Scale sensitivities in model precipitation skill scores

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

- **Threat Score** = $\frac{\text{Hits}}{\text{Hits} + \text{Misses} + \text{False Alarms}}$

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

- **ETS** is the threat score corrected for a chance forecast...

  $$\text{ETS} = \frac{\text{Hits} - \text{chance}}{\text{Hits} + \text{Misses} + \text{False Alarms} - \text{chance}}$$

- **Bias** = $\frac{\text{Area Forecast}}{\text{Area Observed}}$

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

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>0.157,</td>
<td>0.159,</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.254,</td>
<td>0.309,</td>
</tr>
<tr>
<td>Bias</td>
<td>0.980,</td>
<td>0.980,</td>
</tr>
<tr>
<td>CSI</td>
<td>0.214,</td>
<td>0.161,</td>
</tr>
<tr>
<td>ETS</td>
<td>0.170,</td>
<td>0.102,</td>
</tr>
</tbody>
</table>

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.

\[
\text{ETS}_{10\text{km}} = -0.03 \quad \text{ETS}_{20\text{km}} = 0.20
\]
### IHOP Real-time Modeling

<table>
<thead>
<tr>
<th>Model</th>
<th>Native (km)</th>
<th>Coarsened (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUC10</td>
<td>10</td>
<td>20  40  80</td>
</tr>
<tr>
<td>RUC20</td>
<td>20</td>
<td>20  40  80</td>
</tr>
<tr>
<td>ETA12</td>
<td>12</td>
<td>20  40  80</td>
</tr>
<tr>
<td>LMM12</td>
<td>12</td>
<td>20  40  80</td>
</tr>
<tr>
<td>Stage4 verif</td>
<td>4 (10)</td>
<td>20  40  80</td>
</tr>
</tbody>
</table>

This study is not a model bake off!

- **Experimental**
  - RUC 10-km (GD ensemble convection)
  - LAPS MM5 12-km (KF convection)

- **Operational**
  - RUC 20-km (GD ensemble convection)
  - Eta 12-km (BMJ convection)
Observations: NCEP Stage IV Analysis

Mosaic of regional hourly and 6-hourly multi-sensor (radar+gauges) precipitation analysis at 4km.
Bias = 0.63
ETS = 0.16

Bias = 0.95
ETS = 0.42

Bias = 1.41
ETS = 0.37

Bias = 1.38
ETS = 0.32

Skill scores for 0.25 inches, 6h fcst., 6h accumulation, valid 18z 13 June 2002
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

Upscale forecasts and observations
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

Stage IV

Native

10-km

20-km

40-km

80-km

Model

Native

10-km

20-km

40-km

80-km

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
Stage IV verification

Wavelength (km)

log[\text{S}(k)]

--- comm10
--- comm20
--- comm40
--- comm80

comm80
comm40
comm20
comm10
log[S(k)]

Wavelength (km)

--- native12
--- comm20
--- comm40
--- comm80

--- native10
--- comm10
--- comm20
--- comm40
--- comm80

ETA12 forecast

RUC10 forecast
Expt. 1 Results: Upscale model and verification
Expt. 1: ETS % change relative to native grid

(a) RUC20
(b) RUC10
(c) LMM12
(d) ETA12

Degradation

+40% =
+30% =
+20% =
+10% =
0% ==
-10% =
-20% =
-30% =
-40% =

Improvement

Verification Resolution (km)

Precipitation Threshold (in)

Expt. 1: ETS % change relative to native grid

(a) RUC20
(b) RUC10
(c) LMM12
(d) ETA12

Degradation

+40% =
+30% =
+20% =
+10% =
0% ==
-10% =
-20% =
-30% =
-40% =

Improvement

Verification Resolution (km)

Precipitation Threshold (in)

Expt. 1: ETS % change relative to native grid

(a) RUC20
(b) RUC10
(c) LMM12
(d) ETA12

Degradation

+40% =
+30% =
+20% =
+10% =
0% ==
-10% =
-20% =
-30% =
-40% =

Improvement

Verification Resolution (km)

Precipitation Threshold (in)
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

- **RUC20**
  - Expt. 2 (Smooth model only): ETS % change
  - Improvements:
    - +40%
    - +30%
    - +20%
    - +10%
  - Degradations:
    - -10%
    - -20%
    - -30%
    - -40%

- **RUC10**
  - Expt. 2 (Smooth model only): ETS % change
  - Improvements:
    - +40%
    - +30%
    - +20%
    - +10%
  - Degradations:
    - -10%
    - -20%
    - -30%
    - -40%

- **LMM12**
  - Expt. 2 (Smooth model only): ETS % change
  - Improvements:
    - +40%
    - +30%
    - +20%
    - +10%
  - Degradations:
    - -10%
    - -20%
    - -30%
    - -40%

- **ETA12**
  - Expt. 2 (Smooth model only): ETS % change
  - Improvements:
    - +40%
    - +30%
    - +20%
    - +10%
  - Degradations:
    - -10%
    - -20%
    - -30%
    - -40%
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*
Native % coverage DECREASES for large thresholds.

Native % coverage INCREASES for large thresholds.

(b) Change in % domain with precip > threshold

- DECREASE
- INCREASE
For Expt. 2, verif is fixed cascade only for fcst.

Cascade differences between fcst, verif determine bias change.
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

---

Slope gives measure of small-scale detail in verif

**“PERFECT”**
Fcst: ETS gets WORSE with smoothing

**“REAL”**
Fcst: ETS gets BETTER with smoothing

0.25” threshold

---

Dry model

Smooth model
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!