Scale sensitivities in model precipitation skill scores

Andrew.Loughe@noaa.gov

Stephen Weygandt
Stan Benjamin

Andy Loughe*
Jennifer Mahoney

NOAA Forecast Systems Laboratory
*CIRES, University of Colorado, Boulder, CO

The problem of verifying convective precipitation forecasts

- **Thunderstorms** produce precipitation patterns with **significant small-scale detail**
- **High-resolution numerical models** are increasingly able to produce **similar small-scale detail**

But….

- Detailed model fields often have **small phase errors** compared to observations
- Traditional **skill scores are often worse** for detailed models even though they produce more realistic forecasts

The present situation

- Realize there is no single perfect verification score
- Active research on many new verification approaches
  - Spatial structures measures
  - Object oriented techniques
  - Scale dependent techniques

However….

- Operational precipitation verification still frequently relies upon ETS, bias
- Models with different grid resolution and different resolvable-scales are still being compared

Goals of this Study

- **Systematically *document* the scale-sensitivities known to exist for traditional skill scores by…**

  - Comparing **equitable threat** and **bias scores** for models verified on different resolution grids

  - Examining spectra from various models and observations on different resolution grids

It is not our purpose to:

- **Develop a "new" verification skill score**
- **Decide how much small-scale detail is acceptable in mesoscale models**

Key Questions

- How is **Equitable Threat Score** affected by the amount of small-scale detail in the:
  - forecast field?
  - verification field?

- How does **model bias** affect this scale dependency?

Specifically….

- Are ETS values from models with **different grid spacing** and **different biases** directly comparable?

- Does a **smoother precipitation field** yield a higher ETS value when compared with a highly **detailed verification field?**

Threat Score and Bias

- **<u>Threat Score</u> = Hits / (Hits + Misses + False Alarms)**

  Highlights events that actually occur, rather than those which do not

  ETS is the threat score corrected for a chance forecast…

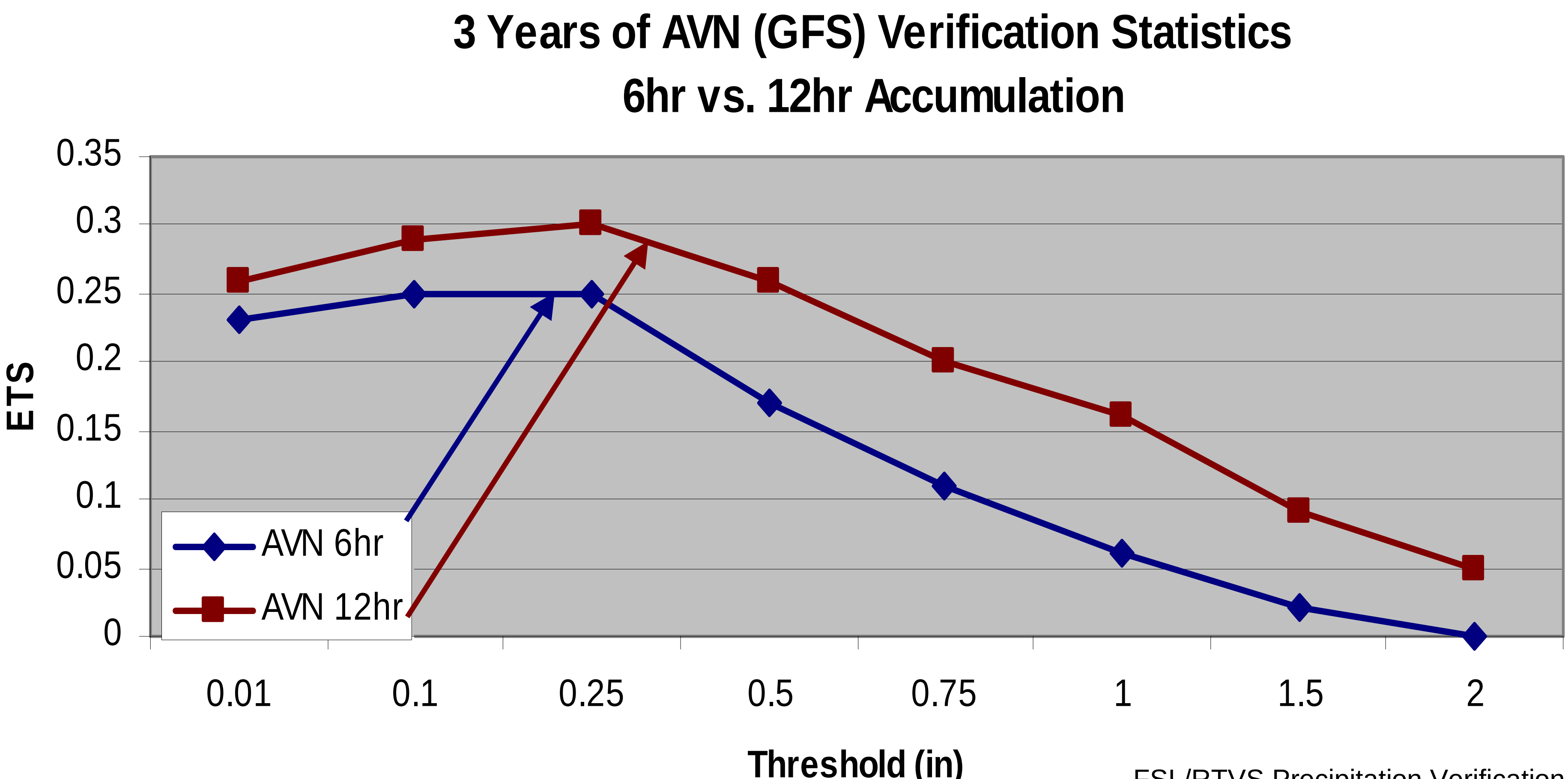
  ETS = (Hits – chance) / (Hits + Misses + False Alarms - chance)

- **<u>Bias</u> = Area Forecast / Area Observed**

  No dependence upon "hits!"

Smoothing of forecast fields over time

QPF verification statistics computed over a longer accumulation period are shown to be better than those computed over a shorter period

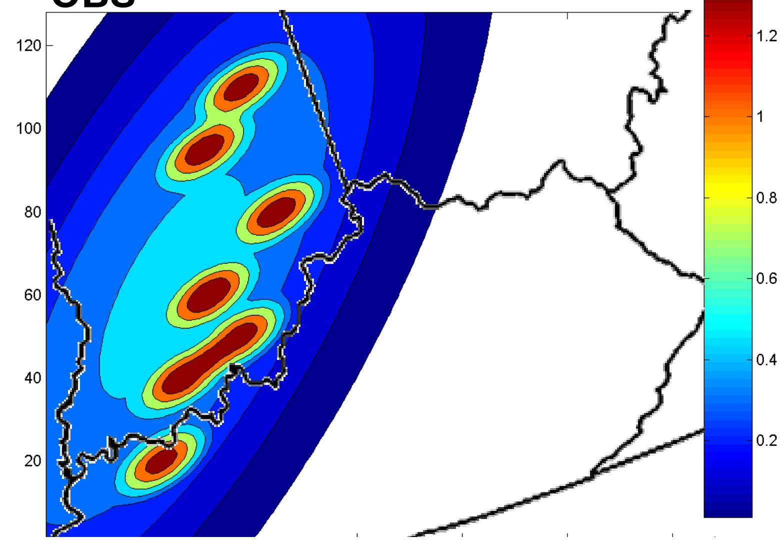


FSL/RTVS Precipitation Verification

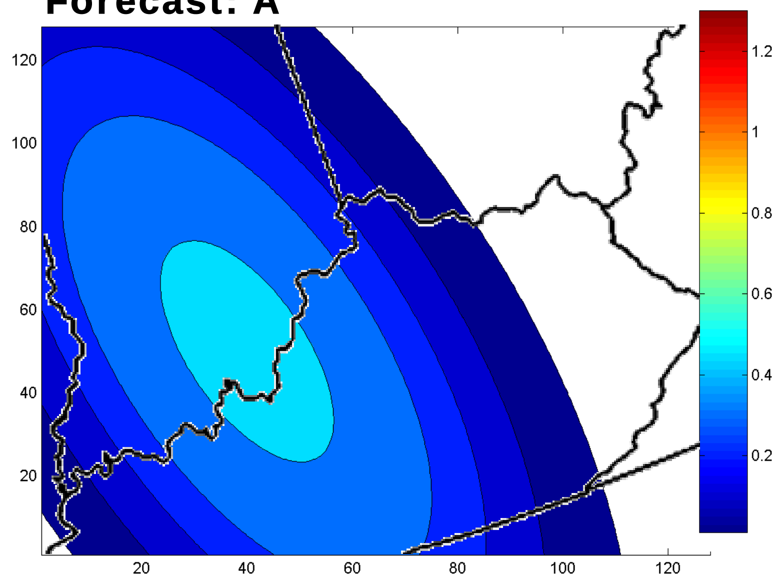
Spatial smoothing of forecast fields also has been shown to result in higher skill scores…

	A	B
MAE:	0.157,	0.159
RMSE:	0.254,	0.309
Bias:	0.980,	0.980
CSI:	0.214,	0.161
ETS:	0.170,	0.102

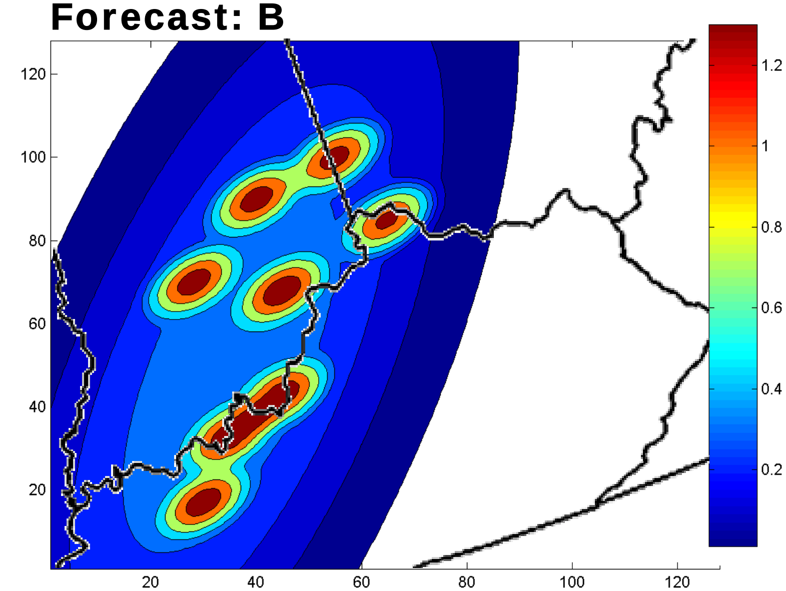
OBS



Forecast: A



Forecast: B



From Mike Baldwin of NOAA/NWS/SPC OU/CIMMS

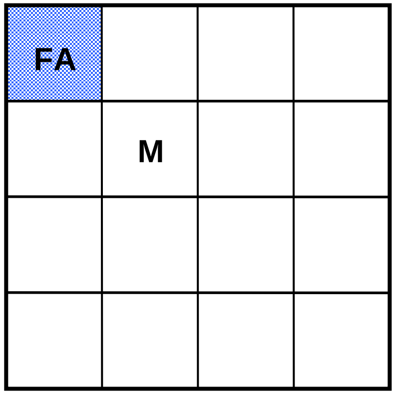
Double penalty

When forecast models resolve very small precipitation detail, they often suffer a **double penalty** when verified categorically on the observational grid.

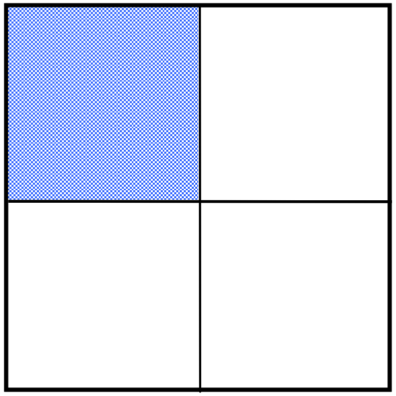
In this example, the **10 km forecast** is penalized twice: once for not placing rain in the correct place (a miss), and once for placing rain in the wrong place (a false alarm).

The **20 km forecast** receives one hit and 3 false alarms, giving a higher ETS and bias.

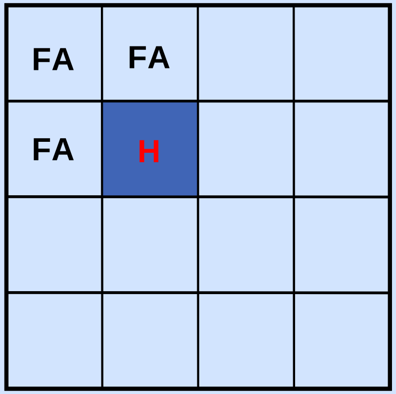
$ETS_{10km} = -0.03$ $ETS_{20km} = 0.20$



10 km forecast



20 km forecast



10 km observational grid

IHOP Real-time Modeling

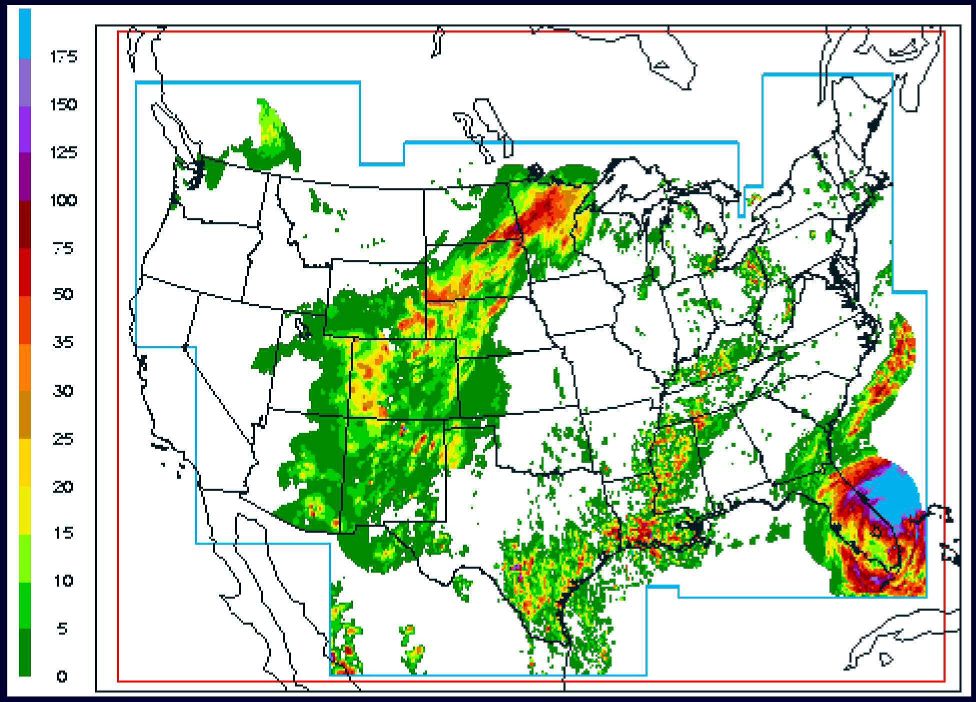
Experimental	**RUC 10-km**	(GD ensemble convection)
Operational	**RUC 20-km**	(GD ensemble convection)
Operational	**Eta 12-km**	(BMJ convection)
Experimental	**LAPS MM5 12-km**	(KF convection)

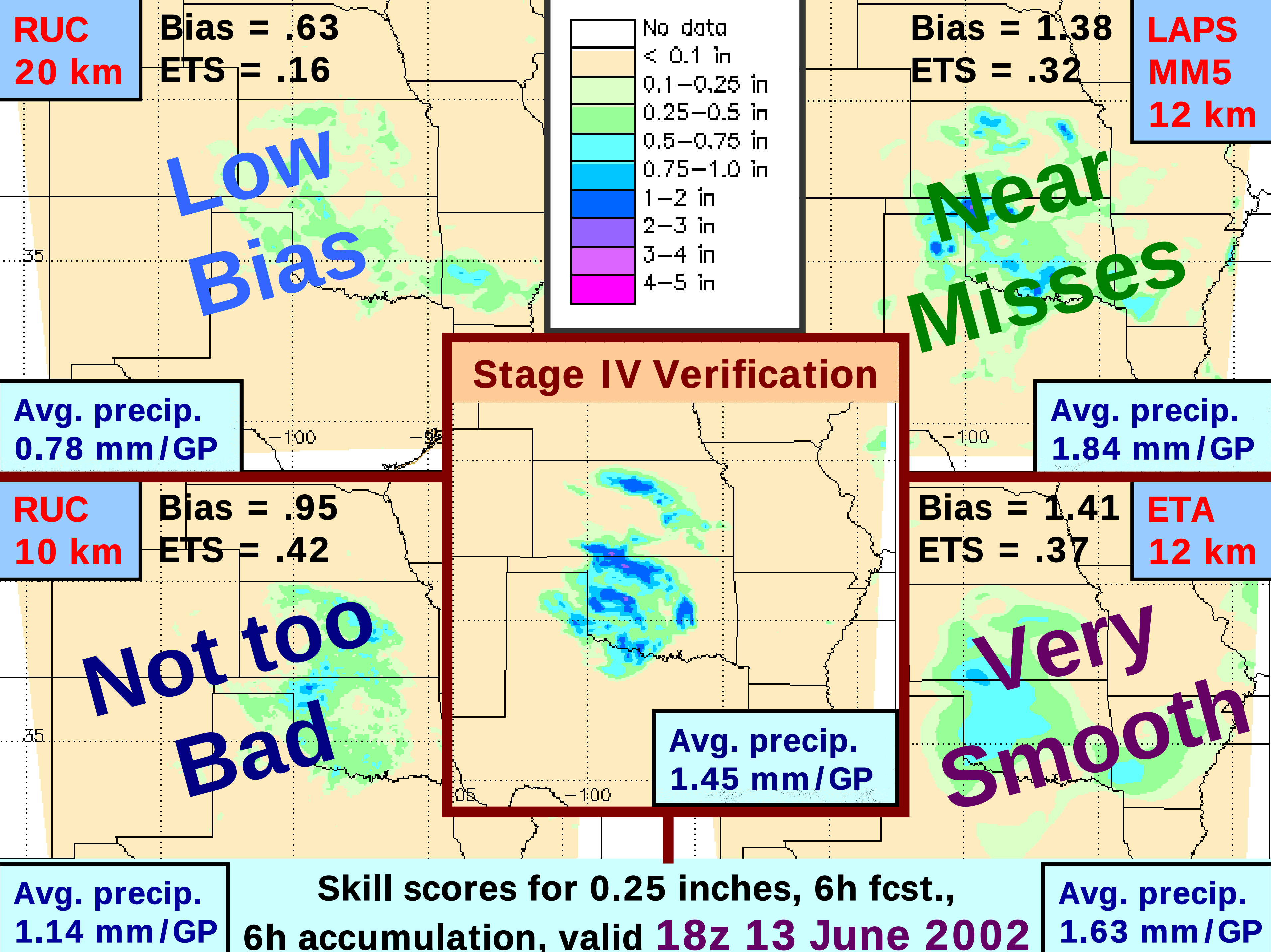
Model	**Native (km)**	**Coarsened (km)**		
RUC10	**10**	20	40	80
RUC20	**20**	20	40	80
ETA12	**12**	20	40	80
LMM12	**12**	20	40	80
Stage4 verif	4 (10)	20	40	80

This study is not a model bake off!

Observations: NCEP Stage IV Analysis

Mosaic of regional hourly and 6-hourly multi-sensor (radar+gauges) precipitation analysis at 4km.



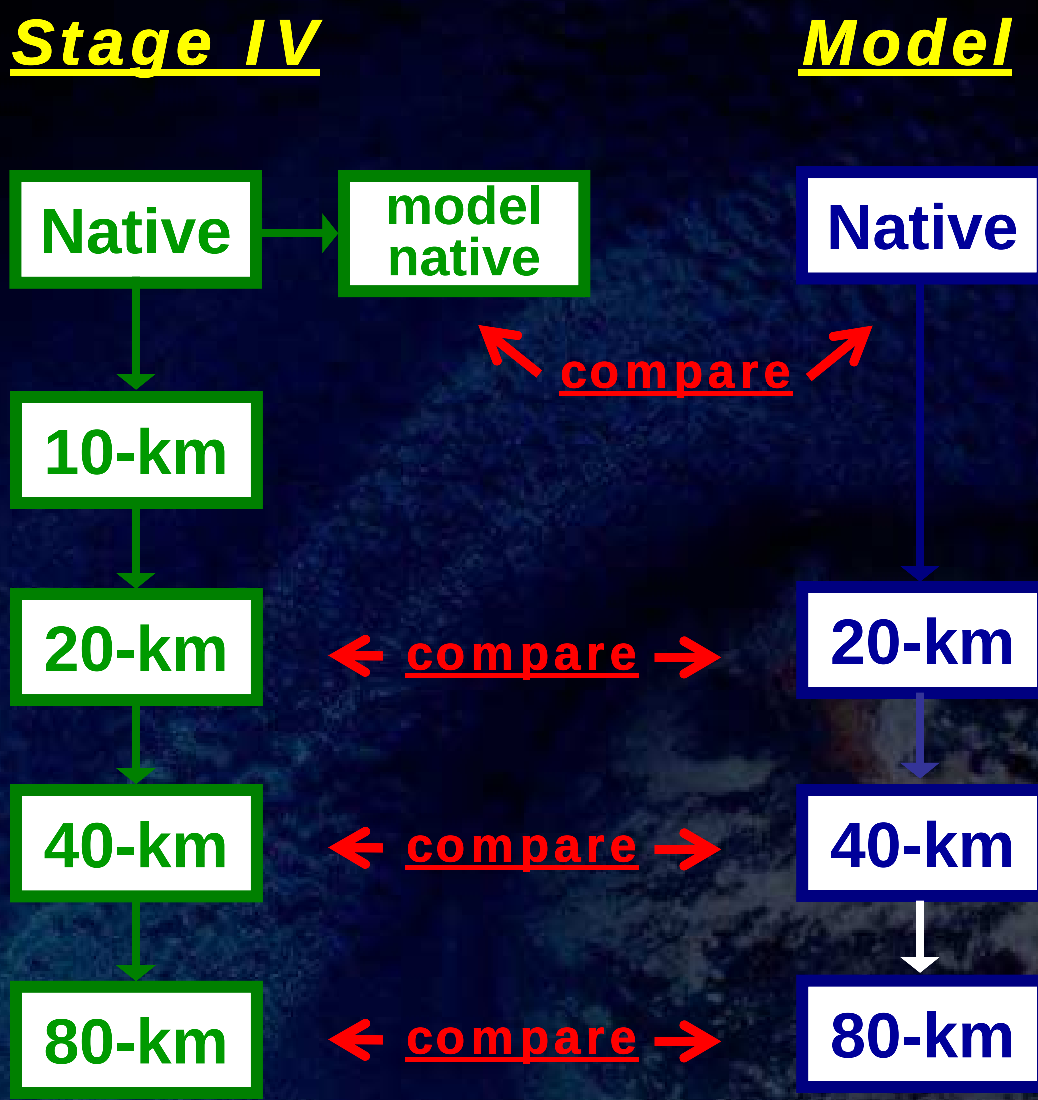

RUC
20 km
Bias = .63
ETS = .16
Low Bias
Avg. precip.
0.78 mm/GP
No data
< 0.1 in
0.1–0.25 in
0.25–0.5 in
0.5–0.75 in
0.75–1.0 in
1–2 in
2–3 in
3–4 in
4–5 in
Bias = 1.38
ETS = .32
LAPS
MM5
12 km
Near Misses
Avg. precip.
1.84 mm/GP
Stage IV Verification
Avg. precip.
1.45 mm/GP
RUC
10 km
Bias = .95
ETS = .42
Not too Bad
Avg. precip.
1.14 mm/GP
Bias = 1.41
ETS = .37
ETA
12 km
Very Smooth
Avg. precip.
1.63 mm/GP
Skill scores for 0.25 inches, 6h fcst.,
6h accumulation, valid 18z 13 June 2002

Upscale forecasts and observations

EXPT. 1

Remap Stage IV and model data to common, coarser resolution grids

Compare scores from forecasts with different precipitation detail verified on their native grid

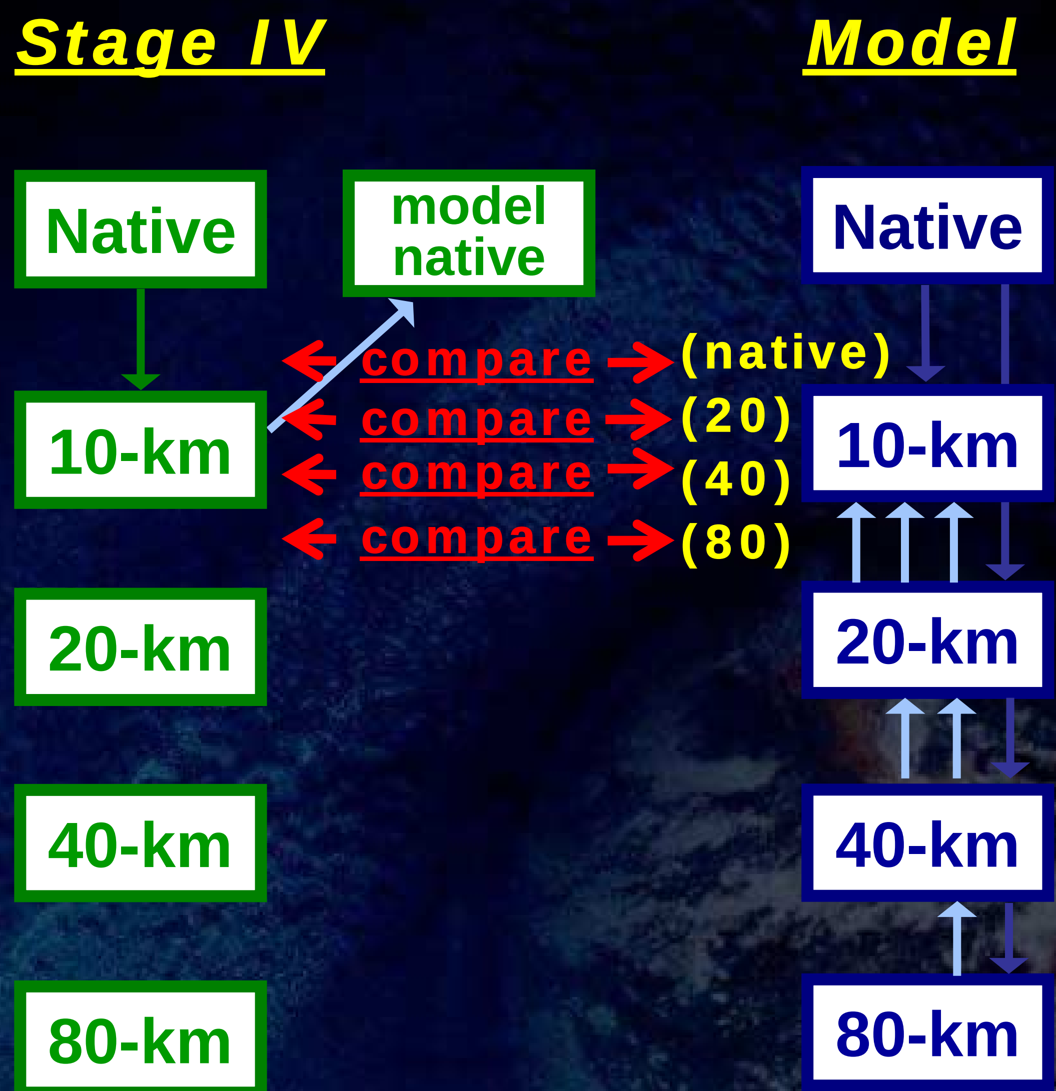


Smooth forecasts only

EXPT. 2

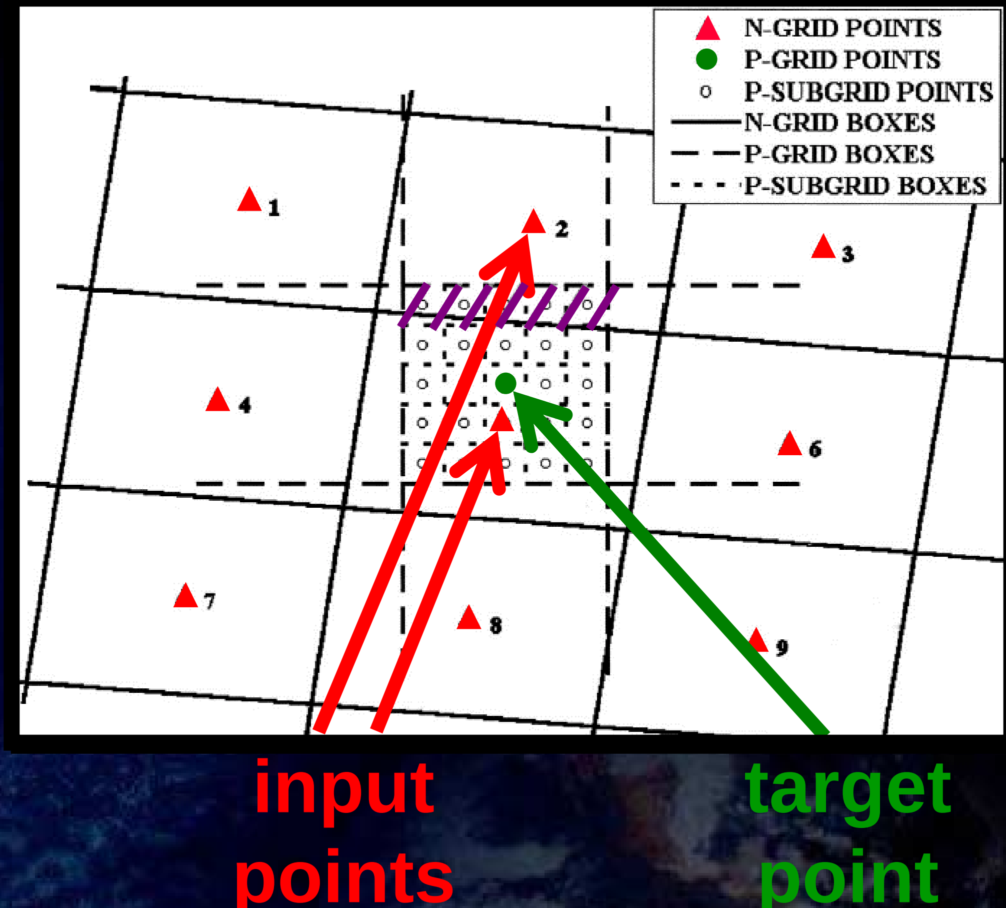
Remap native and coarsened forecast fields to 10-km Stage IV grid

Compare scores from forecasts with different precipitation detail verified against detailed Stage IV data

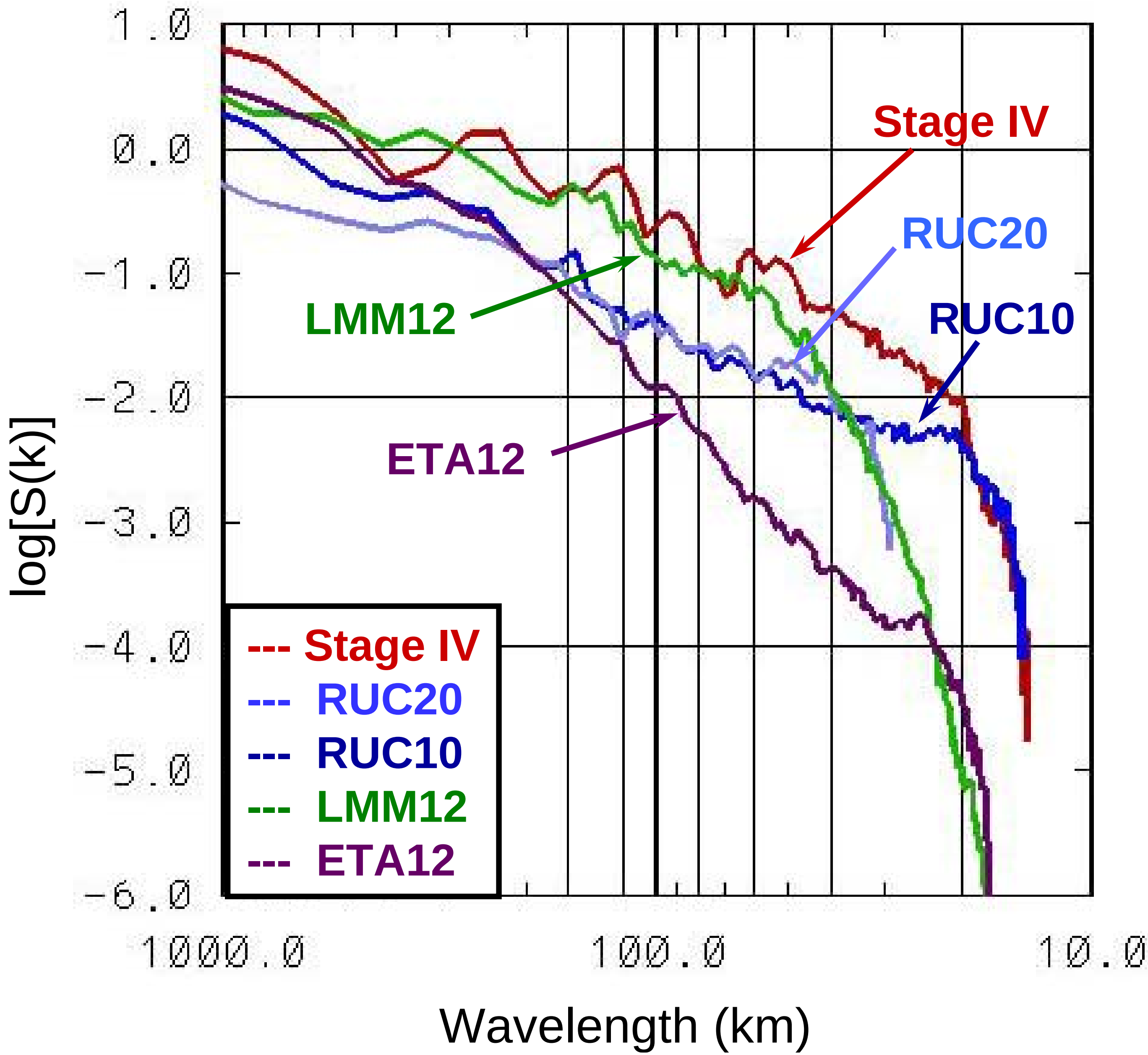


Grid Transformations

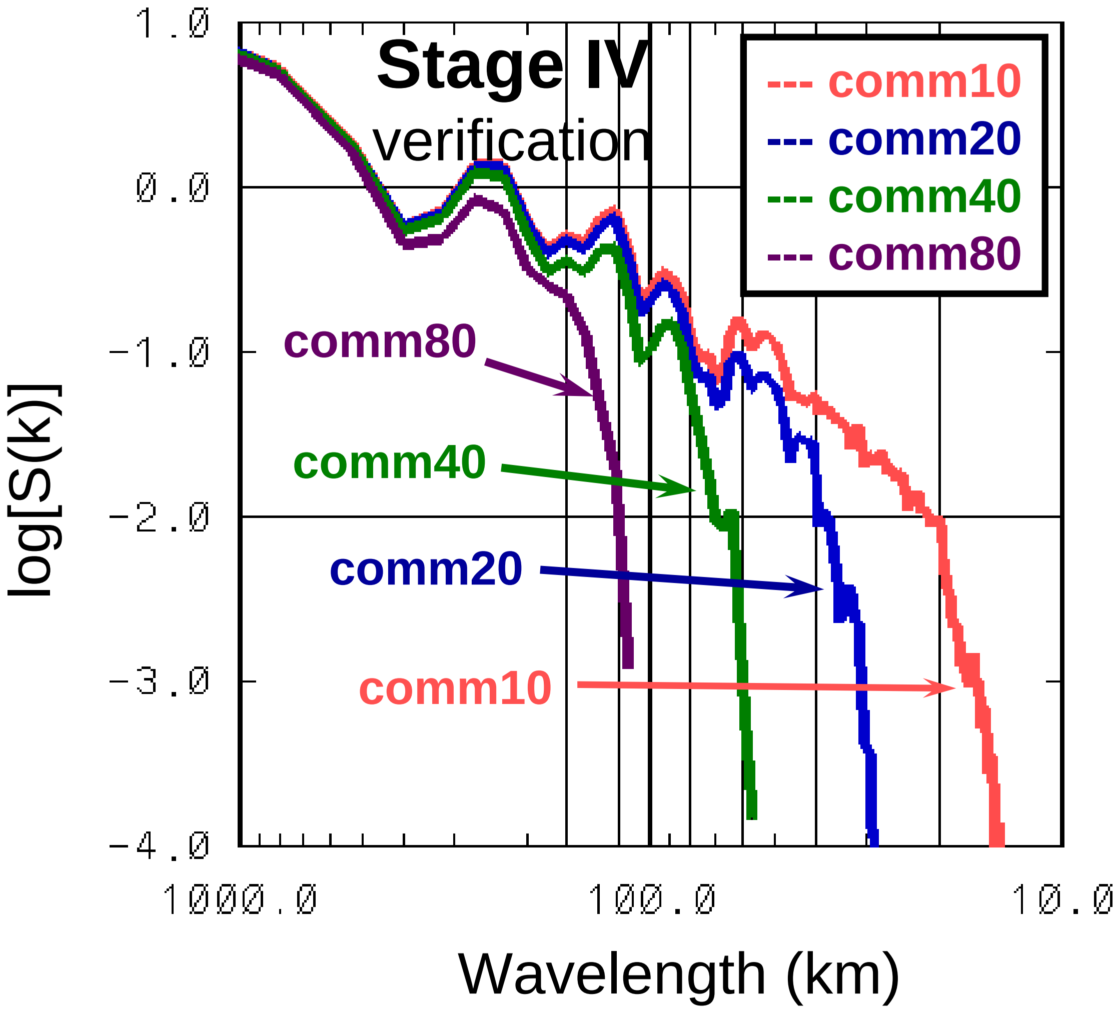
- NCEP "neighbor budget" (Baldwin 2000) used for all grid remappings
- Preserves total precip, minimizes edge smearing
- Less impact on skill scores than bilinear interp (Accadia et al., 2003)



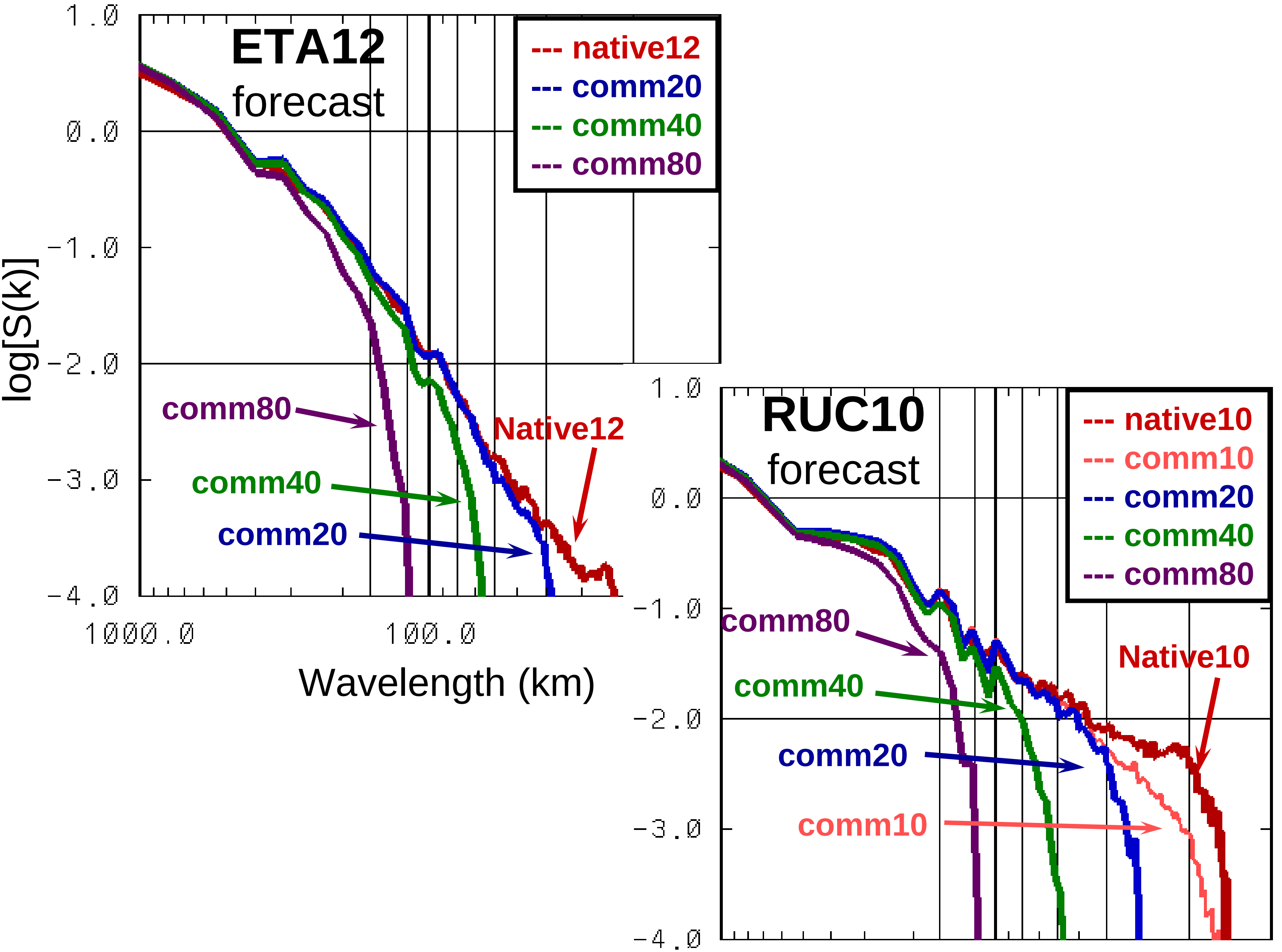
- Sub-divide each target grid-box into 25 sub-boxes (5x5)
- Nearest neighbor from input grid to each sub-box point
- Target values = simple average of 25 sub-box values



1.0
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
log[S(k)]
1000.0
100.0
10.0
Wavelength (km)
Stage IV
RUC20
RUC10
LMM12
ETA12
--- Stage IV
--- RUC20
--- RUC10
--- LMM12
--- ETA12

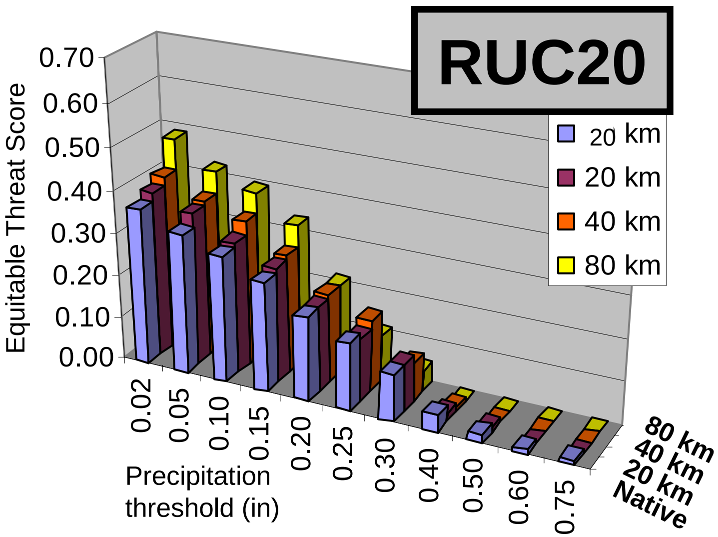


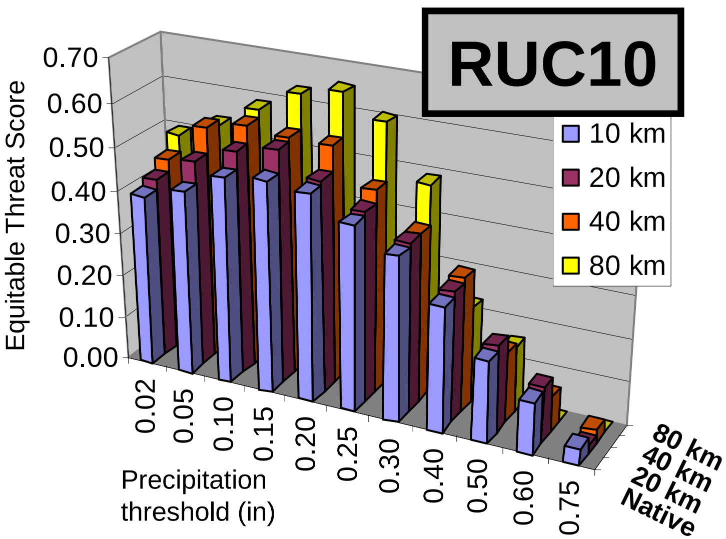
Stage IV
verification
--- comm10
--- comm20
--- comm40
--- comm80
comm80
comm40
comm20
comm10
1.0
0.0
-1.0
-2.0
-3.0
-4.0
1000.0
100.0
10.0
log[S(k)]
Wavelength (km)

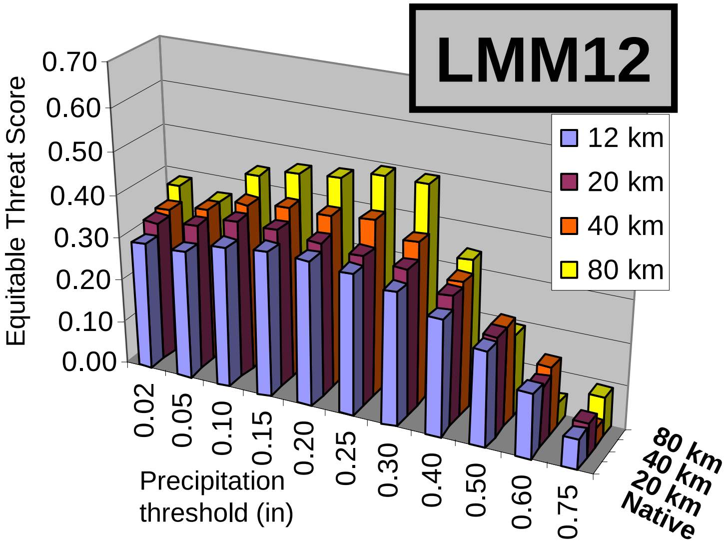


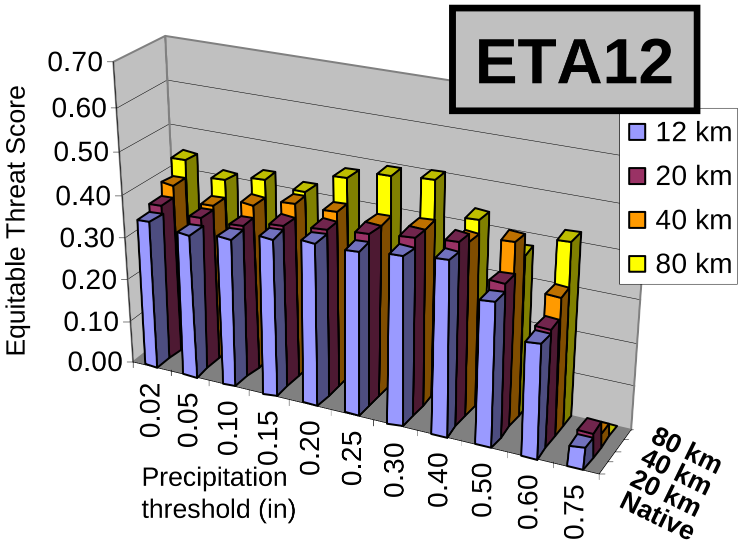
ETA12
forecast
--- native12
--- comm20
--- comm40
--- comm80
1.0
0.0
-1.0
-2.0
-3.0
-4.0
log[S(k)]
comm80
Native12
comm40
comm20
1000.0
100.0
Wavelength (km)
RUC10
forecast
--- native10
--- comm10
--- comm20
--- comm40
--- comm80
1.0
0.0
-1.0
-2.0
-3.0
-4.0
comm80
Native10
comm40
comm20
comm10

Expt. 1 Results: Upscale model and verification

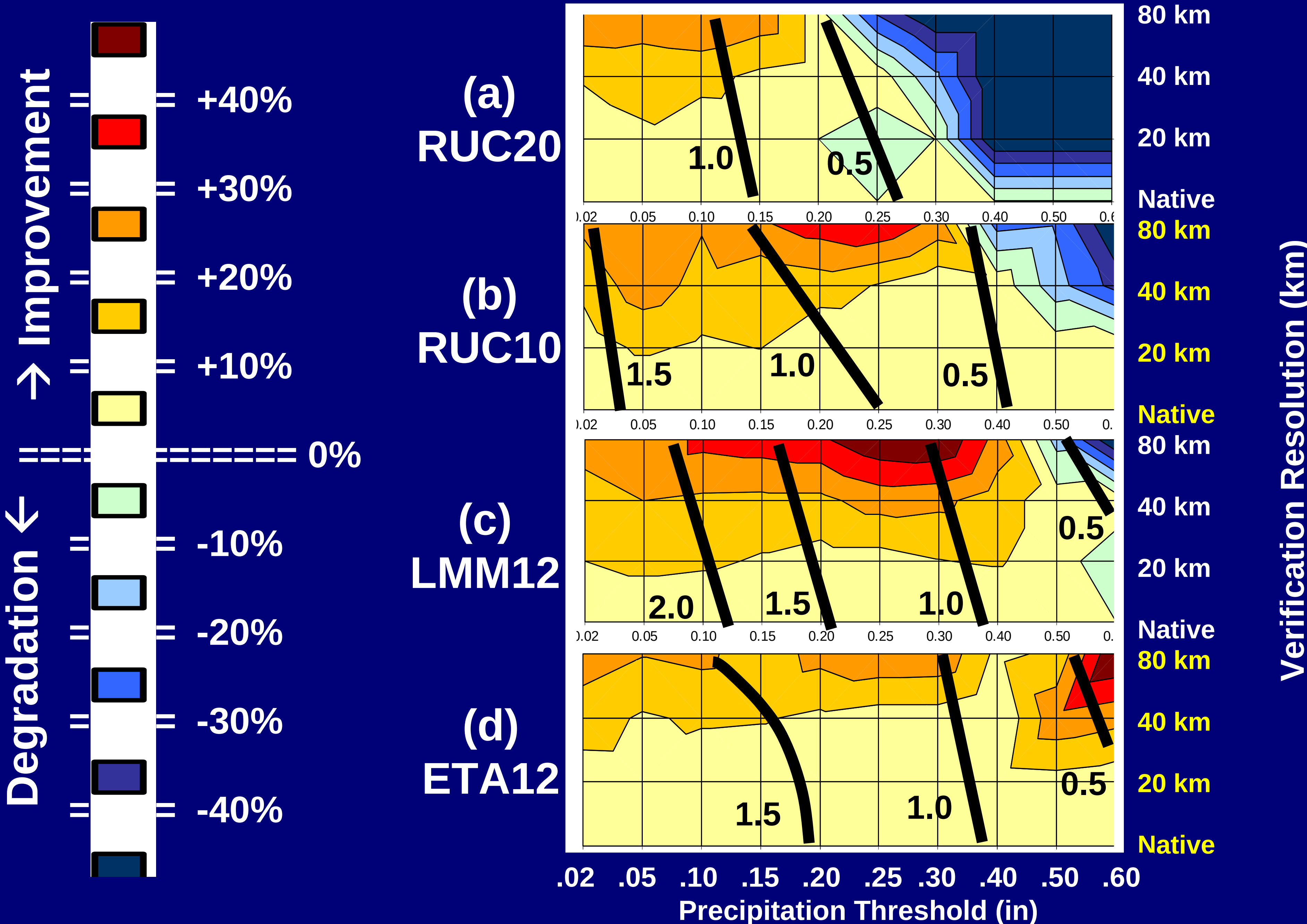




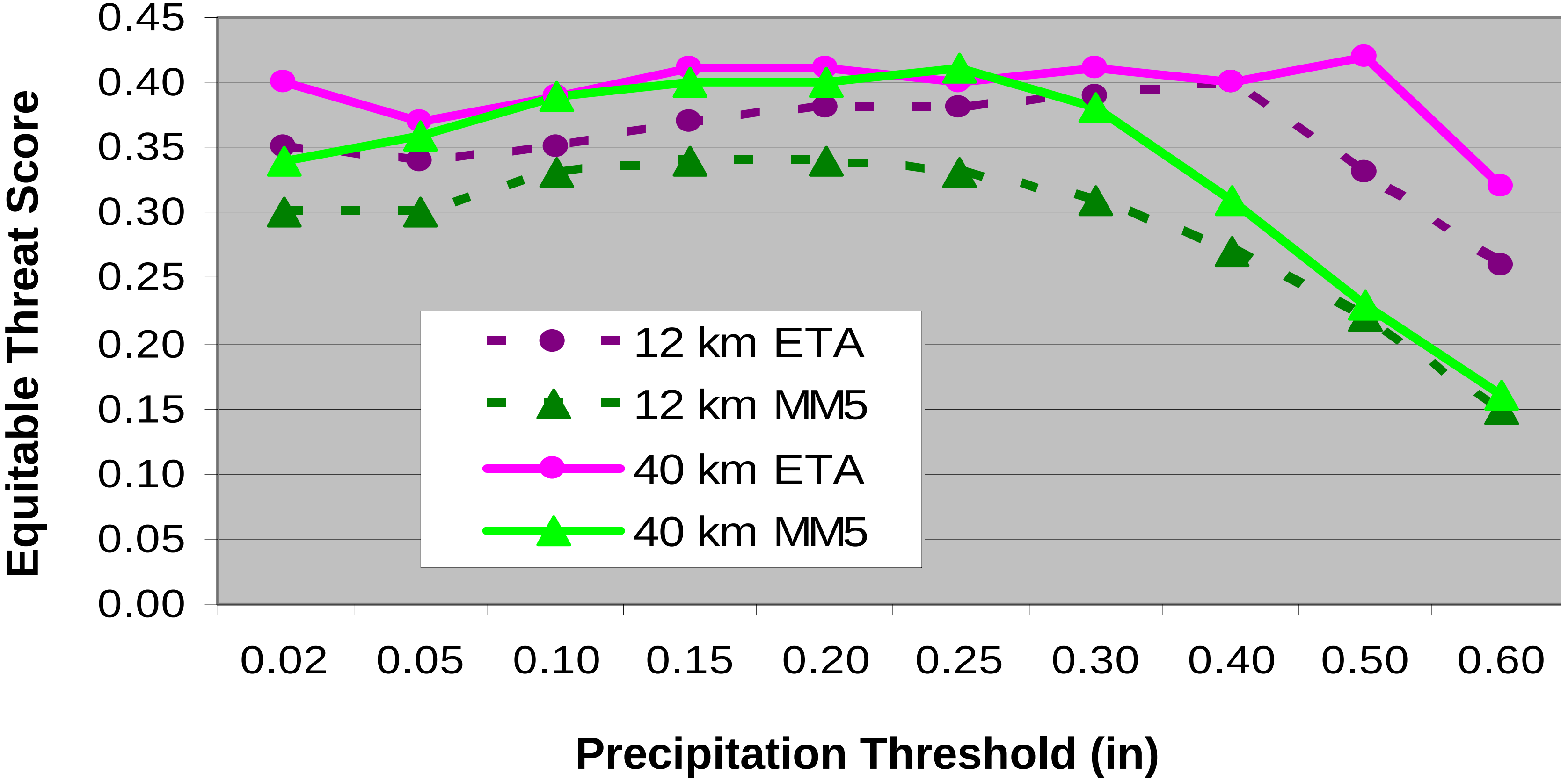




Expt. 1: ETS % change relative to native grid



Expt. 1 Results: Upscale model and verification



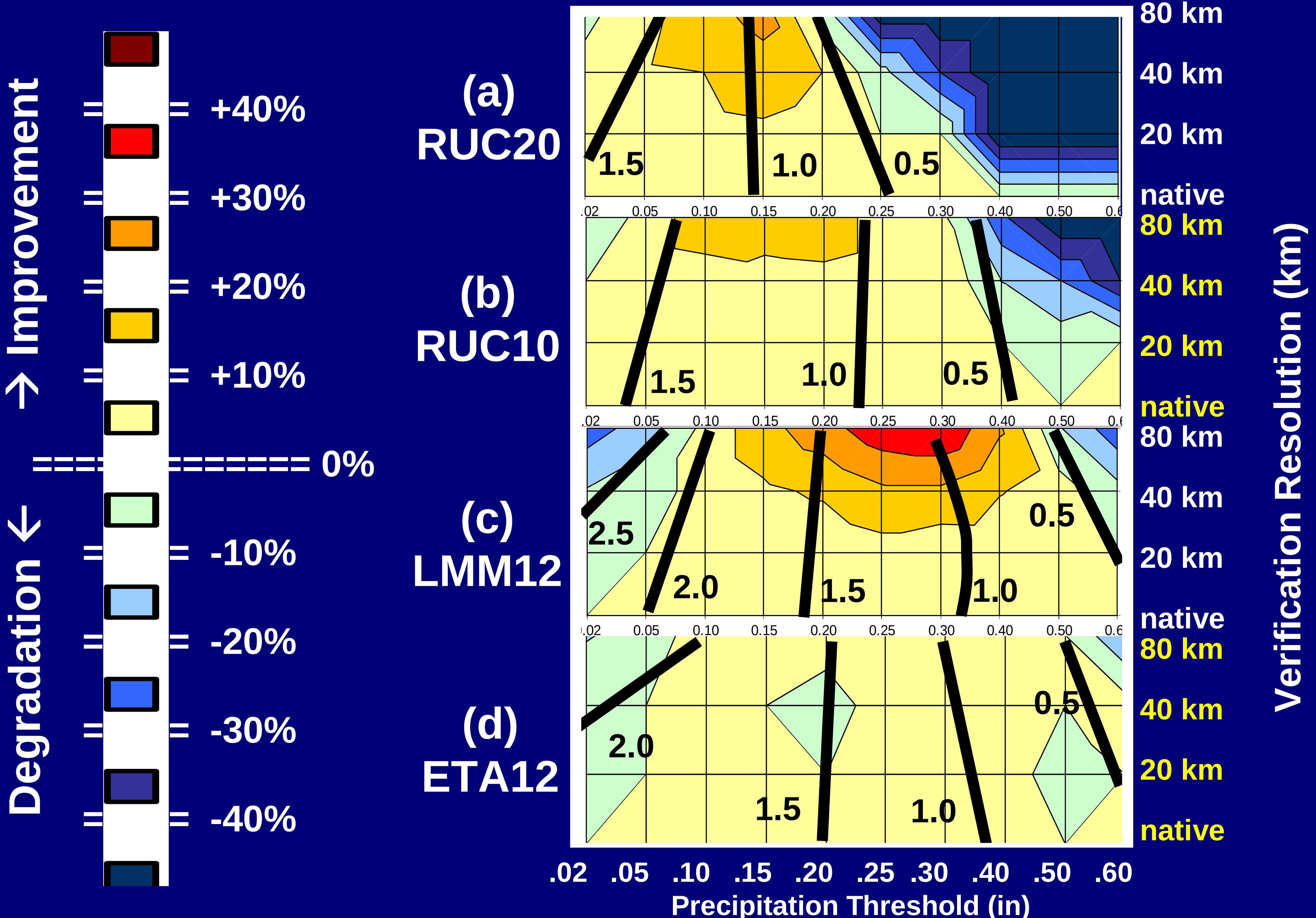
LMM12 (near misses) improves more with upscaling than ETA12 (very smooth)

Summary of Expt. 1 Results

- ETS improves for all models and most thresholds as forecast and verification fields are upscaled
- For detailed forecasts, a precipitation threshold cutoff exists above which forecast degradation occurs with upscaling
- The cutoff threshold shifts to lower amounts with further upscaling, and is correlated with bias ~0.5
- For smooth forecasts, less ETS improvement with upscaling occurs and no cutoff threshold exists

How do these results change, when only the forecast is smoothed?

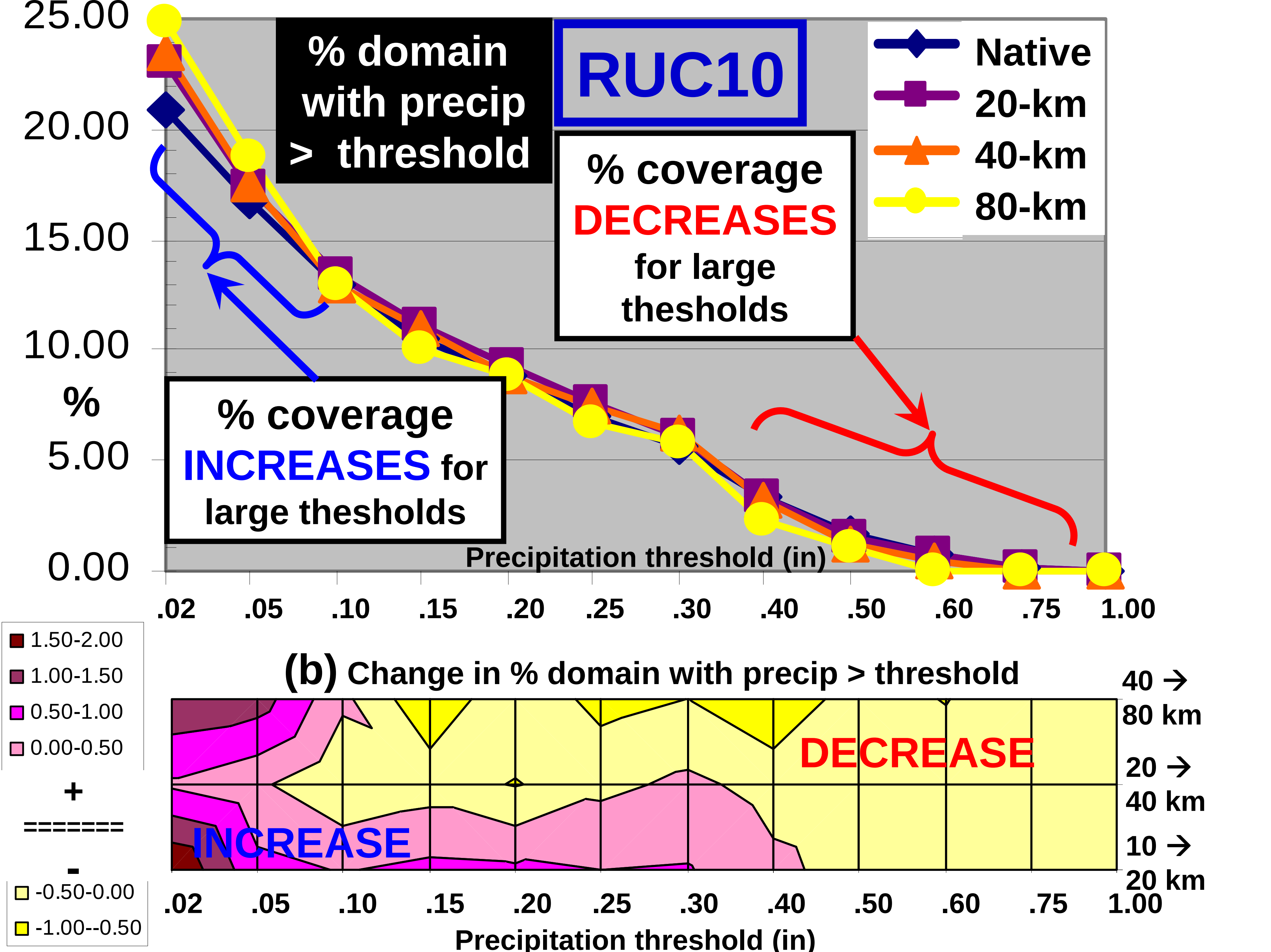
Expt. 2 (Smooth model only): ETS % change



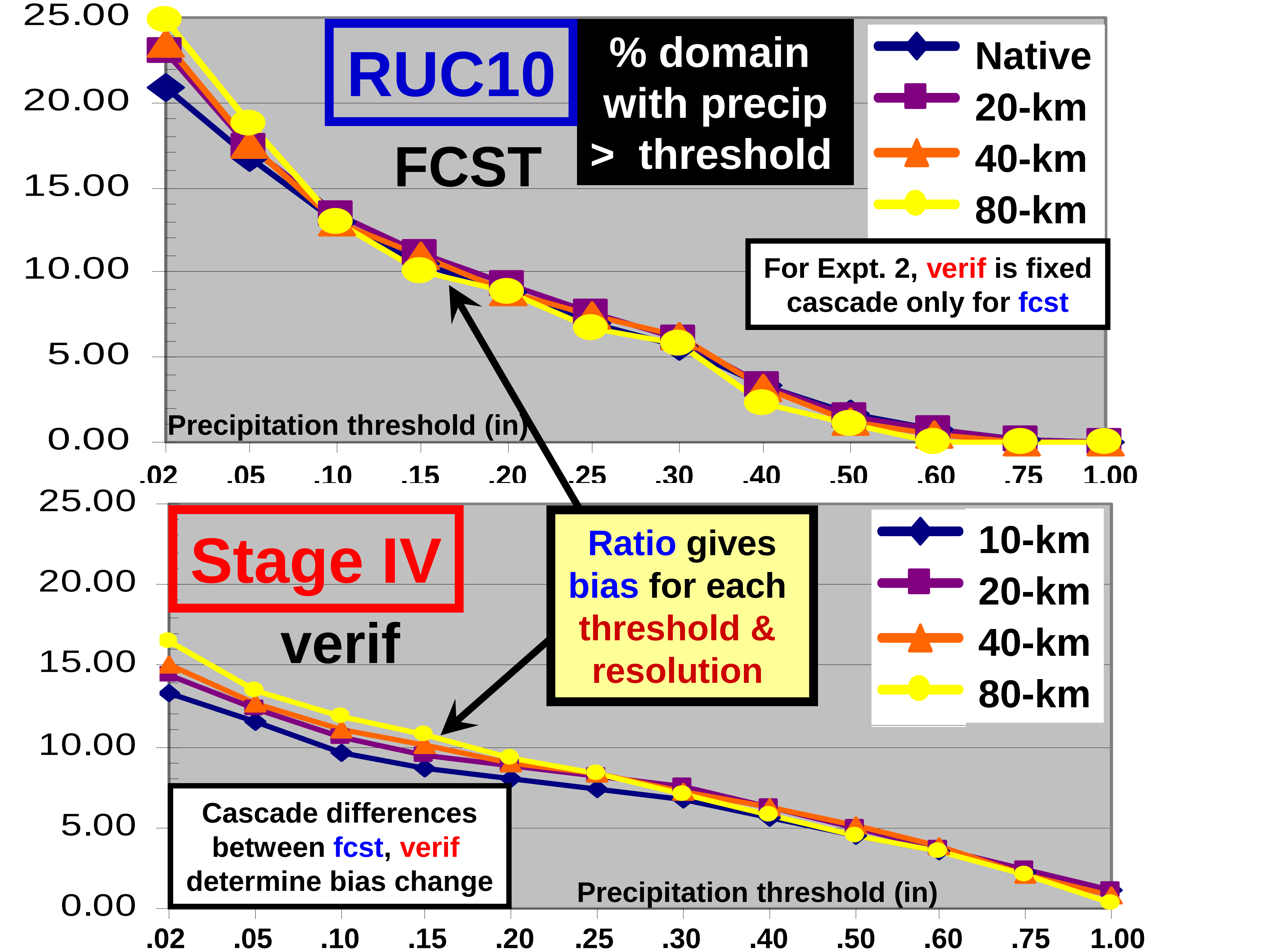
Summary of Expt. 2 Results

- Even when verified against a fixed detailed field, smoothing the forecast improves the ETS score
- Bias decreases for the highest thresholds and increases for the lowest thresholds
- Upper cutoff threshold (bias ~ 0.5) remains, ETS falls for low thresholds as bias exceeds 2.0 for smoothed fields
- For smooth forecasts, very little change in ETS (no changes for either forecast or observations)

For ETS, smoother is better (either forecast or observations), with current model skill*



25.00
20.00
15.00
10.00
5.00
0.00
%
% domain with precip > threshold
RUC10
Native
20-km
40-km
80-km
% coverage DECREASES for large thesholds
% coverage INCREASES for large thesholds
Precipitation threshold (in)
.02 .05 .10 .15 .20 .25 .30 .40 .50 .60 .75 1.00
1.50-2.00
1.00-1.50
0.50-1.00
0.00-0.50
+
=======
-
-0.50-0.00
-1.00--0.50
(b) Change in % domain with precip > threshold
DECREASE
INCREASE
40 → 80 km
20 → 40 km
10 → 20 km
.02 .05 .10 .15 .20 .25 .30 .40 .50 .60 .75 1.00
Precipitation threshold (in)



RUC10
FCST
% domain with precip > threshold
Native
20-km
40-km
80-km
For Expt. 2, verif is fixed cascade only for fcst
Precipitation threshold (in)
25.00
20.00
15.00
10.00
5.00
0.00
.02
.05
.10
.15
.20
.25
.30
.40
.50
.60
.75
1.00
Stage IV
verif
Ratio gives bias for each threshold & resolution
10-km
20-km
40-km
80-km
Cascade differences between fcst, verif determine bias change
Precipitation threshold (in)

What controls ETS and bias changes?

- As forecast and observations are smoothed, local maxima are reduced, and larger precipitation amounts spread to nearby points
- Result is an overall **cascade** of precipitation from higher thresholds to lower thresholds
- **ETS:** Small-scale near misses suddenly become hits!
- **Bias:** Increase in coverage for low thresholds, decrease in coverage for high thresholds

Precipitation cascade is largely controlled by:

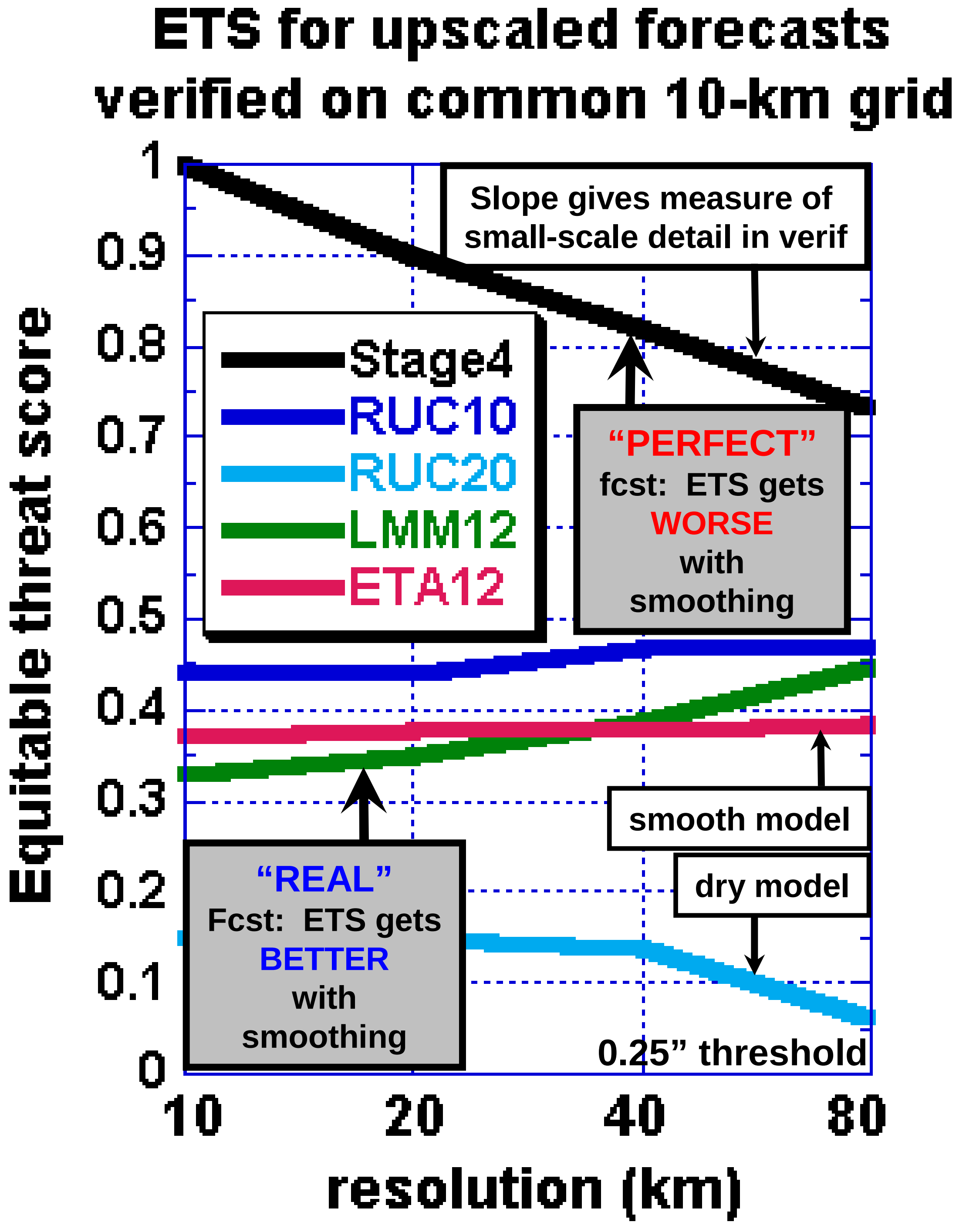
-- Small scale detail (spectra)

-- Total precipitation volume

* What if model skill was better ?

- ETS rewards gridpoint matches
- Details must be in the correct location
- **Models are not that good yet!**

ETS for coarsened "perfect" forecast gives upper-bound on ETS for a given amount of detail



Conclusions

- Forecasts on different native grids are **not** directly comparable (coarser grid has the advantage)
- Forecasts with different degrees of small-scale detail, even if on the same grid, are **not** comparable (smoother field has the advantage)

ETS comparisons should only be made for precipitation fields with similar spectra and bias, compared on matched grid resolutions (using the same verification field)

Better verification measures?

- Spatial structure measures
- Object Oriented measures
- Scale dependent techniques

There is no:

- one-size fits all verification score

- optimal amount of model detail

Highly detailed forecasts often better duplicate observed spatial and temporal structures, contain more information for use in the model post-processing

That’s all folks!