A few considerations for mesoscale forecast verification

Josh Hacker

NCAR

ICAP, Boulder, 22 Oct. 2014

Objective mesoscale verification

- Verify against observations
- Build samples with a sufficient number of cases
- Make an attempt at significance testing
- Usually several scores needed to understand the story

Fundamentals

NCAR

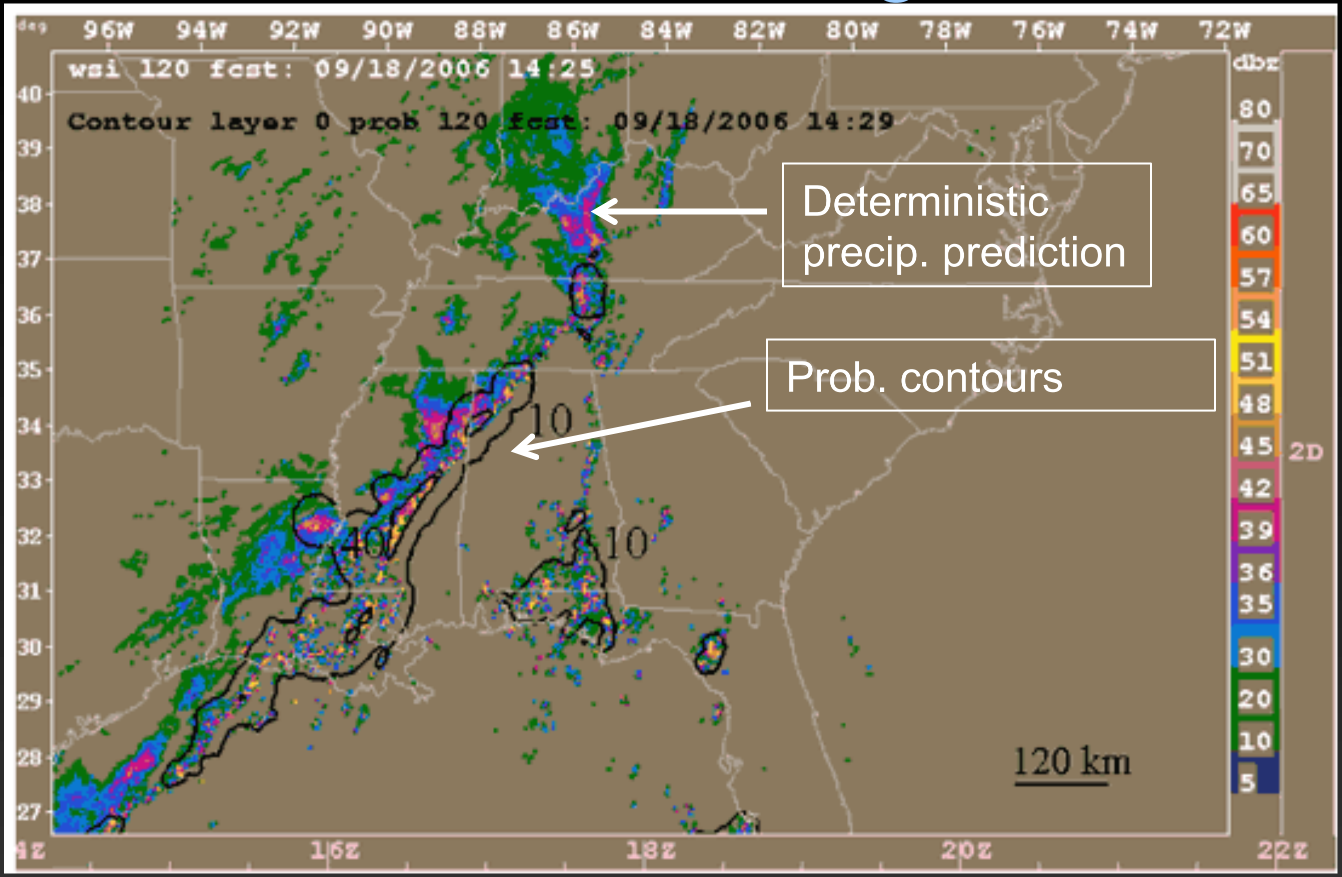
- Forecast errors and observation errors are generally the same order of magnitude
- Model errors (inadequacy) can be as important as initial-condition errors

Consequences

NCAR

- Forecast errors cannot be perfectly known
  - But given a sufficient sample the statistics of the errors can be estimated
  - Requires many forecasts to say anything meaningful
- Analysis errors cannot be perfectly known
  - But given a sufficient sample the statistics of the analysis errors can be estimated
  - Using analyses as a verification reference requires that analysis errors be considered (somewhat defeating the purpose!)
- Observation errors should be considered when possible
  - Biased observations can dominate forecast error statistics
- Large samples are often needed

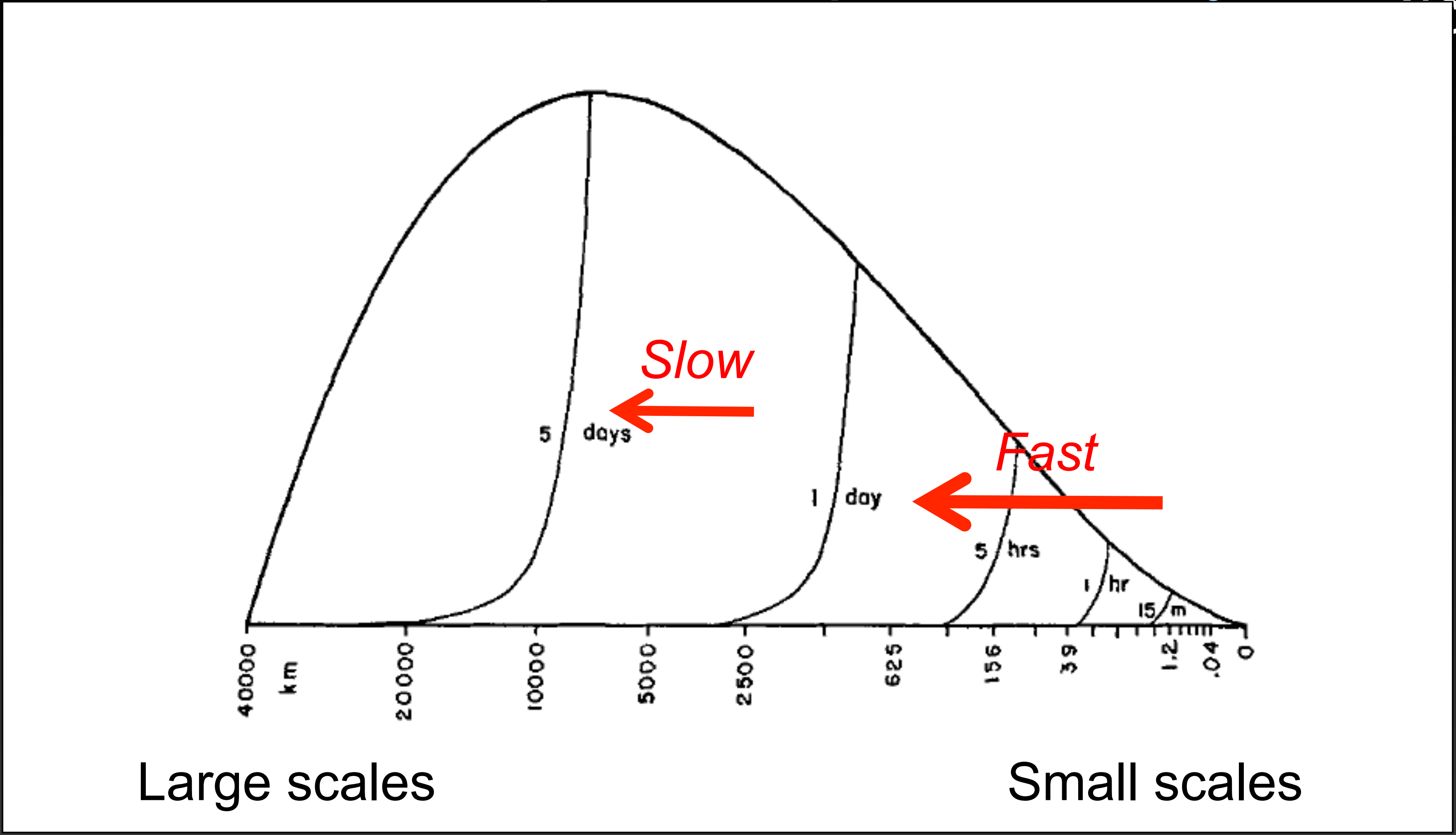
Mesoscale error growth



Highly skillful deterministic predictions of scales of O(1-10 km) are unreasonable to expect under most conditions and most norms.

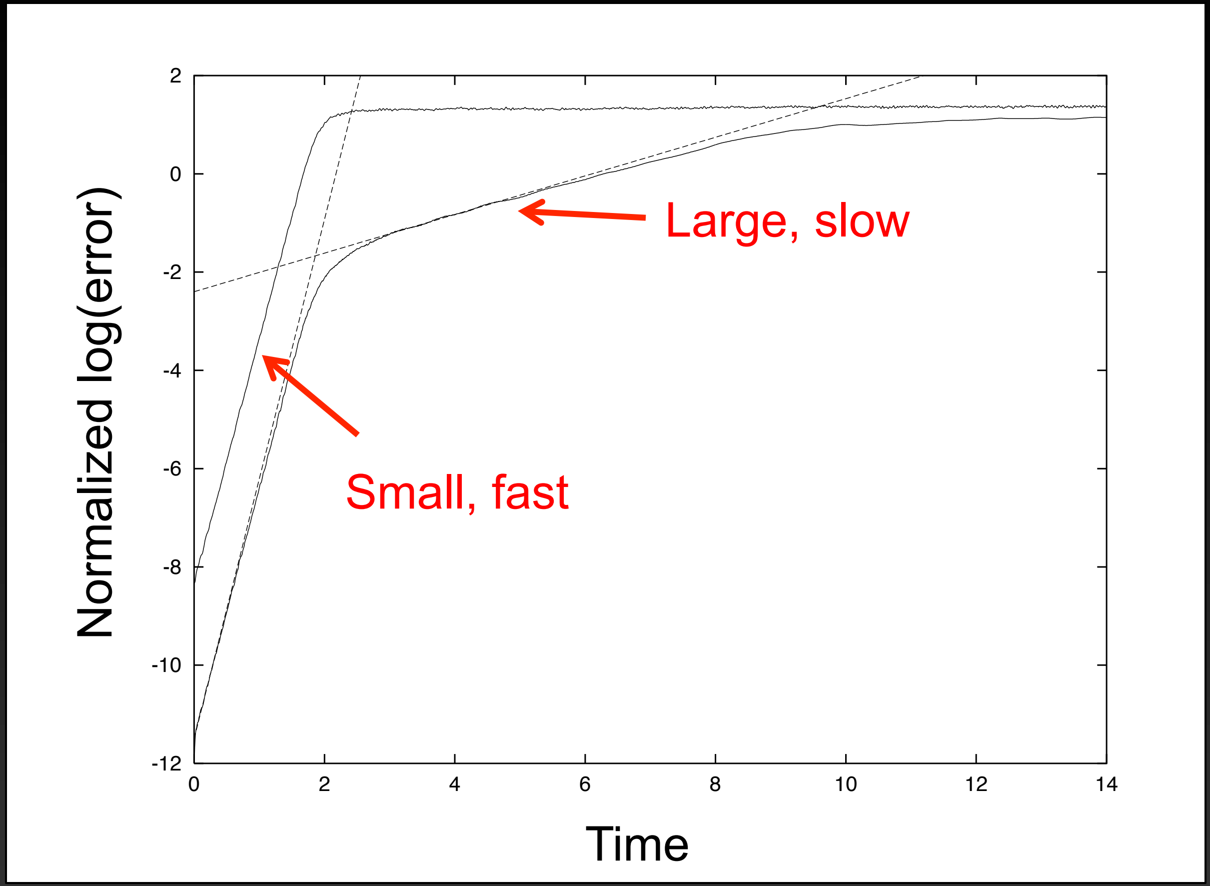
- Deterministic systems behave probabilistically
- Deterministic skill is difficult to detect
- Errors quickly grow to observation error levels

Scale-dependent predictability



From Lorenz (1969): Errors grow up-scale, and small-scale growth is much faster than large-scale growth.

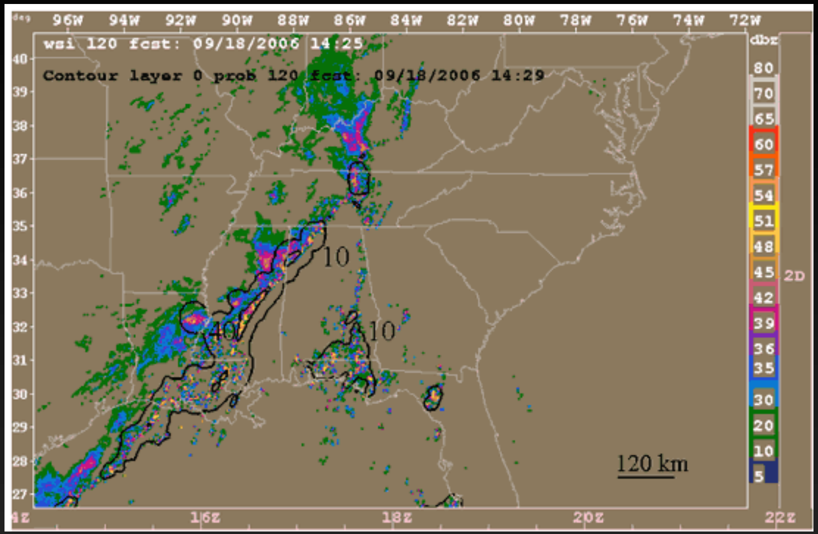
Scale-dependent predictability



Mesoscales are analogous to small, fast scales here

- Rapidly reach saturation
- If not normalized, saturation at much smaller levels of error energy than large (synoptic) scales

Mesoscale error growth



- Given a sufficient sample size:
  - Estimate mean errors (often called bias)
  - Estimate error variances
- Given an even bigger sample:
  - Break domain into sub-regions before averaging scores to avoid over-estimating skill
  - Verify temporal variances and/or spatial variances

NOTE: For now addressing deterministic skill

Verification against observations

- The data assimilation process filters high wavenumbers (analyses are filtered)
  - Filters observational noise
  - Filters background "noise" (really unpredictable scales)
  - Some physical features can be filtered
- Avoids complications from systematic errors in analyses
  - From model used in analysis
  - From data assimilation used in analysis
    - Forward operators
    - Ensemble size
    - Static/stationary error covariances

Systematic errors

NCAR

- Analyses retain at least some part of model bias
- Analyses retain at least some part of observation bias

for $\sigma_b^2 = \sigma_o^2 = \sigma^2$, and an unbiased observation:

$$x_a = \frac{1}{2}\left(x_b + y_o\right)$$

$$E\left(x_a\right) = \frac{1}{2}E\left(x_b + y_o\right) = \frac{1}{2}E\left(x_t + \varepsilon_b + y_t + \varepsilon_o\right) = \frac{1}{2}\left[E\left(x_t\right) + \beta_b + E\left(y_t\right) + \beta_o\right]$$

$$E\left(x_a\right) = \frac{1}{2}\left[E\left(x_t\right) + E\left(y_t\right)\right] + \frac{1}{2}\beta_b$$

Systematic errors

- Analyses retain at least some part of model bias
- Analyses retain at least some part of observation bias

for $\sigma_b^2 = \sigma_o^2 = \sigma^2$, and an unbiased model:

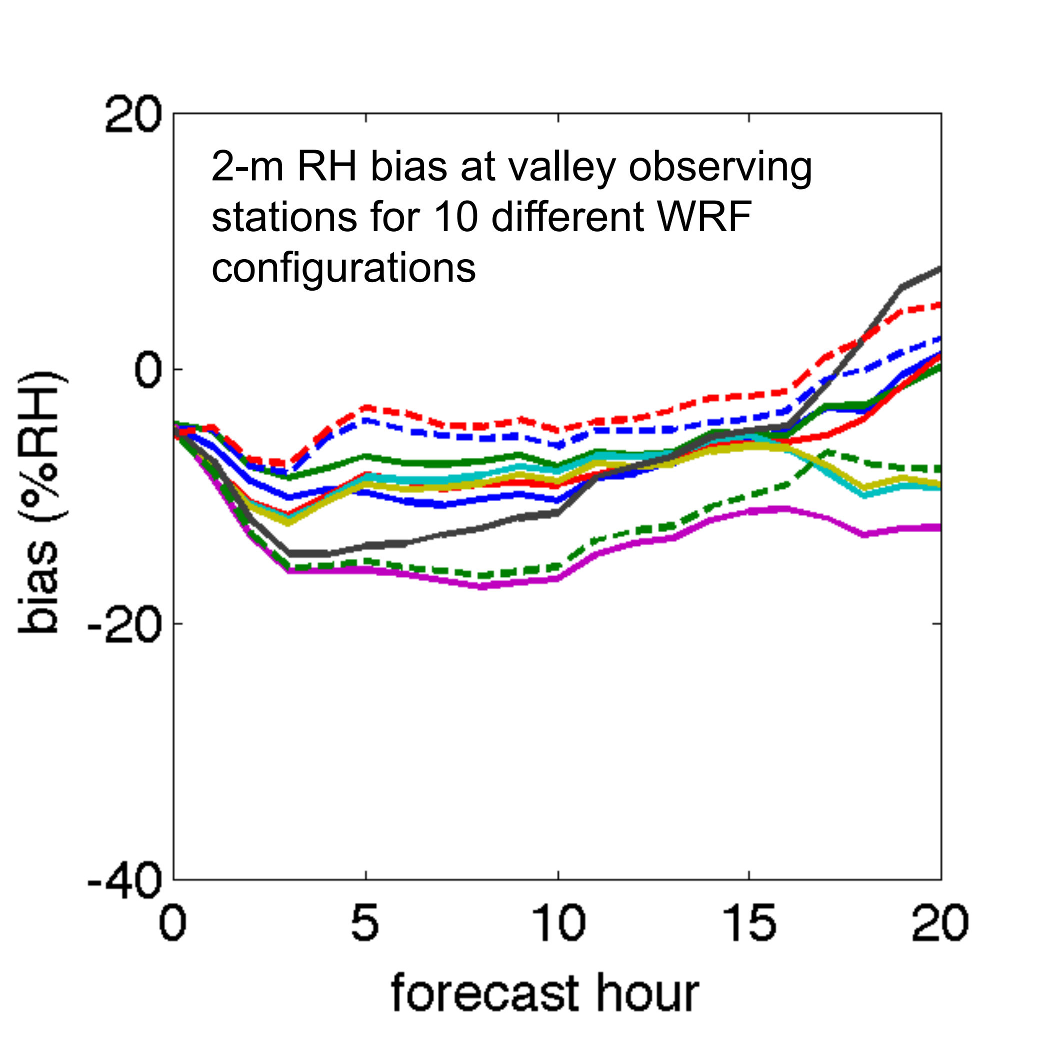
$$x_a = \frac{1}{2}\left(x_b + y_o\right)$$

$$E\left(x_a\right) = \frac{1}{2}E\left(x_b + y_o\right) = \frac{1}{2}E\left(x_t + \varepsilon_b + y_t + \varepsilon_o\right) = \frac{1}{2}\left[E\left(x_t\right) + \beta_b + E\left(y_t\right) + \beta_o\right]$$

$$E\left(x_a\right) = \frac{1}{2}\left[E\left(x_t\right) + E\left(y_t\right)\right] + \frac{1}{2}\beta_o$$

Inconsistent biases

NCAR



- Bias differences can appear quickly in a forecast (within most data assimilation cycling interval lengths)
- Biases can vary widely from model to model
- Bias differences can easily exceed observation error

Observation errors

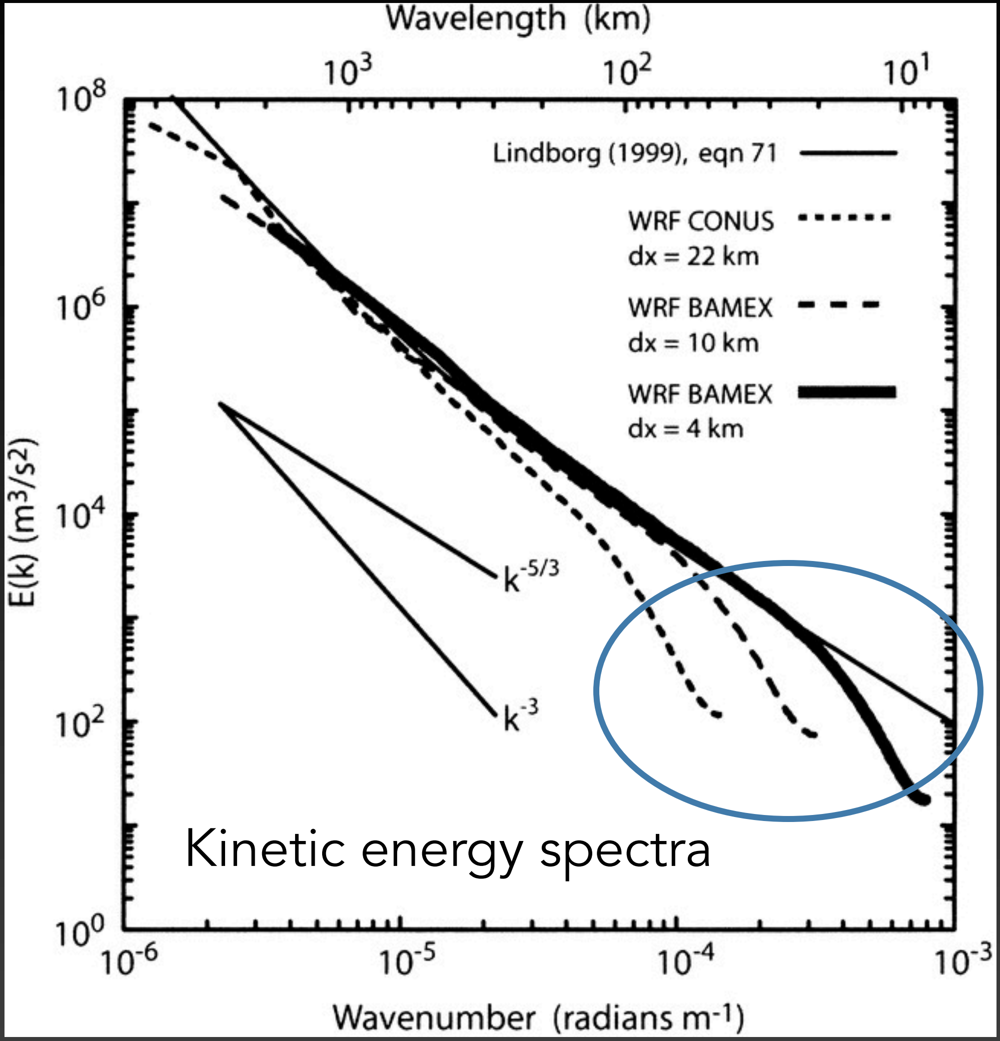
- Instrument error
  - bias and random error
  - may or may not be state dependent
- Random representativeness error
  - difference between modeled scales and observed scales
  - may or may not be state dependent
- Systematic representativeness error
  - constant (bias)
  - state dependent
  - *must be known to do something about it*

Observing scales

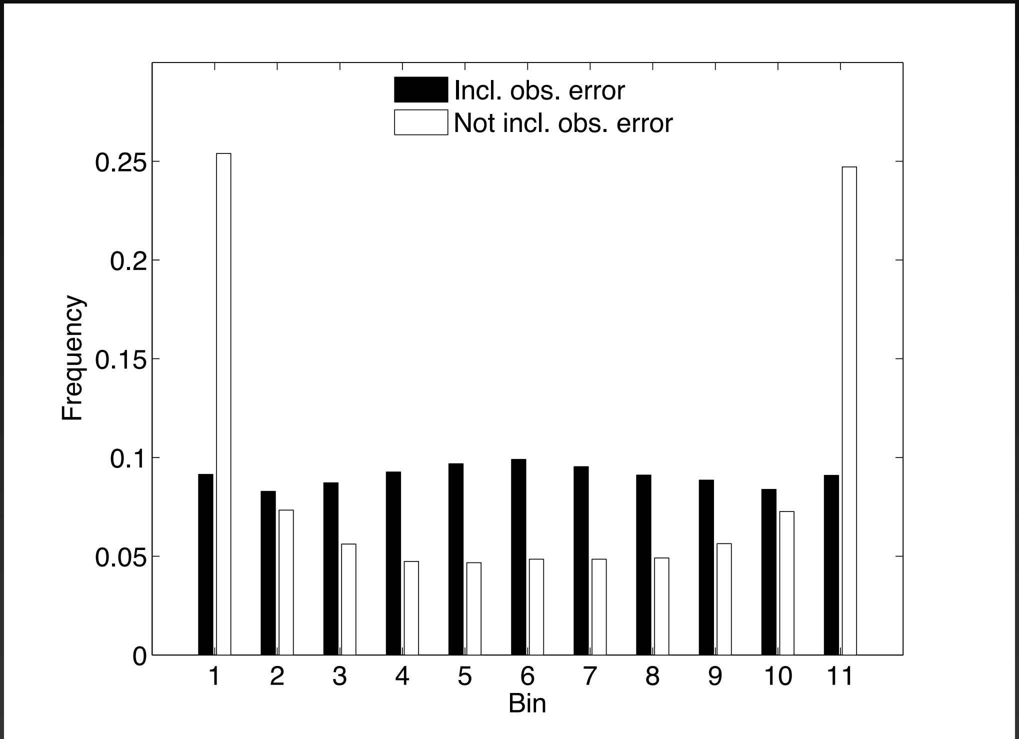
- An observation "sees" all scales of motion slower than its sampling rate
- Difference between variance in model and variance in an observation viewed as representativeness

Time-averaging an observation reduces the representativeness error, but not always clear in what way.

Skamarock (2004)



Including observation uncertainty

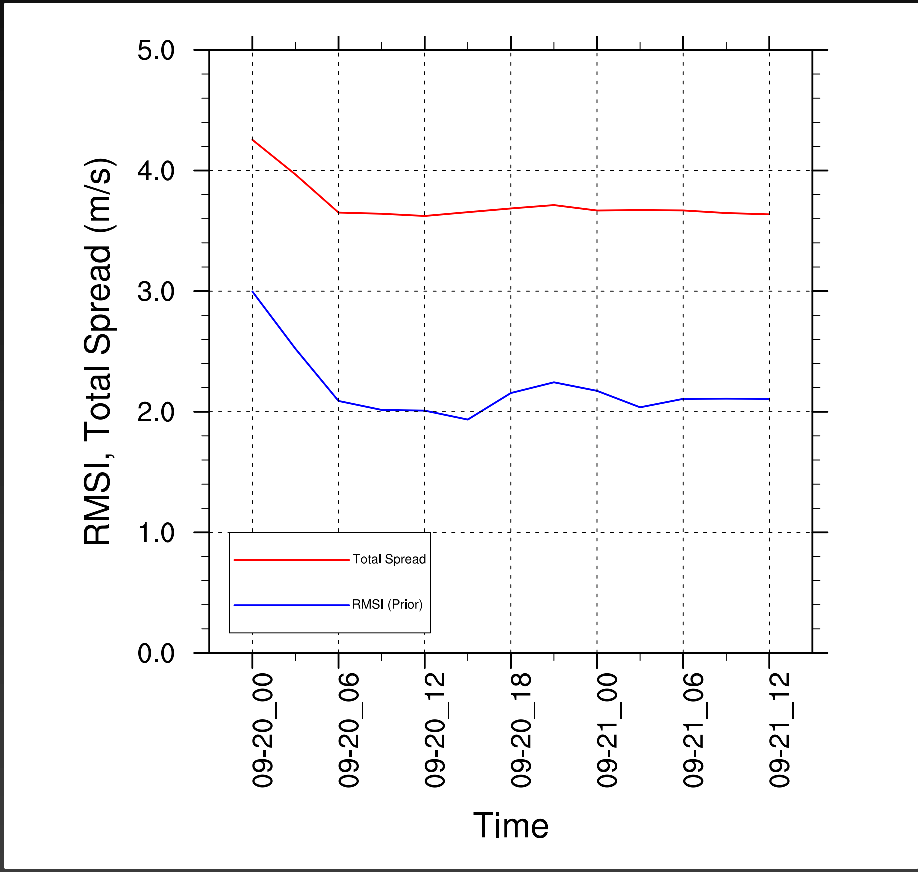


- Random errors in unbiased observations
- If included, the canonical underdispersive ensemble becomes pretty good or possibly overdispersive
- Is this an accurate estimate of the observation error variance? *Probably not in this case…*

Data assimilation to estimate random observation uncertainty

NCAR

forecast error = forecast uncertainty + observation uncertainty



- Derived from estimation theory
- Analogous to statistical consistency in ensemble prediction
- Result is for a particular model and data assimilation system
- Requires a good data assimilation system as a basis for estimation

Systematic observation errors

- Difficult (not impossible) to distinguish between model bias and observation bias in data assimilation

For $\sigma_b^2 = \sigma_o^2 = \sigma^2$ and an unbiased forecast:

$$x_a - x_b = \frac{1}{2}(x_b + y_o) - x_b = \frac{1}{2}(y_o - x_b)$$

$$2E(x_a - x_b) = E(y_o - x_b) = E(y_t - \varepsilon_b - x_b) = \left[E(y_t) + \beta_o - E(x_b)\right]$$

$$= \left[E(x_t) + \beta_o - E(x_t) - E(\varepsilon_b)\right] = \beta_o$$

Given an unbiased model, it is easy to estimate observation bias.

Systematic model errors

- Difficult (not impossible) to distinguish between model bias and observation bias in data assimilation

For $\sigma_b^2 = \sigma_o^2 = \sigma^2$ and an unbiased observation:

$$x_a - x_b = \frac{1}{2}\left(x_b + y_o\right) - x_b = \frac{1}{2}\left(y_o - x_b\right)$$

$$2E\left(x_a - x_b\right) = E\left(y_o - x_b\right) = E\left(y_o - x_t - \varepsilon_b\right) = \left[E\left(y_o\right) - E\left(x_t\right) - \beta_b\right]$$

$$= \left[E\left(x_t\right) + E\left(\varepsilon_o\right) - E\left(x_t\right) - \beta_b\right] = \beta_b$$

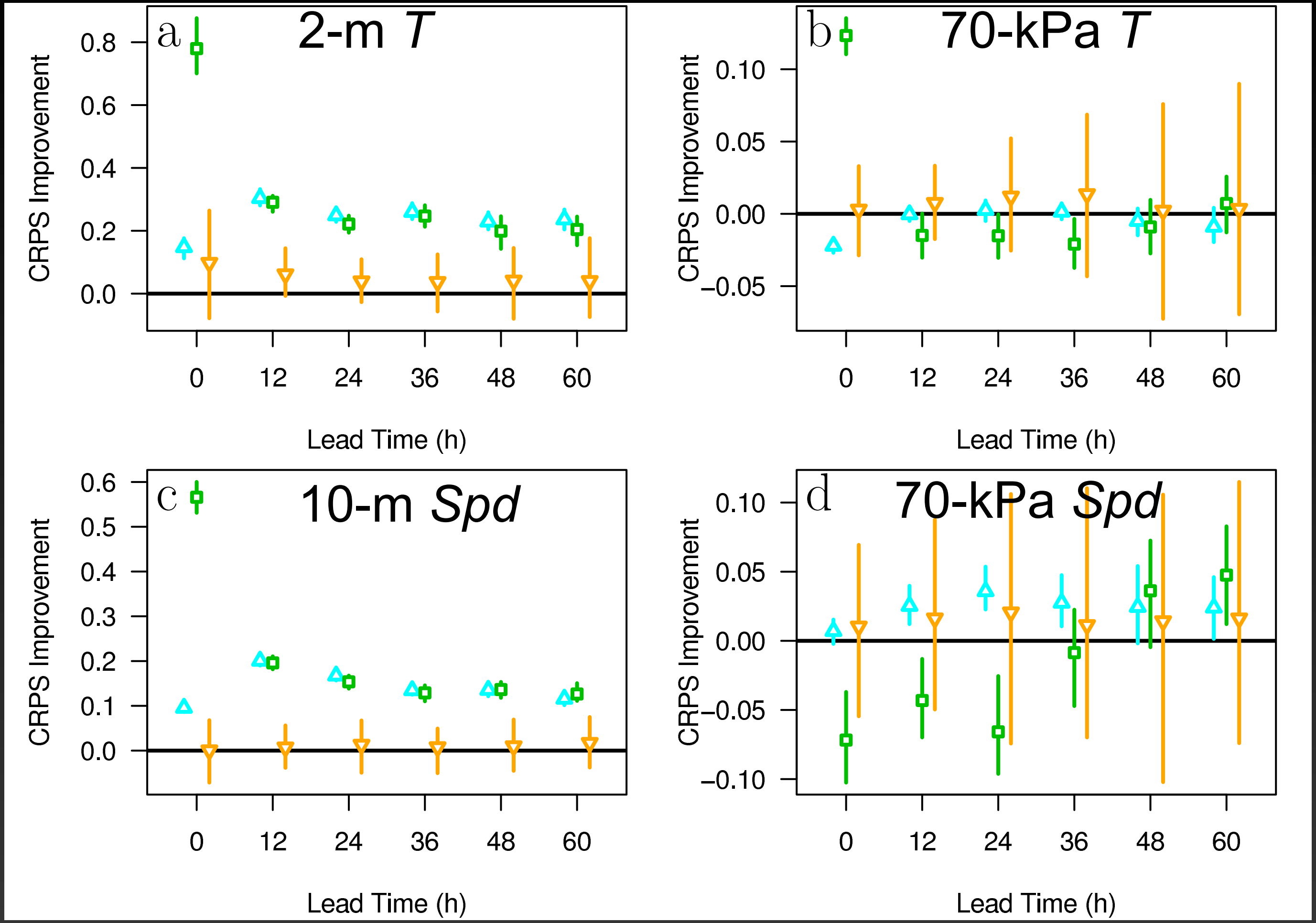
Given an unbiased observation, it is easy to estimate model bias.

It is possible to estimate both both observation and model biases simultaneously in data assimilation; a set of unbiased observations makes life easier.

Verification sample size

- As large as possible
- Necessary sample size depends on how samples are formed
- Key is ability to form independent samples
  - Large spatial distances between observing points
  - Different microclimates (mean conditions and variability)
- Examples:
  - Global prediction systems can provide good statistics with ~2 weeks of forecasts
  - Mesoscale/regional prediction systems may require much more than a month
  - Can be improved by spreading cases out in time

Significance testing



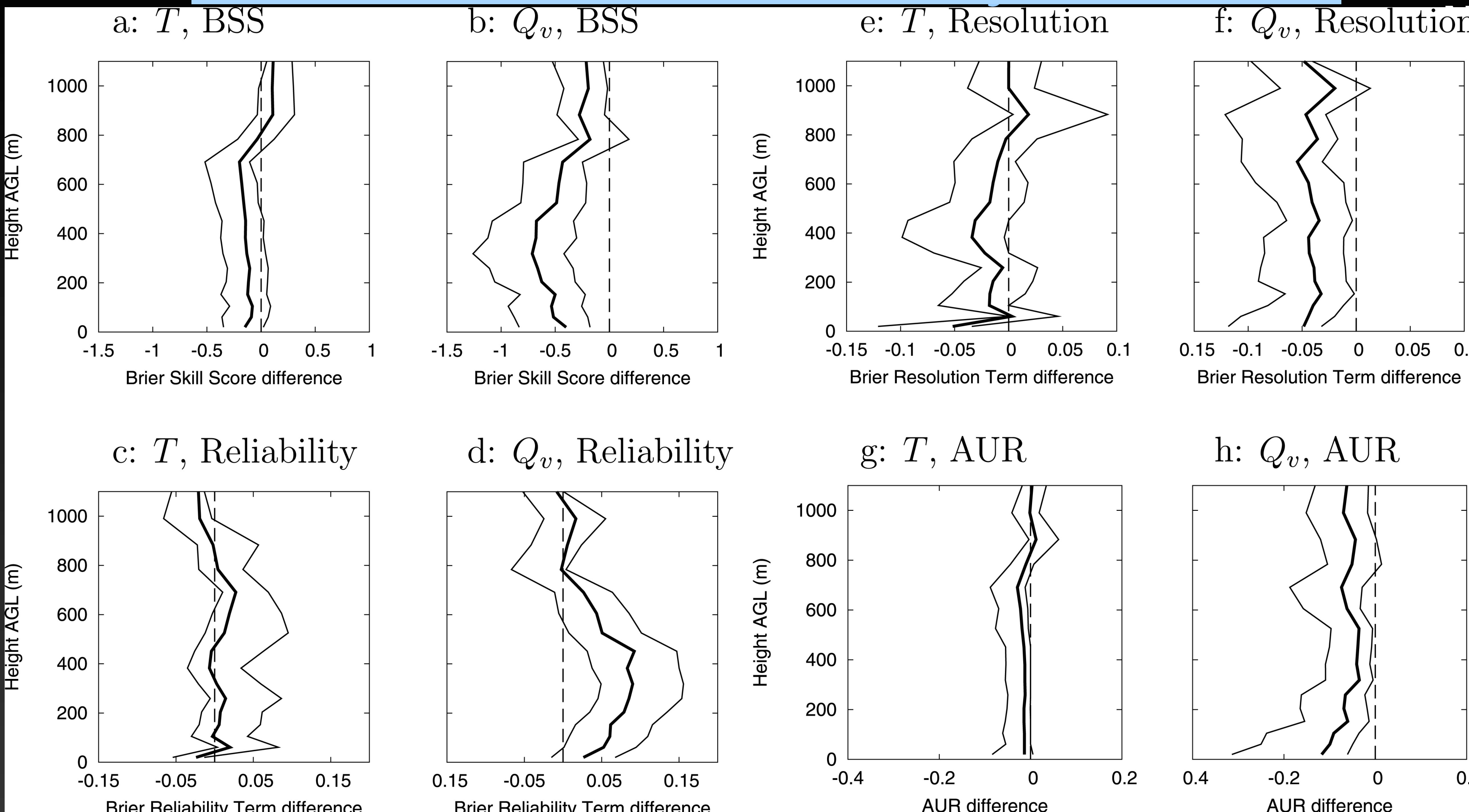
- Test on score differences to avoid under-estimating significance
- Here bootstrapping is a useful approach (not perfect)
- Room for creativity

Probabilistic mesoscale forecasting

NCAR

- No silver bullet
- Need to look at several scores or metrics
- The goal is usually phrased: maximize resolution subject to reliability
  - Reliability: climatological agreement between probabilistic forecasts and observations
  - Resolution: skill in predicting probabilities that are far from the climatological mean probability
  - Discrimination: events versus non-events
- Methods (examples):
  - Rank histograms (reliability)
  - Receiver Operating Characteristic (ROC) curve (discrimination)
  - Attributes diagram (reliability, resolution, discrimination, sharpness, conditional bias)
  - Rank probability score and related (reliability and resolution)

Resolution and reliability tradeoffs



Decomposition of Brier Skill Score differences shows one forecast system has better reliability, resolution, and discrimination. This consistency is not guaranteed.

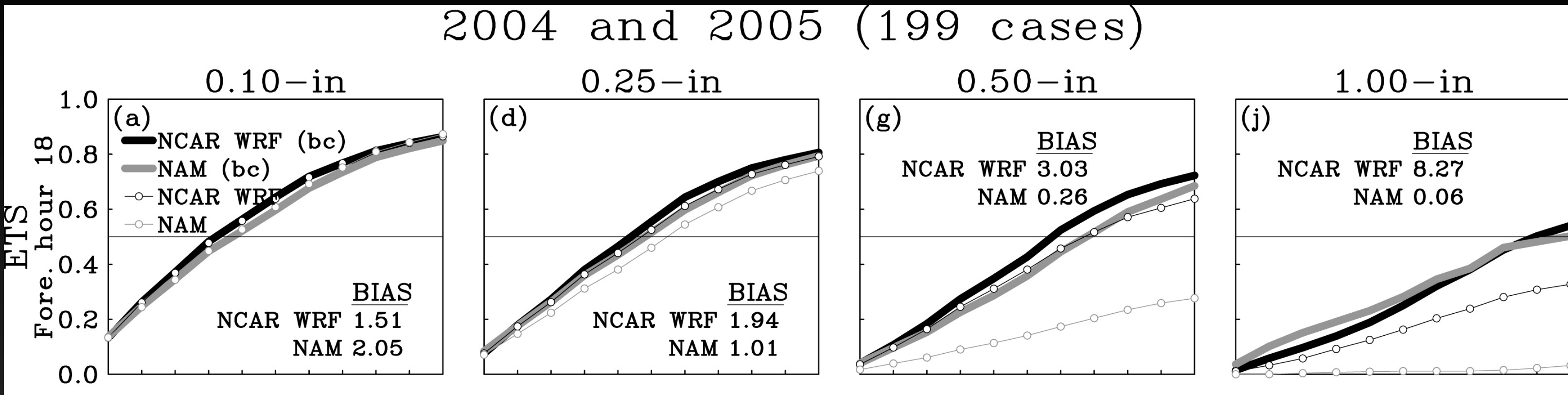
From Rostkier-Edelstein and Hacker, 2013: Impact of Flow Dependence, Column Covariance, and Forecast Model Type on Surface-Observation Assimilation for Probabilistic PBL Profile Nowcasts. *Wea. Forecasting,* **28**, 29–54.

Neighborhood methods

- Rely on:
  - Fact that exact timing and location is not predictable
  - Intuition for what range of spatial or temporal errors are acceptable
- Recognize lack of deterministic skill
- Related to "fuzzy" methods
- Several published methods available
- Observation errors not yet considered in any work using neighborhood methods (that I know of)

Neighborhood methods

NCAR

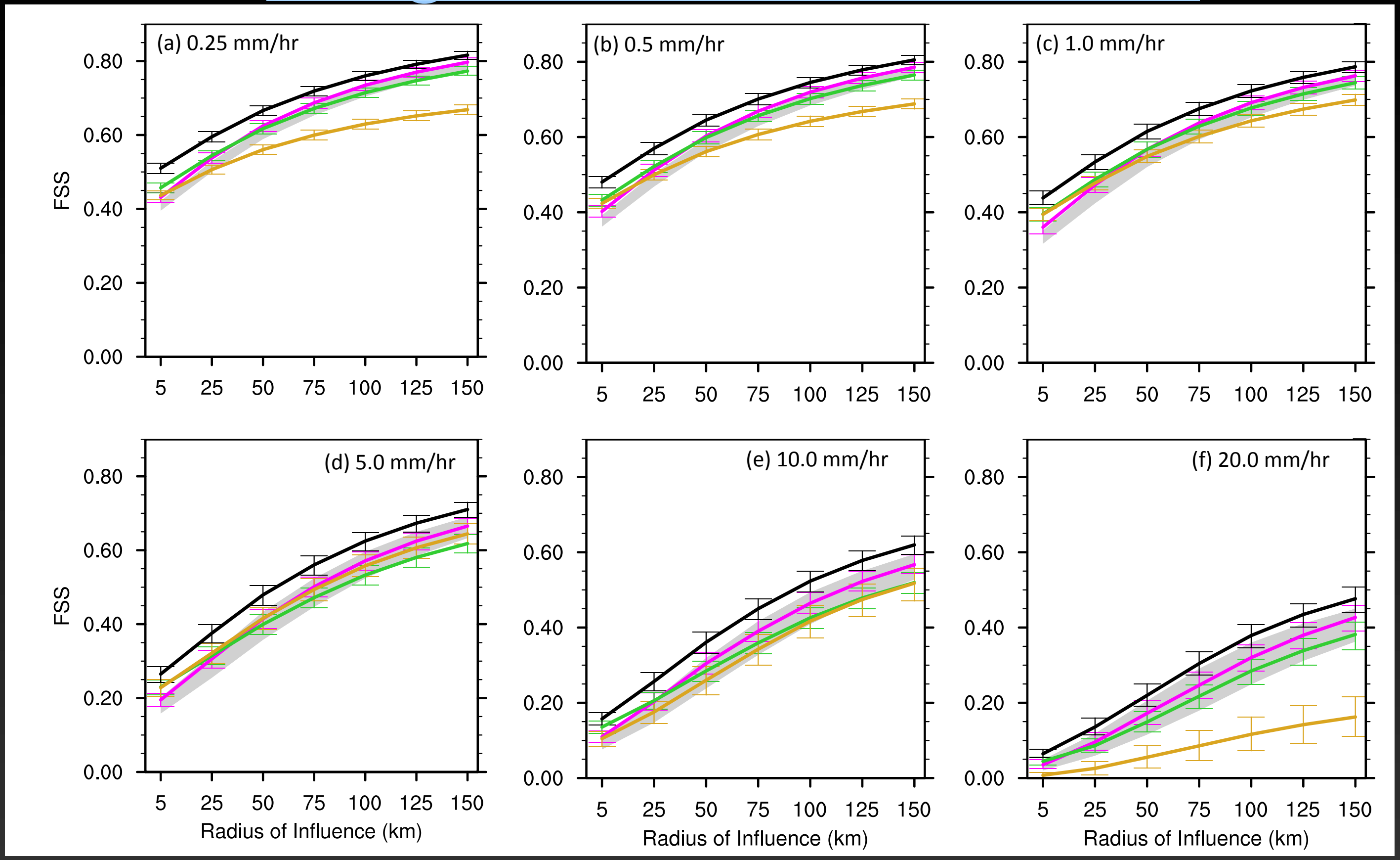


Radius and skill increase →

Equitable Threat Score (ETS) as a function of radius around a grid point. ETS compares hits to hits by chance.

From Clark et al., 2010: Neighborhood-Based Verification of Precipitation Forecasts from Convection-Allowing NCAR WRF Model Simulations and the Operational NAM. *Wea. Forecasting*, **25**, 1495–1509.

Neighborhood methods



Fractional Skill Score for of various forecast methods as a function of radius. 90% confidence intervals from bootstrapping.

From Schwartz et al., 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. Wea. and Forecast., Early Online Release.

Summary

NCAR

- Verification against observations a necessity
- Sample size and observation errors matter for mesoscale forecast verification
- We can learn from data assimilation
- Neighborhood (and related) methods are useful when intuition about forecast utility is available
- No single score/metric can tell the story