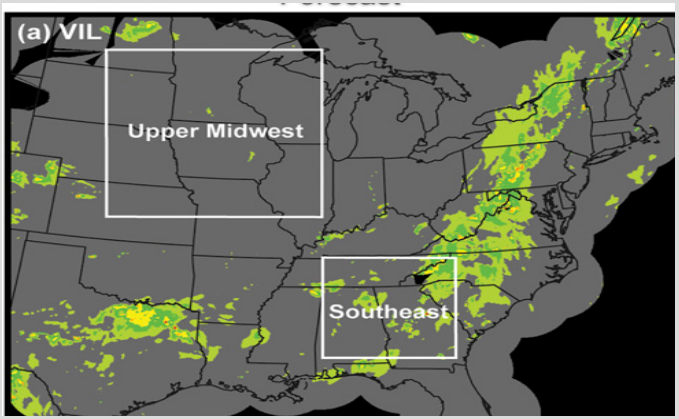
Verification of WRF Simulations

Ming Chen

National Center for Atmospheric Research

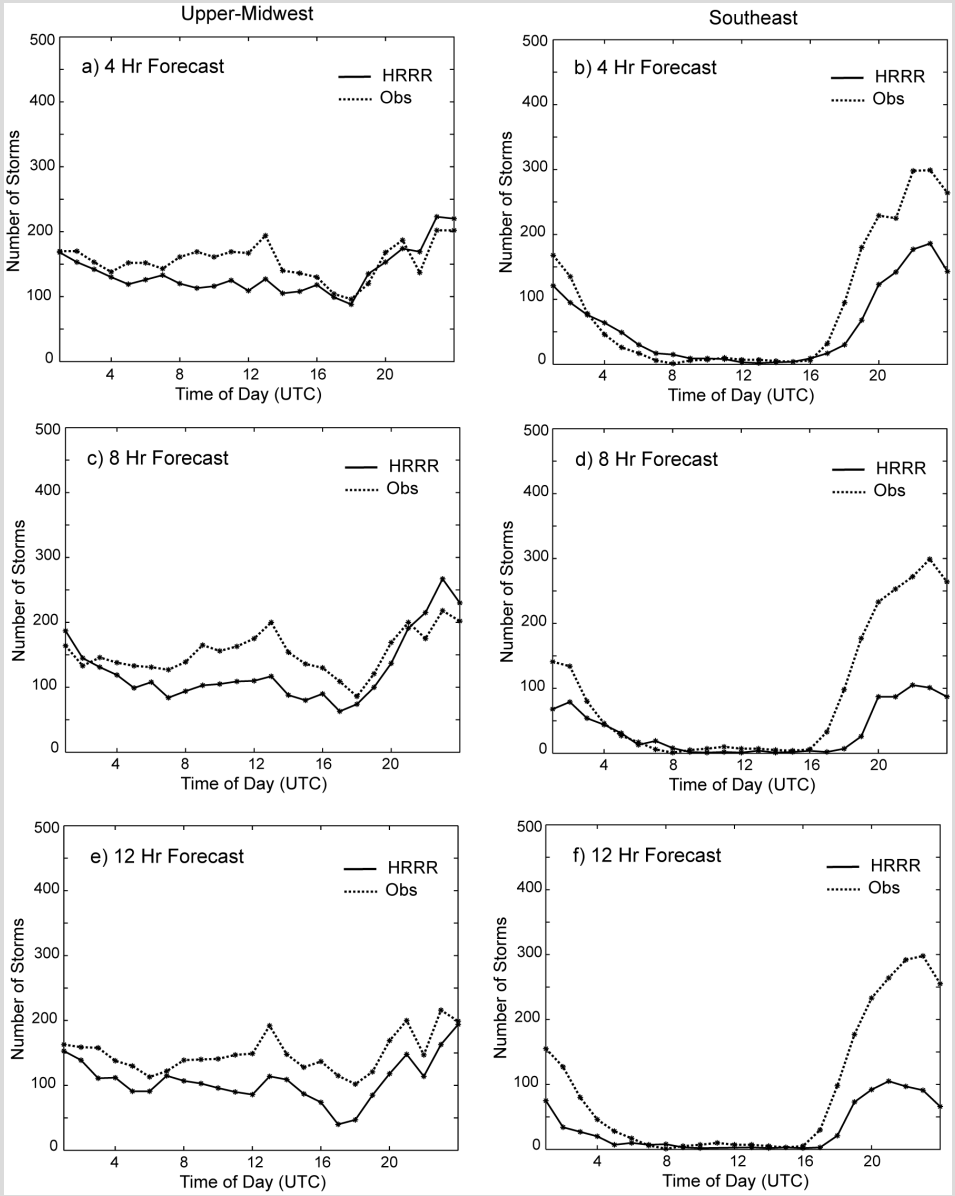
Verification of WRF Simulations

- Operational Forecasting
  - We need to monitor forecast quality – how accurate are the forecasts?
- Research
  - Compare the performance of different schemes/ scheme combinations
  - To what extent does one scheme or one set of scheme combination give better simulation than another, and in what ways is that scheme better?
- Evaluation of WRF performance
  - Help users identify model weaknesses, strengths --- important for further improvement
  - We need to know what is wrong before we can improve



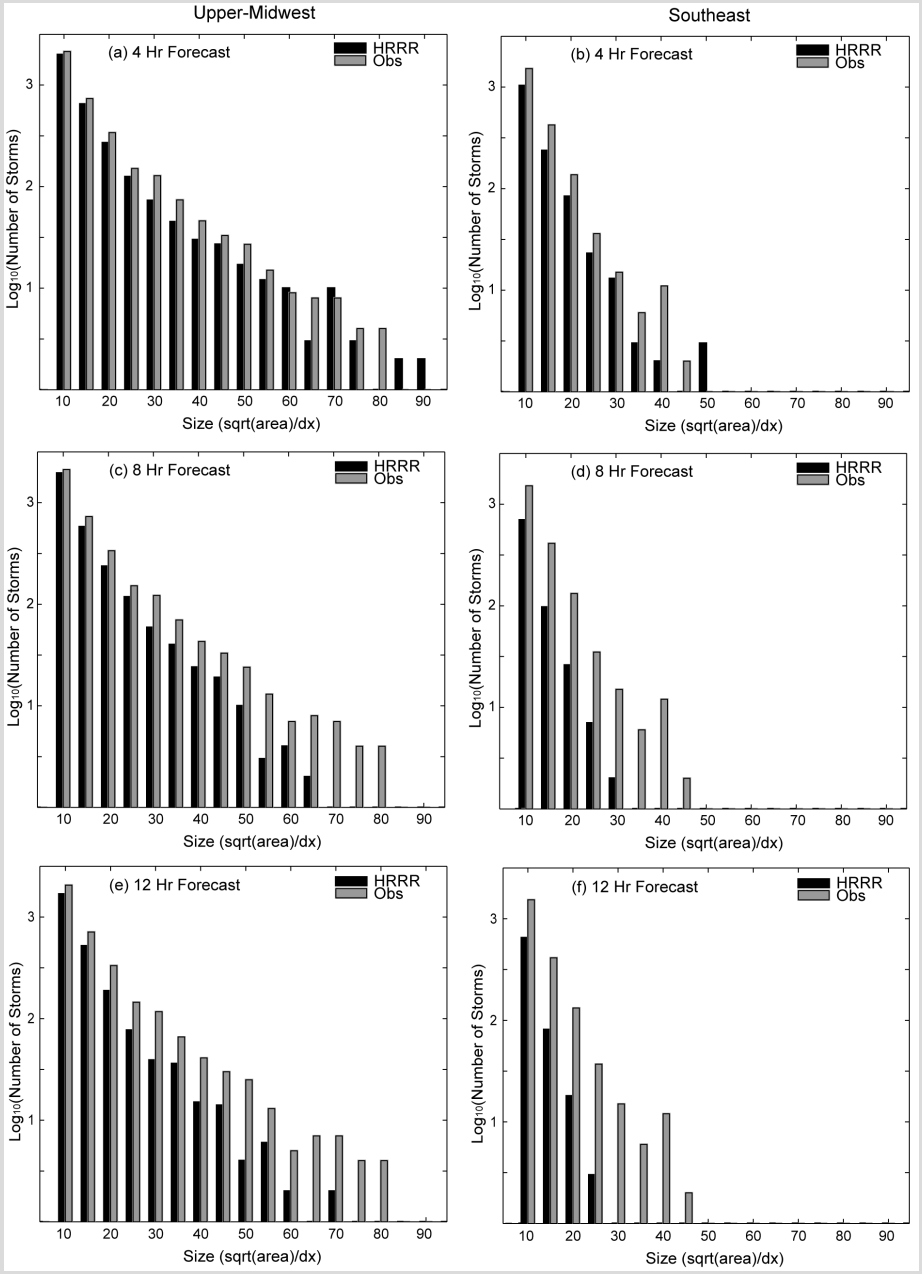
Model domains for convective storm forecasts by the High Resolution Rapid Refresh (HRRR) in the summer of 2010

(HRRR is a WRF- ARW based forecast system. See Weygandt et al. 2009; Benjamin et al.2011)



total number of storms as a function of time of the day

(Cai et al., 2015)



total number of storms as a function of the storm size

(Cai et al., 2015)

What can we find based on the verification?

- The diurnal variation of the total number of storms in the Southeast is stronger than that of the upper Midwest --- different forcing mechanisms are responsible for the storm initiation and evolution in these two subdomains.
- All forecasts for the upper Midwest showed almost simultaneous increases in the total number of storms compared to the observations starting at 1800 UTC --- fairly good timing of storm initiation
- All HRRR forecasts for the Southeast exhibited a significant delay or lack of new storms starting at 1700 UTC --- fewer new storms initialized in the model
- For longer forecast lead times the model tended to have fewer large storms compared with the observations in both the Midwest and the southwest --- large storms were not realistically maintained in the model

(Cai et al., 2015)

Verification of WRF Simulations

- Introduce methods for WRF simulation verification. The methods range from traditional statistics to methods for more detailed verification
- Give examples for each method
- Provide links and references for further information
- Does not provide source codes (details can be found in Model Evaluation Tools http://www.dtcenter.org/met/users/)

Recommendations on the Verification of WRF Simulations

- Types of forecast variable
  - Continuous
    - Temperature,
    - Precipitation
    - Winds, humidity, etc.
  - Categorical:
    - Rain vs no rain;
    - Strong winds vs no strong winds;
    - Fog vs no fog; clouds vs no clouds, etc.

Recommendations on the Verification of WRF Simulations – Continuous Variables

- **Mean Error (Bias):** a simplest and most familiar score to provide average direction of error

$$\text{Mean Error} = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right) = \bar{f} - \bar{o}$$

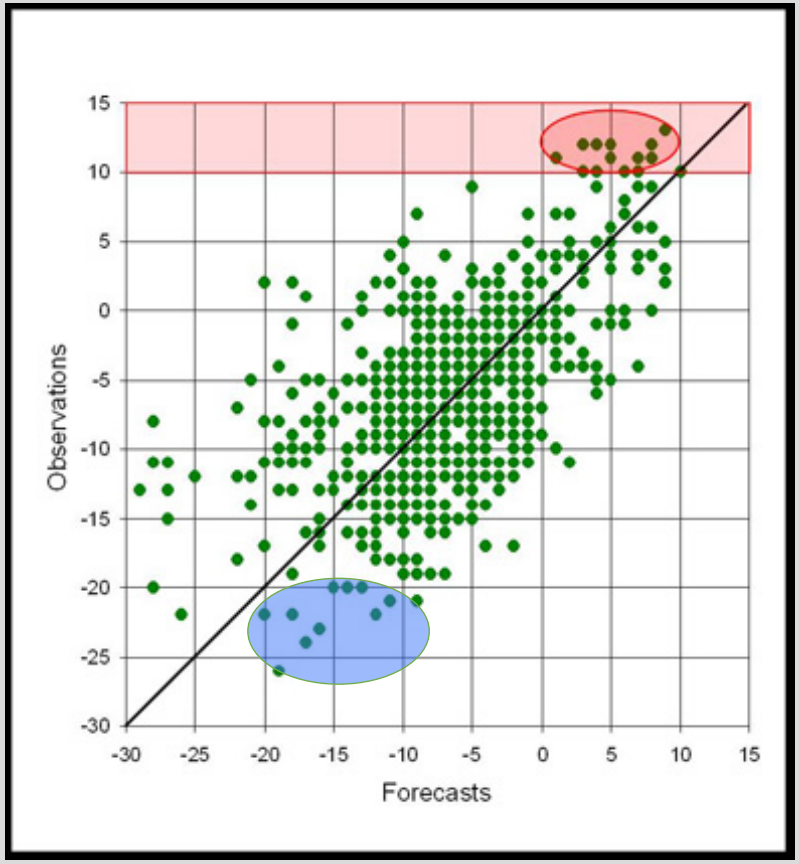
- **MAE:** average of the magnitude of errors (always view the ME and MAE simultaneously)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - o_i\right|$$

- **MSE (RMSE):** sensitive to large errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right)^2 \qquad RMSE = \sqrt{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right)^2$$

Verification of Continuous Variables: Scatterplot



All points with observed temperatures above the diagonal mean they are forecast too cold.

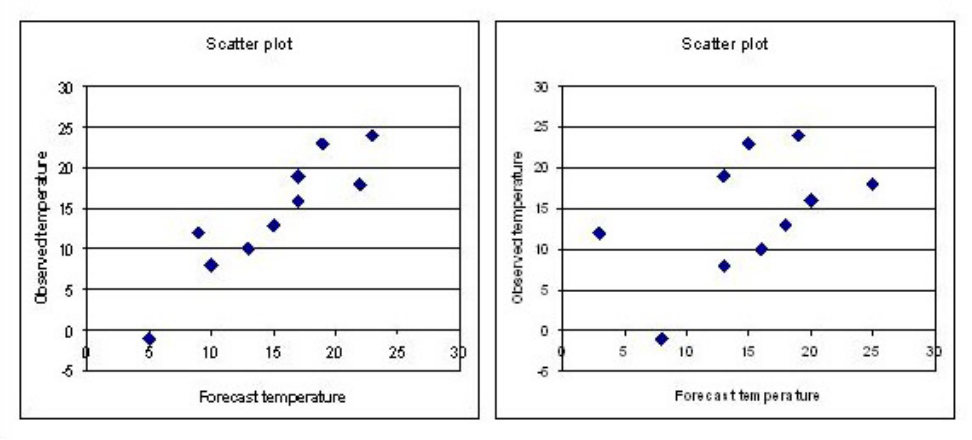
All the forecast is too cold for T above +10?

All the observed T below -20 are forecast too warm except one

(http://www.eumetcal.org/)

Verification of Continuous Variables: Scatterplot

Below are two scatter plots representing two different sets of forecasts. The observations are the same in both cases. Can we say that these two sets of forecasts is positively correlated with the observations?



(http://www.eumetcal.org/)

Verification of Continuous Variables

- Model-generated vertical profiles of variables
  - Profiles of meteorological variables can be extracted from the WRF output files and placed on the desired location and time
    - use a sounding from the nearest grid point (i.e. no interpolation) to the desired location,
    - or use bilinear /inverse distance weight interpolation to horizontally interpolate WRF to the desired location
  - General rule for vertical interpolation: the pressure level intervals shouldn't be too large; for the vertical height levels, the layers can be very thin for close examination and allowed to be be thicker for regions of less detailed study
  - sources of comparison data may come from, for example, radar profilers and lidar for wind, microwave radiometers for temperature and moisture, and radio acoustic sounding systems for virtual temperature. Nevertheless radiosondes have remained the primary source of comparison data above the near surface layer

Sources of Observation Data

- Soundings

  http://www.weather.uwyo.edu/upperair/sounding.html

  This site contains WMO soundings in several formats

  http://www.esrl.noaa.gov/raobs)

  This site provides WMO sounding data, but requires different processing in the input function

- Verifications

  NCEP (http://www.emc.ncep.noaa.gov/gmb/STATS_vsdb/),

  ECMWF (http://www.ecmwf.int/en/forecasts/charts/medium/monthly-wmo-scores-against-radiosondes)

  Worldwide comparisons are available for deterministic forecasts at

  http://apps.ecmwf.int/wmolcdnv/

  and for ensemble forecasts at the Japan Meteorological Agency (JMA)

  (http:/epsv.kishou.go.jp/EPSv/).

72469 DNR Denver Observations at 12Z 04 May 2016

PRES hPa	HGHT m	TEMP C	DWPT C	RELH %	MIXR g/kg	DRCT deg	SKNT knot	THTA K	THTE K	THTV K
1000.0	160									
925.0	832									
850.0	1549									
844.0	1625	6.4	-0.6	61	4.36	175	5	293.4	306.4	294.2
842.0	1644	7.6	-1.4	53	4.12	180	6	294.9	307.3	295.6
834.0	1722	9.2	-2.8	43	3.75	199	10	297.4	308.8	298.1
824.0	1822	12.4	-2.6	35	3.85	223	15	301.8	313.8	302.5
823.3	1829	12.4	-2.7	35	3.84	225	15	301.9	313.8	302.6
802.0	2046	12.4	-4.6	30	3.40	204	6	304.1	314.9	304.8
793.5	2134	11.6	-4.3	33	3.52	195	3	304.2	315.3	304.8
765.0	2435	8.8	-3.2	43	3.97	210	4	304.4	316.8	305.1
764.8	2438	8.8	-3.2	43	3.96	210	4	304.4	316.8	305.1
749.0	2608	7.8	-5.2	39	3.48	268	3	305.1	316.2	305.8
736.7	2743	6.7	-4.8	44	3.65	315	2	305.4	316.9	306.1
724.0	2884	5.6	-4.4	49	3.83	313	4	305.7	317.8	306.4
708.0	3064	4.8	-8.2	38	2.92	311	7	306.8	316.2	307.3
700.0	3156	4.0	-9.0	38	2.78	310	9	306.9	315.8	307.4
678.0	3415	1.8	-10.2	41	2.61	305	12	307.2	315.7	307.7
657.8	3658	0.7	-13.6	33	2.05	300	15	308.7	315.5	309.1
643.0	3841	-0.1	-16.1	29	1.70	309	18	309.8	315.5	310.1
633.3	3962	-0.8	-20.5	21	1.19	315	20	310.4	314.4	310.6
620.0	4132	-1.7	-26.7	13	0.70	321	16	311.2	313.7	311.3
609.4	4267	-2.8	-27.0	14	0.69	325	13	311.4	313.9	311.6
586.0	4572	-5.3	-27.7	15	0.67	335	14	312.0	314.4	312.1
563.6	4877	-7.9	-28.5	17	0.65	330	16	312.5	314.8	312.6
533.0	5313	-11.5	-29.5	21	0.63	316	15	313.2	315.4	313.3
521.1	5486	-12.6	-34.1	15	0.41	310	14	313.9	315.4	313.9
516.0	5561	-13.1	-36.1	13	0.34	312	14	314.2	315.4	314.2
500.0	5800	-15.3	-35.3	16	0.38	320	12	314.3	315.7	314.4
480.5	6096	-17.9	-35.6	20	0.38	325	12	314.8	316.2	314.8
457.0	6470	-21.1	-36.1	25	0.39	329	14	315.2	316.7	315.3
438.0	6782	-22.5	-40.5	18	0.26	332	16	317.3	318.3	317.4
424.5	7010	-24.6	-41.1	20	0.25	335	17	317.5	318.5	317.6
411.0	7245	-26.7	-41.7	23	0.24	327	16	317.7	318.7	317.8
400.0	7440	-27.5	-45.5	16	0.16	320	16	319.2	319.8	319.2
392.0	7585	-28.5	-48.5	13	0.12	312	17	319.7	320.2	319.7
390.0	7620	-28.8	-48.7	13	0.12	310	17	319.8	320.3	319.8
357.4	8230	-33.7	-52.0	14	0.09	295	15	321.3	321.7	321.4
327.5	8839	-38.5	-55.2	16	0.07	310	12	322.8	323.0	322.8
315.0	9110	-40.7	-56.7	16	0.06	292	13	323.4	323.6	323.4
313.4	9144	-40.9	-56.9	16	0.06	290	13	323.5	323.7	323.5
300.0	9440	-42.9	-58.9	16	0.05	280	15	324.8	325.0	324.8
299.6	9449	-43.0	-59.0	16	0.04	280	15	324.8	325.0	324.8
281.0	9880	-45.5	-61.5	15	0.03	278	15	327.2	327.3	327.2
250.0	10650	-52.1	-71.1	8	0.01	275	14	328.5	328.5	328.5
242.0	10860	-53.7	-72.7	8	0.01	277	15	329.1	329.2	329.1
226.6	11278	-57.5	-71.1	16	0.01	280	18	329.6	329.6	329.6
225.0	11324	-57.9	-70.9	17	0.01	279	18	329.6	329.7	329.6
207.0	11847	-60.9	-70.9	26	0.01	266	15	332.9	332.9	332.9
205.7	11887	-61.3	-71.3	25	0.01	265	15	332.9	333.0	332.9
200.0	12060	-62.9	-72.9	25	0.01	270	19	333.0	333.1	333.0
189.0	12405	-65.3	-74.3	28	0.01	285	14	334.6	334.6	334.6
186.2	12497	-65.5	-74.6	27	0.01	285	14	335.8	335.8	335.8

Sources of Observation Data

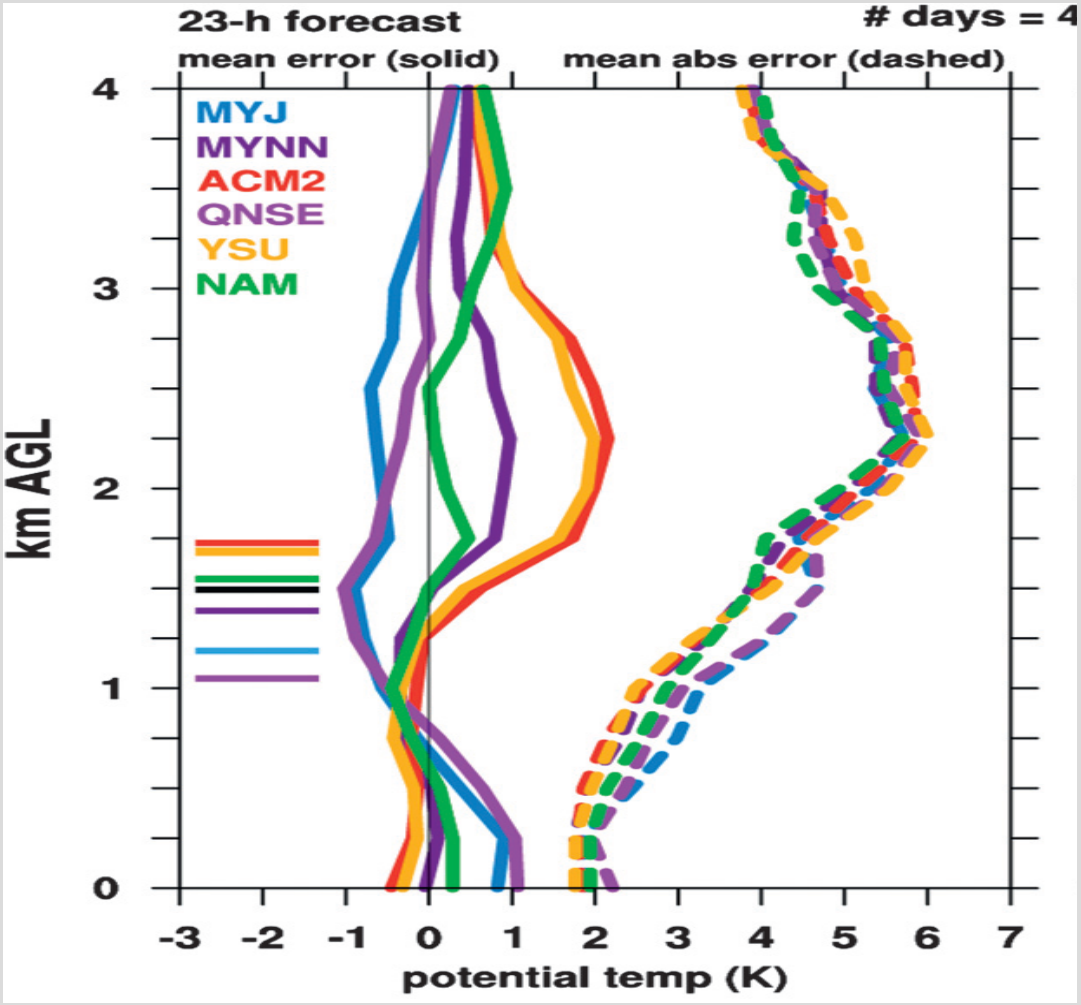
- Station Observations: GDAS prebufr format data

  NCEP FTP Site: ftp://ftpprd.ncep.noaa.gov/pub/data/nccf/com/gfs/prod

  BUFRLIB User Guide: http://www.nco.ncep.noaa.gov/sib/decoders/BUFRLIB/)

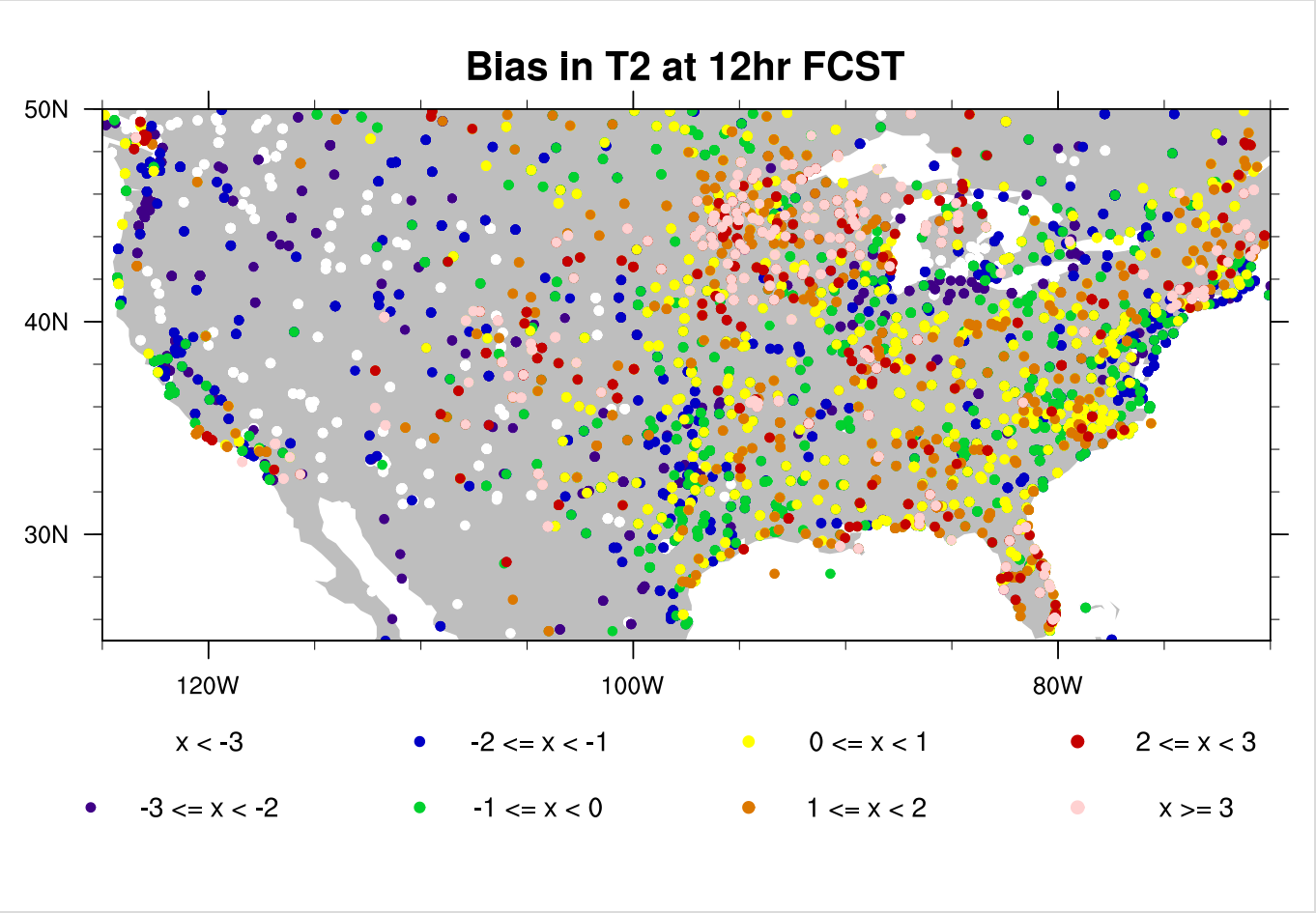
  - UPPER-AIR
  - AIRCRAFT REPORTS
  - SATELLITE-DERIVED WIND REPORTS
  - WIND PROFILER AND ACOUSTIC SOUNDER (SODAR) REPORTS
  - SURFACE LAND (SYNOPTIC, METAR) AND SURFACE MARINE (SHIP, BUOY, C-MAN PLATFORM) REPORTS

Example: vertical profile verification against radiosondes



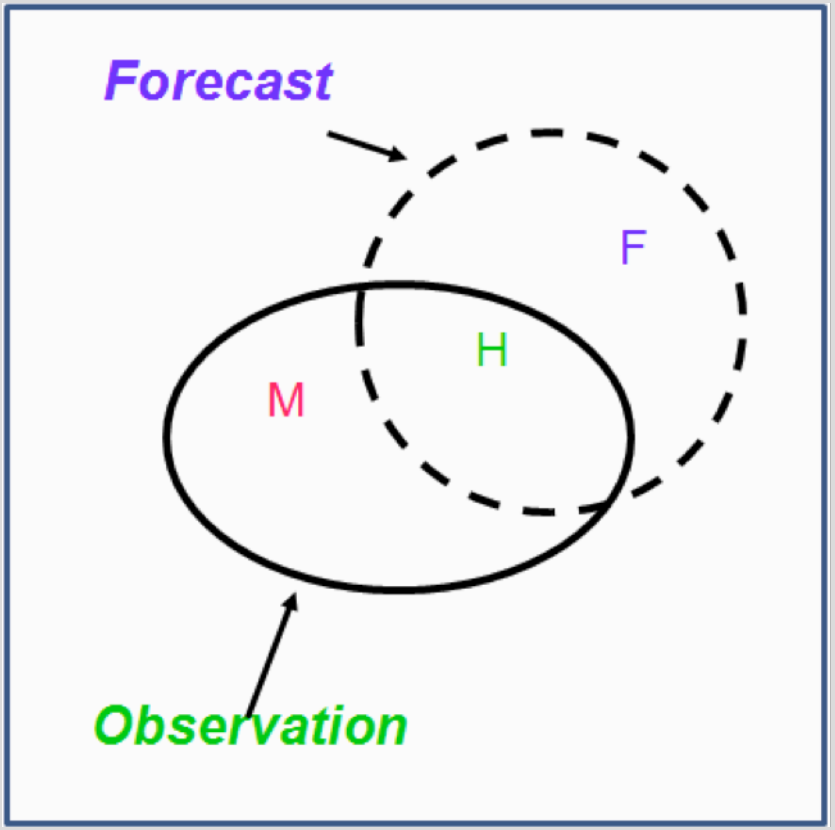
Forecasts at evening time --- (Coniglio et al., 2013, Weather and Forecasting)

Example : verification against station observations



Verification of WRF Simulations – Categorical Variables

- Contingency table
- Several commonly used measures:
  - Accuracy
  - Frequency bias
  - Probability of detection
  - False alarm rate
  - Critical success index (Threat Score)
  - Gilbert Skill Score (ETS)
  - Heidke Skill Score



H: Hit M: Missed F: False Alarm

(NSSL 2012 Spring Forecast Experiment)

Verification of WRF Simulations – Categorical Variables

Contingency table in terms of counts: precipitation

Forecast	Observation		Total
	Yes	No	
Yes	Hits (YY)	False Alarm (YN)	**YY+YN**
No	Misses (NY)	Correct (NN)	**NY+NN**
total	YY+NY	YN+NN	**T=YY+YN+NY+NN**

Categorical Variables

Forecast	Observation		Total
	Yes	No	
Yes	Hits (YY)	False Alarm (YN)	YY+YN
No	Misses (NY)	Correct (NN)	NY+NN
total	YY+NY	YN+NN	T=YY+YN+NY+NN

Accuracy= (YY+NN)/(YY+YN+NY+NN)
what fraction of the forecasts were correct
Range: 0 to 1. Perfect score: 1

Threat Score (Critical Success Index)
TS=YY/(YY+NY+YN)
How well did the forecast "yes" events correspond to the observed "yes" events
Range: 0-1, 0 indicates no skill, 1 represents perfect score

Equitable Threat Score (Gilbert Skill Score)
$ETS=(YY - YY_{random})/(YY + NY + YN - YY_{random})$
How well did the forecast "yes" events correspond to the observed "yes" events (accounting for hits that would be expected by chance
Range: -1/3 – 1, 0 indicates no skill, 1 is perfect score
Where
$YY_{random}=(YY+YN)*(YY+NY)/(YY + YN + NY + NN)$
It is the number of hits for random forecasts

Bias (Or frequency Bias):
Bias=(YY+YN)/(YY+NY)
How similar were the frequencies of Yes forecasts and Yes observations? **Range:** 0 to infinity. **Perfect score:** 1
When Bias is greater than 1, the event is overforecast; less than 1, underforecast

Verification of WRF Simulations – Categorical Variables

Forecast	Observation		Total
	Yes	No	
Yes	Hits (YY)	False Alarm (YN)	YY+YN
No	Misses (NY)	Correct (NN)	NY+NN
total	YY+NY	YN+NN	T=YY+YN+NY+NN

Probability of Detection (Hit Rate):
POD=YY/(YY+NY) (hits/(hits+misses))

False Alarm Ratio:
FAR=YN/(YY+YN) (False Alarm/(Hits+False Alarm))

False Alarm Rate (Probability of False Detection):
PODF=YN/(YN + NN) (False Alarm/(False Alarm+Correct))

Recommendations on the Verification of WRF Simulations –Categorical Variables

Example: daily rain forecasts and observations over 1-year period

Forecast	Observation		Total
	Yes	No	
Yes	82	38	**120**
No	23	222	**245**
total	105	260	**365**

(WCRP 2015)

Example:

Accuracy = (82+222)/365 = 0.83

Bias=(82+38)/(82+23)=1.14

POD=82/(82+23)=0.78

FAR=38/(82+38)=0.32

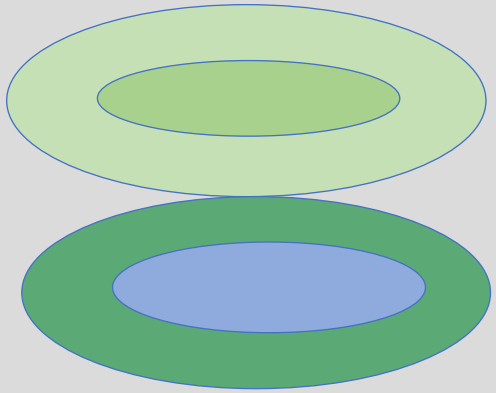
TS=82/(82+23+38)=0.57

ETS=(82-34)/(82+23+38-34)=0.44

Forecast	Observation		Total
	Yes	No	
Yes	82	38	120
No	23	222	245
total	105	260	365

Verification of WRF Simulations – Categorical Variables

- Problems in traditional statistical measures -- scale-dependent
  - Warm season precipitation has significant small-scale variability
  - High-resolution models are becoming practical
  - Traditional scores are worse for detailed forecast --- double penalty

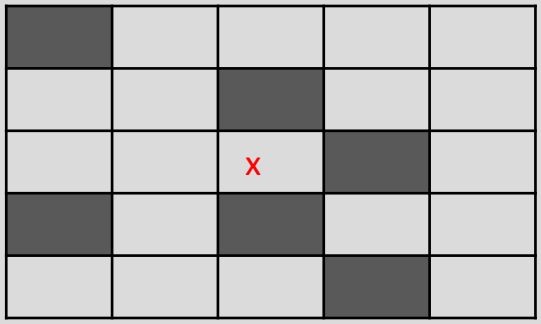


Verification of WRF Simulations – Categorical Variables

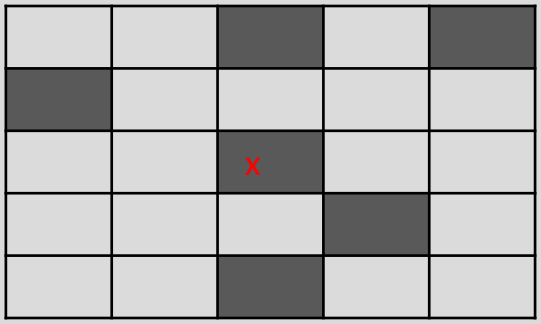
- A more sophisticated metrics to accurately quantify the realism of detailed forecast --- continuous, neighborhood method
  - Stage I: model forecast and observational fields are transformed into fraction grids
  - Stage II: Fractions are compared using the fractions skill score (FSS)

➔ The result is a measure of forecast skill against spatial scale for each selected threshold.

Recommendations on the Verification of WRF Simulations – Categorical Variables



Forecast



Observation

A schematic example of fractional creation for a forecast and the corresponding observation. The precipitation exceeds the accumulation threshold in the shaded boxes.

At the central grid: NP_F=0, NP_O=1 ➔ FCST wrong
Over 3 x 3 grids: NP_F=3/9, NP_O=2/9 ➔ FCST over-forecast
Over 5 x 5 grids: NP_F=6/25, NP_O=6/25 ➔ FCST correct

Recommendations on the Verification of WRF Simulations – Categorical Variables

- Fraction of occurrences within a sample area:

$$\mathrm{FBS} = \frac{1}{N_v}\sum_{i=1}^{N_v}[\mathrm{NP}_{F(i)} - \mathrm{NP}_{O(i)}]^2$$

(Fraction Brier Score)

$$\mathrm{FBS}_{\mathrm{worst}} = \frac{1}{N_v}\left[\sum_{i=1}^{N_v}\mathrm{NP}_{F(i)}^2 + \sum_{i=1}^{N_v}\mathrm{NP}_{O(i)}^2\right]$$

(The Worst FBS: no overlap of nonzero fractions)

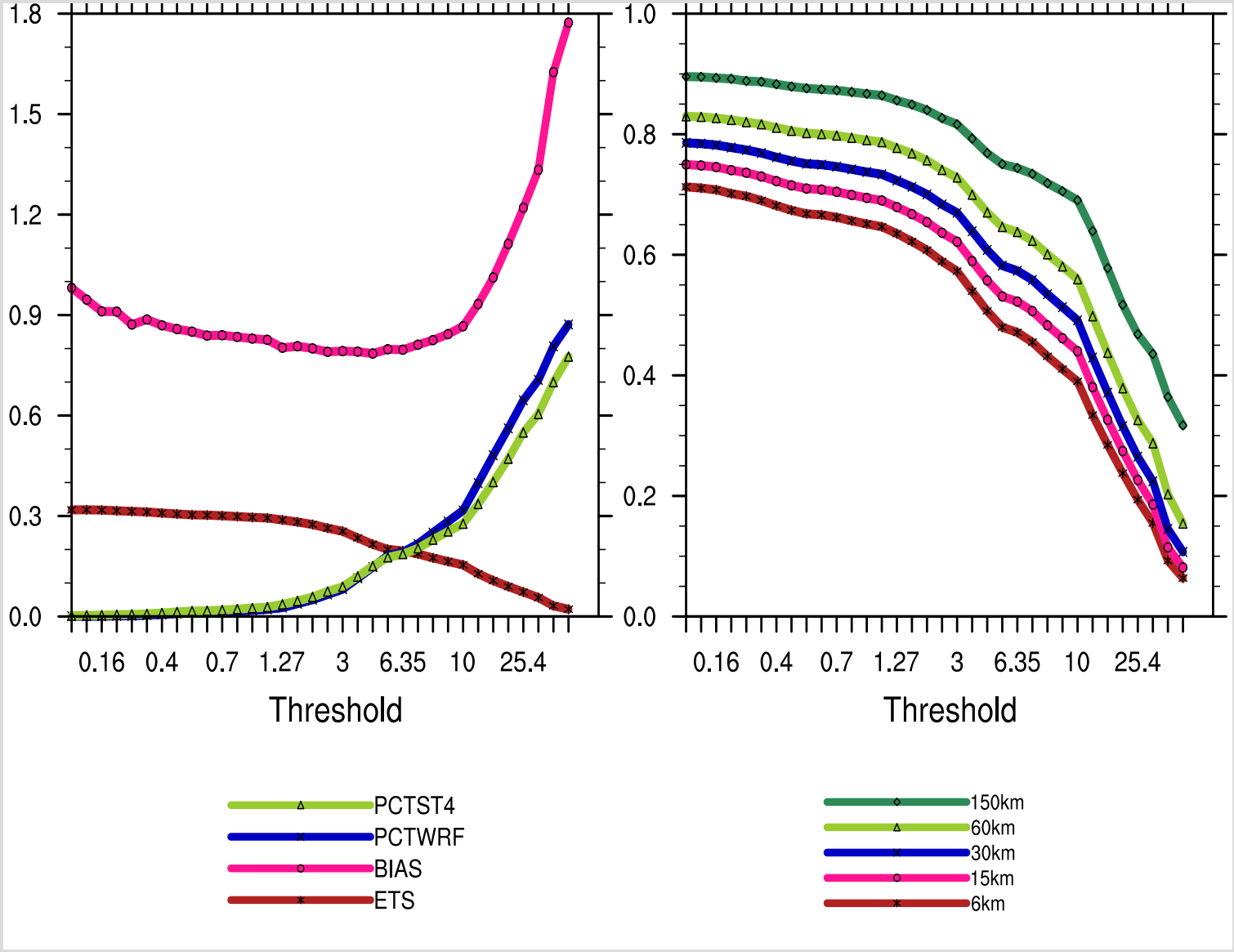
$$\mathrm{FSS} = 1 - \frac{\mathrm{FBS}}{\mathrm{FBS}_{\mathrm{worst}}}$$

(Fractions Skill Score)

$NP_{F(i)}$ and $NP_{O(i)}$ are the neighborhood probabilities at the ith grid box in the model forecast and observed fraction fields, respectively. N is the number of grids in the verification area.

Verification of WRF Simulations –Categorical Variables (continue)

- FBS is negatively oriented
  - 0: perfect performance
  - Large FBS: poor correspondence between FCST and OBS
  - FBSworst: no overlap of nonzero fractions
  - FBS strongly depends on the frequency of the event
- FSS is defined to compare the FBS to the low-accuracy reference forecast (FBSworst)
  - FSS range (0,1): 1 for perfect forecast and 0 indicates no skill
  - As the number of grid boxes increases, FSS improves



WRFV3.9, 3km Test Cases in Summer 2016