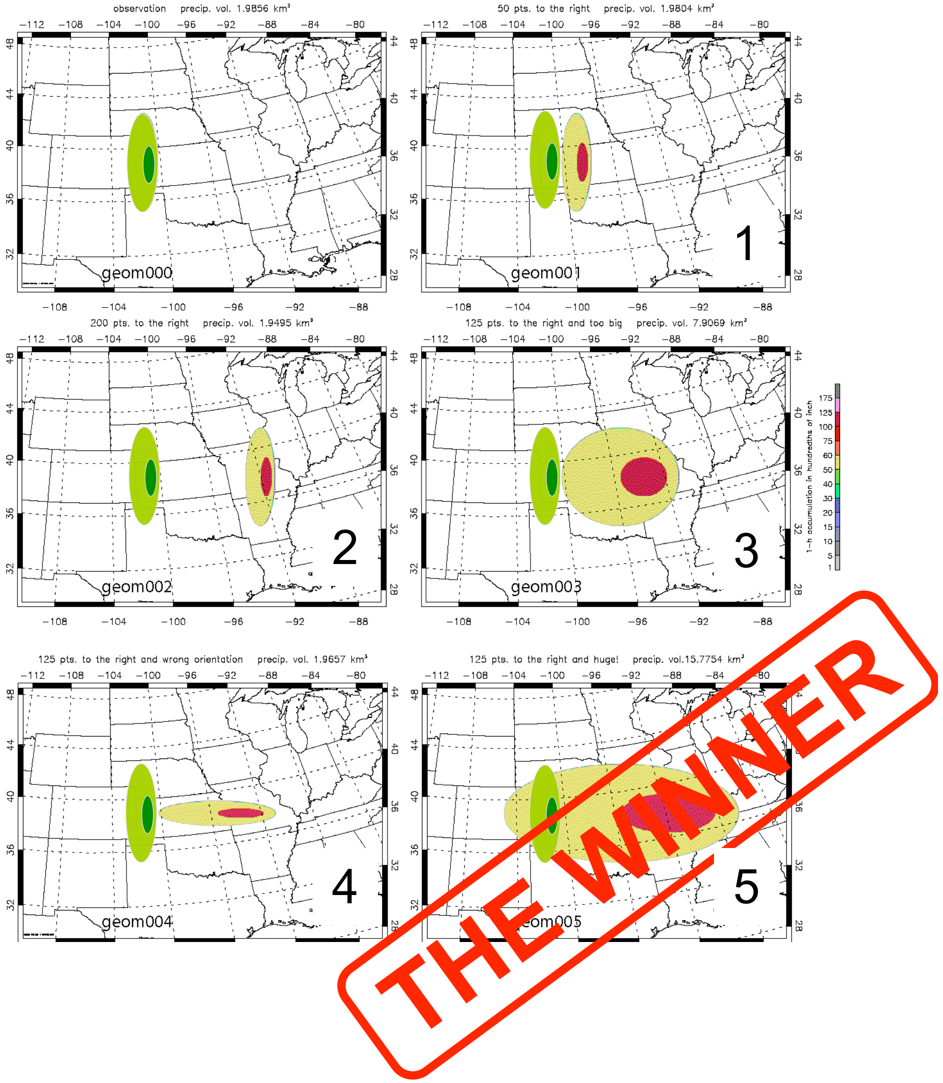
Traditional Verification Scores

- Fake forecasts
  - 5 geometric
  - 7 perturbed

- subjective evaluation
  - expert scores from last year’s workshop
  - 9 cases x 3 models

Geometric

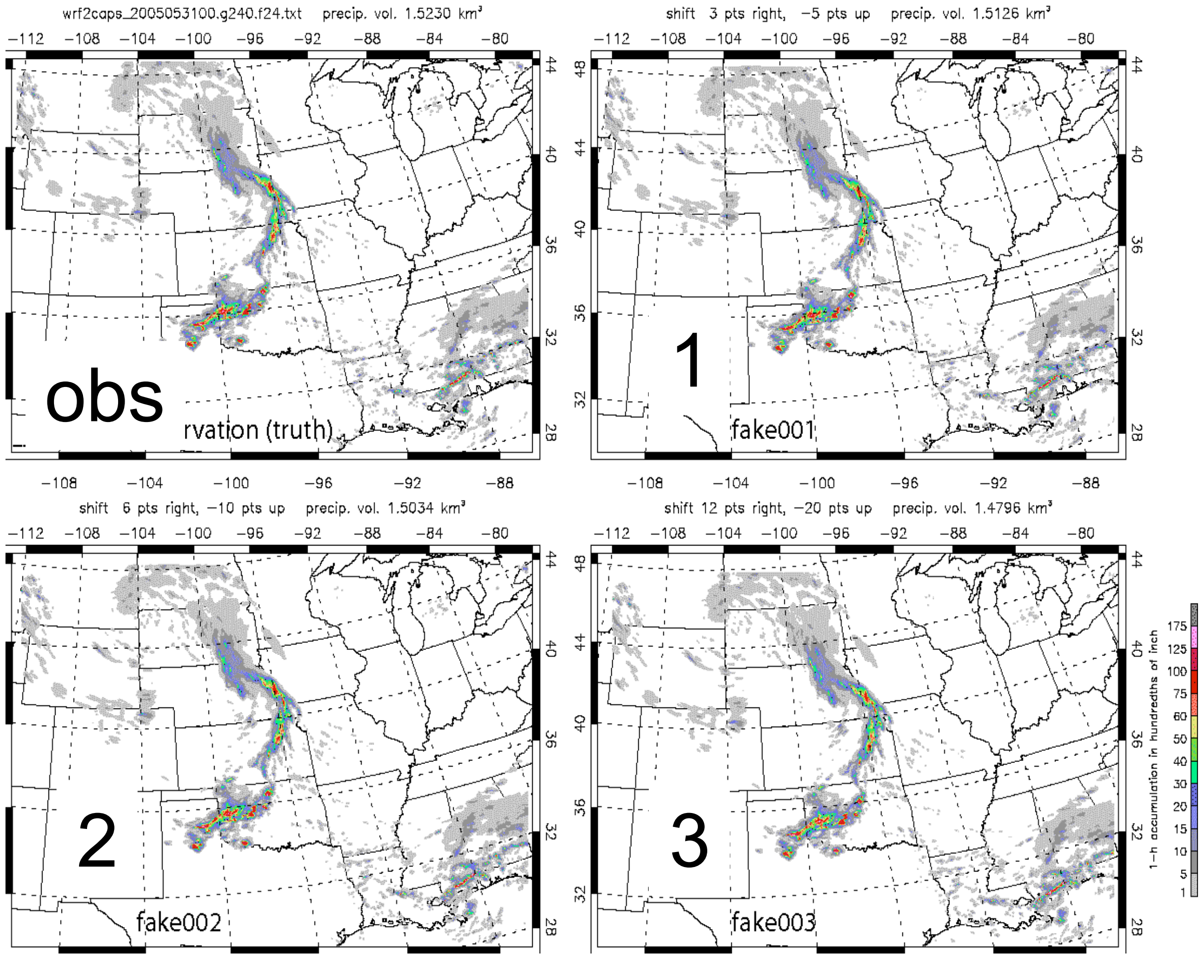
- error/scores for first 4 cases
  - correlation coefficient = -0.02
  - prob of detection = 0.00
  - false alarm ratio = 1.00
  - Hanssen&Kuipers = -0.03
  - equitable threat = -0.01
- case 5
  - correlation coefficient = 0.2
  - prob of detection = 0.88
  - false alarm ratio = 0.89
  - Hanssen&Kuipers = 0.69
  - equitable threat = 0.08

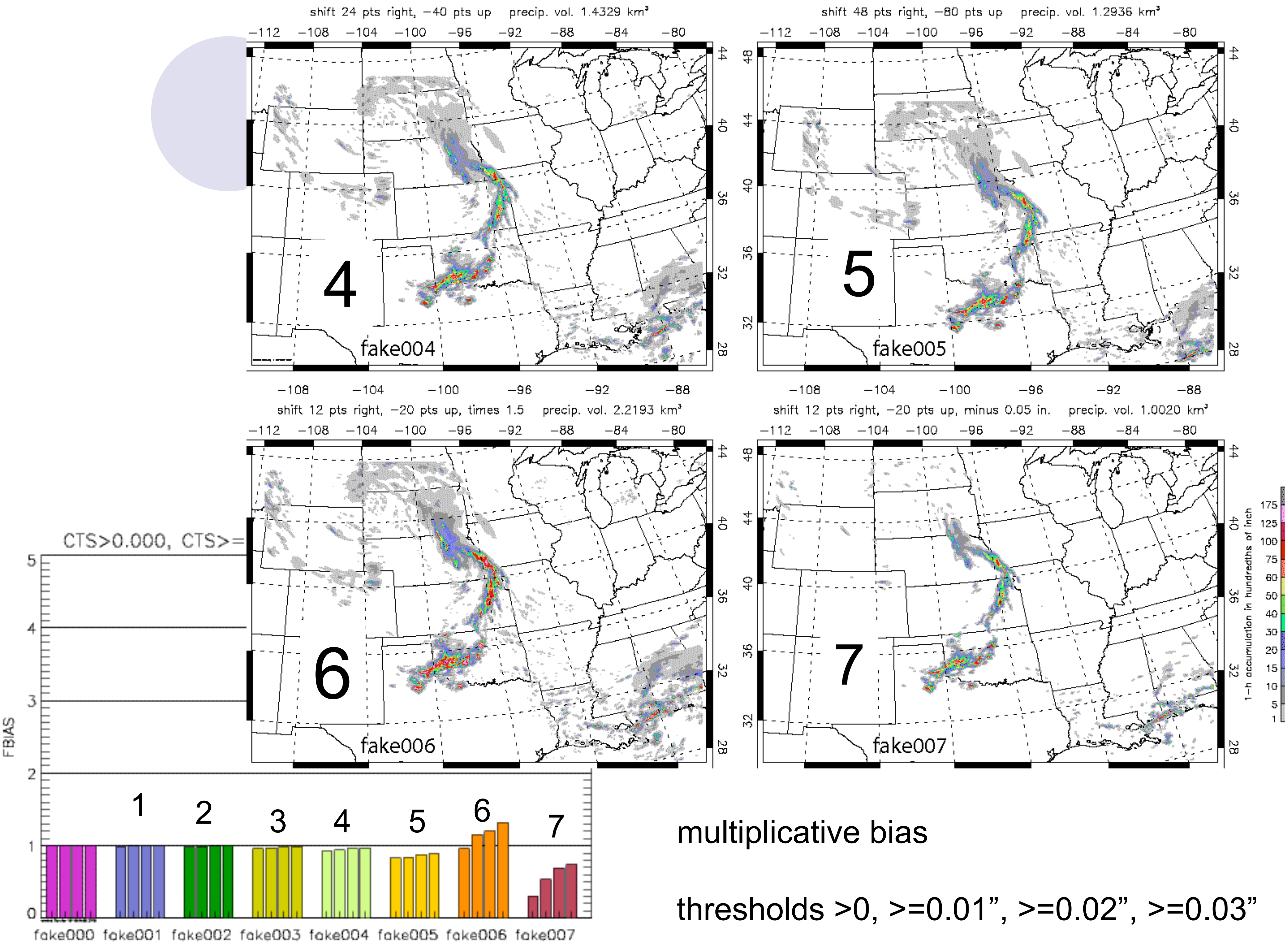

observation precip. vol. 1.9856 km³
geom000
50 pts. to the right precip. vol. 1.9804 km³
geom001
1
200 pts. to the right precip. vol. 1.9495 km³
geom002
2
125 pts. to the right and too big precip. vol. 7.9069 km³
geom003
3
1-h accumulation in hundredths of inch
125 pts. to the right and wrong orientation precip. vol. 1.9657 km³
geom004
4
125 pts. to the right and huge! precip. vol.15.7754 km³
geom005
5
THE WINNER

Perturbed fake cases – known errors

- 3 pts right, 5 pts down
- 6 pts right, 10 pts down
- 12 pts right, 20 pts down
- 24 pts right, 40 pts down
- 48 pts right, 80 pts down
- 12 pts right, 20 pts down, times 1.5
- 12 pts right, 20 pts down, minus 0.05”

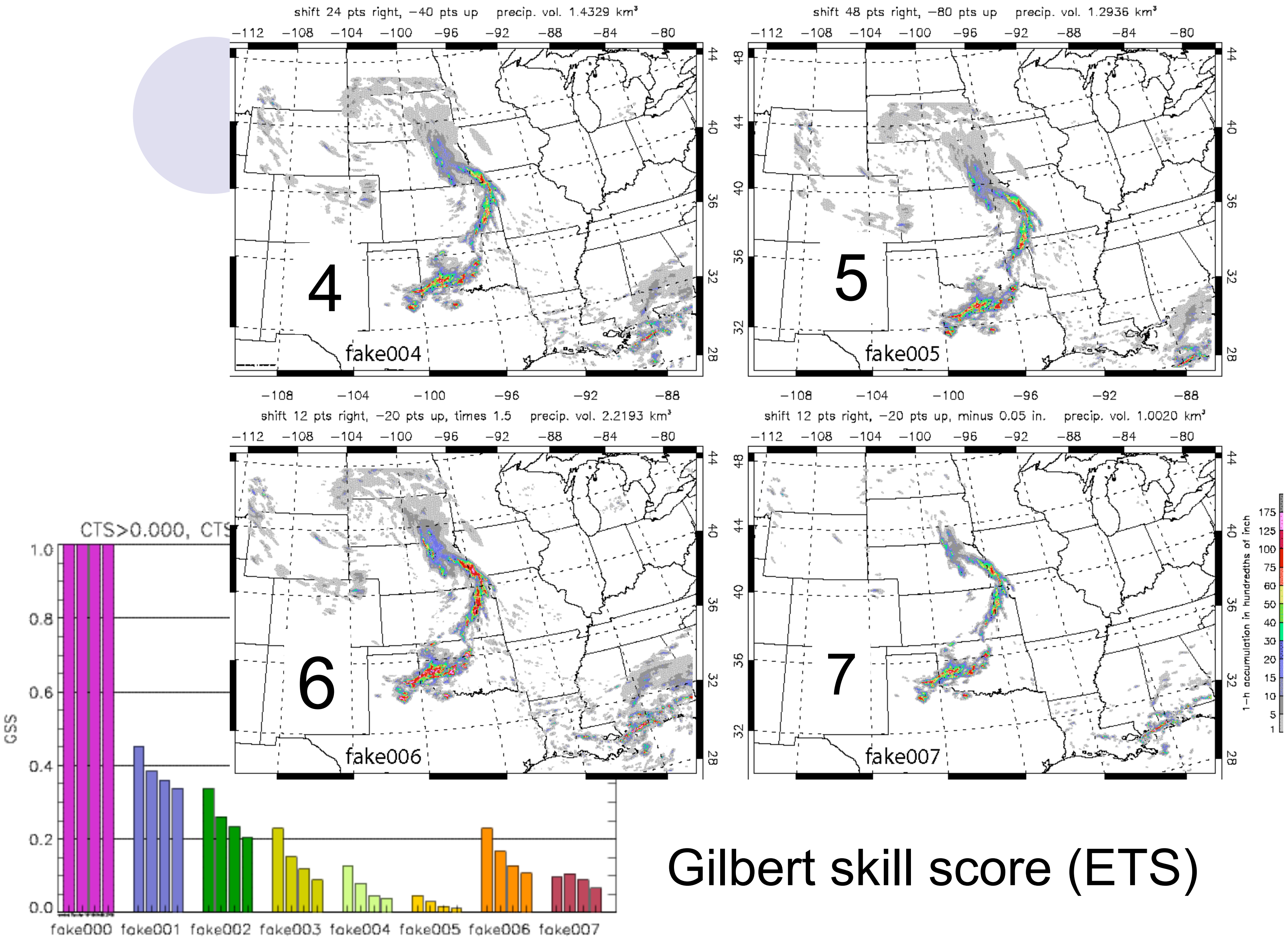
Perturbed fake cases 1-3





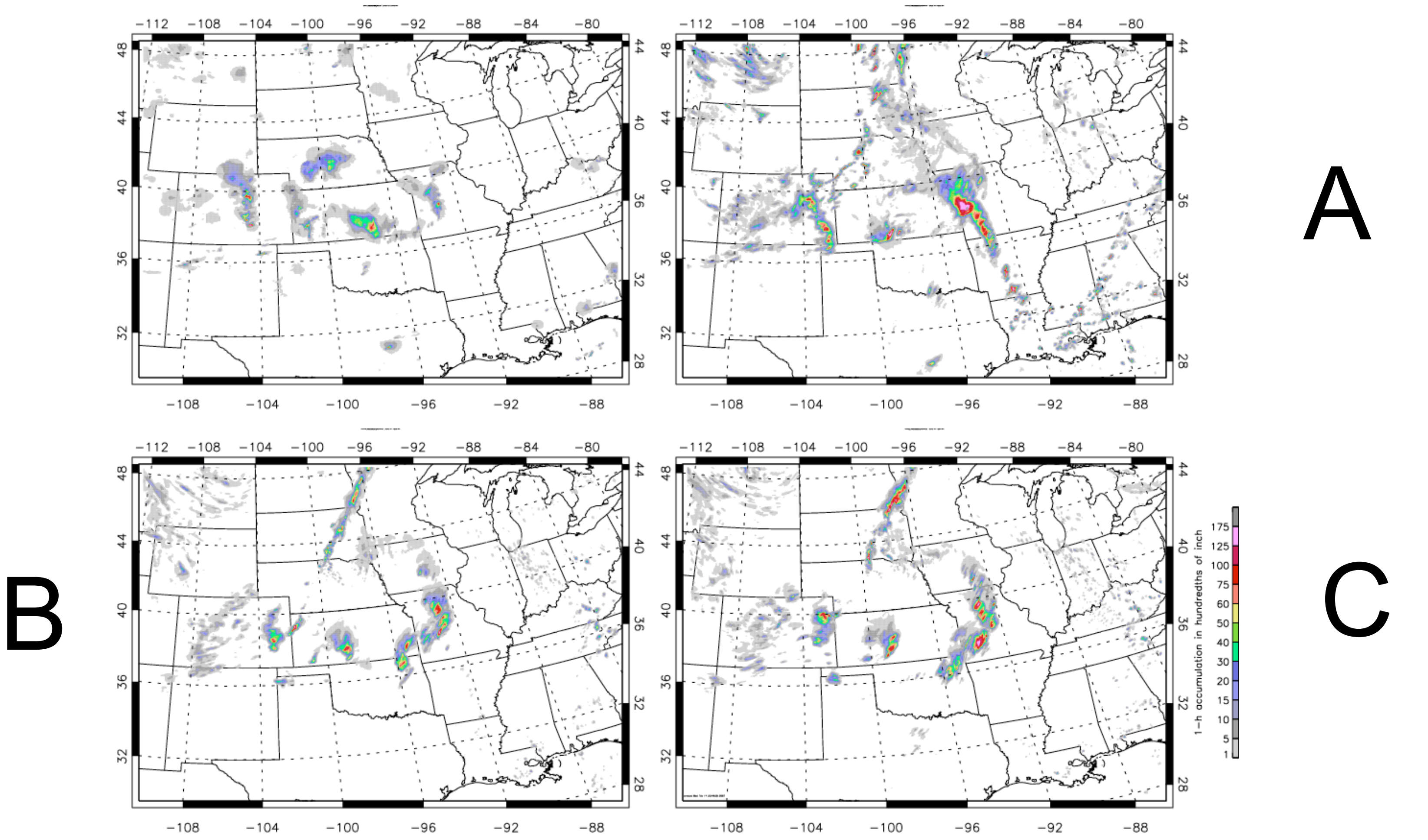
multiplicative bias

thresholds >0, >=0.01”, >=0.02”, >=0.03”

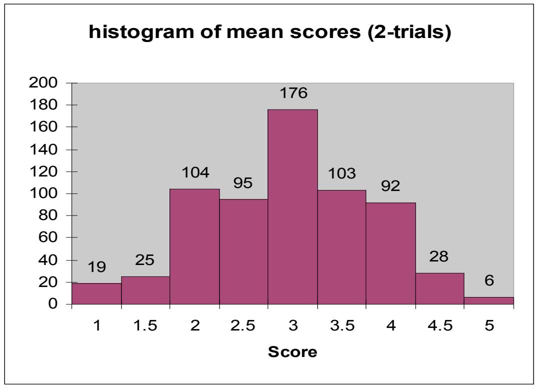


Gilbert skill score (ETS)

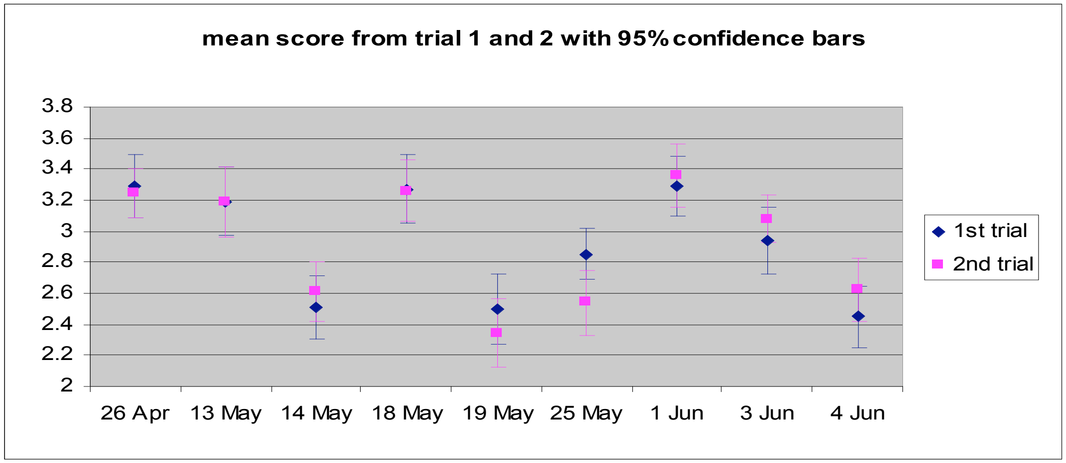
subjective evaluation



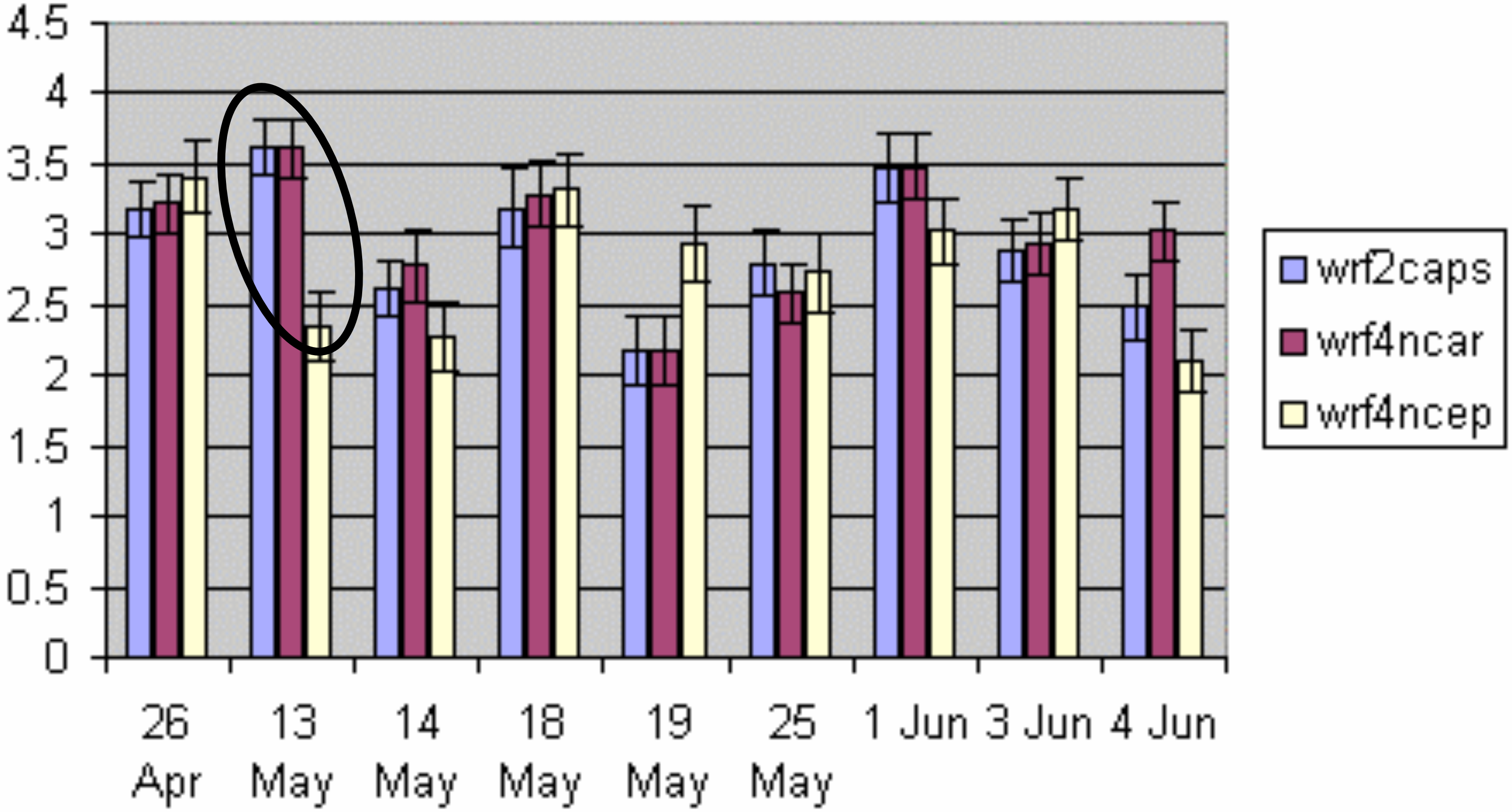
histograms of expert scores

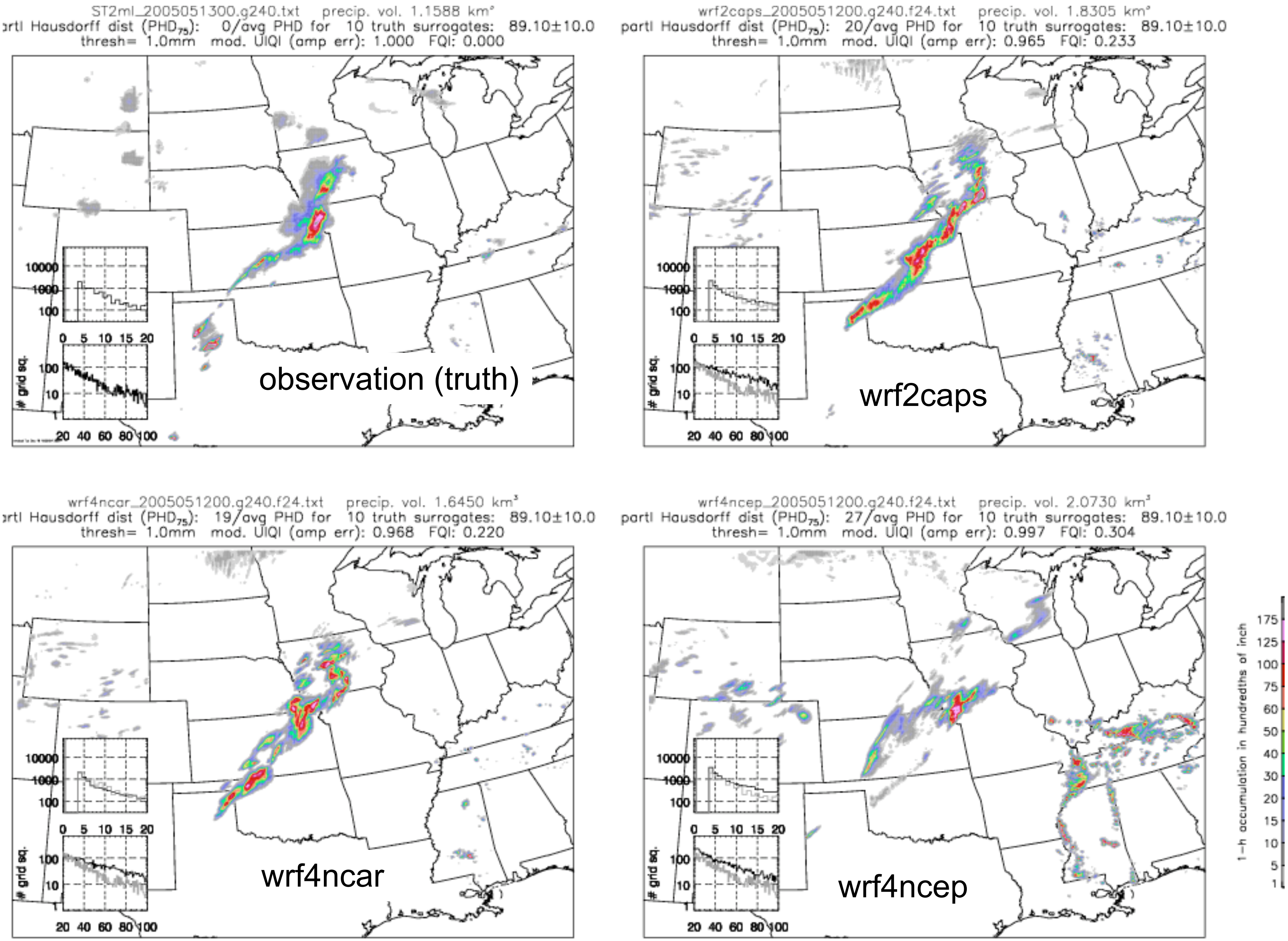


- 24 first-trial scores
- 22 second-trial scores

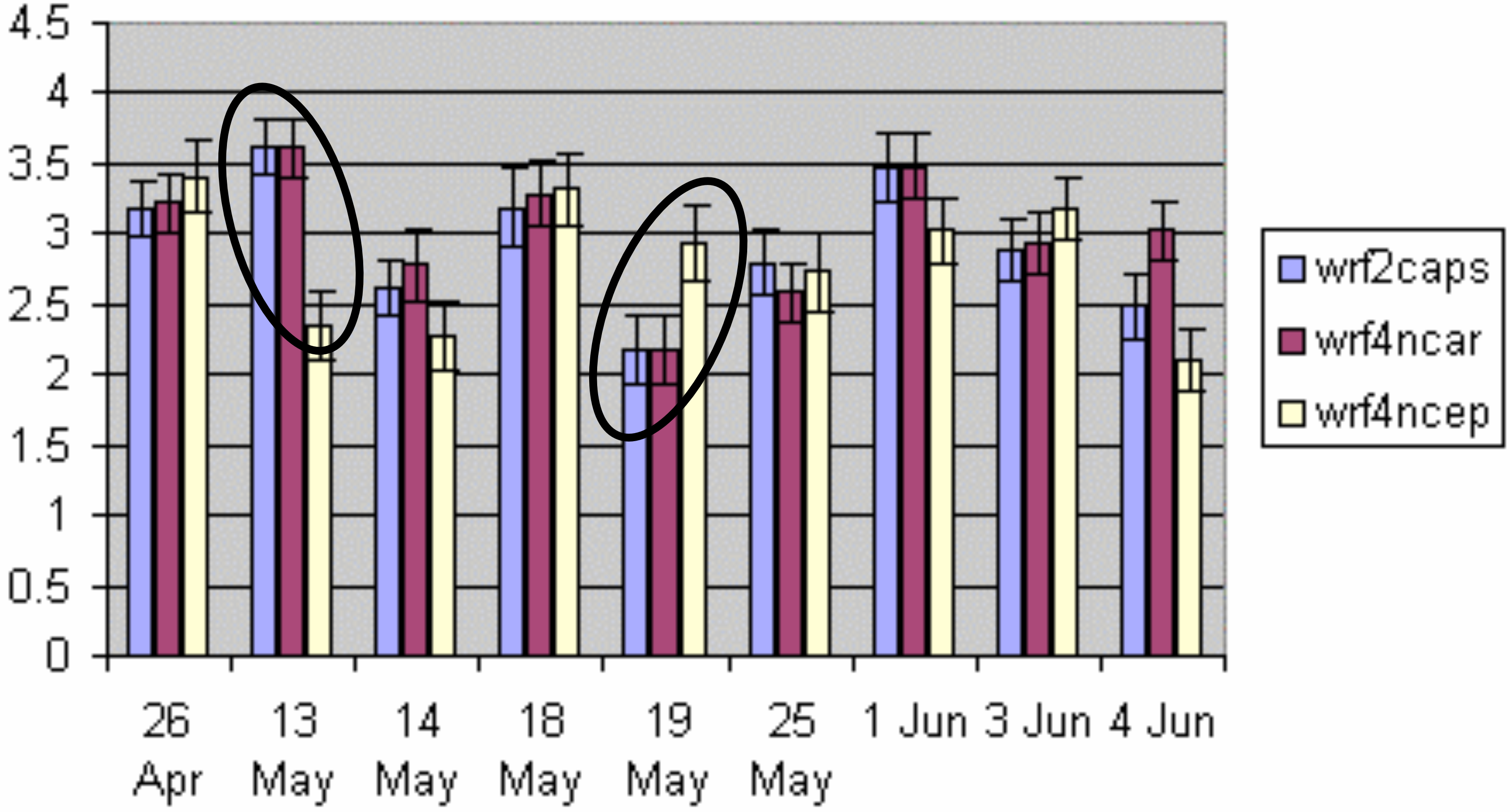


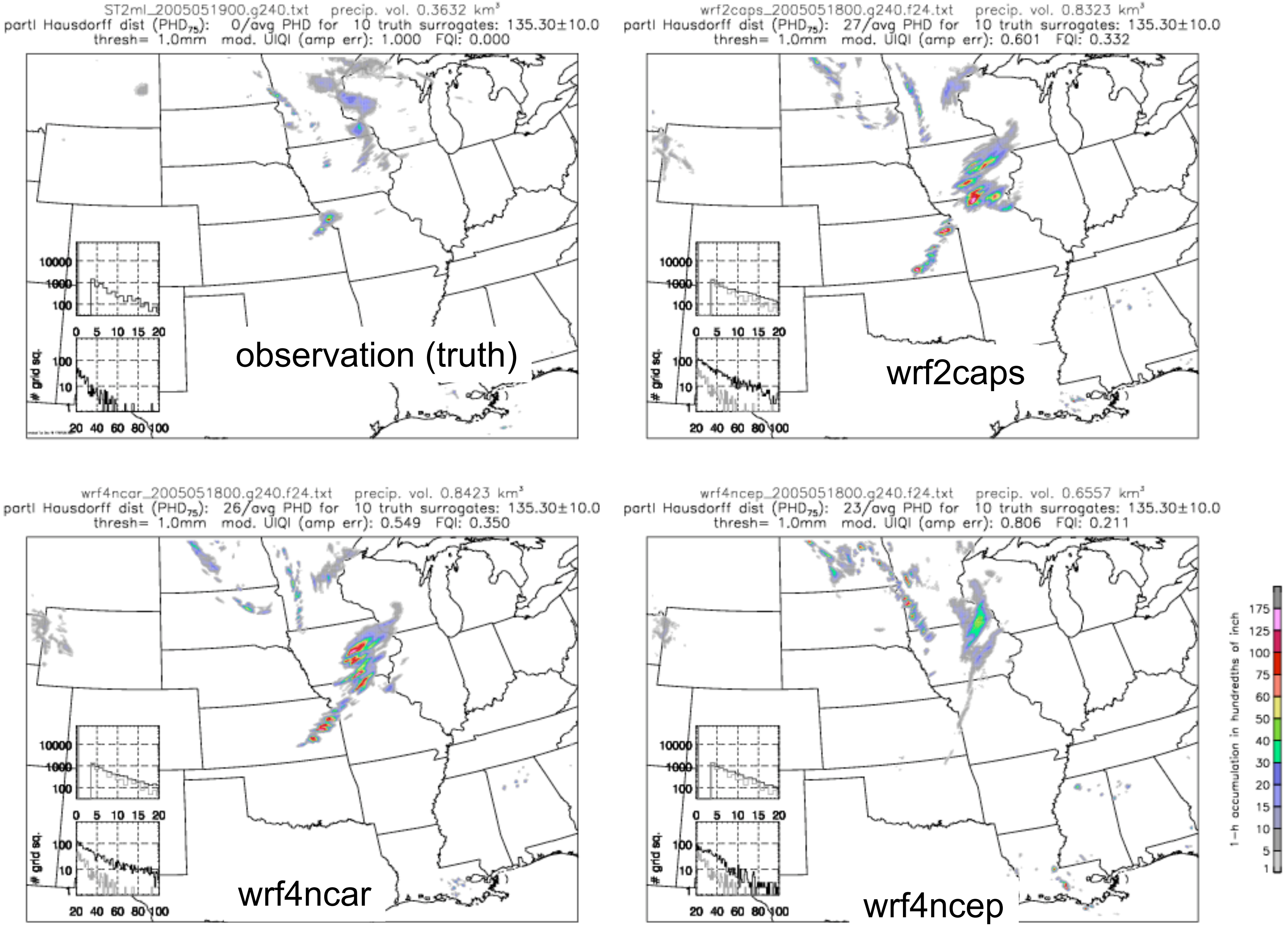
mean score +/- 1.96 std err



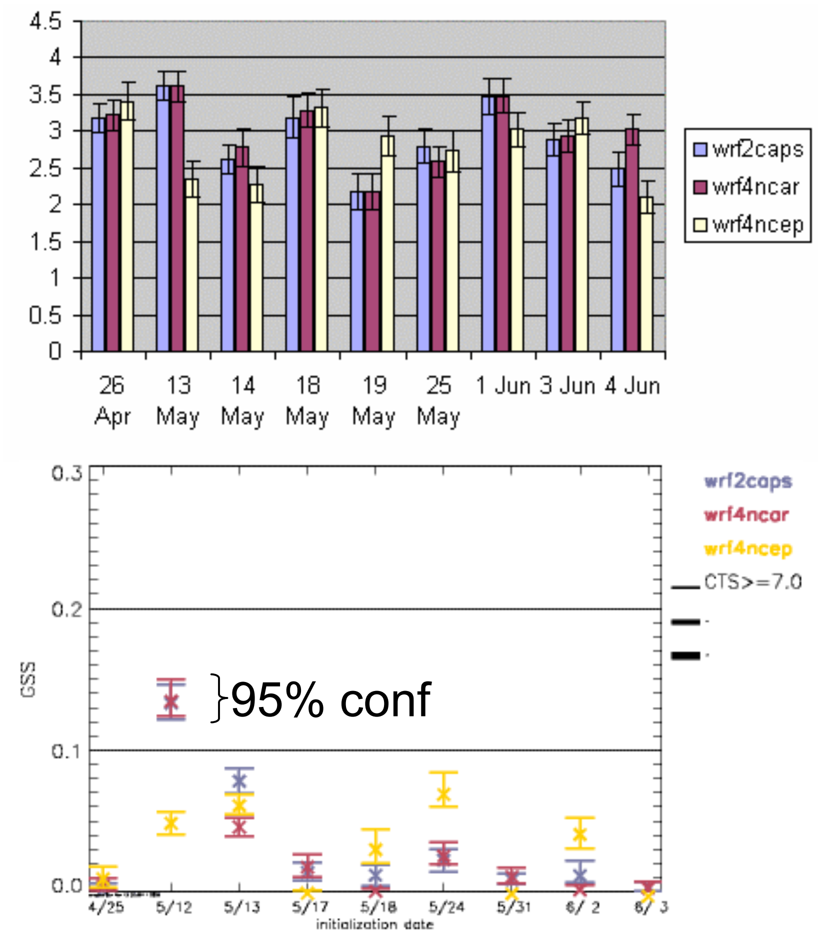

ST2ml_2005051300.g240.txt precip. vol. 1.1588 km³
rtl Hausdorff dist (PHD75): 0/avg PHD for 10 truth surrogates: 89.10±10.0
thresh= 1.0mm mod. UIQI (amp err): 1.000 FQI: 0.000
10000
1000
100
0 5 10 15 20
# grid sq.
100
10
1
20 40 60 80 100
observation (truth)
wrf2caps_2005051200.g240.f24.txt precip. vol. 1.8305 km³
partl Hausdorff dist (PHD75): 20/avg PHD for 10 truth surrogates: 89.10±10.0
thresh= 1.0mm mod. UIQI (amp err): 0.965 FQI: 0.233
10000
1000
100
0 5 10 15 20
# grid sq.
100
10
1
20 40 60 80 100
wrf2caps
wrf4ncar_2005051200.g240.f24.txt precip. vol. 1.6450 km³
rtl Hausdorff dist (PHD75): 19/avg PHD for 10 truth surrogates: 89.10±10.0
thresh= 1.0mm mod. UIQI (amp err): 0.968 FQI: 0.220
10000
1000
100
0 5 10 15 20
# grid sq.
100
10
1
20 40 60 80 100
wrf4ncar
wrf4ncep_2005051200.g240.f24.txt precip. vol. 2.0730 km³
partl Hausdorff dist (PHD75): 27/avg PHD for 10 truth surrogates: 89.10±10.0
thresh= 1.0mm mod. UIQI (amp err): 0.997 FQI: 0.304
10000
1000
100
0 5 10 15 20
# grid sq.
100
10
1
20 40 60 80 100
wrf4ncep
1-h accumulation in hundredths of inch
175
125
100
75
60
50
40
30
20
15
10
5
1

mean score +/- 1.96 std err



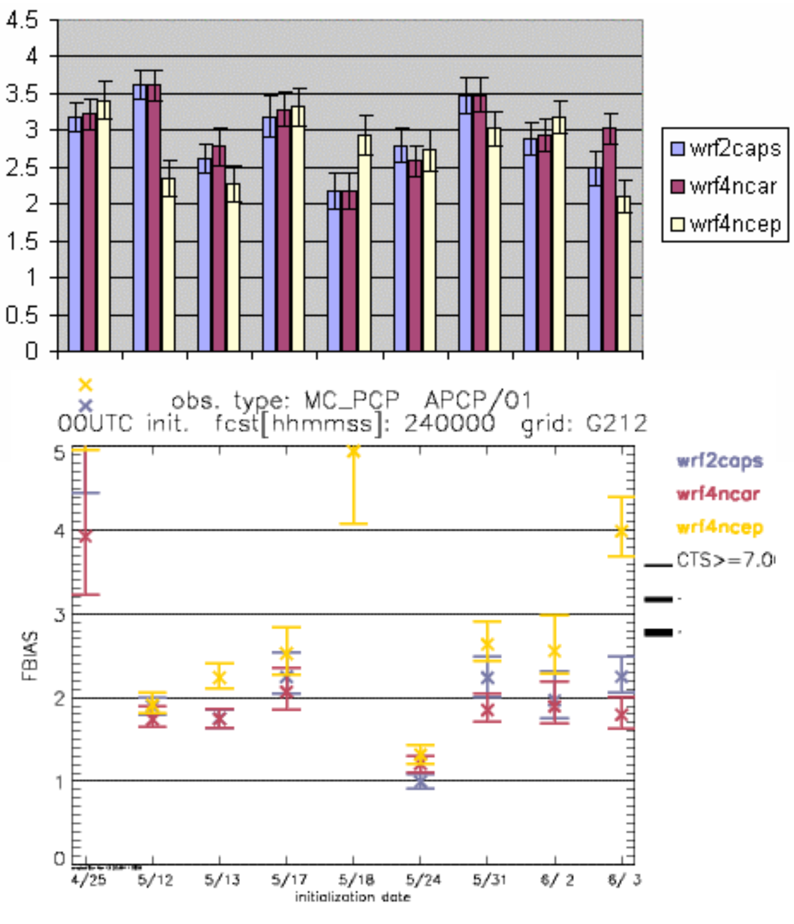

ST2ml_2005051900.g240.txt precip. vol. 0.3632 km³
partl Hausdorff dist (PHD75): 0/avg PHD for 10 truth surrogates: 135.30±10.0
thresh= 1.0mm mod. UIQI (amp err): 1.000 FQI: 0.000
observation (truth)
wrf2caps_2005051800.g240.f24.txt precip. vol. 0.8323 km³
partl Hausdorff dist (PHD75): 27/avg PHD for 10 truth surrogates: 135.30±10.0
thresh= 1.0mm mod. UIQI (amp err): 0.601 FQI: 0.332
wrf2caps
wrf4ncar_2005051800.g240.f24.txt precip. vol. 0.8423 km³
partl Hausdorff dist (PHD75): 26/avg PHD for 10 truth surrogates: 135.30±10.0
thresh= 1.0mm mod. UIQI (amp err): 0.549 FQI: 0.350
wrf4ncar
wrf4ncep_2005051800.g240.f24.txt precip. vol. 0.6557 km³
partl Hausdorff dist (PHD75): 23/avg PHD for 10 truth surrogates: 135.30±10.0
thresh= 1.0mm mod. UIQI (amp err): 0.806 FQI: 0.211
wrf4ncep
10000
1000
100
0 5 10 15 20
# grid sq.
100
10
1
20 40 60 80 100
1-h accumulation in hundredths of inch
175
125
100
75
60
50
40
30
20
15
10
5
1

expert scores vs grid stats



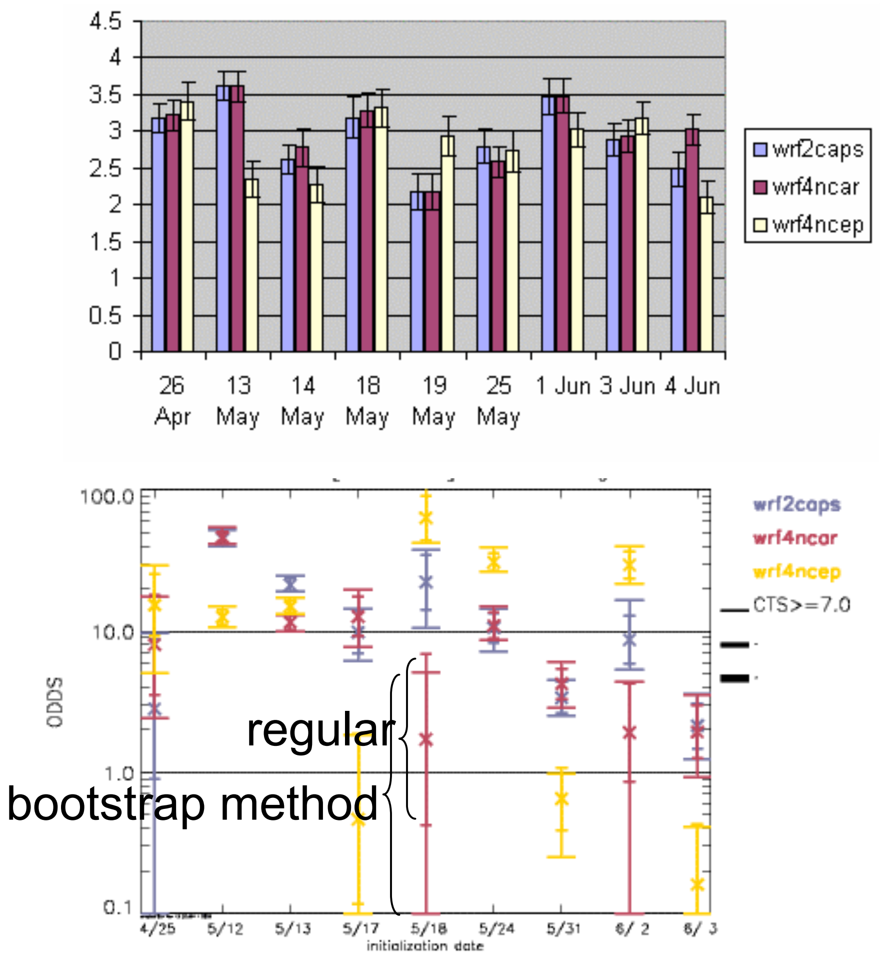
}95% conf

Equitable threat score (Gilbert Skill score)



forecast area bias (thresh=0.07”)

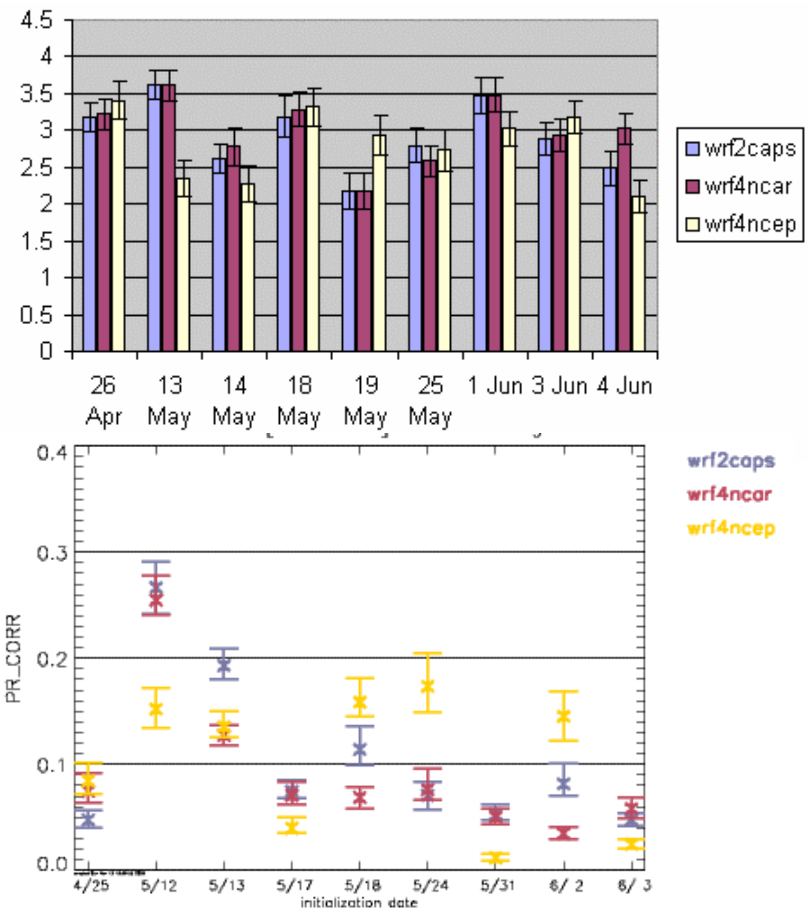
expert scores vs grid stats



regular

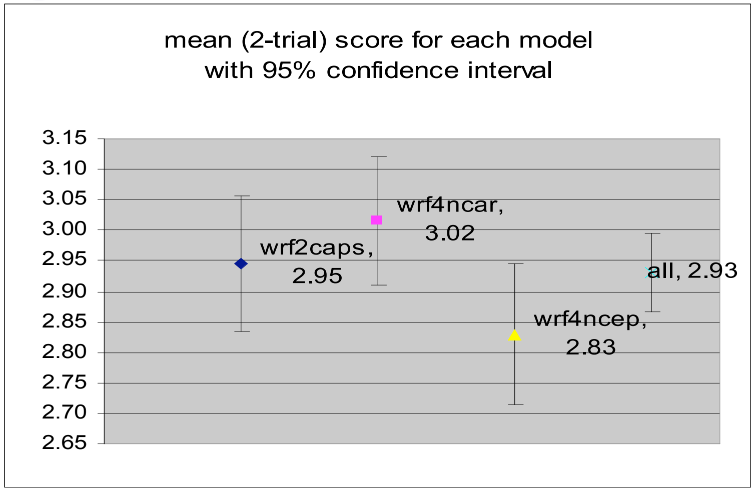
bootstrap method

odds ratio



Pearson product-moment correlation coefficient

do the expert scores show significant differences among the models?



Student's t-Test

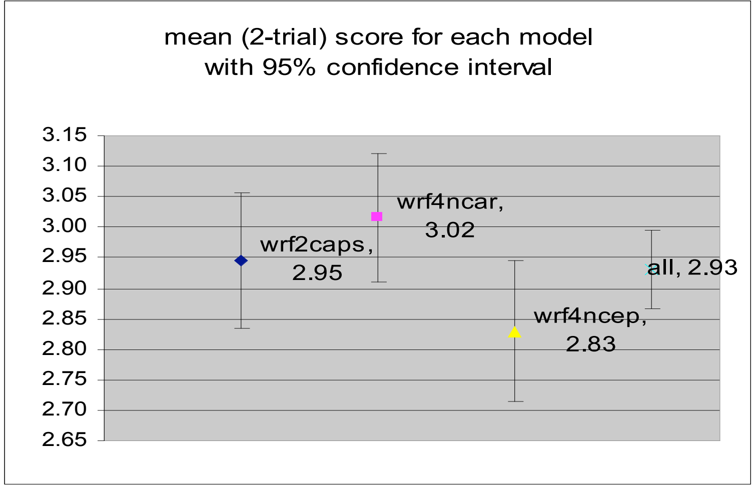
2-tail, paired

2-trial mean	p-value
wrf2caps-wrf4ncar	0.04
wrf2caps-wrf4ncep	0.06
wrf4ncar-wrf4ncep	0.003

Chance null hypothesis is true (i.e. no difference in means)

do the expert scores show significant differences among the models?

Wilcoxon-Mann-Whitney rank-sum test (Wilks, p. 138)	2-tail probability difference in ranks due to chance
wrf2caps-wrf4ncar	0.299
wrf2caps-wrf4ncep	0.148
wrf4ncar-wrf4ncep	0.018



Wilcoxon signed-rank test (Wilks, p. 142)	2-tail
wrf2caps-wrf4ncar	0.737
wrf2caps-wrf4ncep	0.177
wrf4ncar-wrf4ncep	0.152