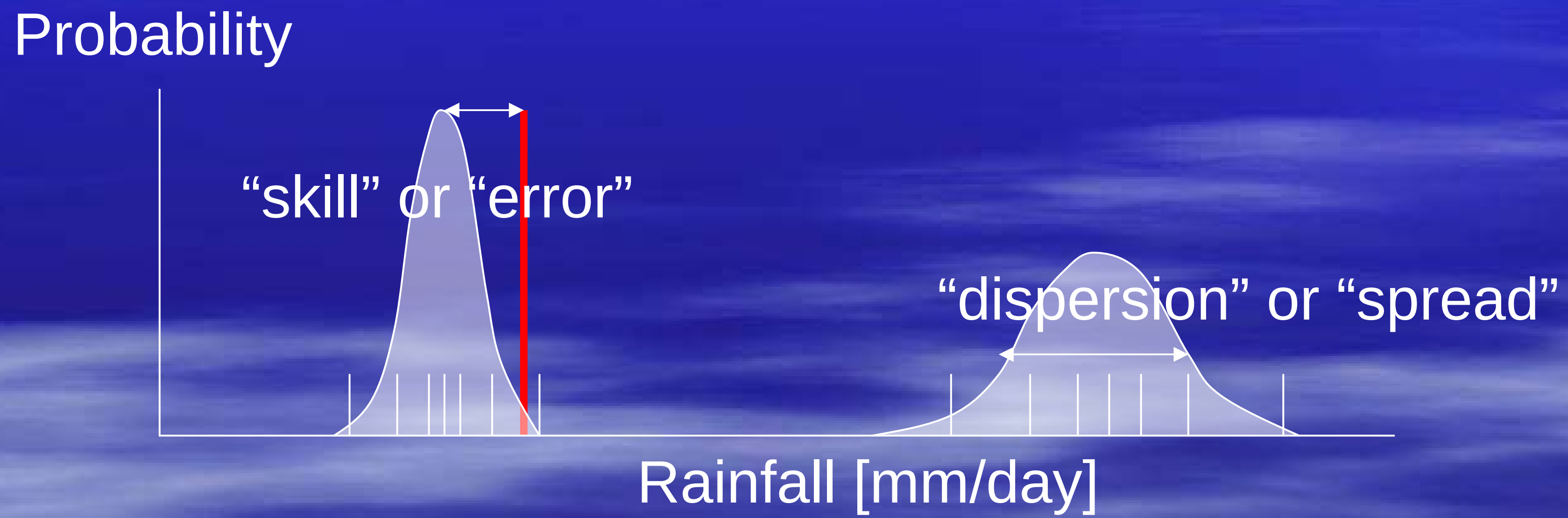
Verifying the Relationship between Ensemble Forecast Spread and Skill

Tom Hopson ASP-RAL, NCAR
Jeffrey Weiss, U. Colorado
Peter Webster, Georgia Instit. Tech.

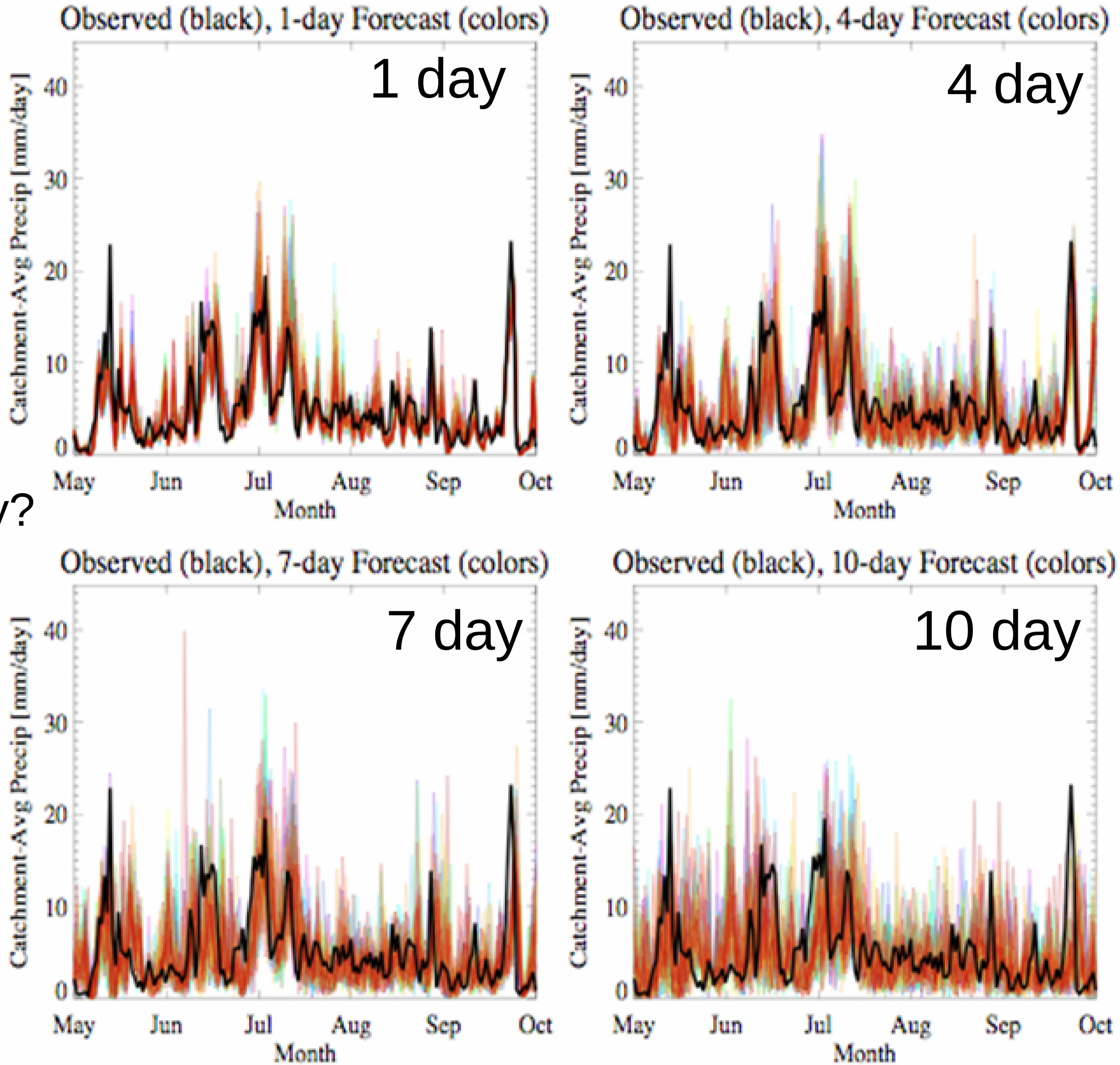
Motivation for generating ensemble forecasts:

1) Greater accuracy of ensemble mean forecast (half the error variance of single forecast)
2) Likelihood of extremes
3) Non-Gaussian forecast PDF’s
4) Ensemble spread as a representation of forecast uncertainty

Ensemble “Spread” or “Dispersion” Forecast “Skill” or “Error”



ECMWF Brahmaputra catchment Precipitation Forecasts vs TRMM/CMORPH/CDC-GTS Rain gauge Estimates



Points:

-- ensemble dispersion increases with forecast lead-time

-- dispersion variability within each lead-time

-- Provide information about forecast certainty?

How to Verify?

-- rank histogram? No. (Hamill, 2001)

-- ensemble spread-forecast error correlation?

Overview -- Useful Ways to Measure Ensemble Forecast System's Spread-Skill Relationship:

- Spread-Skill Correlation misleading (Houtekamer, 1993; Whitaker and Loughe, 1998)
- Propose 3 alternative scores

  1) "normalized" spread-skill correlation

  2) "binned" spread-skill correlation

  3) "binned" rank histogram
- Considerations:

  -- sufficient variance of the forecast spread? (outperforms ensemble mean forecast dressed with error climatology?)

  -- outperform heteroscedastic error model?

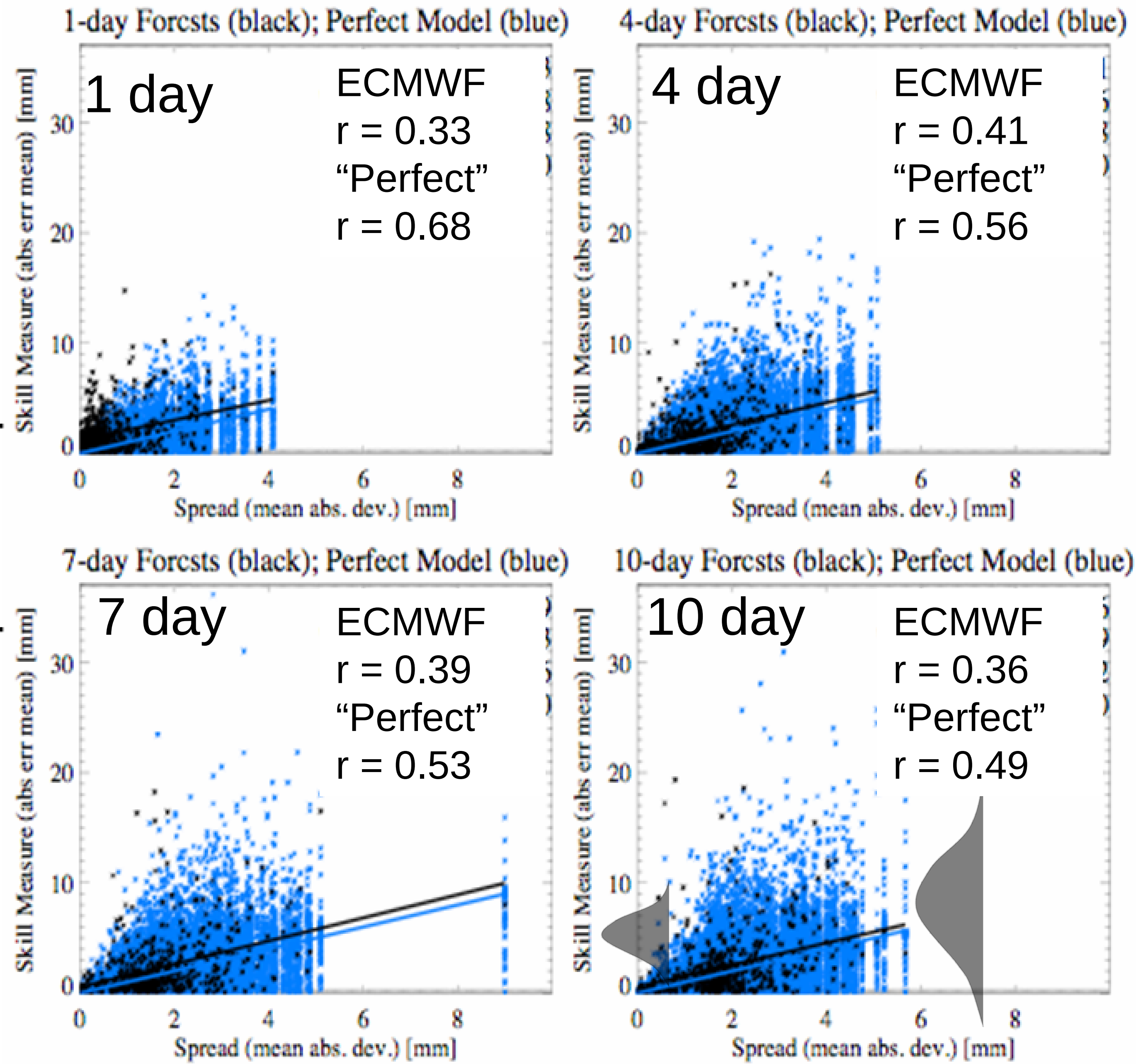
  -- account for observation uncertainty and under-sampling

Naturally Paired Spread-skill measures:

- Set I (L1 measures):
  - Error measures:
    - absolute error of the ensemble mean forecast
    - absolute error of a single ensemble member
  - Spread measures:
    - ensemble standard deviation
    - mean absolute difference of the ensembles about the ensemble mean
- Set II (squared moments; L2 measures):
  - Error measures:
    - square error of the ensemble mean forecast
    - square error of a single ensemble member
  - Spread measures:
    - ensemble variance

Spread-Skill Correlation …

- ECMWF spread-skill (black) correlation << 1
- Even “perfect model” (blue) correlation << 1 and varies with forecast lead-time



Limits on the spread-skill Correlation for a “Perfect” Model

Governing ratio, g:

(s = ensemble spread: variance, standard deviation, etc.)

$$g = \frac{\langle s \rangle^2}{\langle s^2 \rangle} = \frac{\langle s \rangle^2}{\langle s \rangle^2 + \mathrm{var}(s)}$$

Limits:

Set I

$$g \to 1, \quad r \to 0$$

$$g \to 0, \quad r \to \sqrt{2/\pi}$$

Set II

$$g \to 1, \quad r \to 0$$

$$g \to 0, \quad r \to \sqrt{1/3}$$

What’s the Point?
-- correlation depends on how spread-skill defined
-- depends on stability properties of the system being modeled
-- even in “perfect” conditions, correlation much less than 1.0

How can you assess whether a forecast model's varying ensemble spread has utility?

- Positive correlation? Provides an indication, but how close to a "perfect model".
- Uniform rank histogram? No guarantee.

1) One option -- "normalize" away the system's stability dependence via a skill-score:

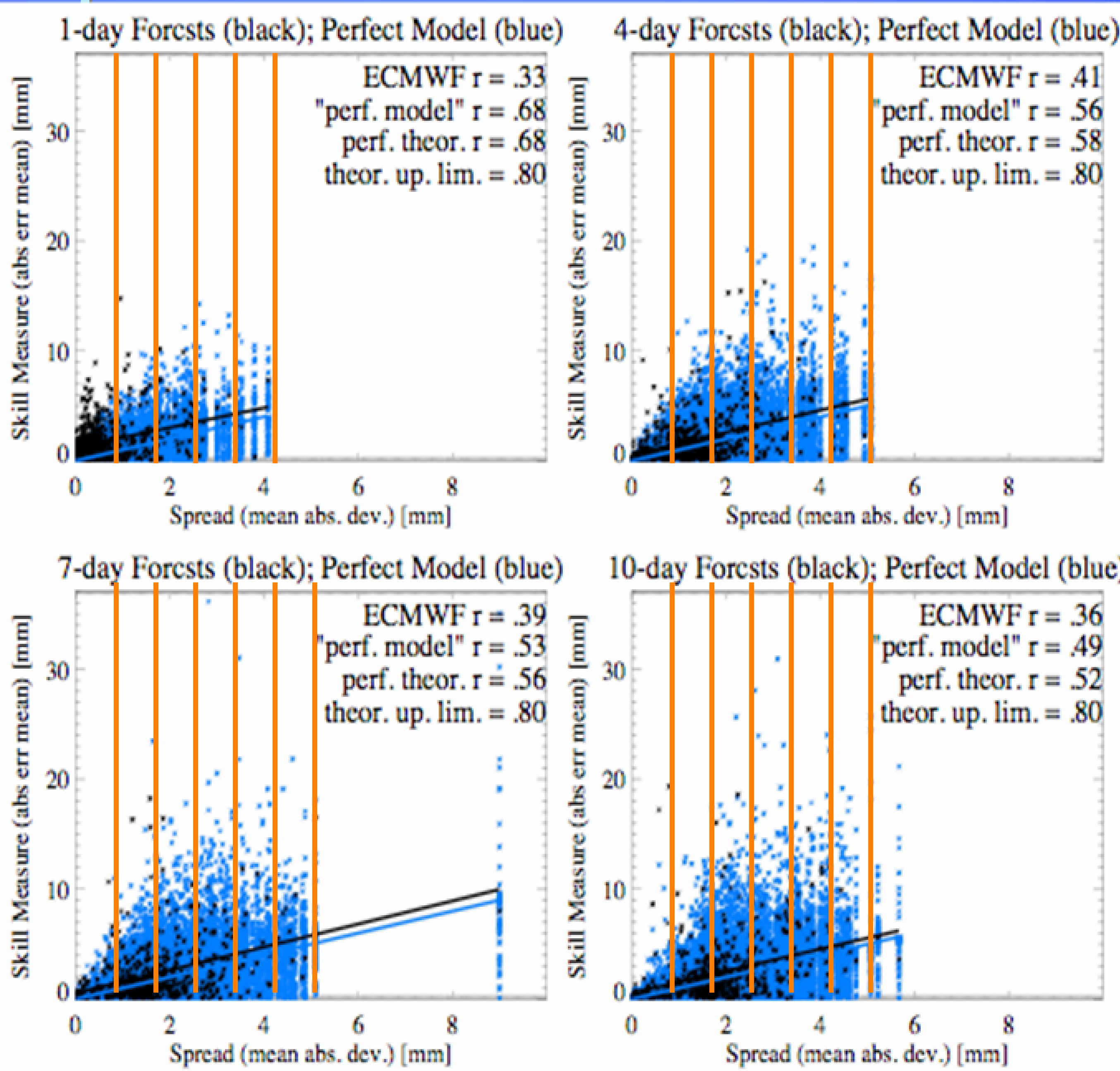
$$SS_r = \frac{r_{frcst} - r_{ref}}{r_{perf} - r_{ref}} X100\%$$

two other options …

Assign dispersion bins, then:

2) Average the error values in each bin, then correlate

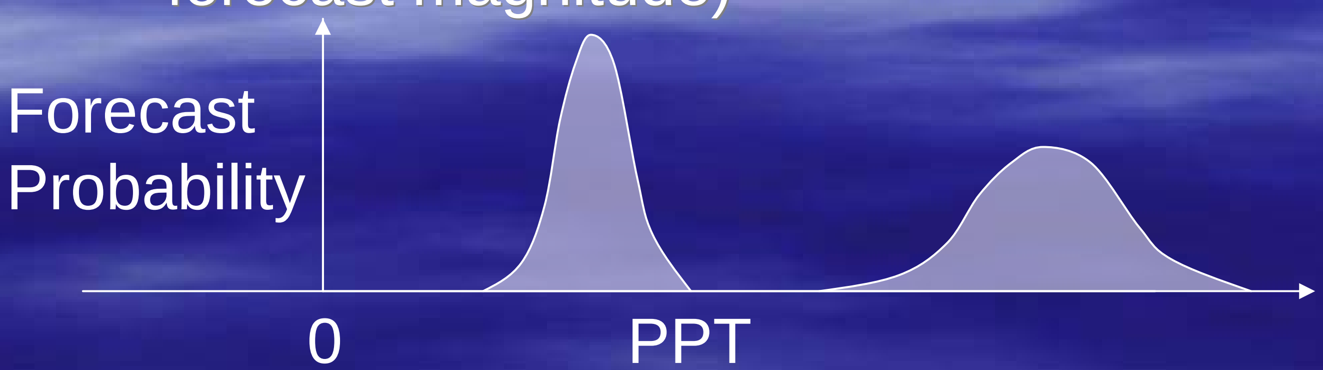
3) Calculate individual rank histograms for each bin, convert to a scalar measure



Skill Score approach

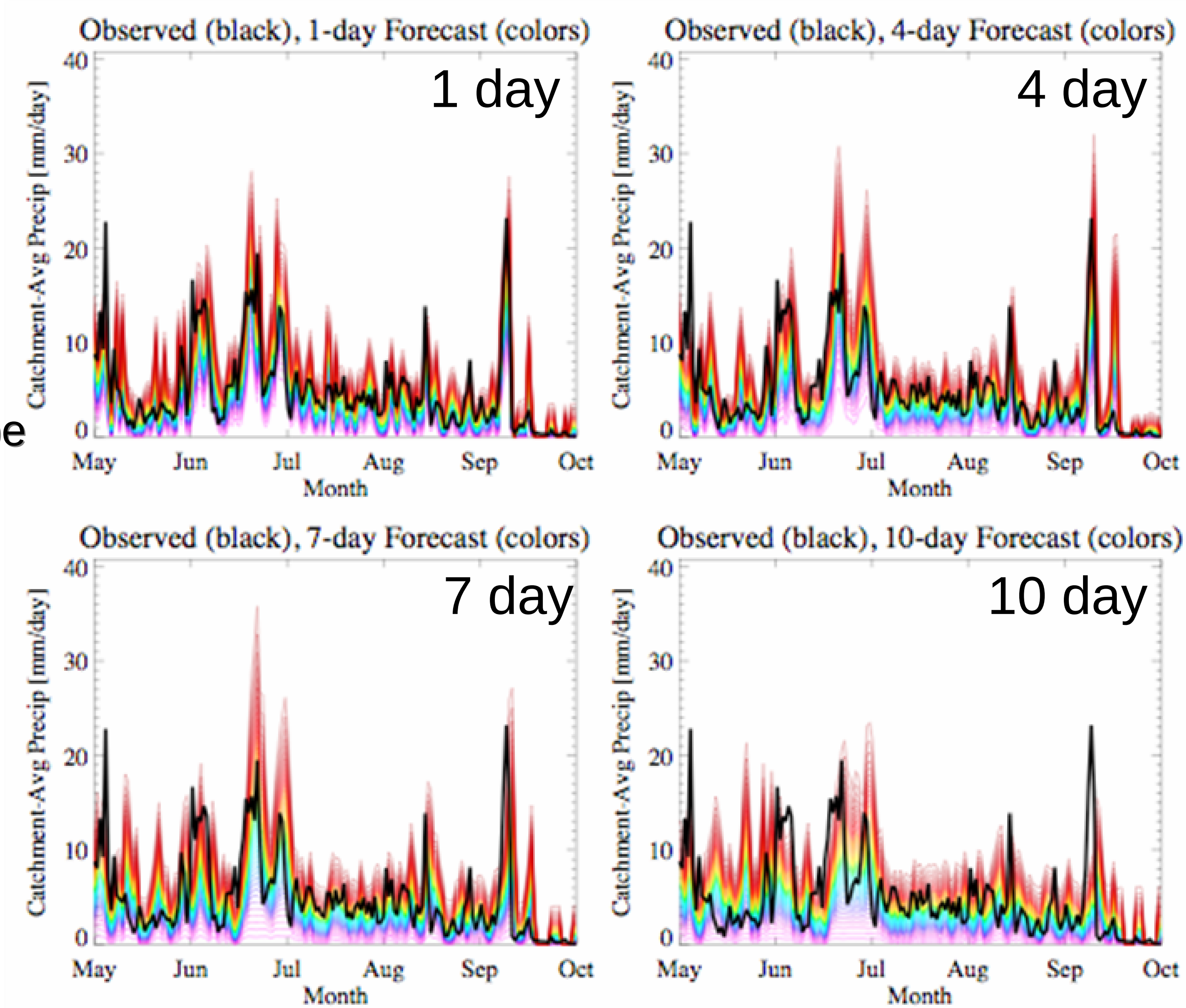
$$SS_r = \frac{r_{frcst} - r_{ref}}{r_{perf} - r_{ref}} X100\%$$

- r_{perf} -- randomly choose one ensemble member as verification
- r_{ref} -- three options:
  1) constant “climatological” error distribution (r --> 0)
  2) “no-skill” -- randomly chosen verification
  3) heteroscedastic model (forecast error dependent on forecast magnitude)

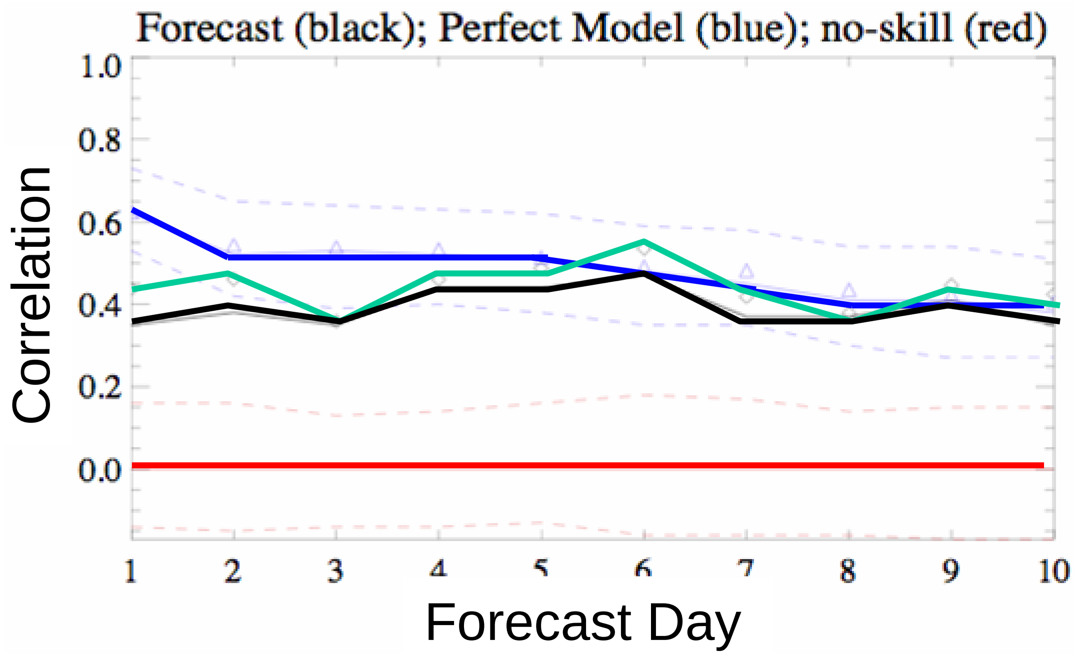


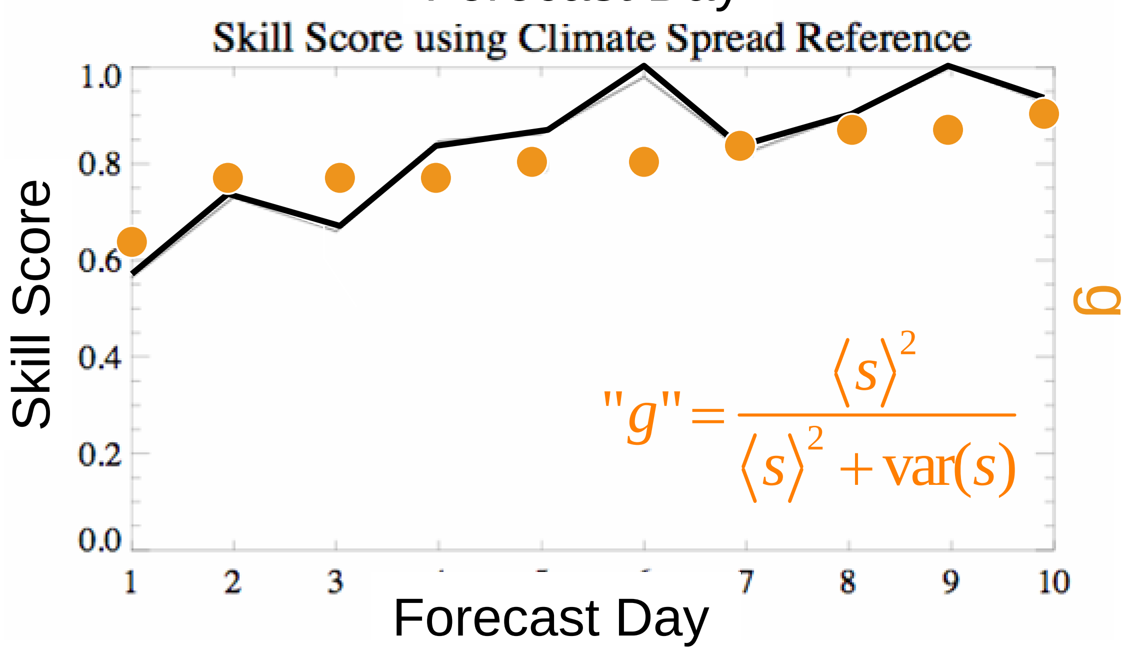
<u>Heteroscedastic Error model</u> dressing the Ensemble Mean Forecast (ECMWF Brahmaputra catchment Precipitation)

- From fit heteroscedastic error model, ensembles can be generated (temporally uncorrelated for clarity)



Option 1: "Normalized" Spread-skill Correlation



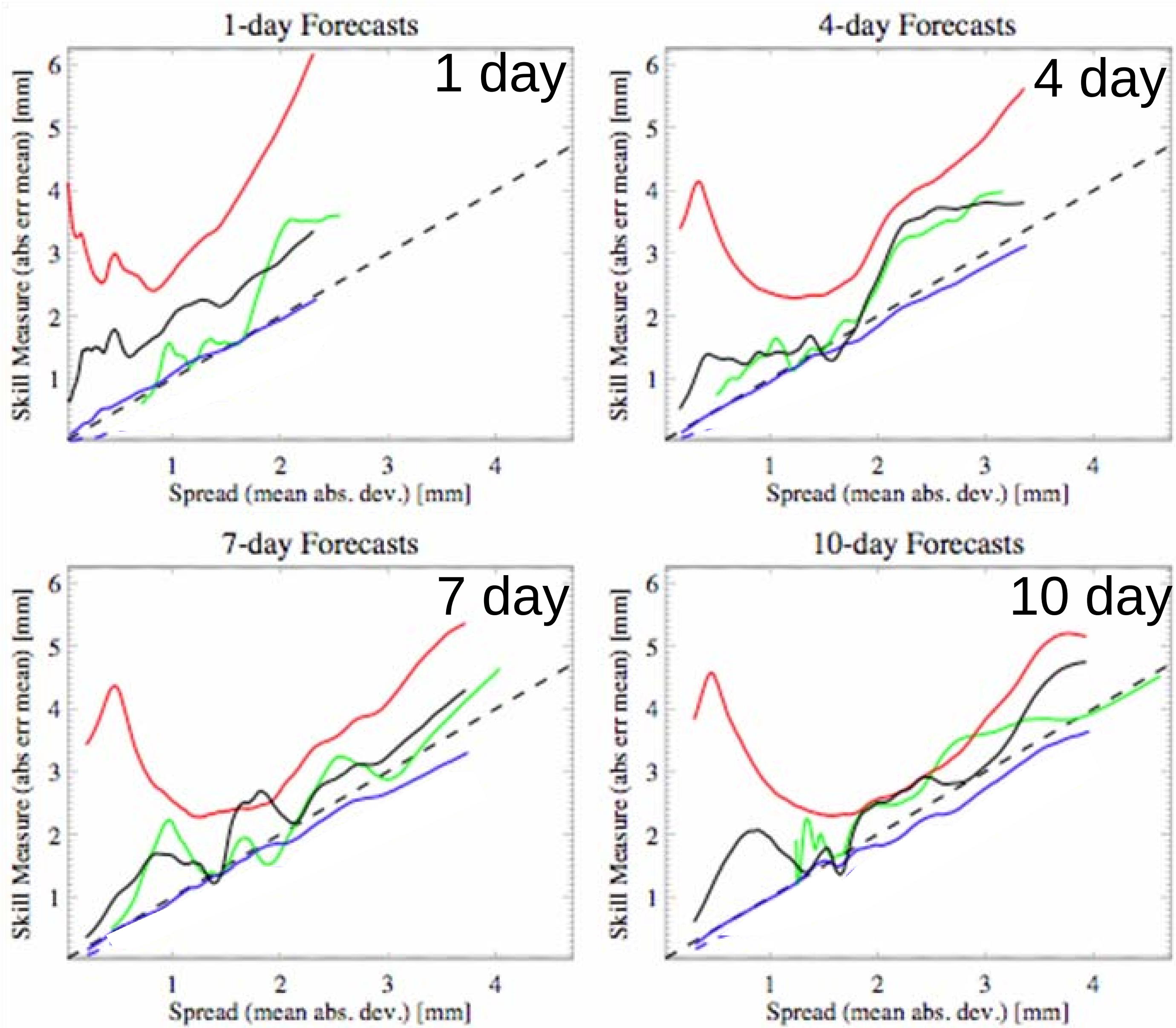


$$"g" = \frac{\langle s \rangle^2}{\langle s \rangle^2 + \mathrm{var}(s)}$$

- Operational Forecast spread-skill approaches "perfect model"
- However, heteroscedastic model outperforms

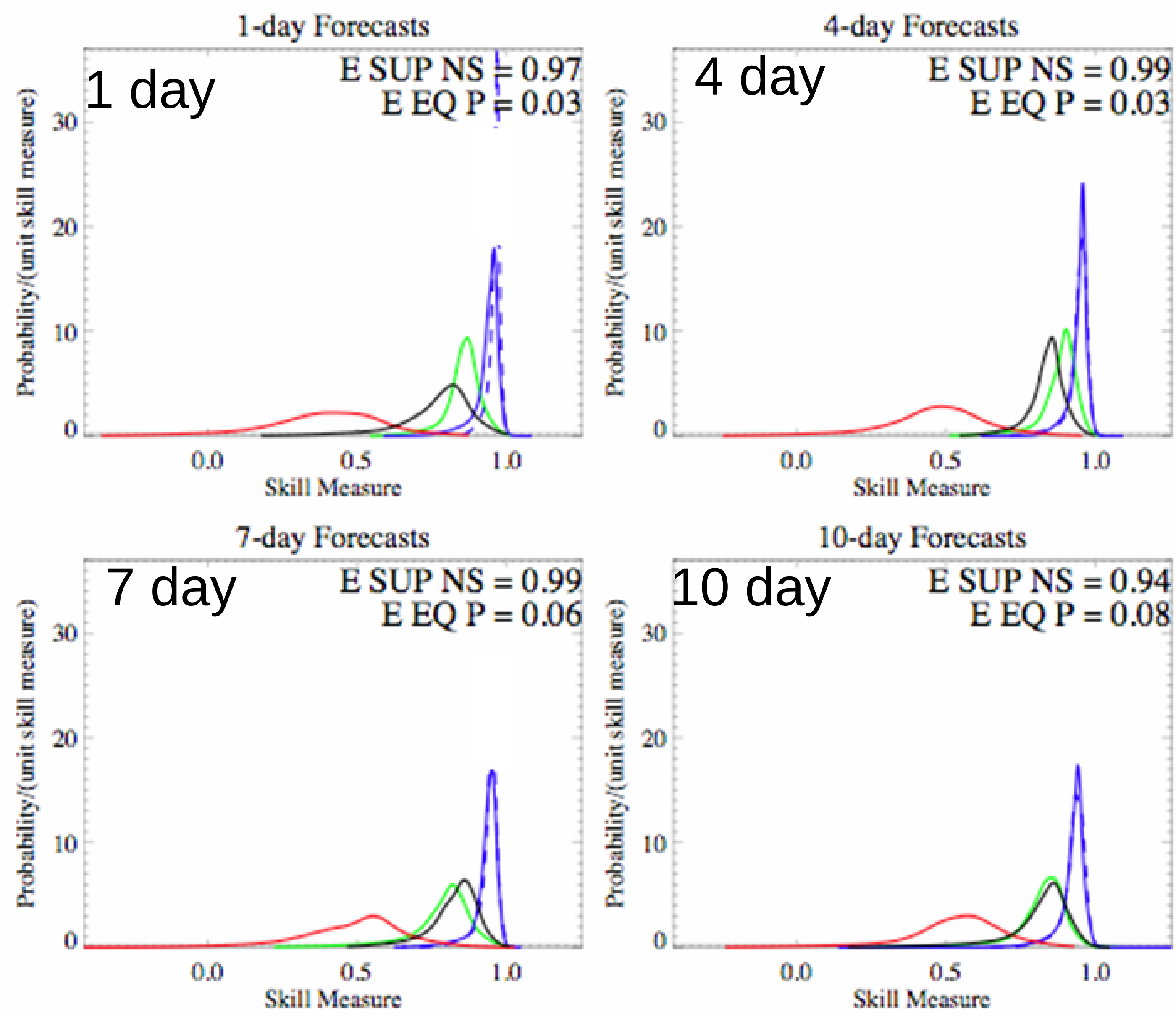
- Skill-scores show utility in forecast ensemble dispersion improves with forecast lead-time
- However, "governing ratio" shows utility diminishing with lead-time

Option 2: "binned" Spread-skill Correlation



- "perfect model" (blue) approaches perfect correlation
- "no-skill" model (red) has expected under-dispersive "U-shape"
- ECMWF forecasts (black) generally under-dispersive, improving with lead-time
- Heteroscedastic model (green) slightly better(worse) than ECMWF forecasts for short(long) lead-times

Option 2: PDF's of "binned" spread-skill correlations -- accounting for sampling and verification uncertainty



- "perfect model" (blue) PDF peaked near 1.0 for all lead-times
- "no-skill" model (red) PDF has broad range of values
- ECMWF forecast PDF (black) overlaps both "perfect" and "no-skill" PDF's
- Heteroscedastic model (green) slightly better(worse) than ECMWF forecasts for short(long) lead-times

Conclusions

- Spread-skill correlation can be misleading measure of utility of ensemble dispersion
  - Dependent on "stability" properties of environmental system
- 3 alternatives:

  1) "normalized" (skill-score) spread-skill correlation

  2) "binned" spread-skill correlation

  3) "binned" rank histogram
- ratio of moments of "spread" distribution also indicates utility

  -- if ratio --> 1.0, fixed "climatological" error distribution may provide a far cheaper estimate of forecast error
- Truer test of utility of forecast dispersion is a comparison with a heteroscedastic error model => a statistical error model may be superior (and cheaper)
- Important to account for observation and sampling uncertainties when doing a verification

Contact hopson@ucar.edu for more information and publications