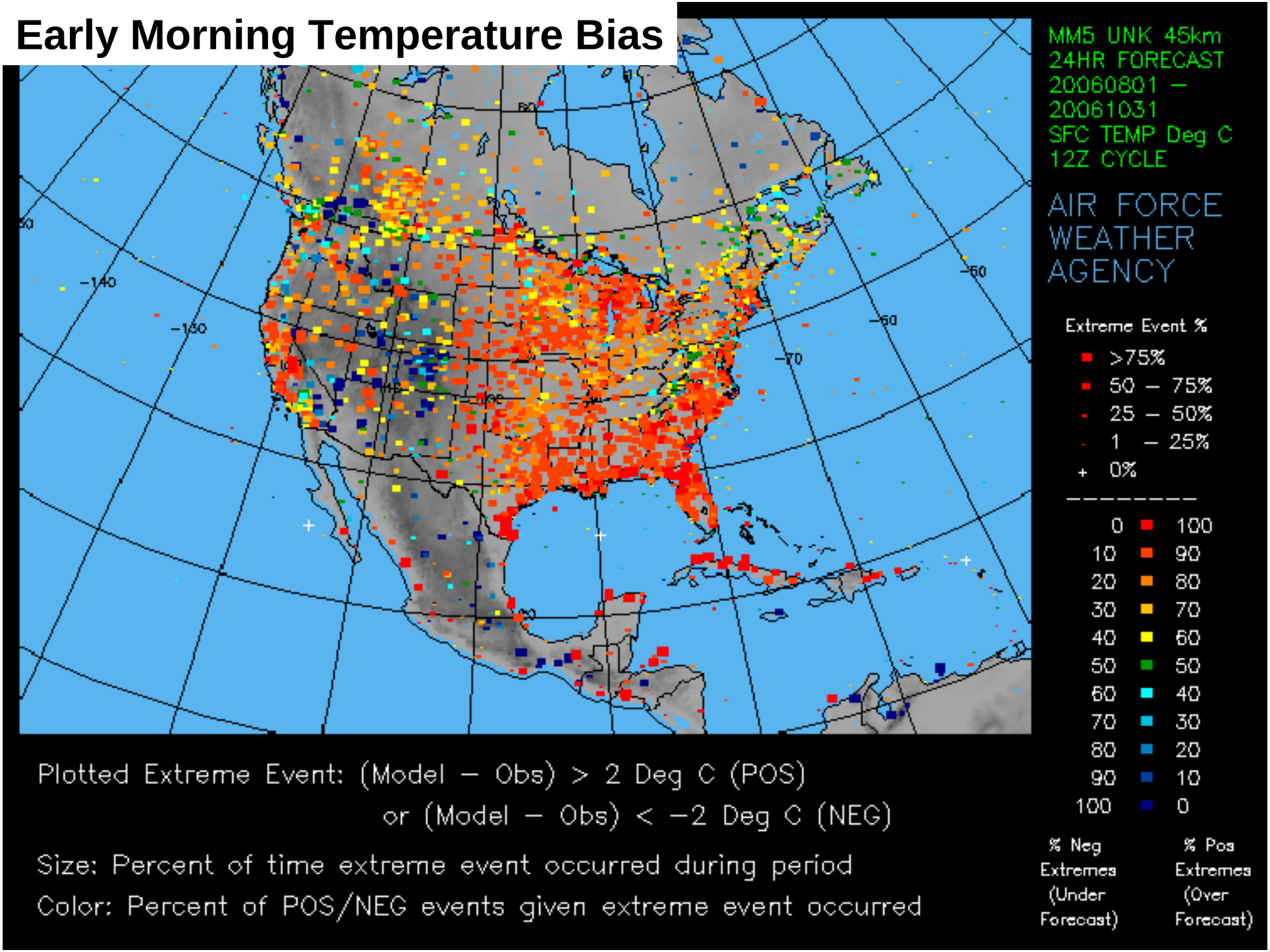
- **Major influences on model biases**
  - **Model Design**
    - **Core, physics, parameterizations, etc.**
  - **Analysis (model initial condition)**
  - **Weather "regime"**
    - **Location, season, weather pattern, etc.**
  - **Time of day and forecast lead time**



Late Afternoon Temperature Bias

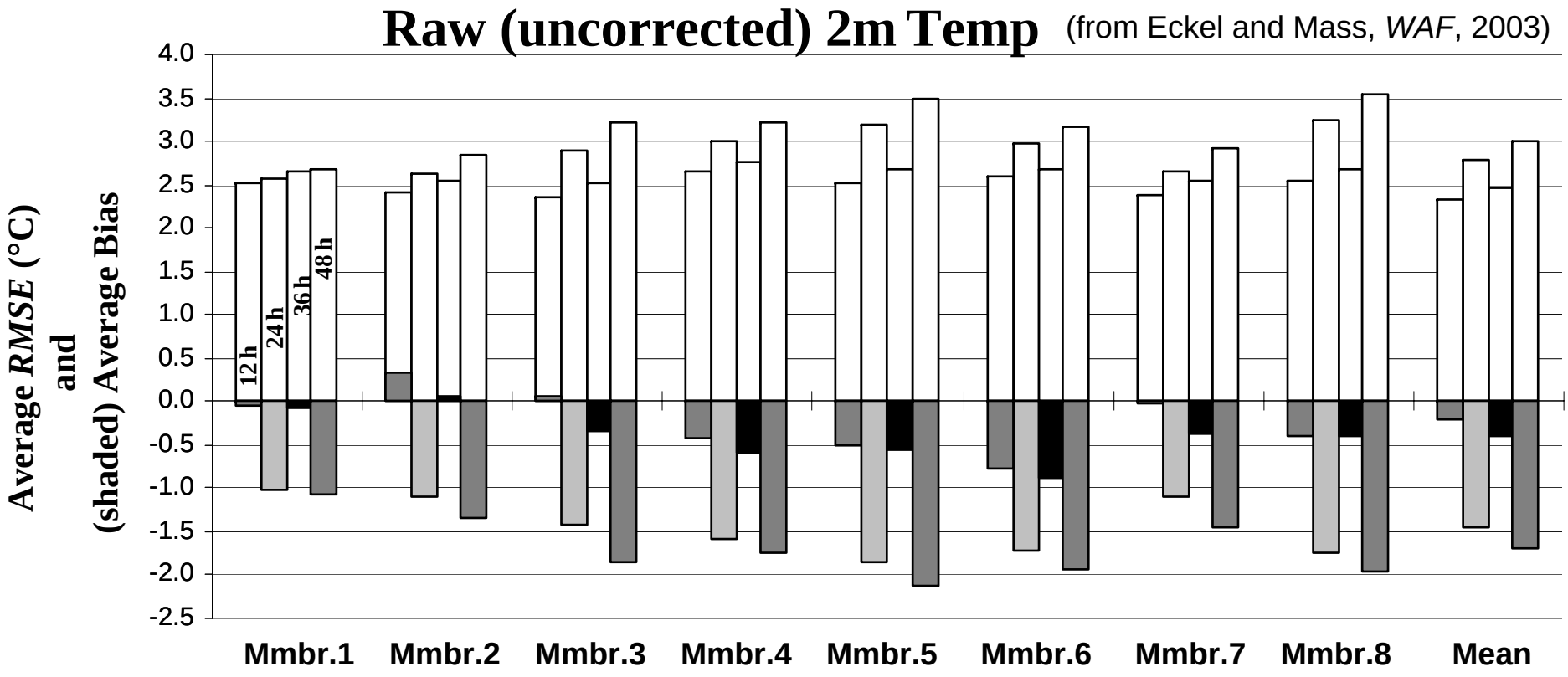


Early Morning Temperature Bias



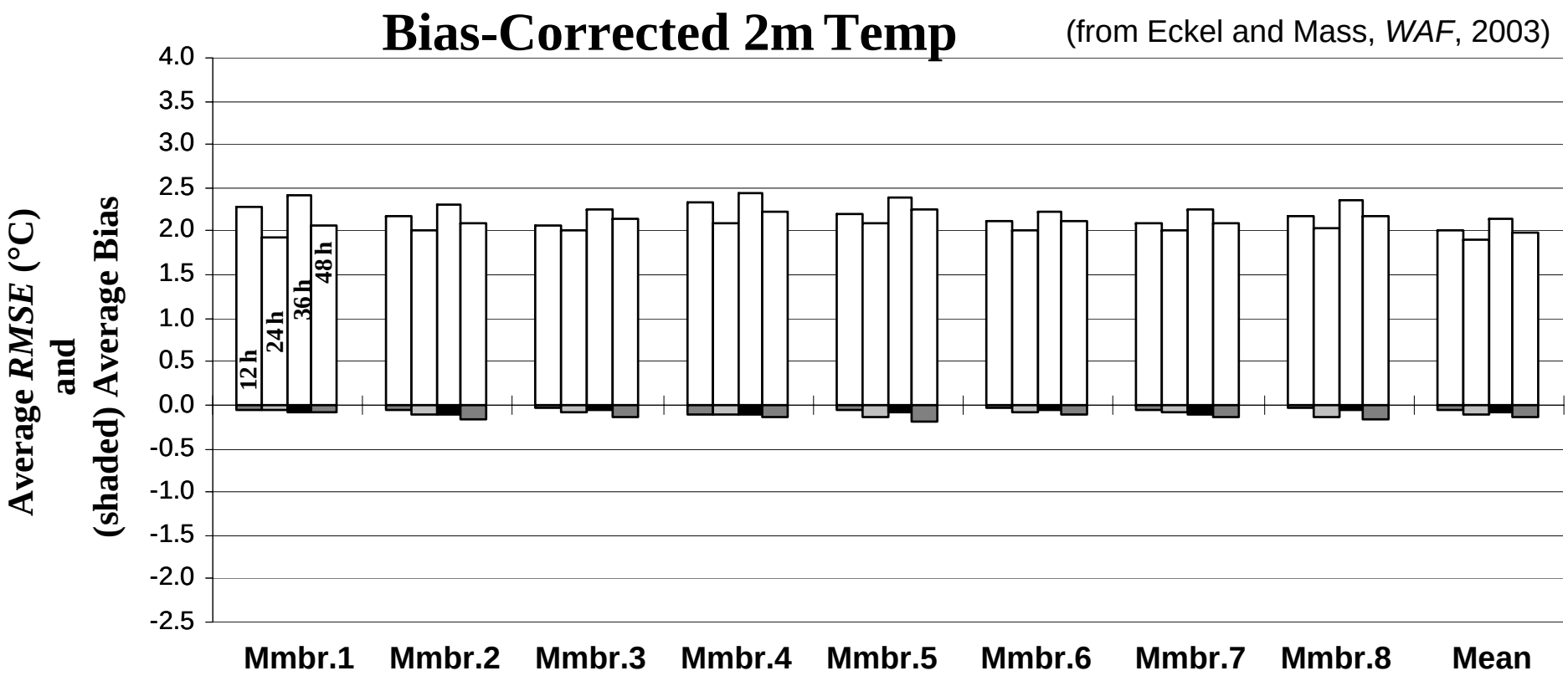
Univ. of WA Mesoscale Ensemble

- http://www.atmos.washington.edu/~ens/uwme.cgi
- 8 12-km MM5 Members over Pacific NW
- Initial Condition and Model Perturbations

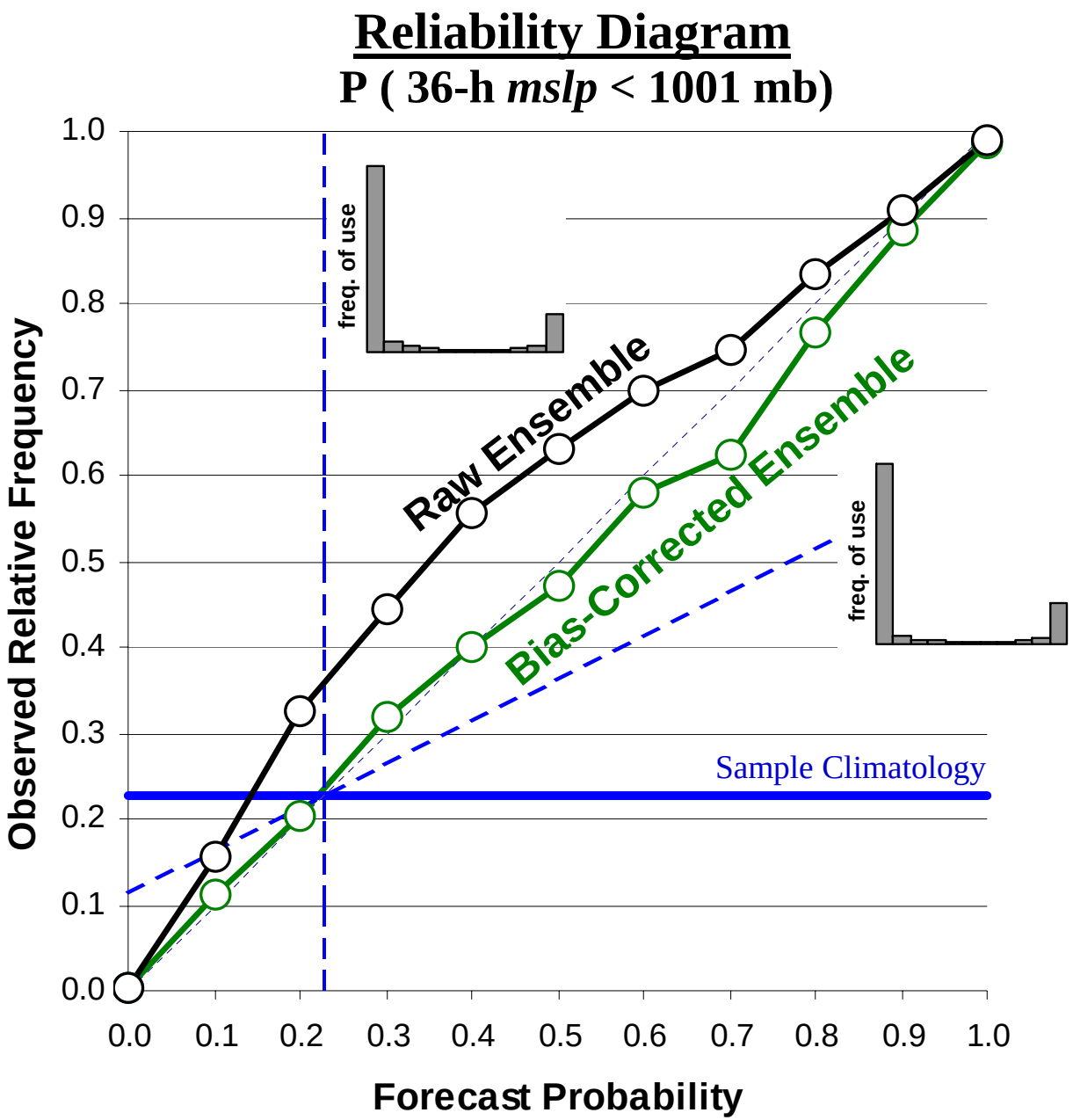


Univ. of WA Mesoscale Ensemble

- http://www.atmos.washington.edu/~ens/uwme.cgi
- 8 12-km MM5 Members over Pacific NW
- Initial Condition and Model Perturbations



Univ. of WA Mesoscale Ensemble



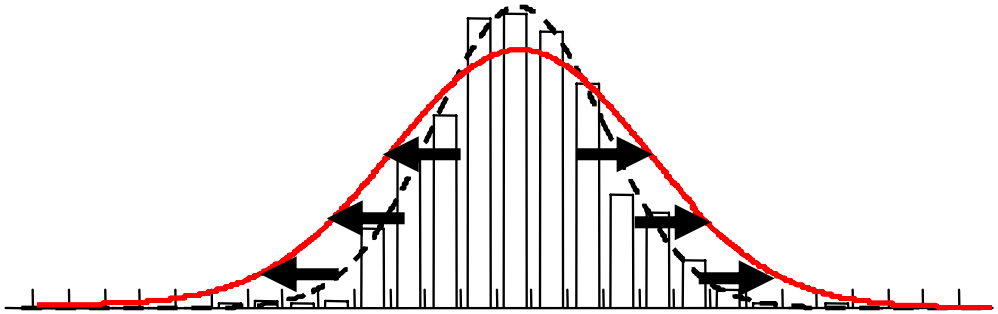
2nd Moment Calibration

(Ensemble Spread Correction)

2nd Moment Calibration,
(Ensemble Spread Correction)

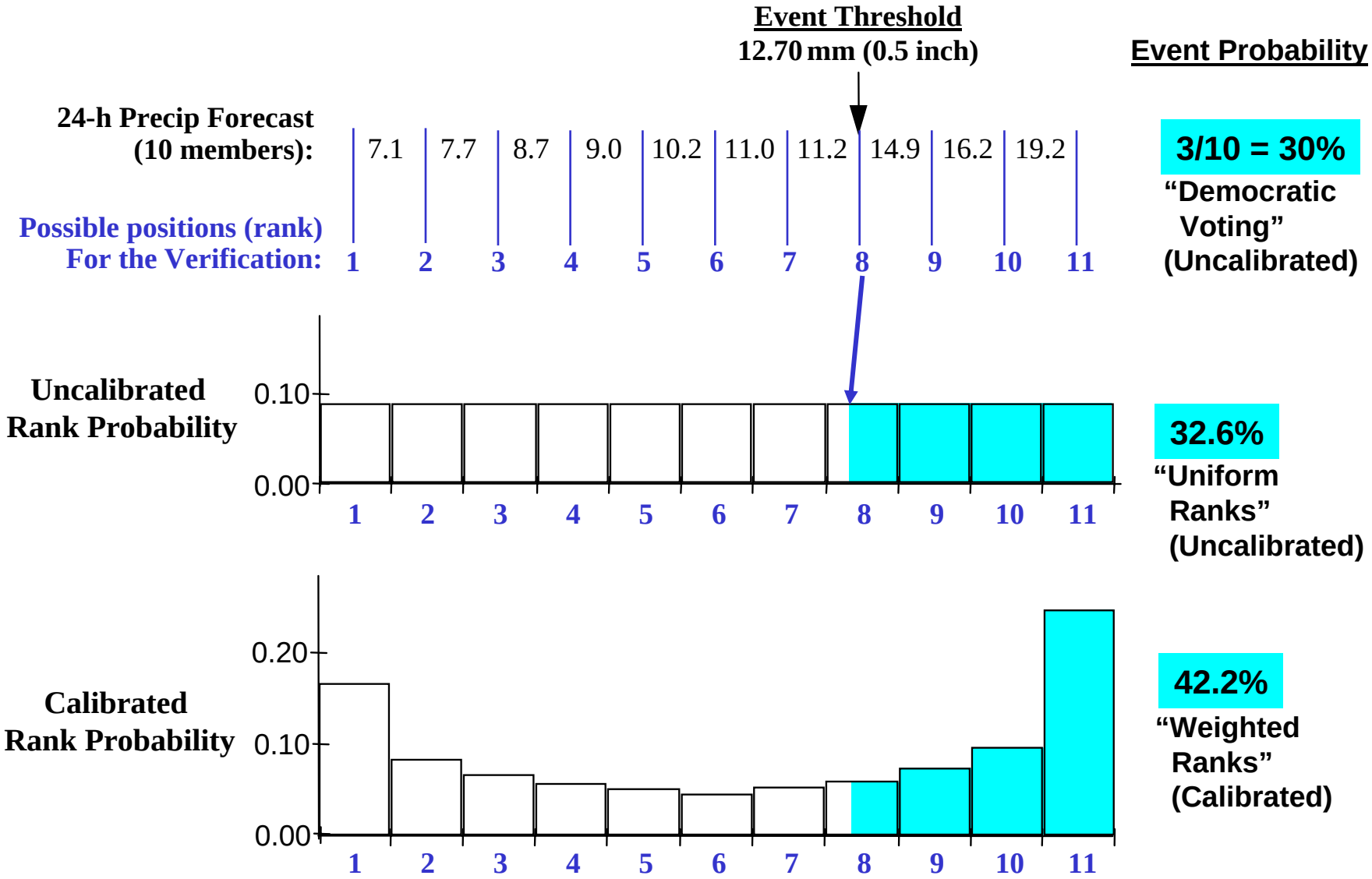
FOCUS: Increase reliability by adjusting width of forecast PDF

Most commonly, this means increasing the PDF width to account for insufficient dispersion (spread) in the raw ensemble.



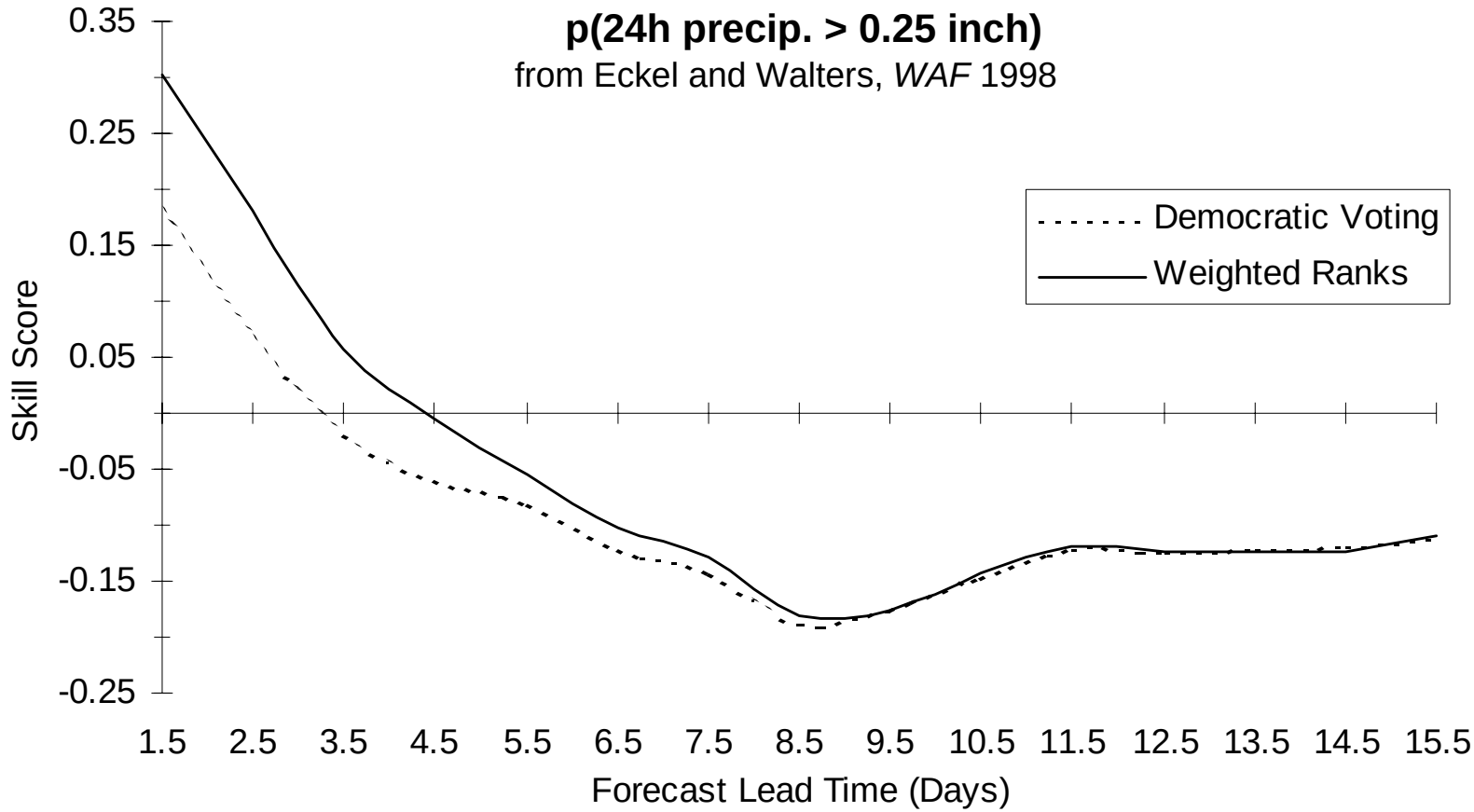
Tricky part: ***tails*** of the PDF

Calibration using Rank Histograms



References: Hamill and Colucci (*MWR*, 1997, 1998; Eckel and Walters, *WAF*, 1998; used at UKMO)

Calibration using Rank Histograms



Advantages:

(1) Demonstrated gain in skill from improved reliability

(2) Applies readily to different variables, regardless of pdf shape

Disadvantages:

(1) Odd pdf's, especially when two ensemble members close in value.

(2) Requires large training period to build robust histograms

(3) May reduce resolution

“Reality” Calibration

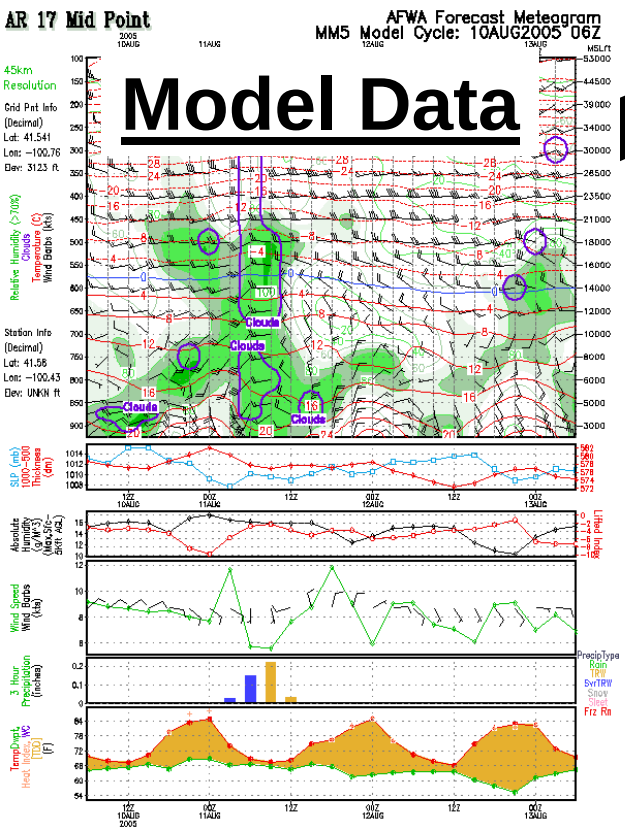
(Down Scaling)

Reality Calibration
(Down Scaling)

FOCUS: Increase ensemble sharpness by incorporating reality

- Refine spatial and/or temporal resolution from coarse model output

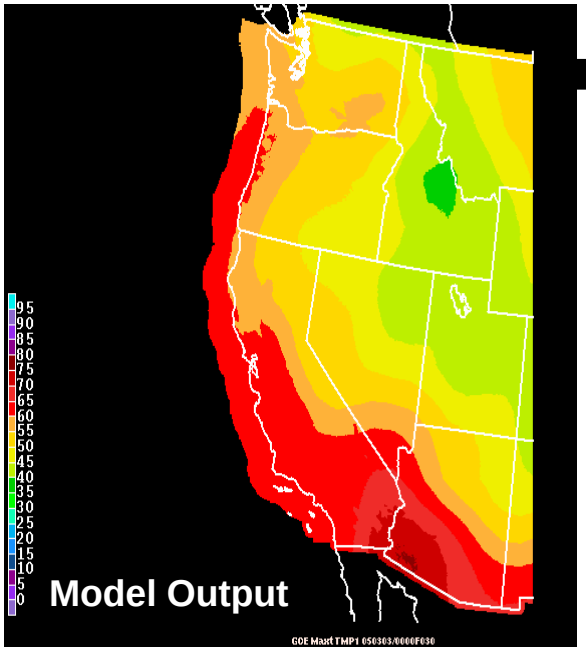
- Attempt to predict phenomena outside model's attractor
(e.g., wind in complex terrain, tornado, snow on roads, etc.)

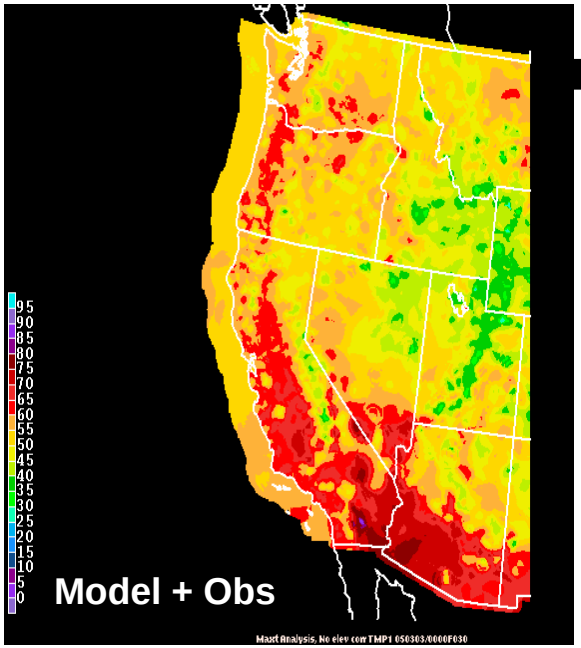


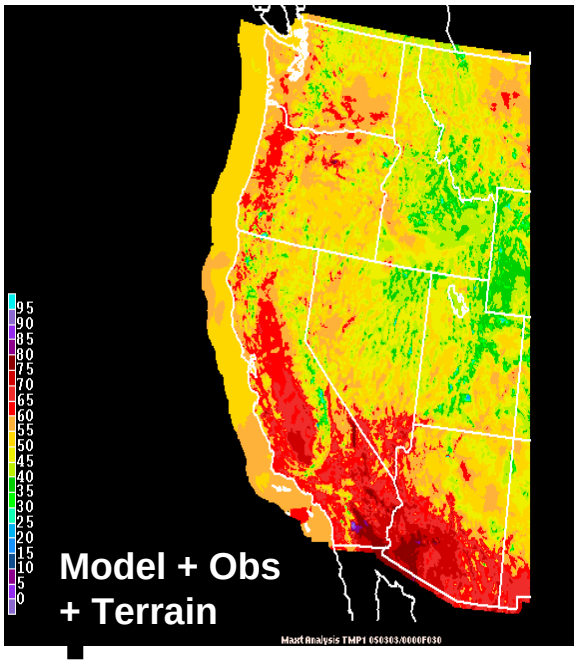
*Example Down Scaling: Gridded GFS MOS

"Blend" first guess and high-density station forecasts

Add further detail with high-resolution geophysical data and "smart" interpolation



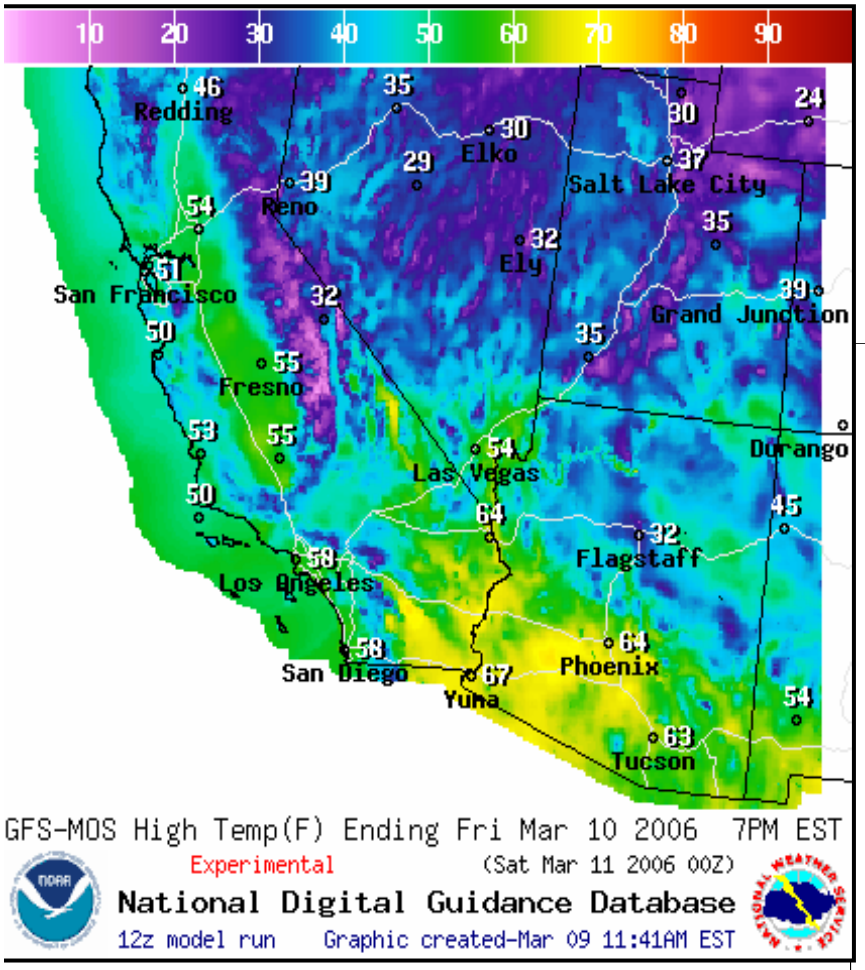


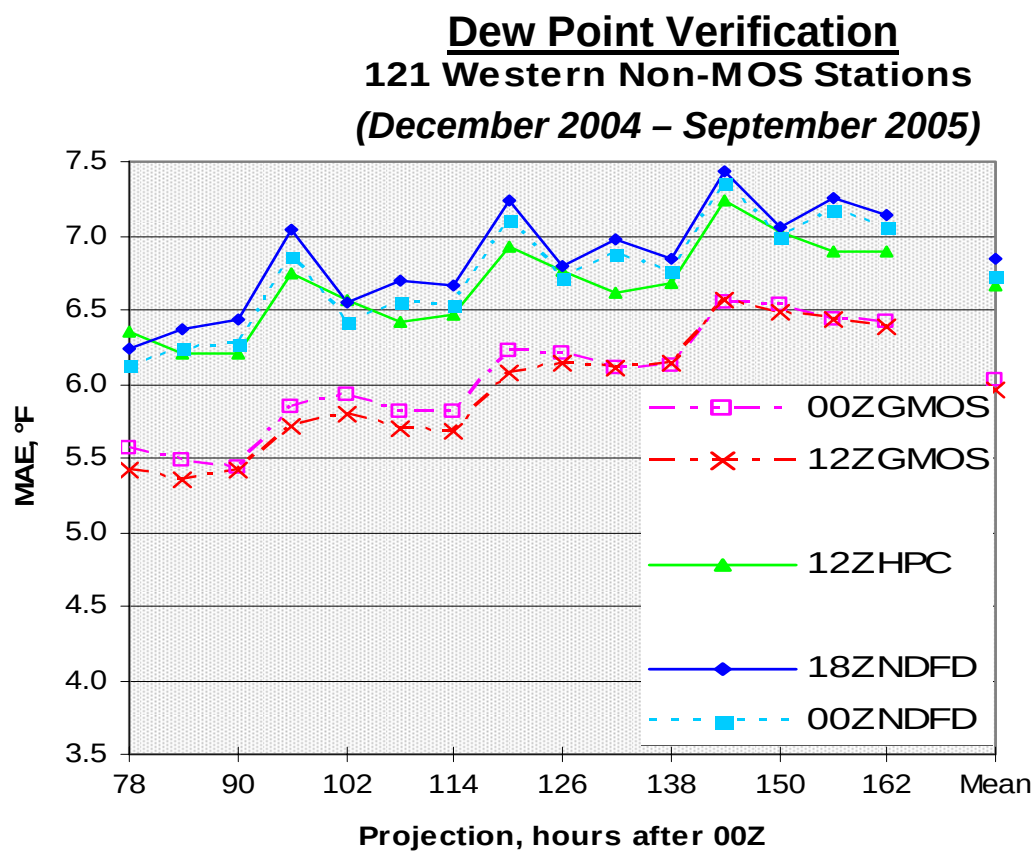


* Mark S. Antolik
Meteorological Development Laboratory
Statistical Modeling Branch
NOAA/National Weather Service
Silver Spring, MD

Use this fine-scale analysis as ground truth to train and down-scale gridded forecasts.

Example Down Scaling: Gridded GFS MOS

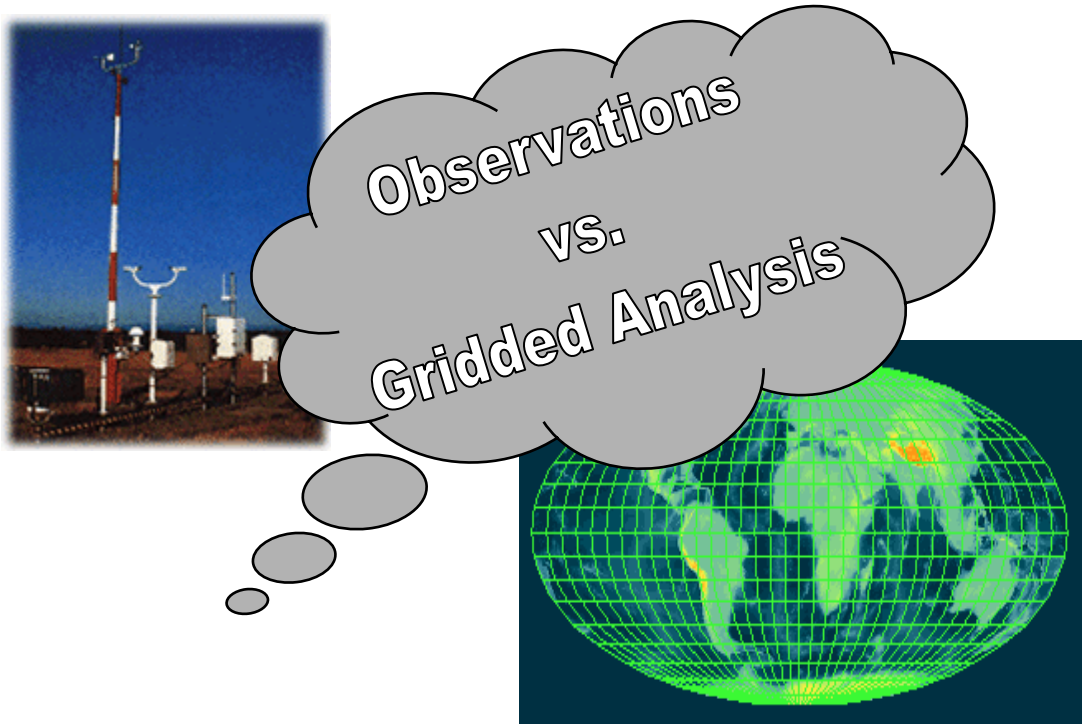




Ground Truth Issues

Ground Truth Issues

Calibration is only as good as what is used for truth!

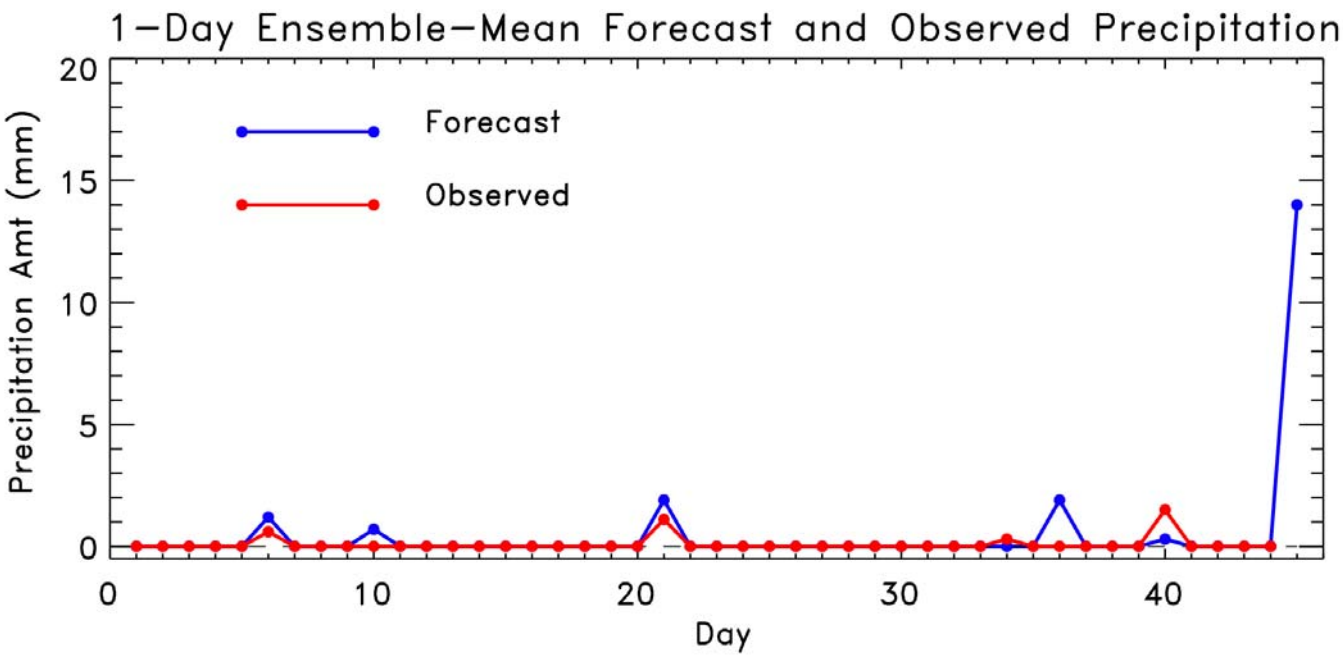


<u>Ideal Ground Truth</u>:

- Accurate & Precise
- Unbiased
- Independent
- Thorough
  - -- Represent all scales of interest
  - -- Encompass full range of climatologic possibilities
  - -- Encompass bulk of systematic forecast errors

Need for Reforecast Dataset

Example: *Attempting to calibrate precip with a just a short sample of recent events can be very problematic.*



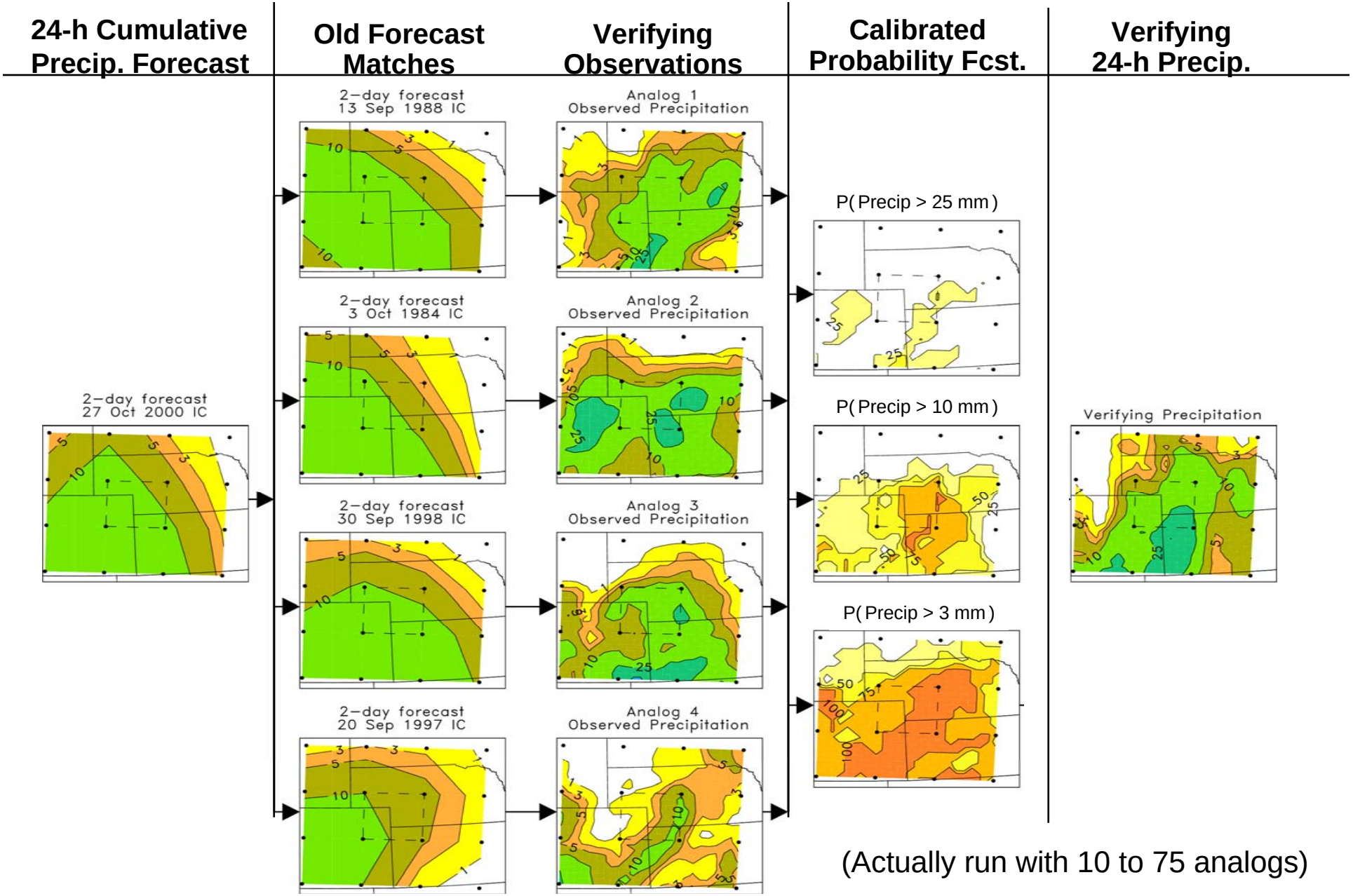
You'd like enough training data to have some similar events at a similar time of year to this one.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434-1447.
http://www.cdc.noaa.gov/people/tom.hamill/reforecast_mwr.pdf

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33-46.
http://www.cdc.noaa.gov/people/tom.hamill/refcst_bams.pdf

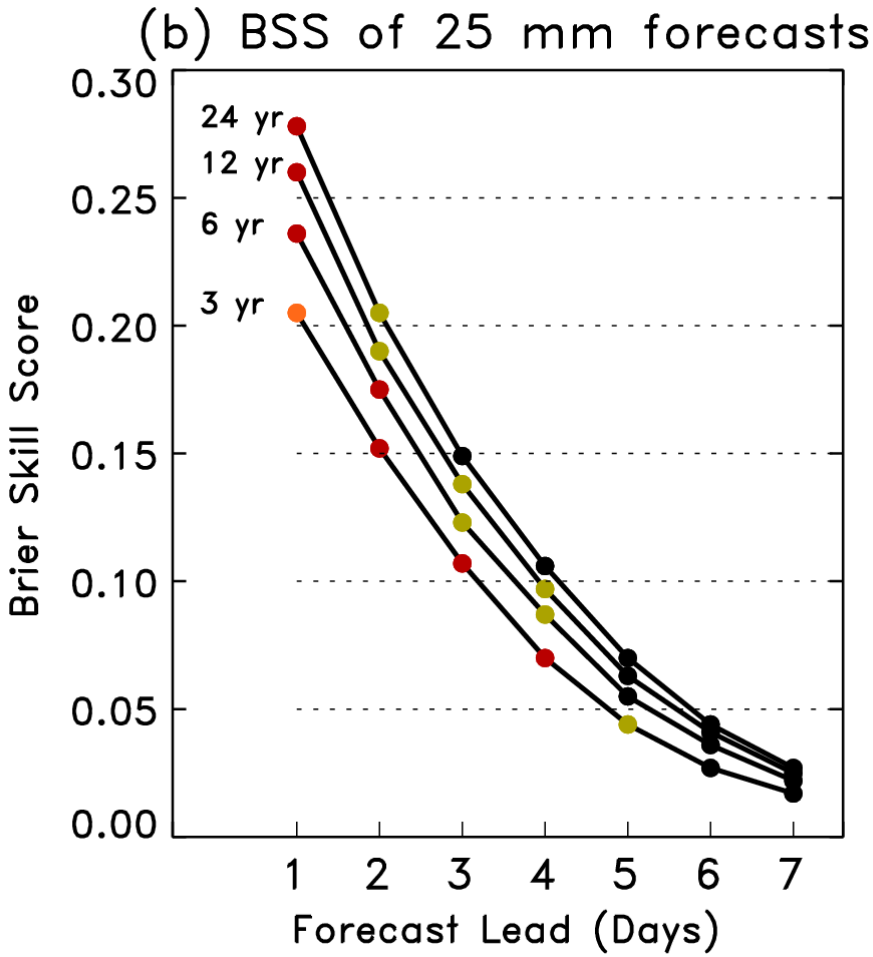
Example Application: Calibration by Analogs

(Hamill et al., *MWR*, 2004, *BAMS*, 2006)



(Actually run with 10 to 75 analogs)

Effect of Training Sample Size



(from Hamill et al., *BAMS*, 2006)

Colors of dots indicate which size analog ensemble provided the largest amount of skill.

- n=75
- n=50
- n=25
- n=10

Q: Do the benefits exceed the cost of building the reforecast dataset?

What to Know

Before using an ensemble, seek to answer:

1) How available are the data & products?
2) What is the model resolution & general skill?
3) How many members?
4) Technique for accounting for analysis uncertainty?
5) Technique for accounting for model uncertainty?

Cake

6) How well is the ensemble calibrated?
   a) Basics of technique(s)
      -- Correction for Bias, Spread, and Reality
   b) Extent of calibration
      -- Variables, lead times, etc.
   c) Ground truth quality
      -- Accuracy, scale, length of period, etc.

Icing

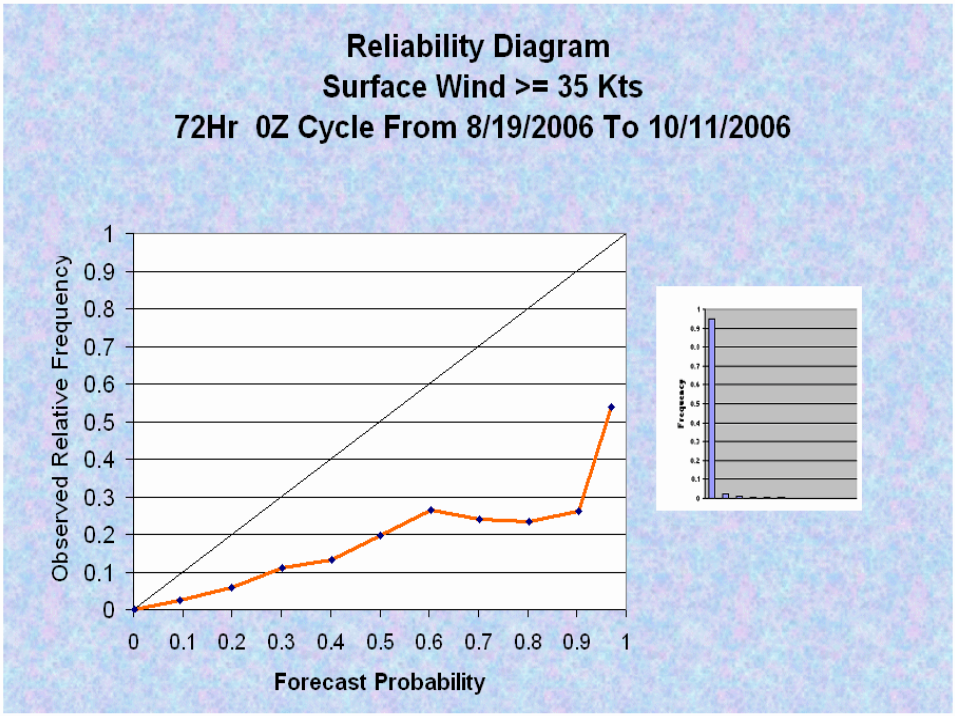
7) What strengths and weaknesses does V&V show?

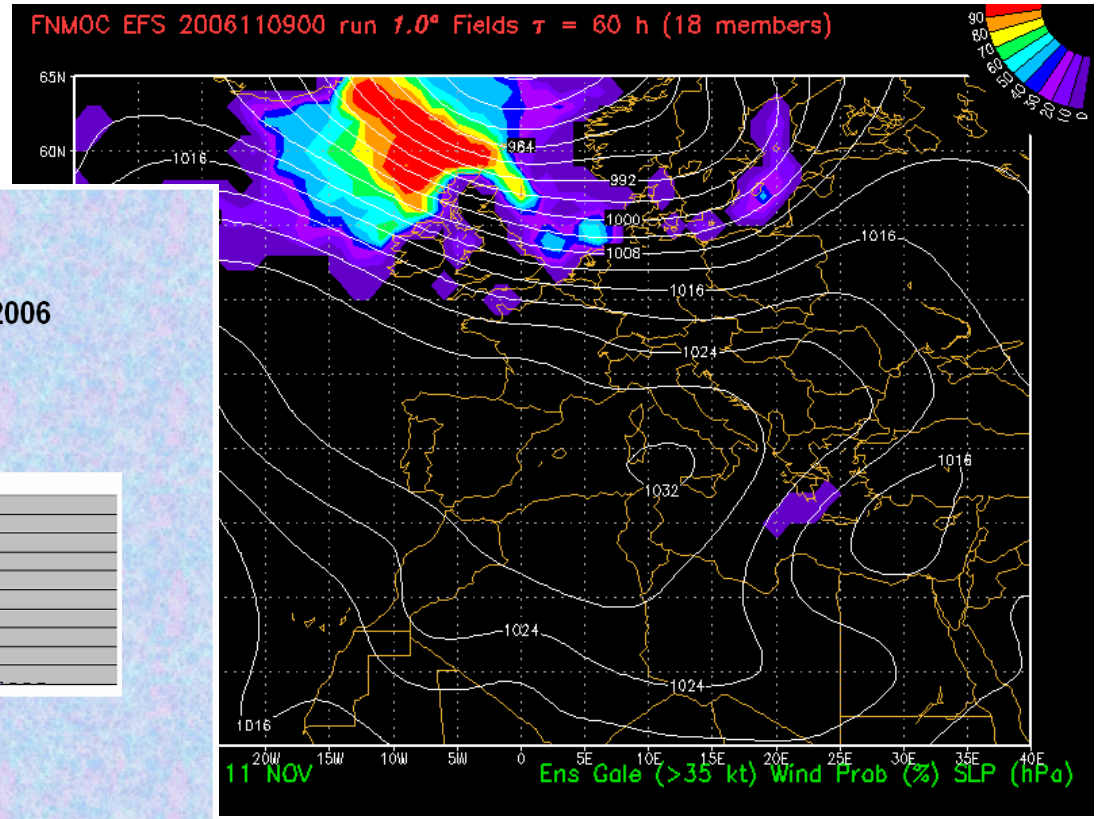
Fleet Numerical Meteorology and Oceanography Center (FNMOC)

NOGAPS Global Ensemble

https://www.fnmoc.navy.mil/PUBLIC/EFS/efs.html

1) Availability: 1 run per day (00Z), 10-day forecast
2) Resolution: T119/L24 (~120 km)
3) Members: 18
4) Analysis Perturbation: Bred Mode
5) Model Perturbation: *None*
6) **Calibration: *None***
7) *V&V: ???*





What is the utility of this product?

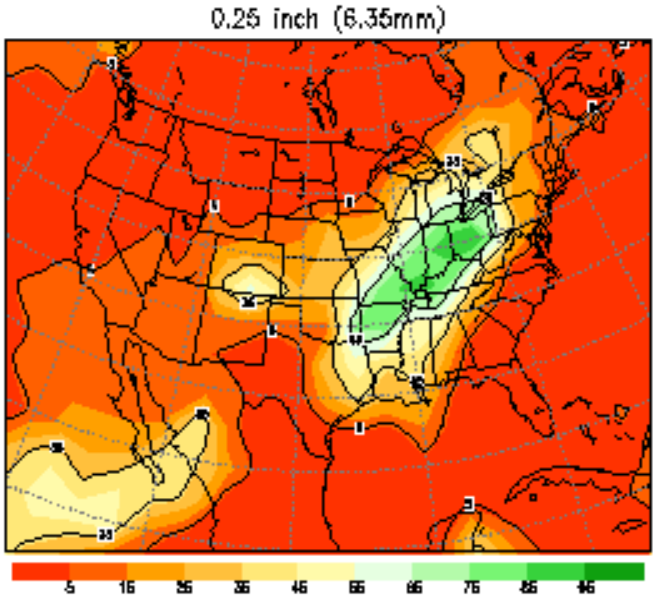
National Centers for Environmental Prediction (NCEP)

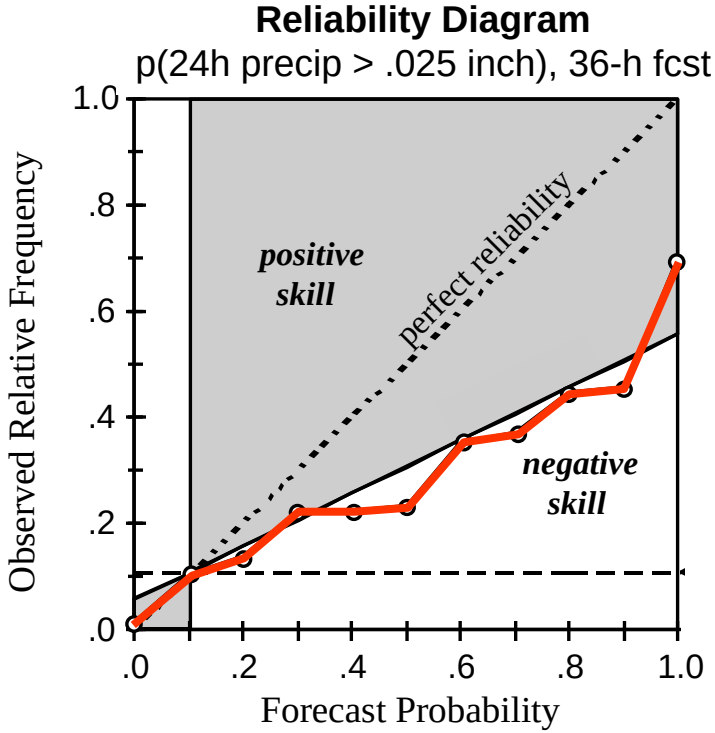
GFS Global Ensemble

http://wwwt.emc.ncep.noaa.gov/gmb/ens/

1) Availability: 4 runs per day (09Z, 21Z)), 15-day forecast
2) Resolution: T126/L28 (~110km)
3) Members: 15
4) Analysis Perturbation: Ensemble Transform Breeding
5) Model Perturbation: *None*
6) **Calibration:** *investigating and developing*
   - ***Bias correction***
   - ***2nd moment***

   } *Using GFS high-res control*
7) V&V: *only 500Hpa GPH posted*

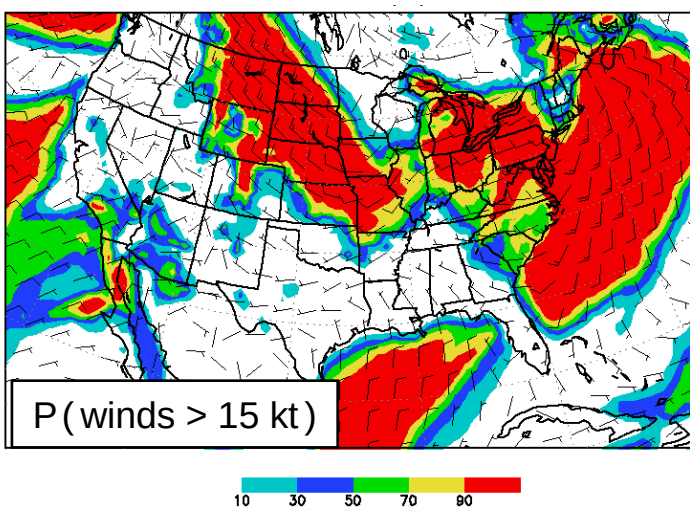
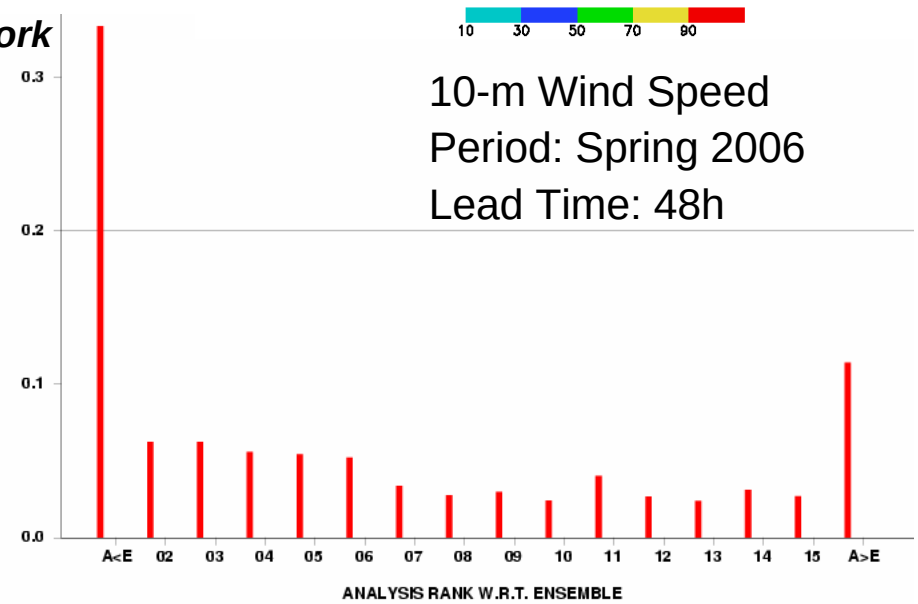




National Centers for Environmental Prediction (NCEP)

CONUS Short Range Ensemble Forecast (SREF)

http://wwwt.emc.ncep.noaa.gov/mmb/SREF/SREF.html

1) Availability: 09Z & 21Z, 87-hour fcst
2) Resolution: 32km
3) Members: 21
4) Analysis Perturbation: Regional Breeding
5) Model Perturbation: 4 Models, some multi-physics
6) **Calibration:** *investigating and developing*
   - ***Bias correction coming soon***
     - ***-- Summer 2007: Basic variables using "decaying average" with regional reanalysis as ground truth***
     - ***-- Winter 2007: Precipitation using Neural Network and Stage-IV precip analysis***
   - ***Spread Correction by BMA in design***
   - ***Down scaling planned***
     - ***-- Neural Network using 5km RTMA (real time mesoscale analysis)***

- V&V: Lots posted on website!