Assessment of Ensemble Forecasts

S. L. Mullen

Univ. of Arizona

HEPEX Workshop, 7 March 2004

Talk Overview

- Ensemble Performance for Precipitation
  - Global EPS and Mesoscale 12 km RSM
    - Biases, Event Discrimination
    - Regional Assessment
- Calibration of Ensemble Output
  - What it can and can't do well
- Analysis Uncertainty
  - Effect on Verification Scores
- Fields Needed by Hydro. Runoff Model
  - Ensemble Validation Issues
    - What remains to be done?

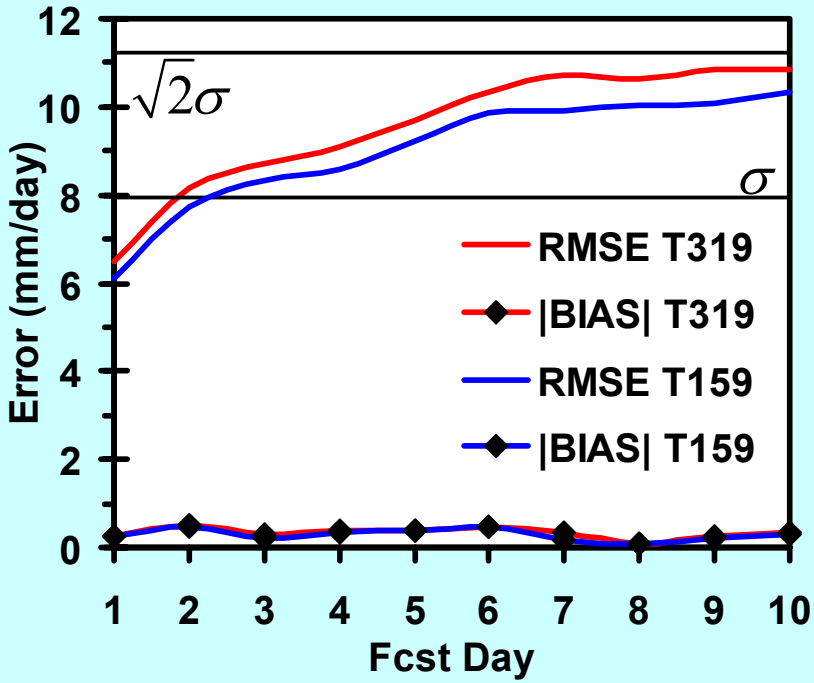
RMSE Deposition

Murphy (1988)

$$MSE = \underbrace{\left(\bar{f} - \bar{o}\right)^2}_{\text{Bias}} + \underbrace{s_f^2}_{\substack{\text{Model} \\ \text{Variance}}} + \underbrace{s_o^2}_{\substack{\text{Observed} \\ \text{Variance}}} - \underbrace{2 s_f s_o \rho_{fo}}_{\substack{\text{Variance} \\ \text{Weighted} \\ \text{Correlation}}}$$

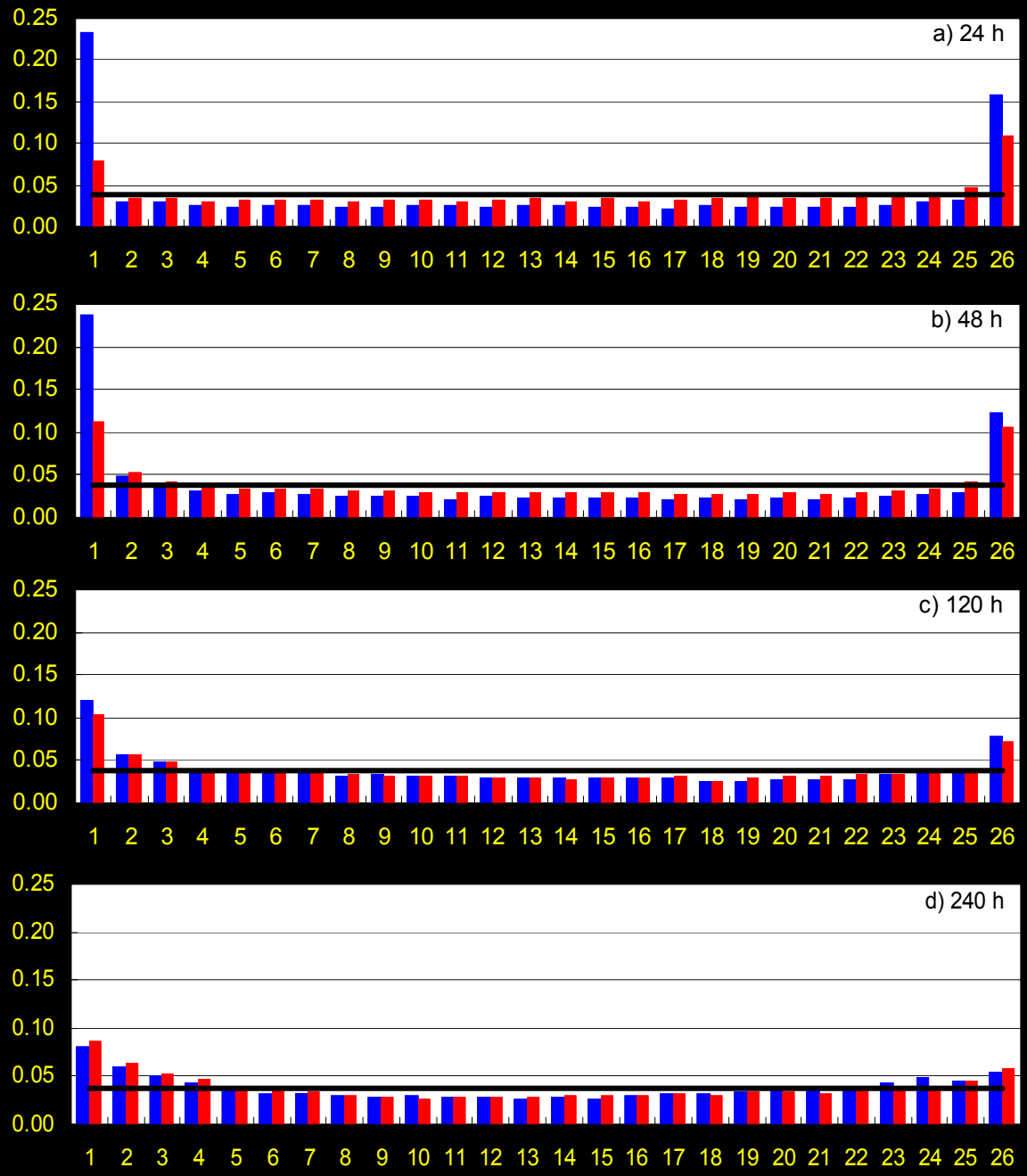
$$SS_{\text{Clim}} = \underbrace{\rho_{fo}^2}_{\text{Phase}} - \underbrace{\left(\rho_{fo} - \frac{s_f}{s_o}\right)^2}_{\text{Reliability}} - \underbrace{\left(\frac{\bar{f} - \bar{o}}{s_o}\right)^2}_{\text{Bias}}$$

Root Mean Square Error
ECMWF Quantitative Precipitation
Cool + Warm Seasons



Forecast error growth

Bias-Spread Decomposition



T159 Rank Histograms

Under Dispersion

Verification lies outside envelope of ensemble too frequently

Related to weak model variance

Brier Score Decomposition

Murphy (1973)

$$BS = BS_{rel} - BS_{res} + BS_{unc}$$

where

$$BS_{rel} = \frac{1}{N}\sum_{i=1}^{I} N_i [f_i - \bar{o}_i]^2$$

$$BS_{res} = \frac{1}{N}\sum_{i=1}^{I} N_i [\bar{o}_i - \bar{o}]^2$$

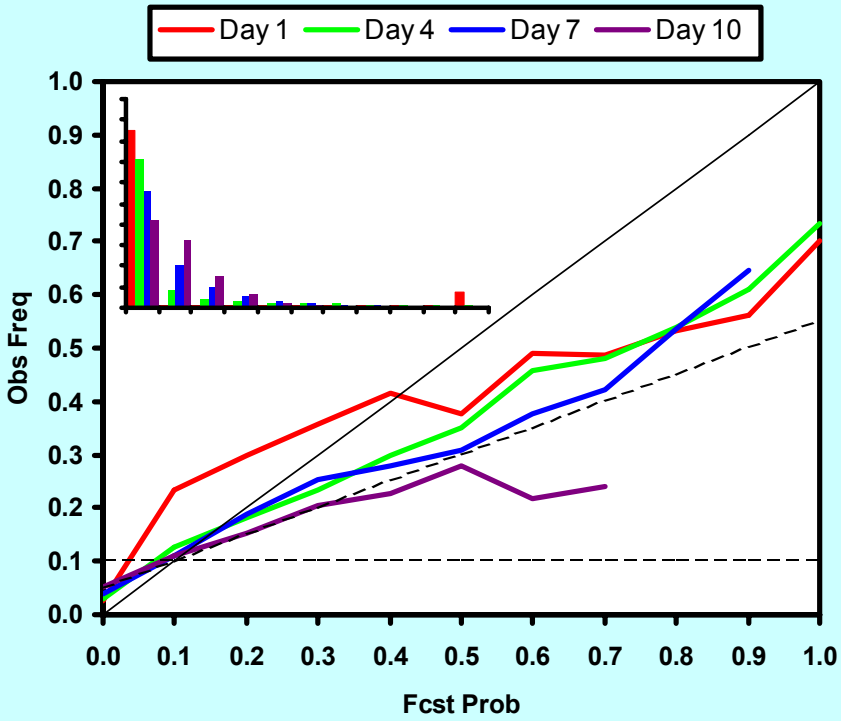
$$BS_{unc} = \bar{o}[1 - \bar{o}]$$

Skill Score

$$BSS = \frac{BS_{cli} - BS}{BS_{cli}} = \frac{BS_{res} - BS_{rel}}{BS_{unc}}$$

Reliability for Old T159

Cool Season 10 mm/day



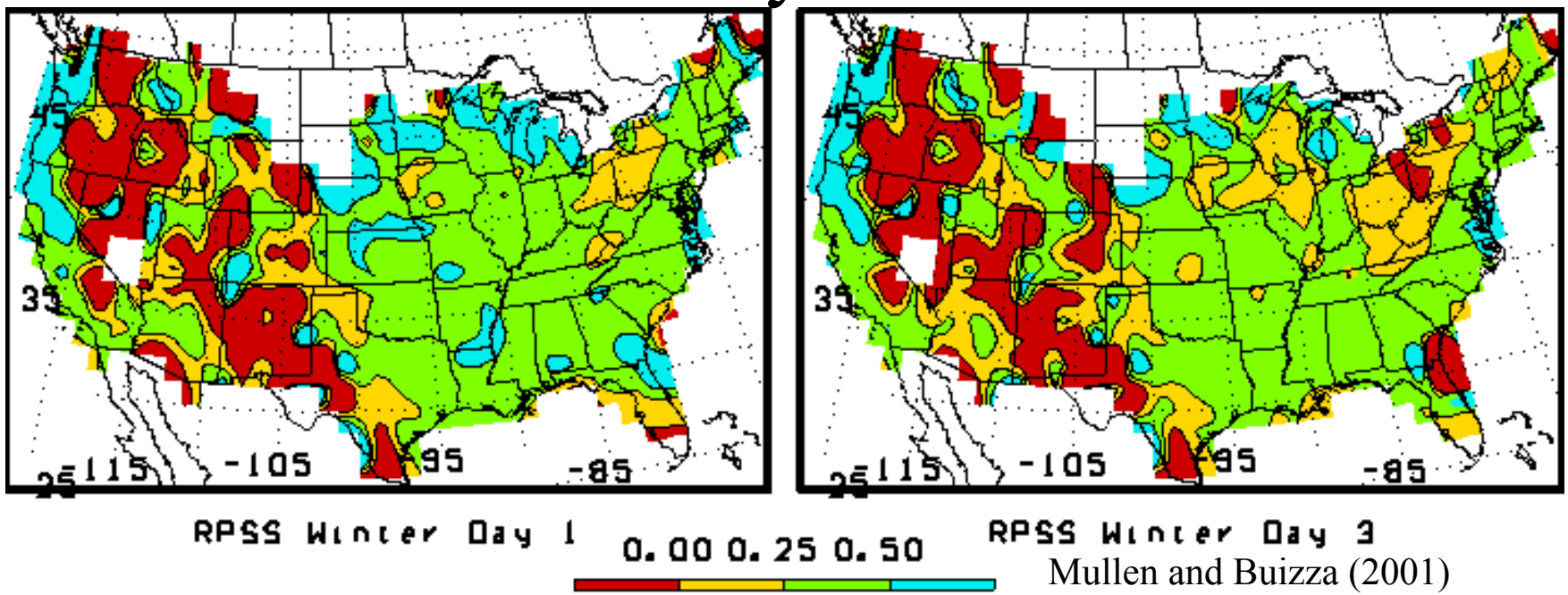
Mullen and Buizza (2001)

Want Forecasts that are

- **Reliable**
- **Discriminating**
- **Sharp**

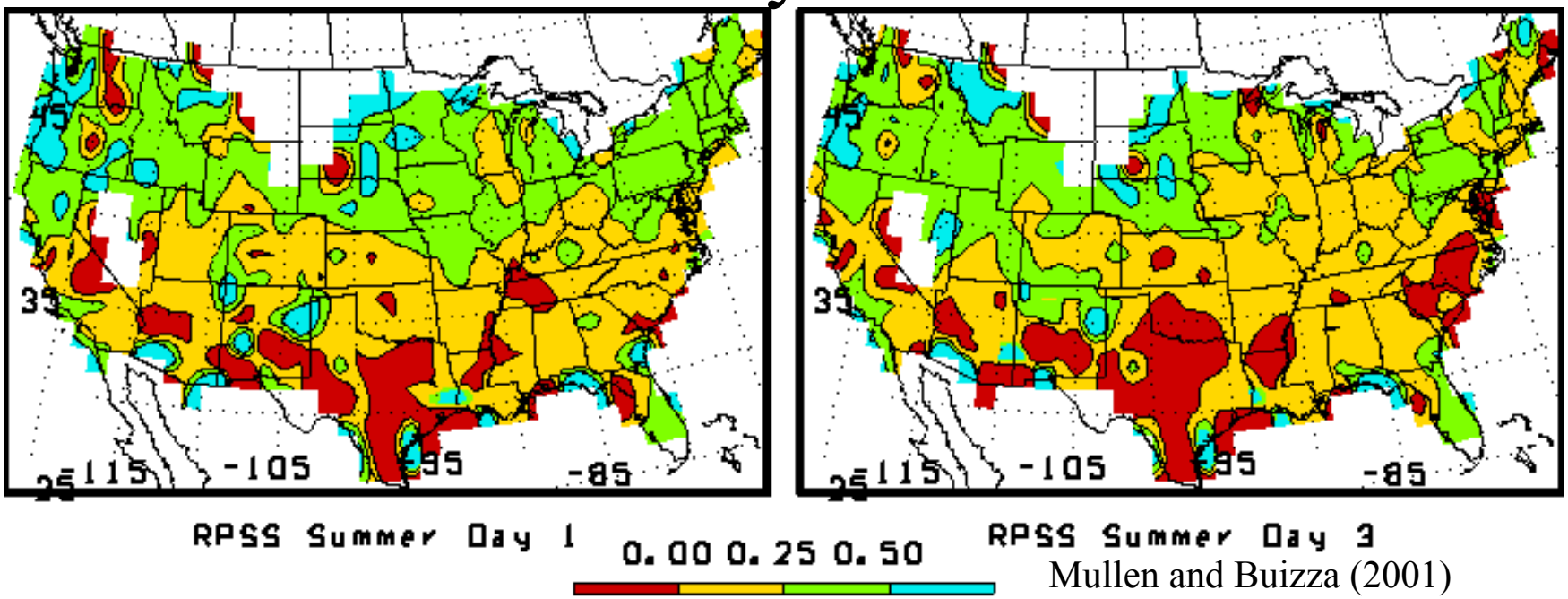
T159 EPS over forecasts likelihood of rain

Ranked Probability Skill Score T159



Mullen and Buizza (2001)

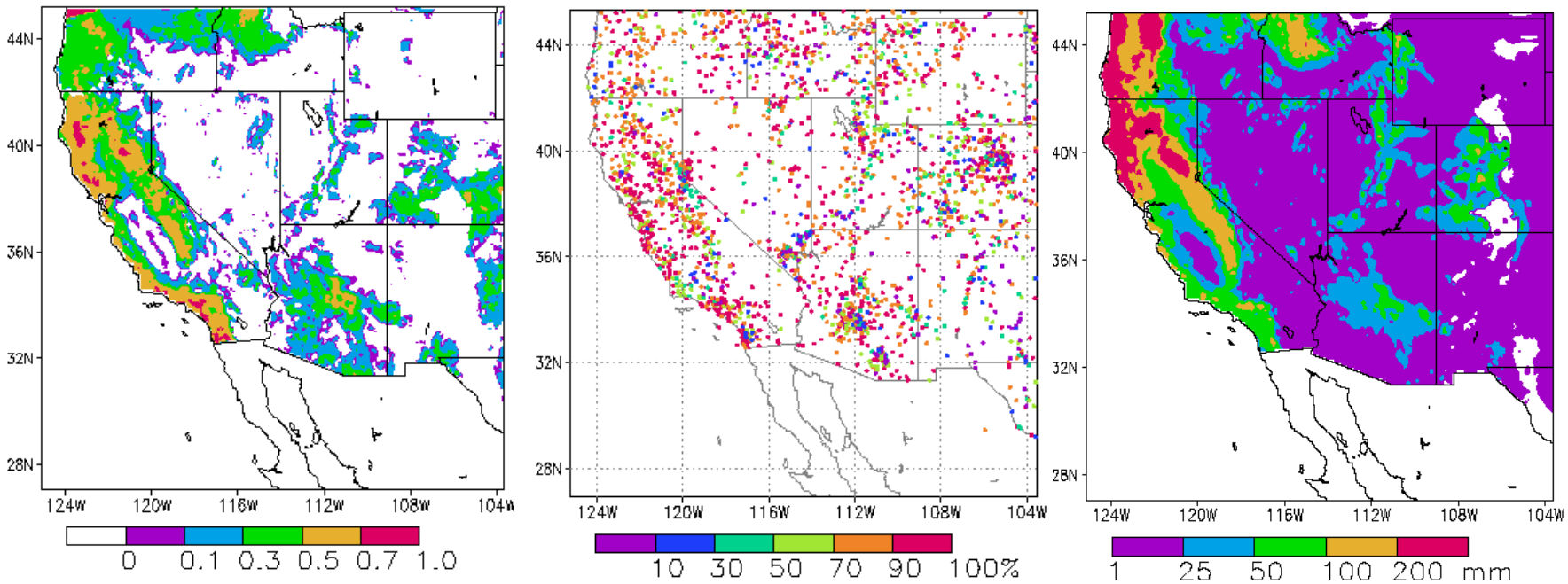
Ranked Probability Skill Score T159



Mullen and Buizza (2001)

24 h RPSS, 12 km RSM, 4 km grid

Overall RPSS | **Gauge stations** | **Average monthly precip**



- RPSS > 0.5 are mainly located along the Pacific Coast, and the windward slopes of Sierra Nevada Mountains and Mogollon Rim of the central Arizona.
- Spatial Correlations

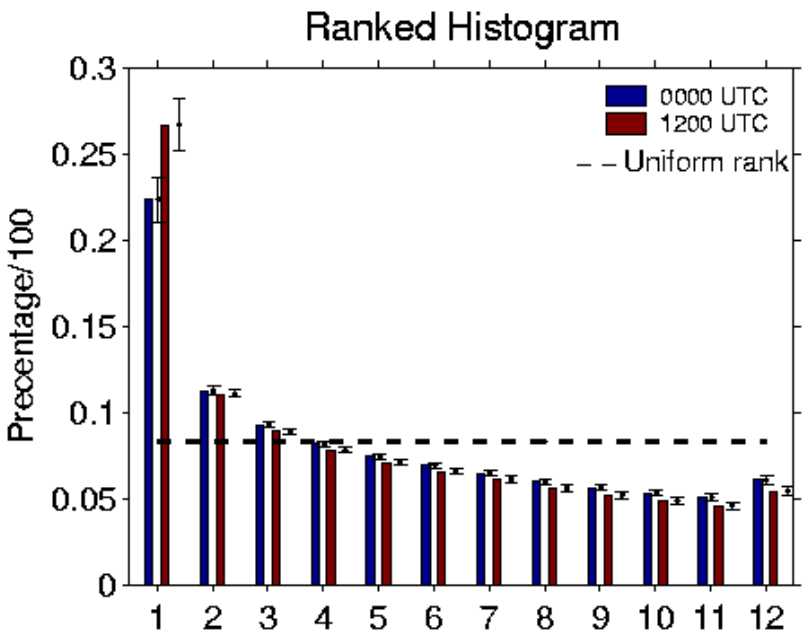
  RPSS and Precipitation: ~0. 60

  RPSS and Gauge Density: ~0.30

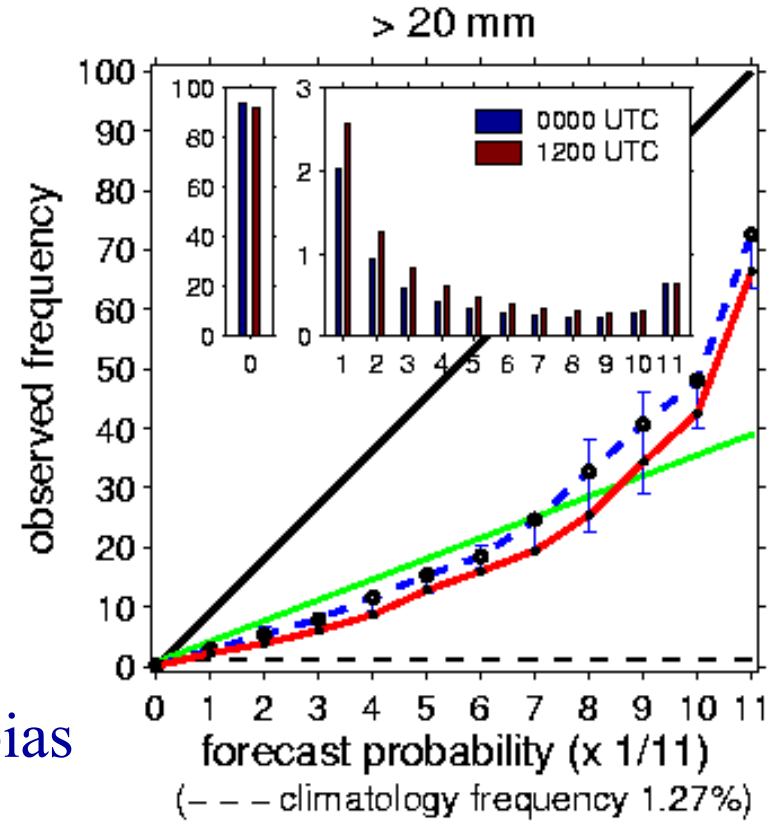
(Yuan et al. 2004, in progress)

24 h Bias for 12 km RSM

Ranked Histogram



Reliability Diagram



"L" shape of RH denotes large wet bias

Wet bias reflected in Reliability Curves

1200 UTC shows stronger wet bias!

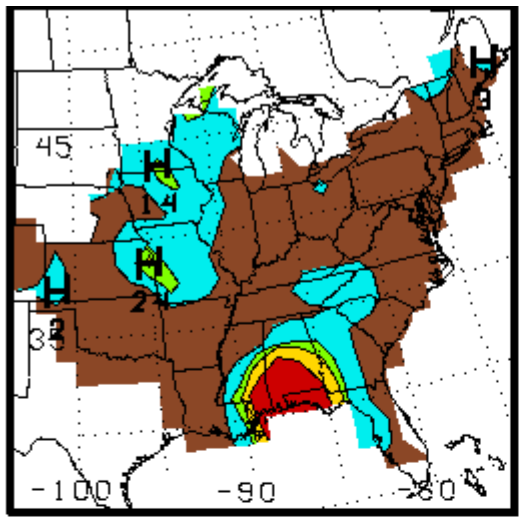
(Yuan et al. 2004, in progress)

UTC	Reliability	Resolution	BSS
0000	0.0035	0.0049	0.1127
1200	0.0057	0.0042	-0.1250

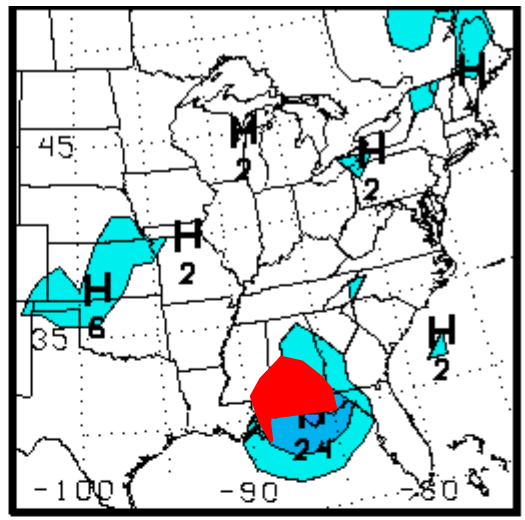
Personal Anecdotal Observation

- Ensemble forecast systems, both global and limit-area, seem to have very similar error characteristics for precipitation
  - Wet conditional bias for 24 h thresholds of 50 mm and lower
  - Under dispersion

29 SEP 98: Pr > 50 mm

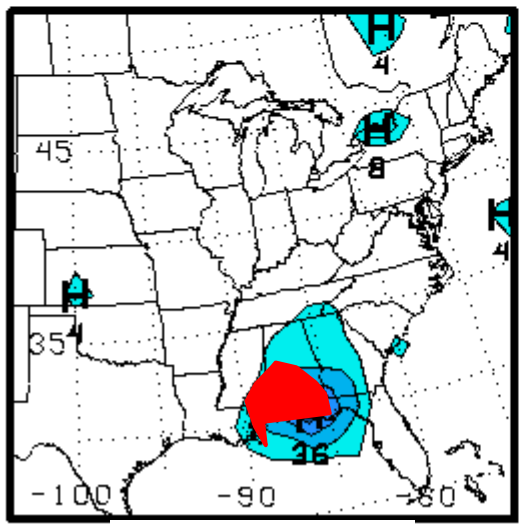


990130/1200 verification (mm)

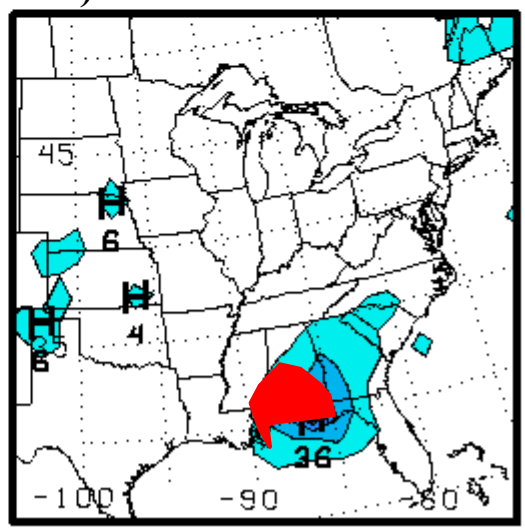


+5 Day T159

Mullen and Buizza (2002)



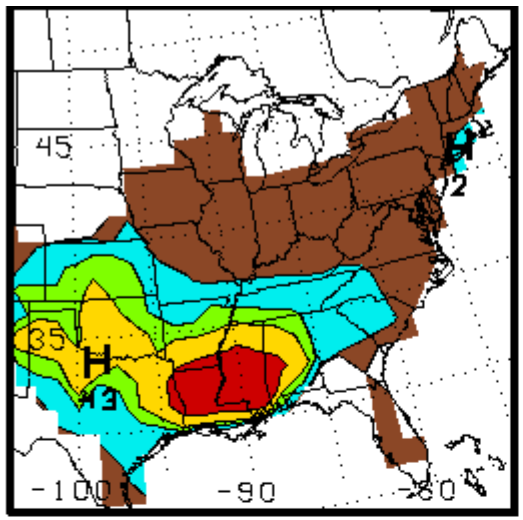
+5 Day T255



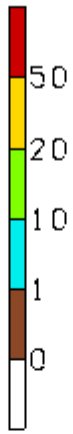
+5 Day T319

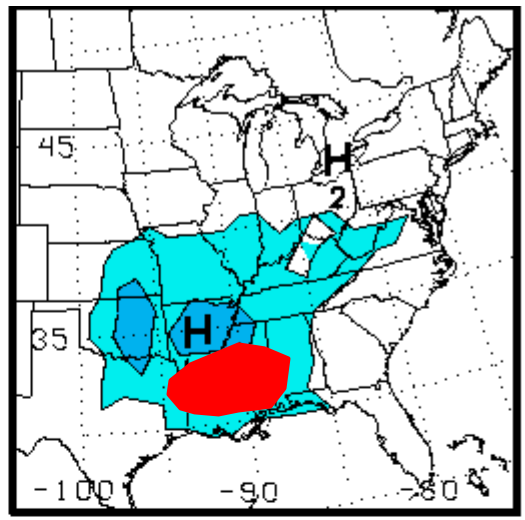
Hurricane Georges - Day 5 Forecast

30 JAN 99: Pr > 50 mm



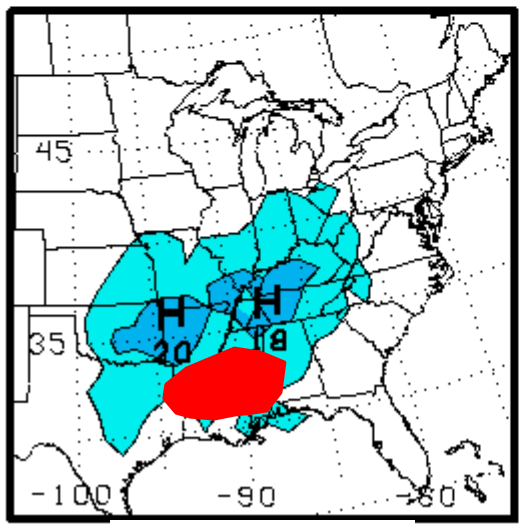
990130/1200 verification (mm)



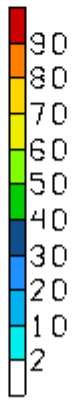


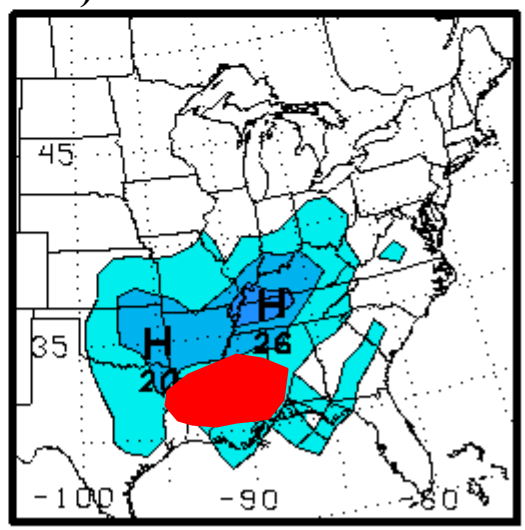
+5 Day T159

Mullen and Buizza (2001)



+5 Day T255





+5 Day T319

Wintertime Severe Thunderstorm Outbreak

Forecast Variations

- Skillful ensemble forecast systems might always yield a few “busts”

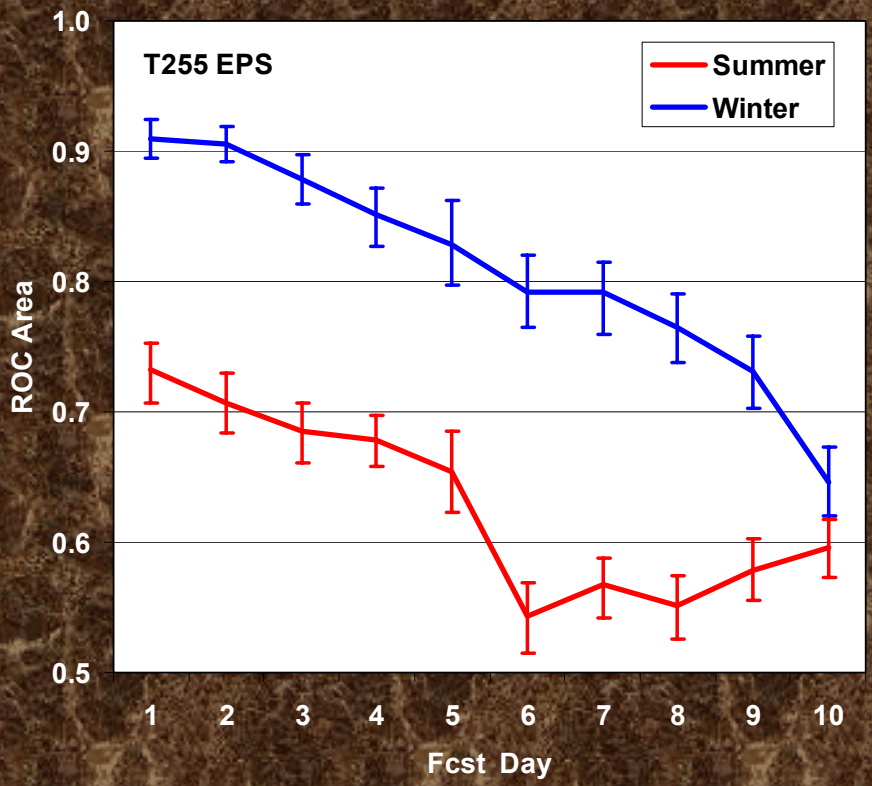
  What are sensitivities of user hydro user community and how do they deal with this situation?

Forecast Discrimination

- How well do ensembles discern precipitation events if biases are removed/ignored?

EPS ROC Areas for Summer-Winter

20 mm Threshold, Model Grid

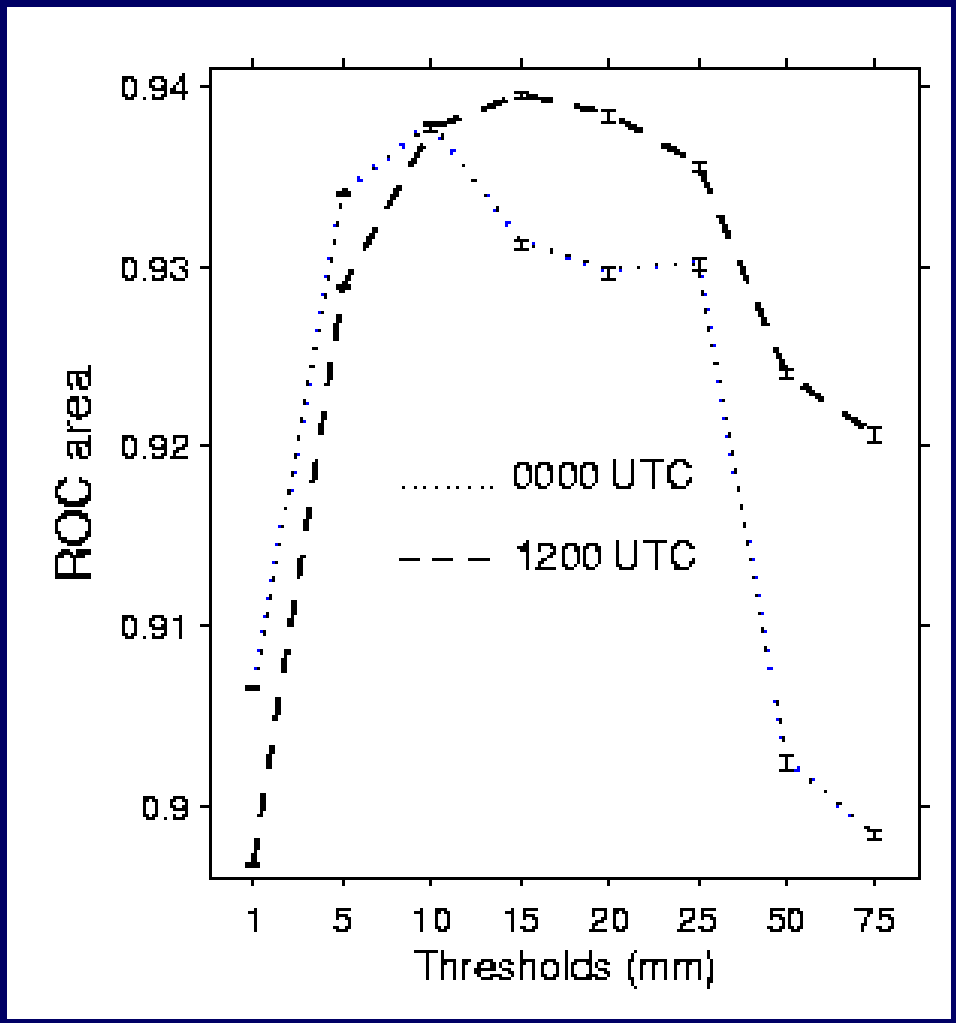


Summer precipitation is tougher to discern than winter ones

Why? CONVECTION!

- **Small sub-grid scale**
- **Intermittency**
- **Weak synoptic forcing**

24 h ROC Areas for 12 km RSM



(Yuan et al. 2004, in progress)

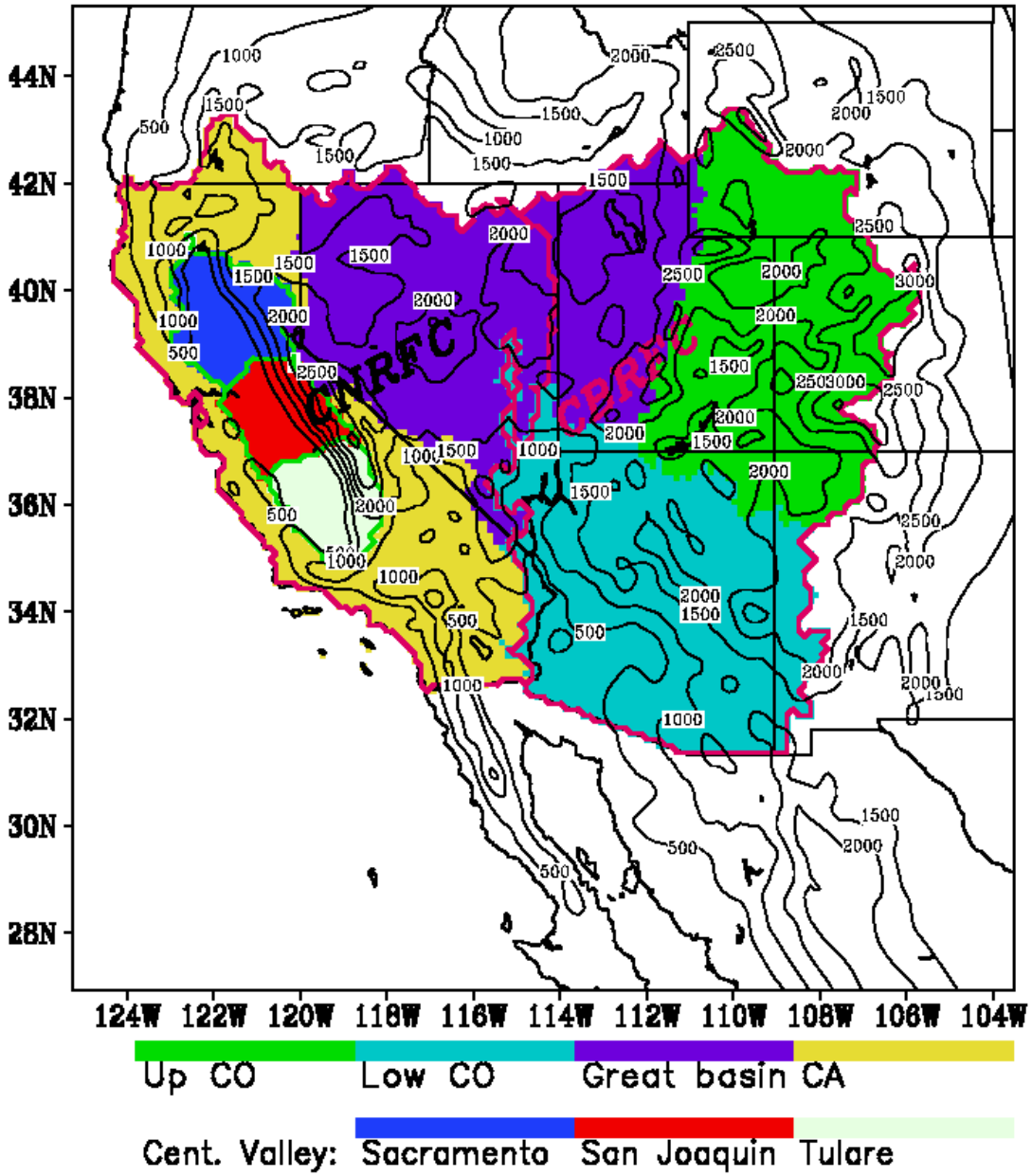
Outstanding ability to discriminate precipitation events

ROC areas ~0.90

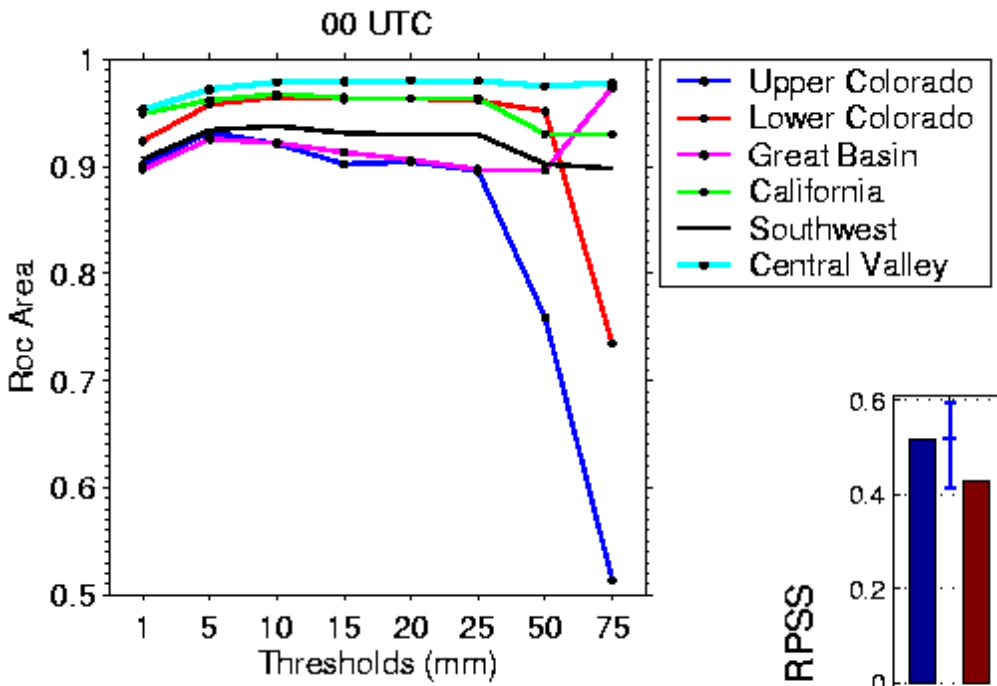
Local regions can show better performance e.g. Sierra Nevada

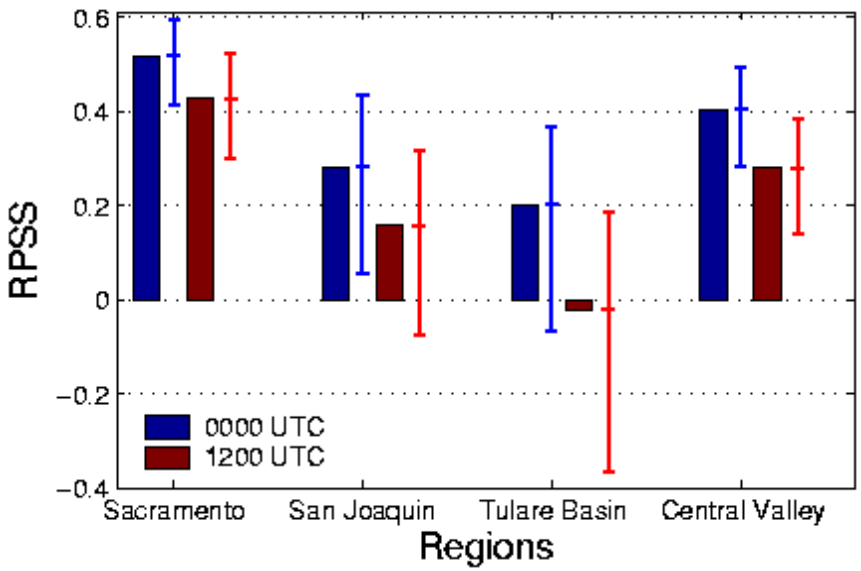
Implication: ensembles contain valuable predictive input to drive runoff models

RSM Verification for River Basins



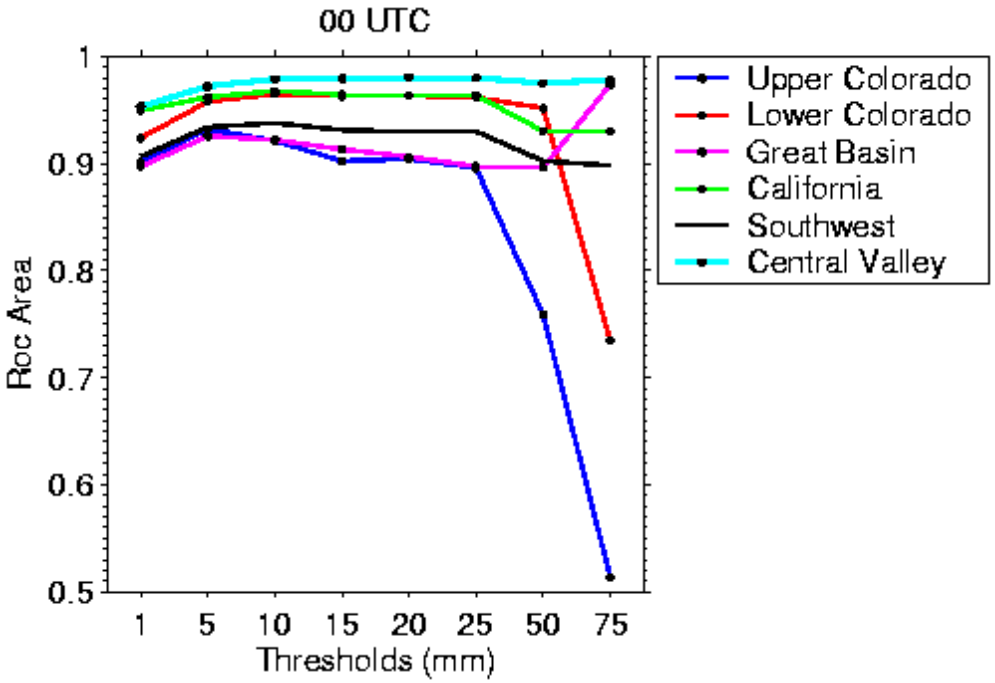
Regional Variations in 12 km RSM Skill

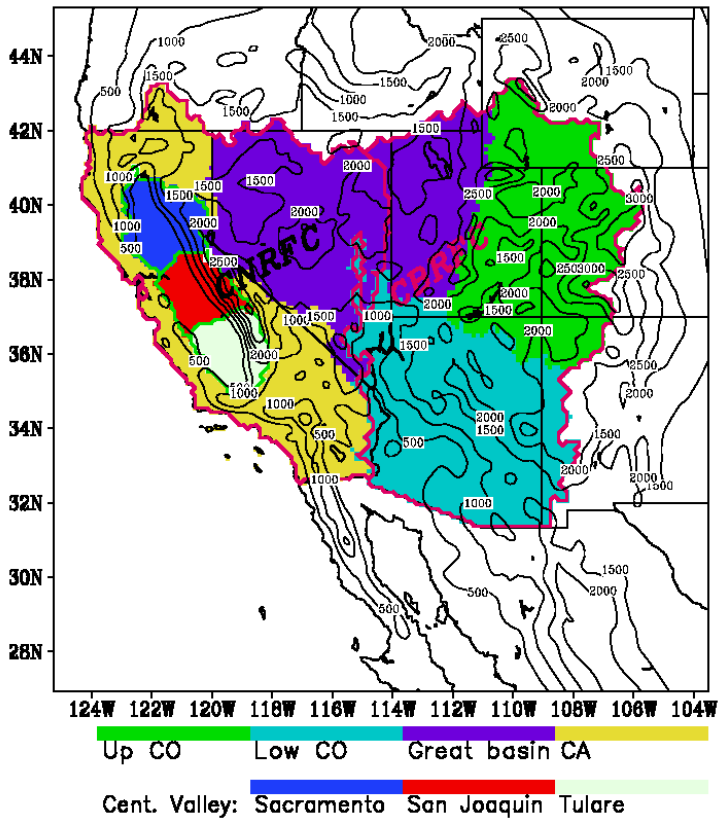




(Yuan et al. 2004, in progress)

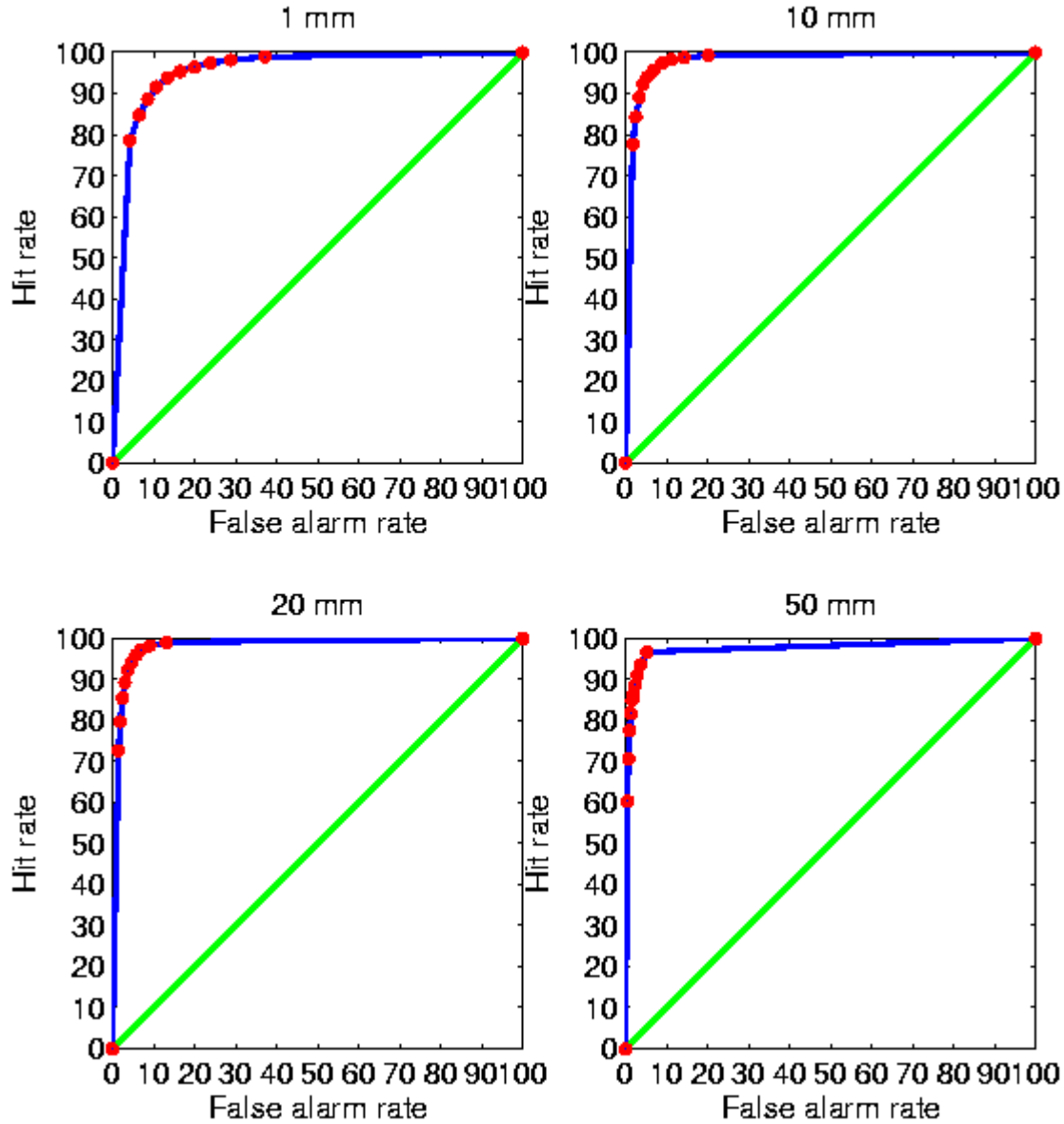
Regional Variations in 12 km RSM Skill



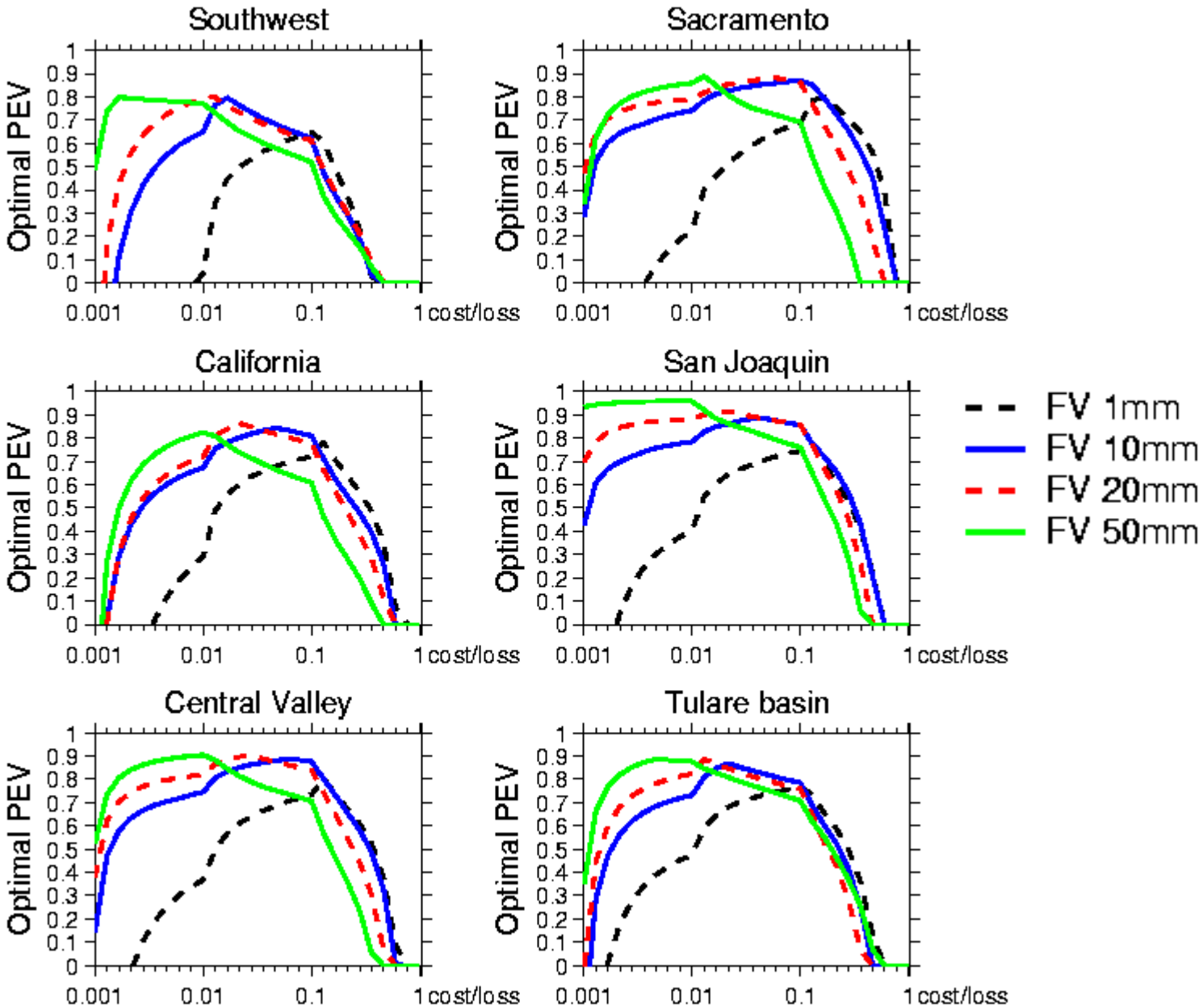


(Yuan et al. 2004, in progress)

12 km RSM ROC Central Valley



Optimal Potential Economic Value (PEV)



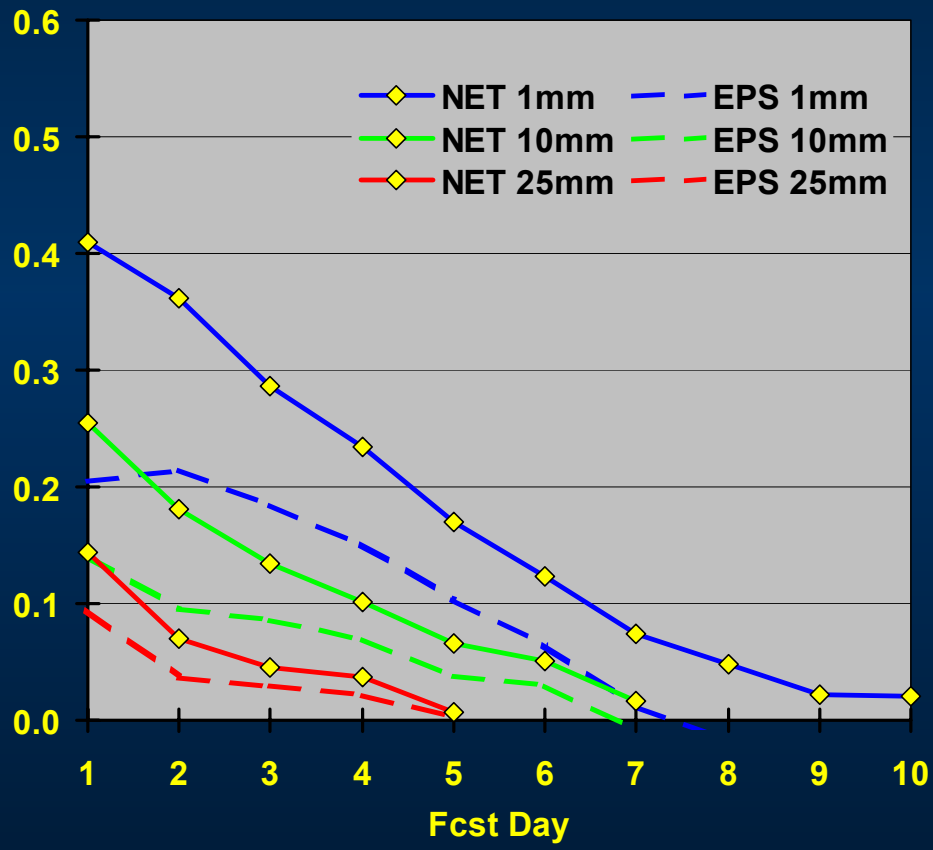
Calibration Questions

- **Calibration of EPS Ensemble Output by Artificial Neural Networks (ANN)**

  How much can calibration improve medium-range QPF skill?

  Can post-processing of just precipitation output from EPS significantly improve more than reliability term?

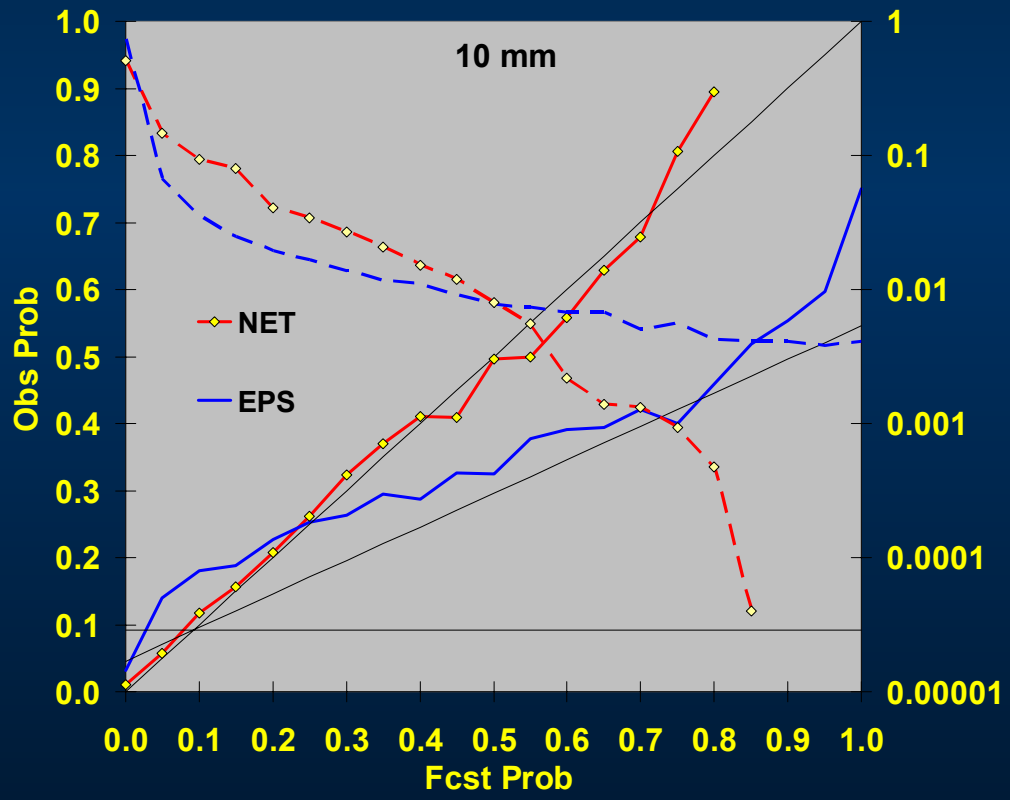
Brier Skill Score
(*4 Summers---DCA, OKC, FL, PNW)*



- **Skill Increases for 1, 10 and 25 mm but not 50 mm**
- **Largest Improvement Early in Forecast**

Attributes Diagram Day 2

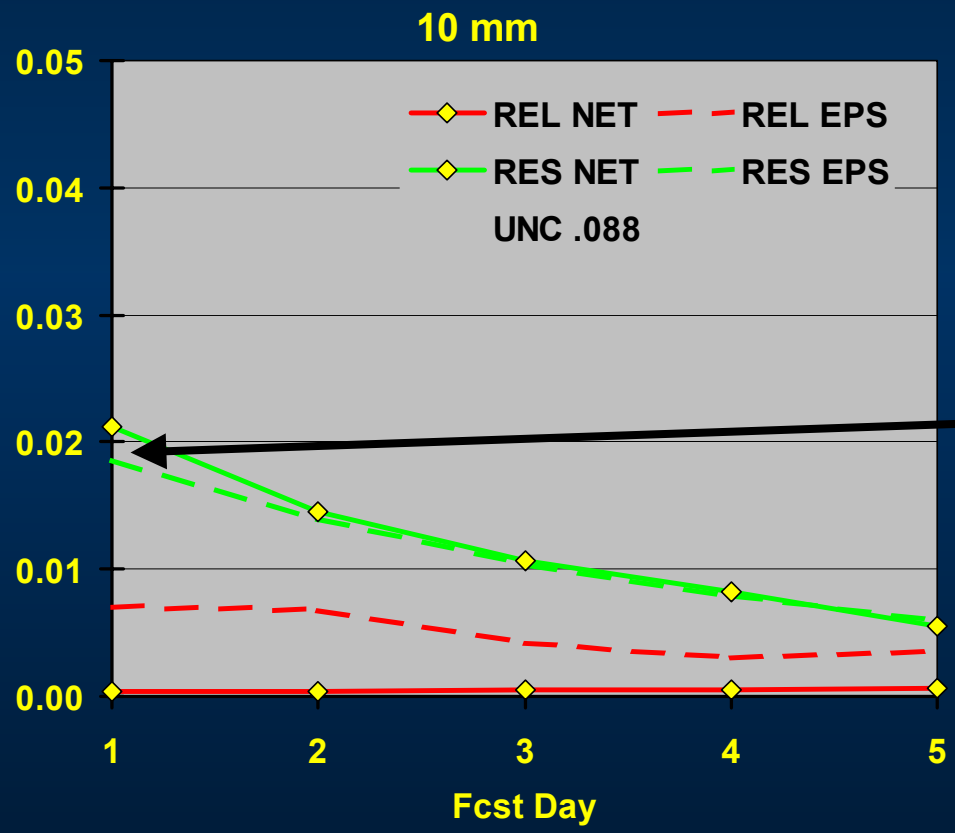
(4 Summers---DCA, OKC, FL, PNW)



- **Excellent Calibration Every Year-Season**
- **No High Probabilities (e.g. No Probs ≥ 90% for 10 mm at D+2)**
- **NET not as Sharp Note Differences in Forecast Frequencies (logarithmic scale)**

Summer Brier Decomposition

Where Does Improvement Come?

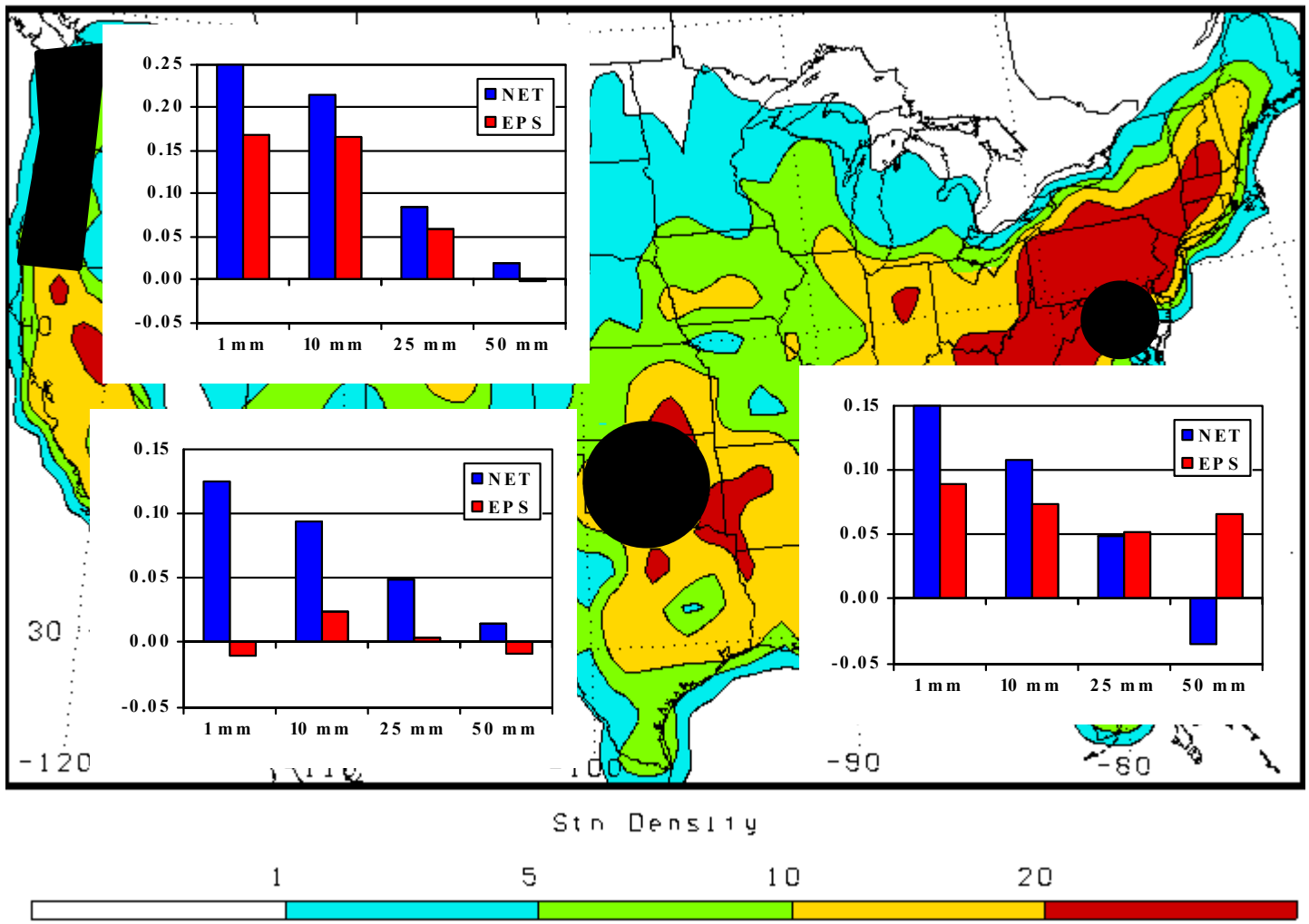


- **REL (Reliability) NET Increases Skill Through D+5**
- **RES (Resolution) *Slight* Increase @ D+1**

Calibration slightly improves ability to discriminate events early in forecast

Average Regional Skill

(Averaged over all forecast projections)



"General" Conclusions

- **Calibration of *Only* QPF Output**
  - **-Improves Brier Skill Score by ~20%**
  - **-Improves *Reliability***
  - **-Improves *Resolution* only small amount**
  - **-Calibrated forecasts lack sharpness**
    few or no extreme probability forecasts

Somewhat Related Questions

- How should ensemble output be calibrated prior to input into hydrological models?
- In what form (state vector for single forecast or state "matrix" for entire ensemble) should forecast fields from ensembles be input into hydro models?

Inclusion of Analysis Uncertainty in Verification

- **Thorough Verification…**

  Should include estimate of observational or analyses uncertainty

  Inclusion can lead to markedly different
  - values for accuracy measures
  - conclusions

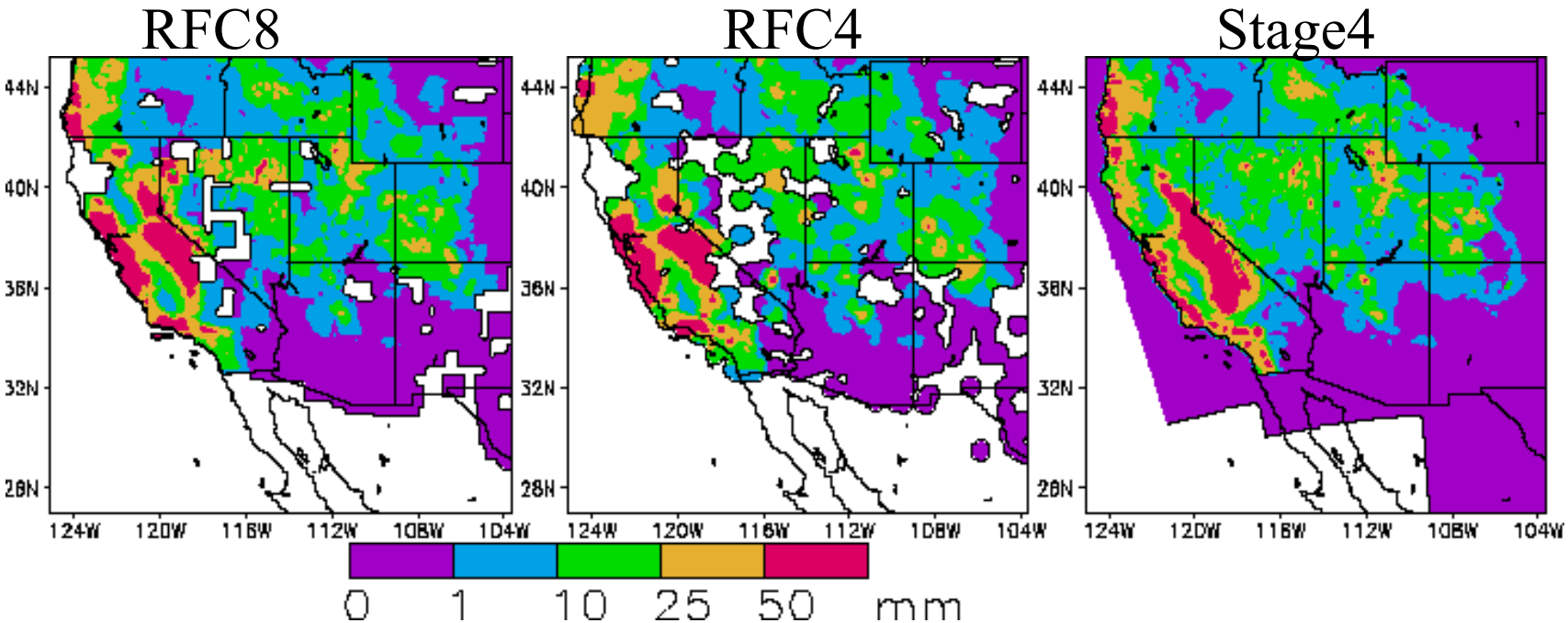
- **Rainfall marked by LARGE uncertainty**

  QPE differences can be comparable to spread at 24-48 h for QPF in localized regions!

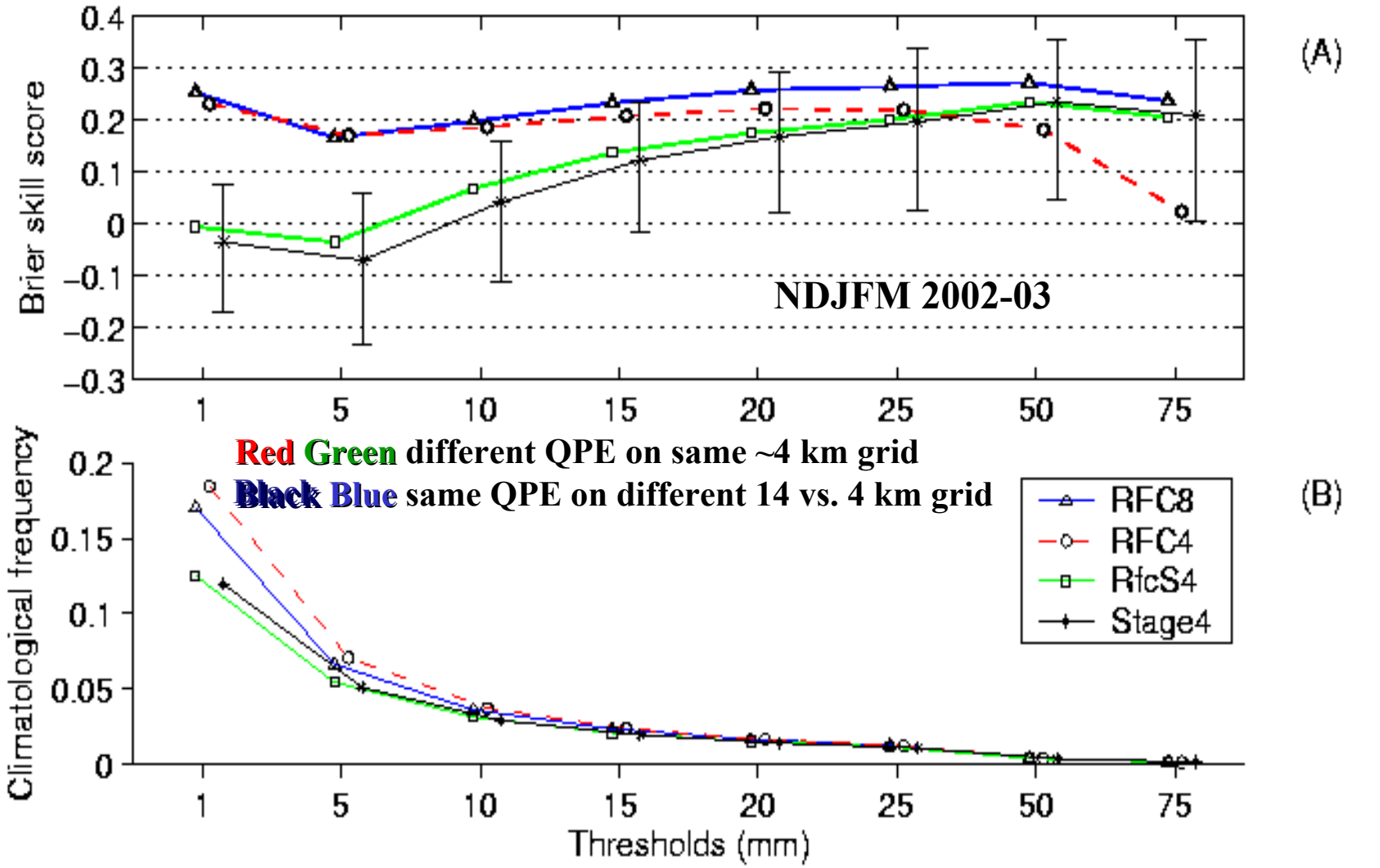
Uncertainty in Verification Analyses

		The NCEP precipitation analyses					
	Resolution	Data source	QC	Interval	Time (UTC)	mask	Gauge
RFC8	1/8th(14km)	Radar+Gauge	Yes	24 h	1200	Yes	7~8000
RFC4	4 km	Gauge only	No	24 h	1200	Yes	7~8000
Stage4	4 km	Radar+Gauge	Yes	6 h/24 h	0000&1200	No	3000
	QC: Quality Control done at RFCs						

- **Accumulated 24-h precipitation for 1200 UTC 8 Nov-9 Nov, 2002**



Different Verifying Analyses



Yuan et al. (2004, in progress)

Analysis-Observational Uncertainty

- What is the impact from uncertainty associate with other variables that might be needed as input by hydrological runoff models?

Driving Hydro Runoff Models with EPS/Ensemble Output

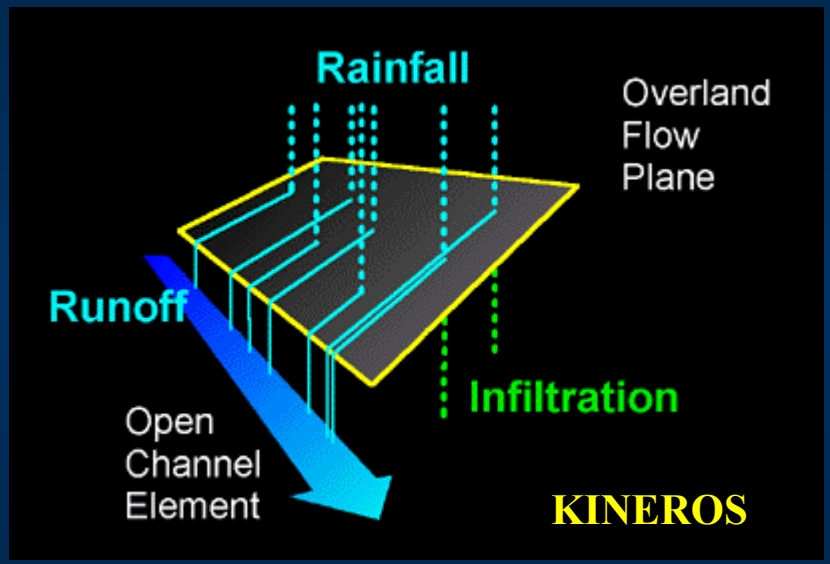
- QPF/NWP Forecasts are NOW Sufficiently Accurate to Use as Forcing for Predictive Hydrological Runoff Models…may require

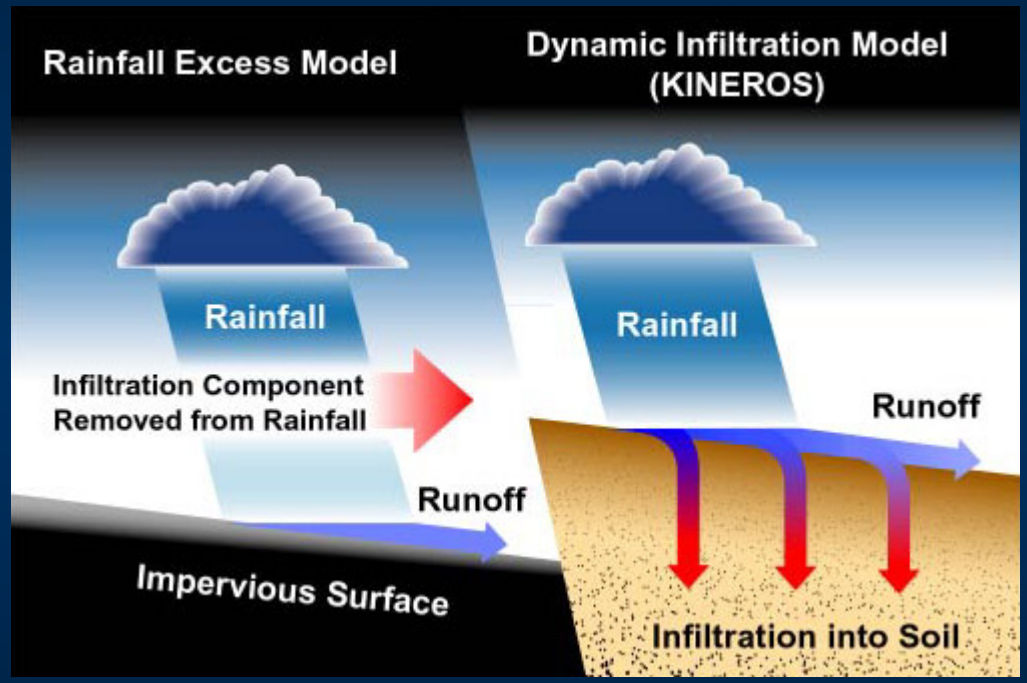
  DOWNSCALING

  Time-Space Scales
- Need to Verify Additional Atmospheric Parameters Not Commonly Examined.

  SFC Fluxes, Radiation, H_2O Vapor, Cloud

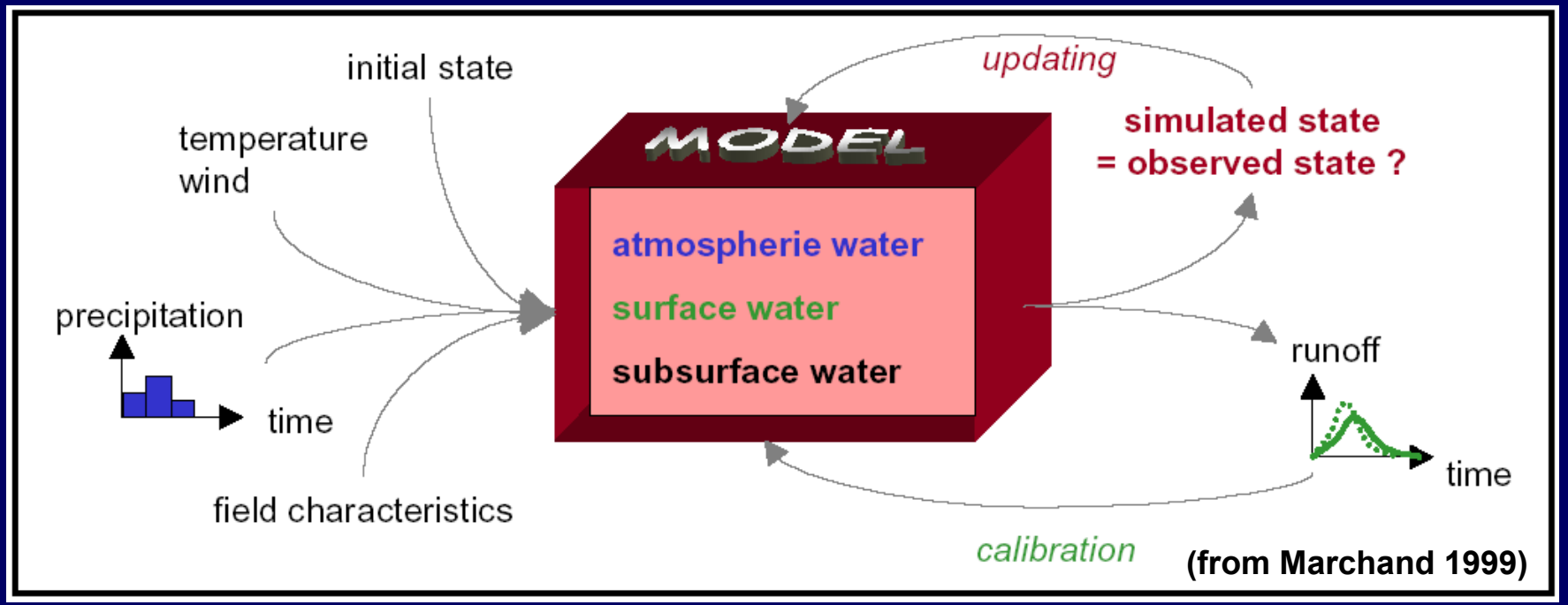
Hydrologic Models





Some Basically Need Precipitation from Atmospheric Ensembles

Hydrologic Models



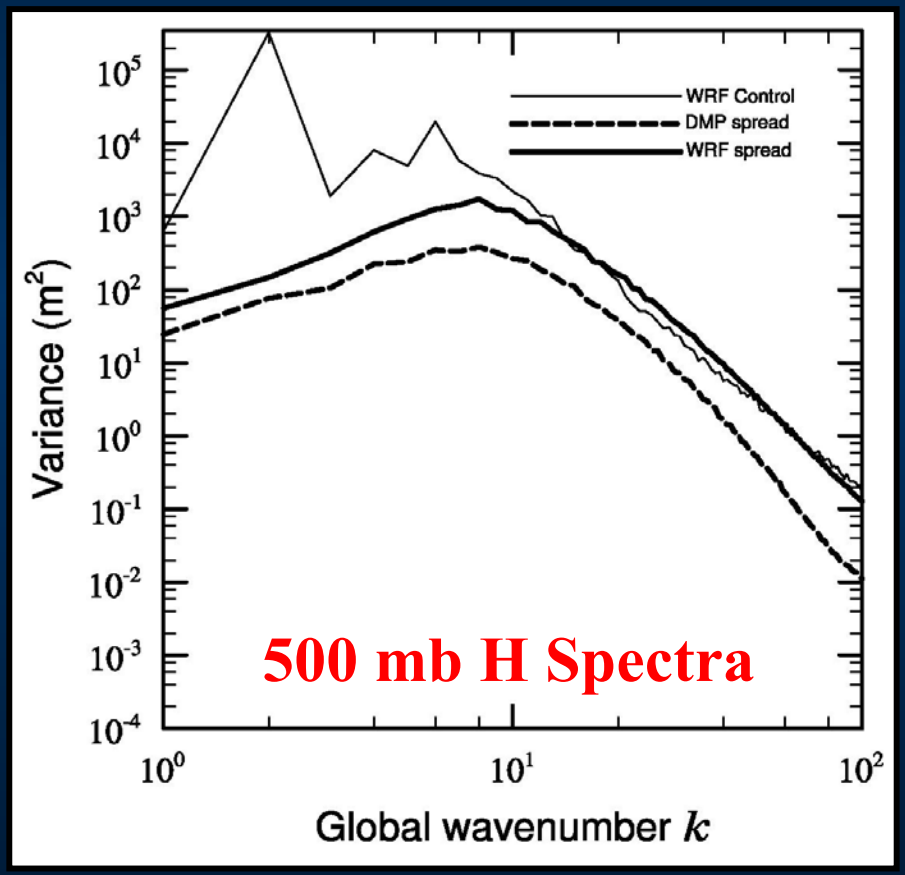
Others Require Atmospheric Input Fields of
Precipitation (amount, type, intensity)
Wind, T, Q, Clouds, Fluxes, Radiation, BC's

Driving Hydro Runoff Models with EPS/Ensemble Output

- What is Ensemble Performance for "Other" Parameters Needed by Hydro Models?

  Fluxes, Radiation, Water Vapor, Cloud
- How Well Are These Fields Observed? What is the uncertainty?
- How Does it Affect Estimates of Skill?

Time-Space Spectra, Intermittency



Hacker and Baumhefner (2004)

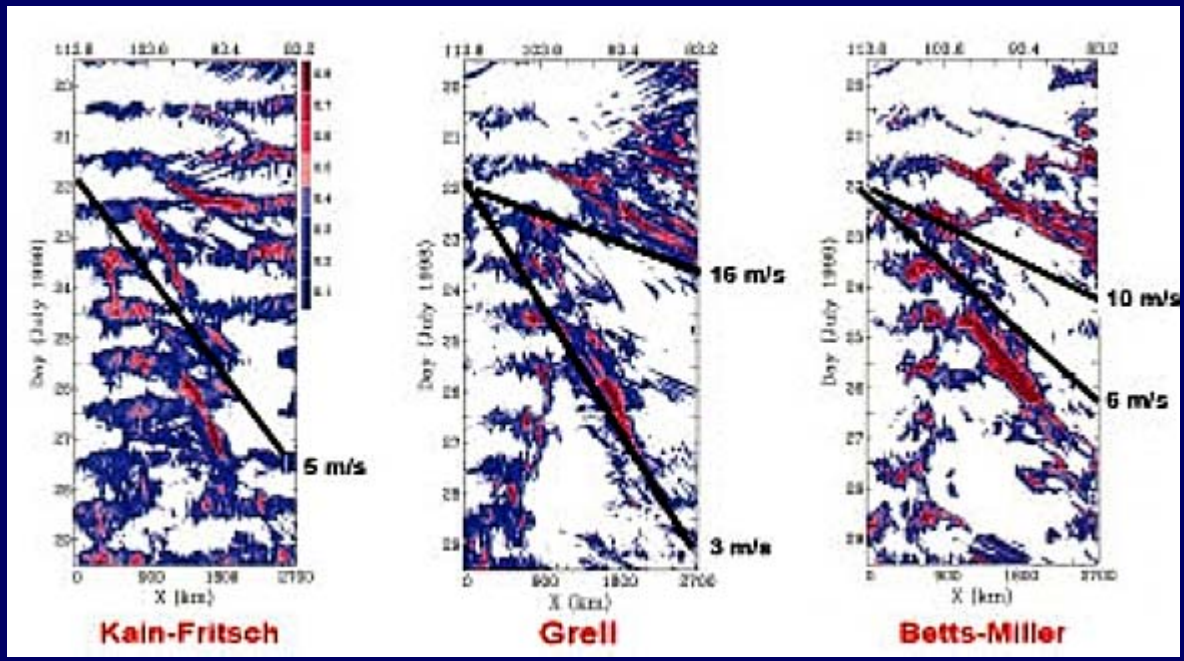
Runoff is Sensitive to Precipitation Intensity

25 mm/1 hr ⇒ runoff
25 mm/24 hr ⇒ none

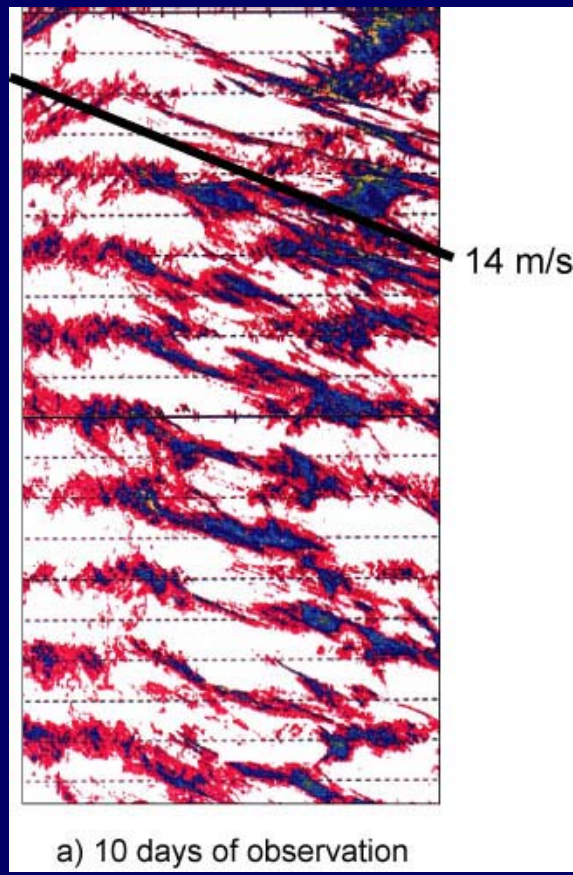
Intermittency Issue of Heavy Precipitation

Better Documentation and Understanding of Ensemble Variances

Feature Based Verification

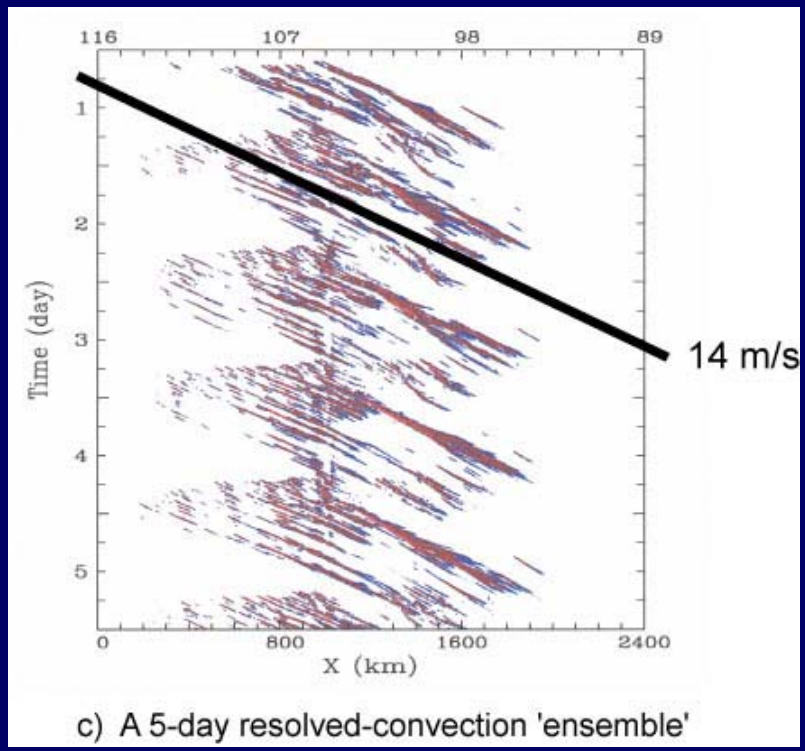


Moncrieff and collaborators (2002 ongoing)

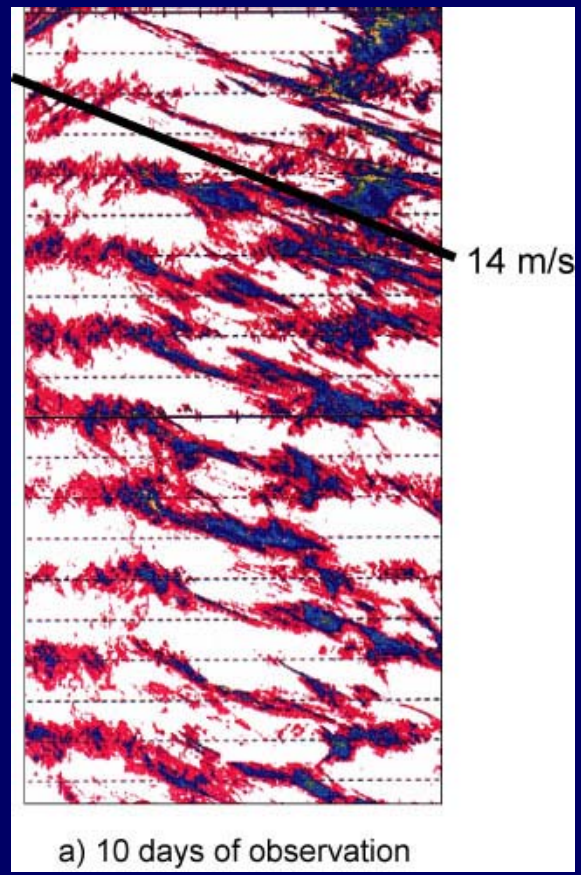


Carbone et al (2002)

Feature Based Verification



Moncrieff and collaborators (2002)



Carbone et al (2002)

Mucho Better Climate!
Simulates Aspect of Feature

Feature Based Verification

- **Heavy Precipitation/Flash Flooding is Often Associated with Features Like:**
  - **Quasi-Stationary Convection**
  - **"Training" Convection**
  - **Topographic Interactions**
  - **Warm Clouds**
  - **Land Falling Hurricanes, Tropical Systems**
  - **Warm Rain over Snow**
- **How well do ensemble systems perform?**

Closing Thoughts

- **Data/Com Requirements**
  - **Analyses/Observations (atmospheric and hydrologic) at requisite time/space scales with uncertainty estimates**
  - **Frequent, full resolution ensemble fields for parameters of relevance to hydrological models for calibration of ensemble and hydrological models**
- **What does Hydro Community need?**