NCAR

An Update on the Community Effort for Convection Allowing Model (CAM) Scorecarding

Tara Jensen, Tina Kalb, Tatiana Burek, Adam Clark, Burkely Gallo, Brett Roberts, Israel Jirak, John Halley Gotway, Jason Levit, Curtis Alexander, Geoff Manikin, Patrick Skinner and the rest of the UFS CAM Working Group

AMS 29th WAF/25th NWP Conferences
Denver, CO June 4-8 2018

National Center for Atmospheric Research

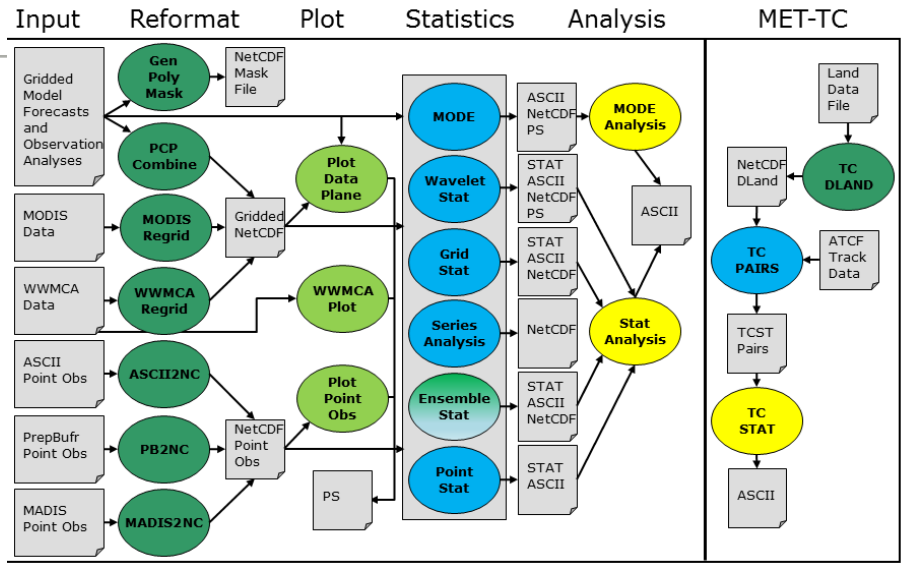
CAM Scorecard Project

- Identify fields, techniques and metrics to measure skill for Convection Allowing Models
- Determine the best set to include on a scorecard
- Set up a system to have this available during the Hazardous Weather Testbed
- Iterate until we get it right
- Funded by the United States Weather Research Program

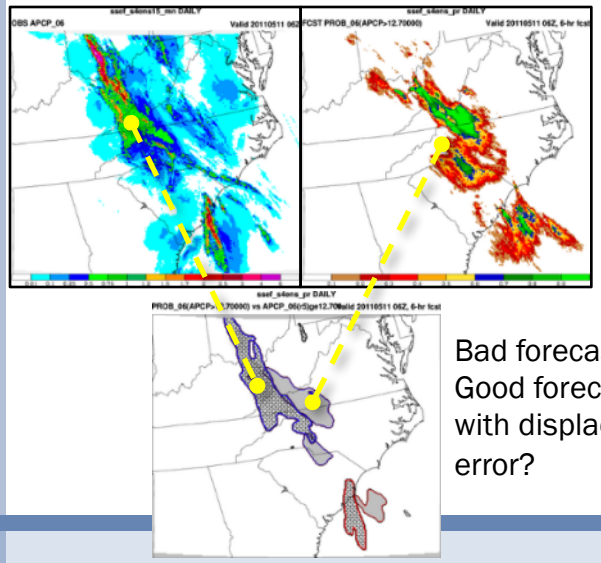
MET
Model Evaluation Tools

A verification toolkit designed for flexible yet systematic evaluation (supported to the community via the DTC)

- Originally developed to replicated the EMC mesoscale verification system
- **Over 85 traditional statistics** using both point and gridded datasets
- Multiple interpolation methods
- Computation of confidence intervals
- Able to read in GRIB1, GRIB2 and CF-compliant NetCDF
- Applied to many spatial and temporal scales
- 3500+ users, both US & Int'l

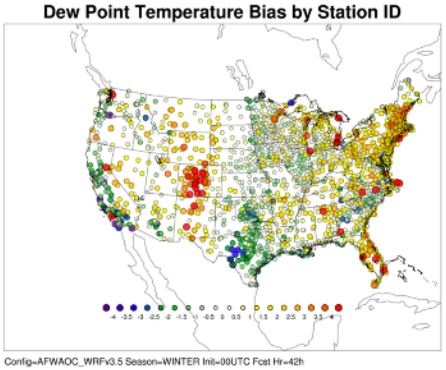


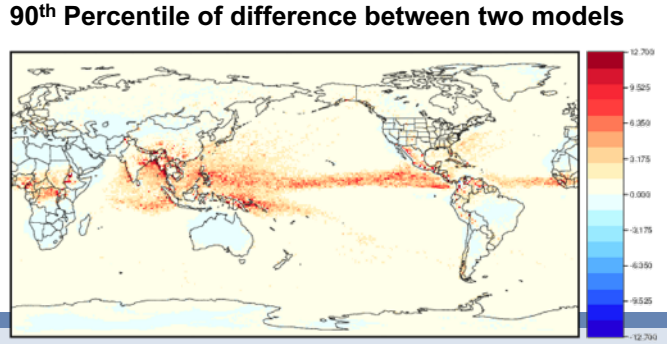
Object Based and Spatial Methods



Bad forecast or Good forecast with displacement error?

Geographical Representation of Errors





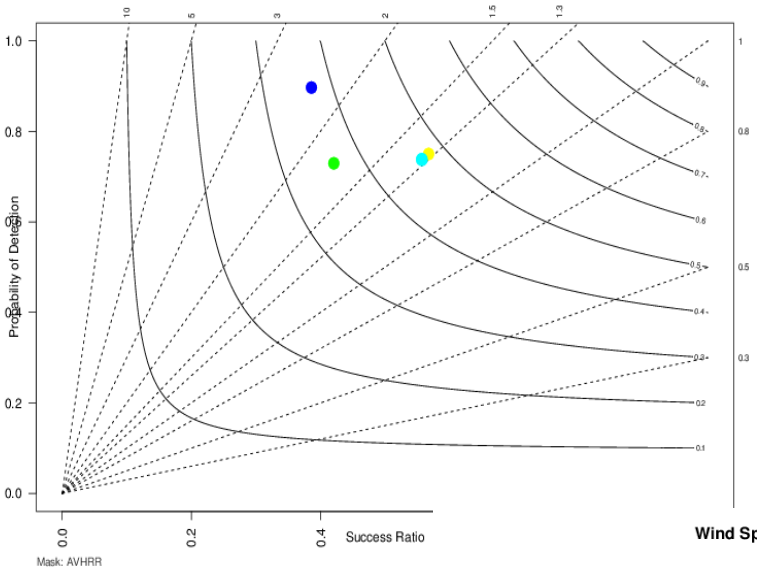
Synthesis Tools

NCAR

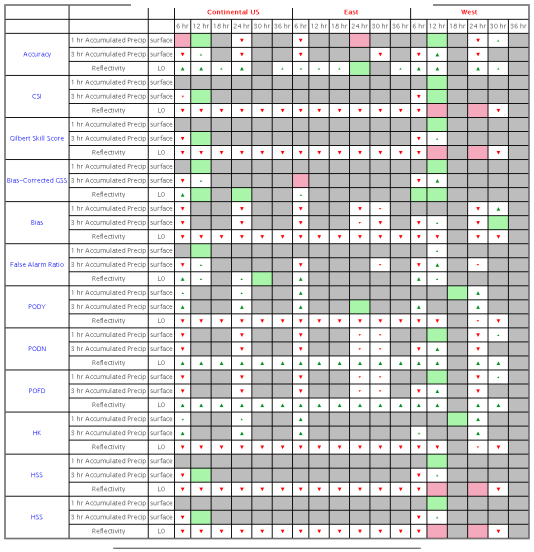
NWP Index

		Forecast Period				
		T+24	T+48	T+72	T+96	T+120
NH	PMSL	10	8	6	4	4
	H500	6	4	2	-	-
	W250	12	-	-	-	-
Tropics	W850	5	3	2	-	-
	W250	6	-	-	-	-
SH	PMSL	5	4	3	2	2
	H500	3	2	1	-	-
	W250	6	-	-	-	-

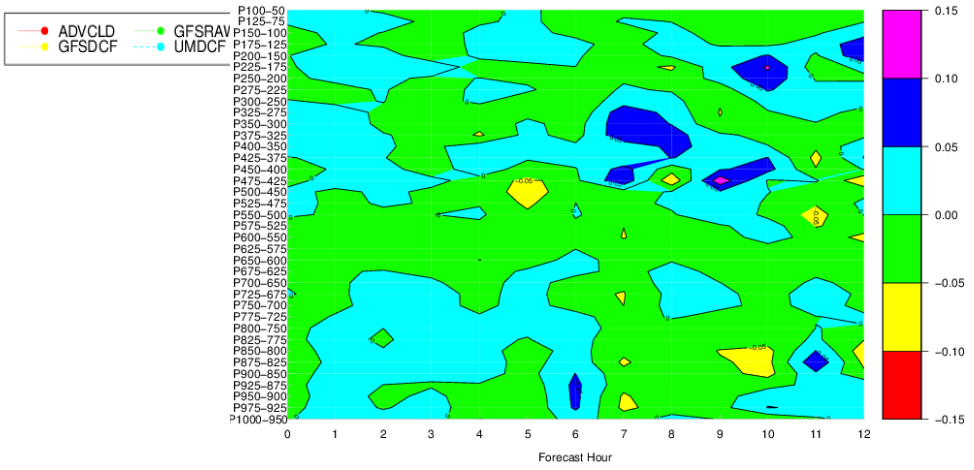
Diagrams
(e.g. Performance – Roebber, 2009)



Scorecards



Contour Plots/ Quilt Plots/ Heat Maps



MET+ Scorecard

NCAR

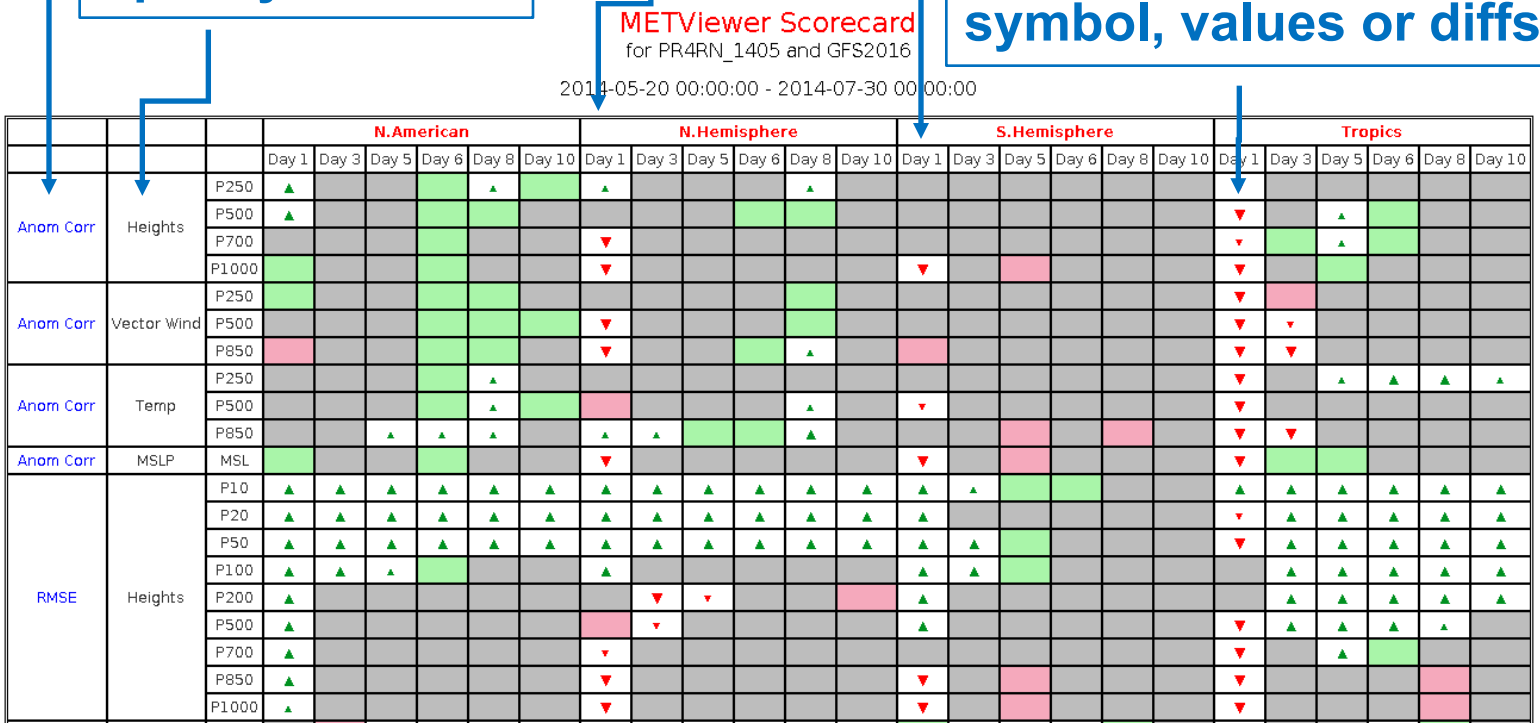
Specify the Statistic

Specify the aggregations

Specify the regions

Specify the Field

Specify whether you have symbol, values or diffs



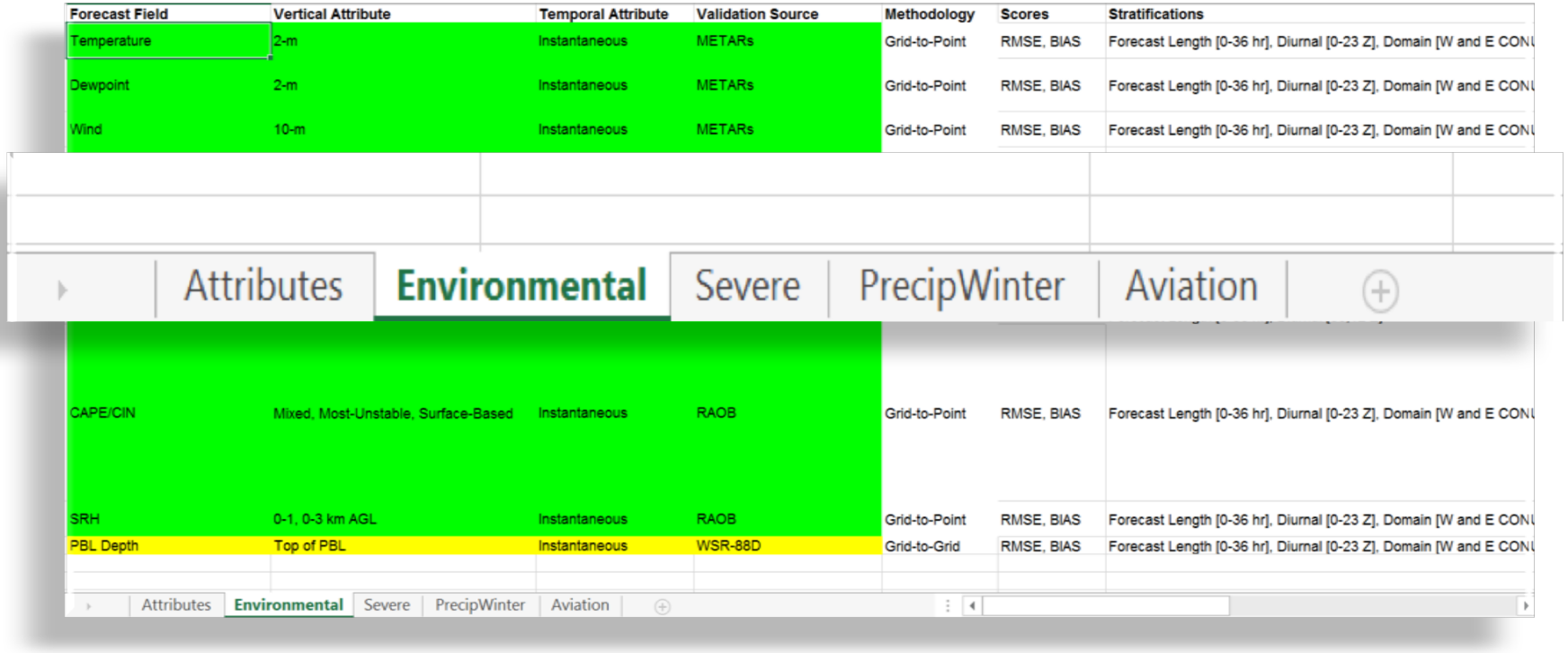
Computing Significance

- Based on Pairwise Differences
- P-value computation
  - Student-T that relaxes to a normal
  - Bootstrapping available and used here

Model 1 Model 2

▲	GFDLFV3 is better than NSSLFV3 at the 99.9% significance level
▲	GFDLFV3 is better than NSSLFV3 at the 99% significance level
	GFDLFV3 is better than NSSLFV3 at the 95% significance level
	No statistically significant difference between GFDLFV3 and NSSLFV3
	GFDLFV3 is worse than NSSLFV3 at the 95% significance level
▼	GFDLFV3 is worse than NSSLFV3 at the 99% significance level
▼	GFDLFV3 is worse than NSSLFV3 at the 99.9% significance level
	Not statistically relevant

Working with UFS CAM WG



Forecast Field	Vertical Attribute	Temporal Attribute	Validation Source	Methodology	Scores	Stratifications
Temperature	2-m	Instantaneous	METARs	Grid-to-Point	RMSE, BIAS	Forecast Length [0-36 hr], Diurnal [0-23 Z], Domain [W and E CONU
Dewpoint	2-m	Instantaneous	METARs	Grid-to-Point	RMSE, BIAS	Forecast Length [0-36 hr], Diurnal [0-23 Z], Domain [W and E CONU
Wind	10-m	Instantaneous	METARs	Grid-to-Point	RMSE, BIAS	Forecast Length [0-36 hr], Diurnal [0-23 Z], Domain [W and E CONU
CAPE/CIN	Mixed, Most-Unstable, Surface-Based	Instantaneous	RAOB	Grid-to-Point	RMSE, BIAS	Forecast Length [0-36 hr], Diurnal [0-23 Z], Domain [W and E CONU
SRH	0-1, 0-3 km AGL	Instantaneous	RAOB	Grid-to-Point	RMSE, BIAS	Forecast Length [0-36 hr], Diurnal [0-23 Z], Domain [W and E CONU
PBL Depth	Top of PBL	Instantaneous	WSR-88D	Grid-to-Grid	RMSE, BIAS	Forecast Length [0-36 hr], Diurnal [0-23 Z], Domain [W and E CONU

Attributes | Environmental | Severe | PrecipWinter | Aviation

DTC UFS Test Plan and Metrics Workshop: Jul 30-Aug 1 in College Park, MD
https://dtcenter.org/news/2018/2018-dtc-community-unified-forecast-system-test-plan-metrics-workshop

HWT/SFE 2018 Operations

- Deterministic CAMs
  - GFDL FV3
  - NSSL FV3
  - HRRRv3
- CAM Ensembles
  - HRRRE
  - HREFv2

Focused on a small subset of HWT guidance

- Fields:
  - Reflectivity at various dBZ thresholds*
  - Probability of reflectivity exceeding a threshold*
  - Accumulated precipitation over 1-h, 3-h, and 6-h
  - Surrogate severe (probabilistic) using different UH thresholds

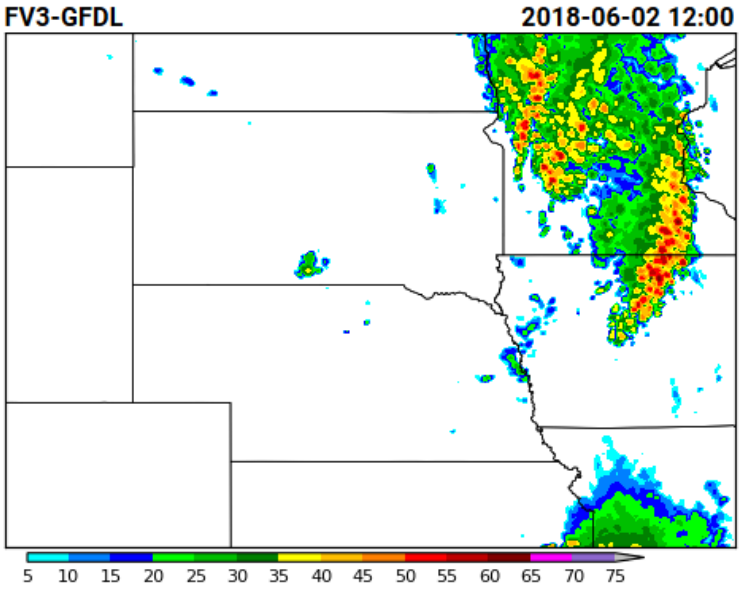
* *Also testing different neighborhood sizes*

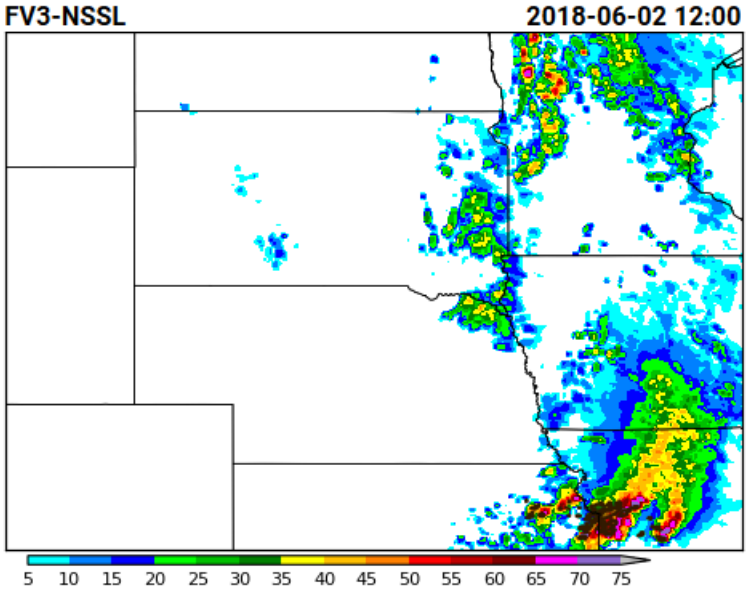
Preliminary Results from Weeks 1-5

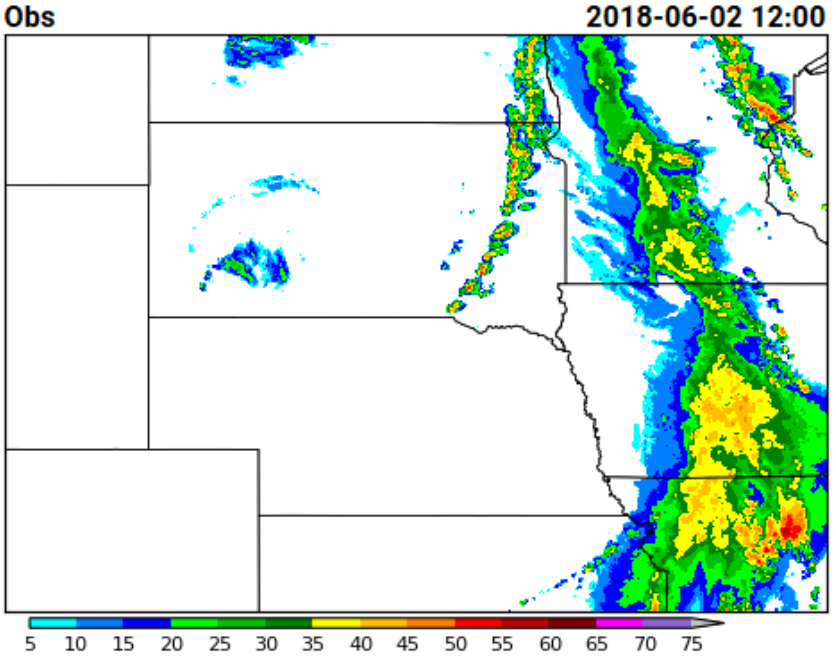
- Images from SFE 2018 homepage under the objective verification tab (https://hwt.nssl.noaa.gov/sfe/2018/)

NCAR

Example Product – Reflectivity







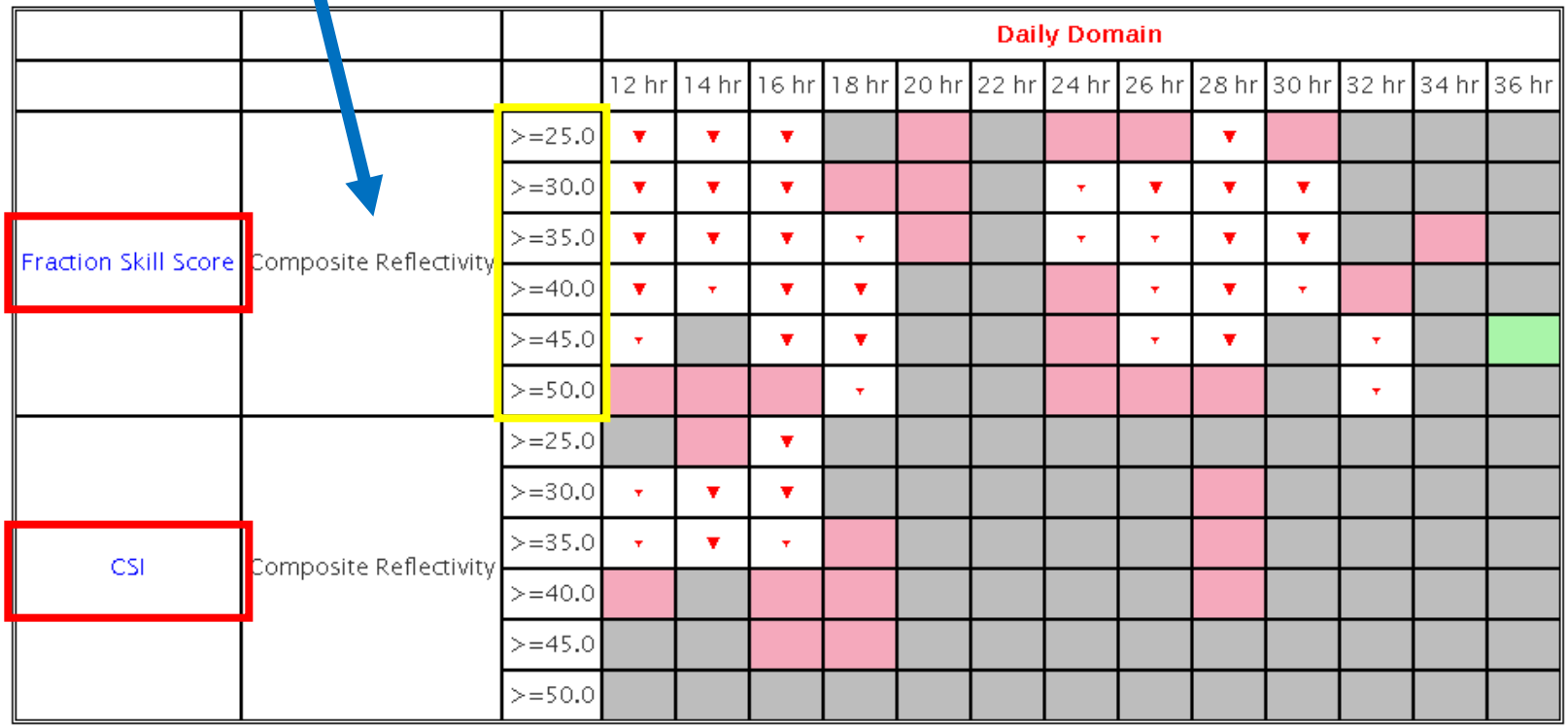
Reflectivity

METViewer CAM Scorecard

for GFDLFV3 and NSSLFV3

2018-04-30 00:00:00 - 2018-06-01 00:00:00

NCAR



▲	GFDLFV3 is better than NSSLFV3 at the 99.9% significance level
▲	GFDLFV3 is better than NSSLFV3 at the 99% significance level
	GFDLFV3 is better than NSSLFV3 at the 95% significance level
	No statistically significant difference between GFDLFV3 and NSSLFV3
	GFDLFV3 is worse than NSSLFV3 at the 95% significance level
▼	GFDLFV3 is worse than NSSLFV3 at the 99% significance level
▼	GFDLFV3 is worse than NSSLFV3 at the 99.9% significance level
	Not statistically relevant

NSSLFV3 generally better

METViewer CAM Scorecard

for GFDLFV3 and HRRR

2018-04-30 00:00:00 - 2018-06-01 00:00:00

NCAR

			Daily Domain											
			12 hr	14 hr	16 hr	18 hr	20 hr	22 hr	24 hr	26 hr	28 hr	30 hr	32 hr	34 hr	36 hr
Fraction Skill Score	Composite Reflectivity	>=25.0	▼	▼			▾		▾	▼	▼	▾			
		>=30.0	▼	▼	▾		▼		▼	▼	▼	▼			
		>=35.0	▼	▼	▼				▼	▼	▼	▾			
		>=40.0	▼	▾	▼	▾			▼	▾	▾				
		>=45.0	▼		▾				▼	▼	▾				
		>=50.0	▾	▾			▾		▼	▼					
CSI	Composite Reflectivity	>=25.0	▾	▾											
		>=30.0	▼	▼											
		>=35.0	▾	▾											
		>=40.0	▼												
		>=45.0													
		>=50.0			▲										

HRRR generally better

▲	GFDLFV3 is better than HRRR at the 99.9% significance level
▴	GFDLFV3 is better than HRRR at the 99% significance level
	GFDLFV3 is better than HRRR at the 95% significance level
	No statistically significant difference between GFDLFV3 and HRRR
	GFDLFV3 is worse than HRRR at the 95% significance level
▾	GFDLFV3 is worse than HRRR at the 99% significance level
▼	GFDLFV3 is worse than HRRR at the 99.9% significance level
	Not statistically relevant

METViewer CAM Scorecard

for NSSLFV3 and HRRR

2018-04-30 00:00:00 - 2018-06-01 00:00:00

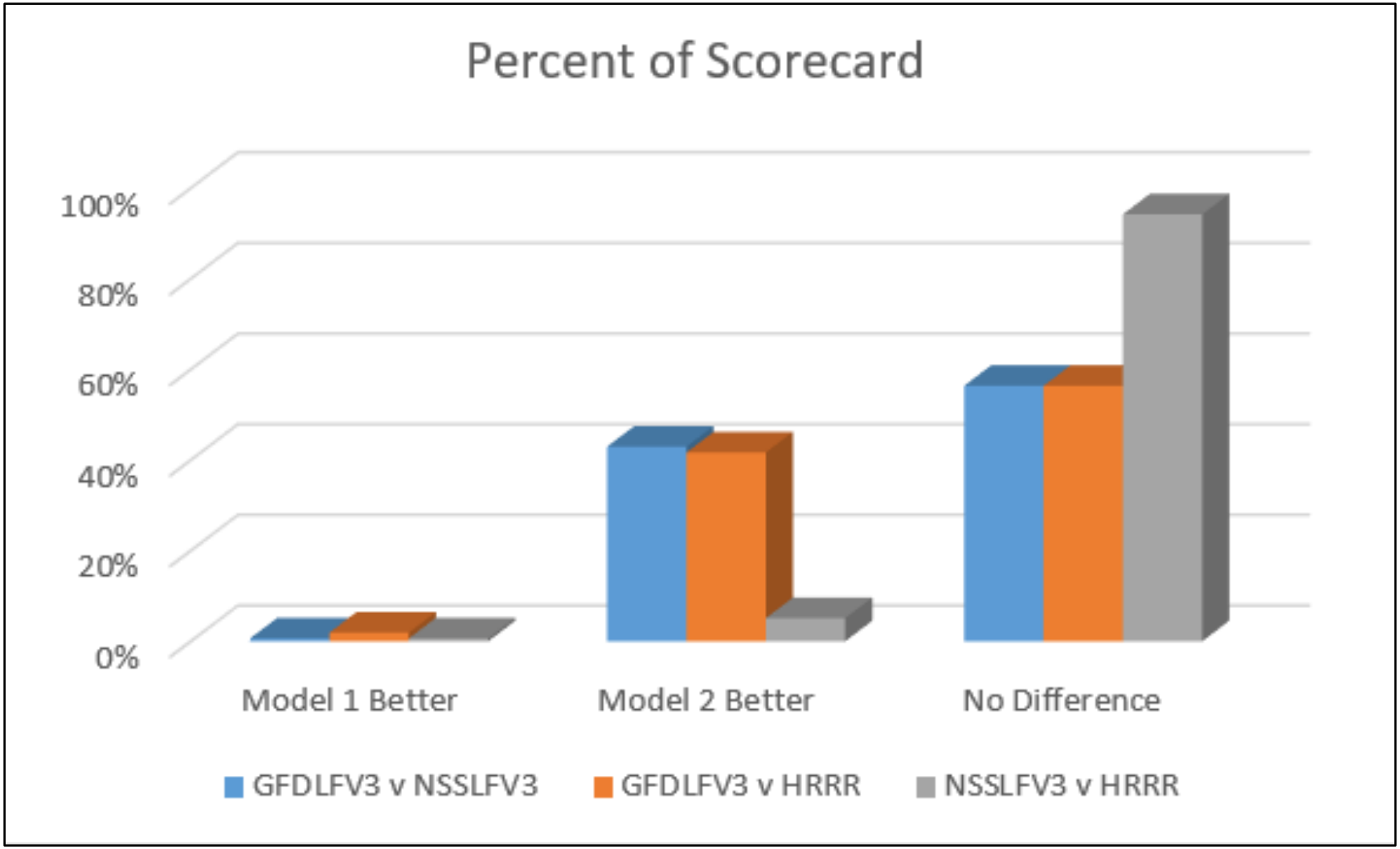
NCAR

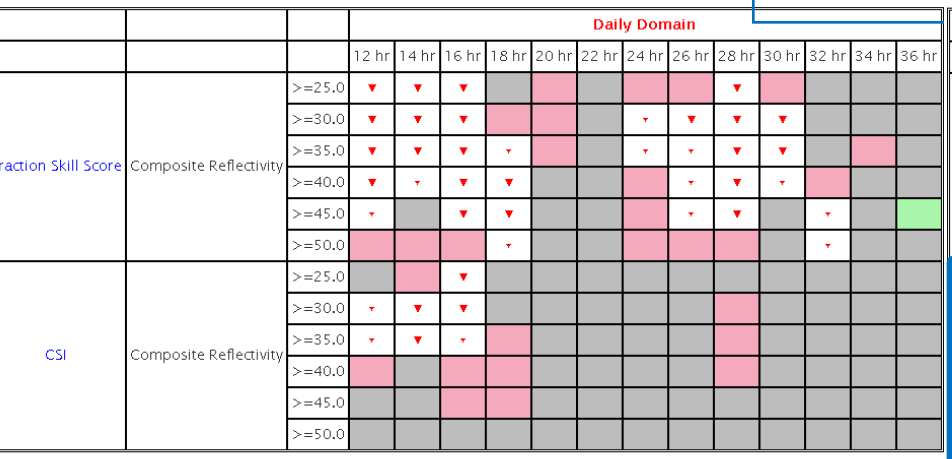
			Daily Domain											
			12 hr	14 hr	16 hr	18 hr	20 hr	22 hr	24 hr	26 hr	28 hr	30 hr	32 hr	34 hr	36 hr
Fraction Skill Score	Composite Reflectivity	>=25.0				▲									
		>=30.0													
		>=35.0													
		>=40.0													▼
		>=45.0													▾
		>=50.0													
CSI	Composite Reflectivity	>=25.0													
		>=30.0													
		>=35.0													
		>=40.0													
		>=45.0													
		>=50.0													

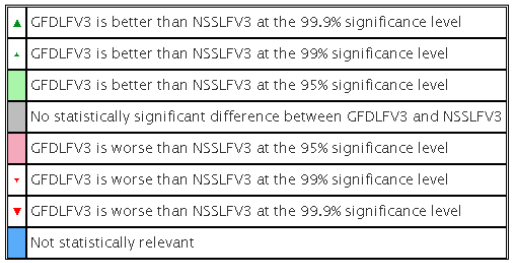
▲	NSSLFV3 is better than HRRR at the 99.9% significance level
▲	NSSLFV3 is better than HRRR at the 99% significance level
	NSSLFV3 is better than HRRR at the 95% significance level
	No statistically significant difference between NSSLFV3 and HRRR
	NSSLFV3 is worse than HRRR at the 95% significance level
▾	NSSLFV3 is worse than HRRR at the 99% significance level
▼	NSSLFV3 is worse than HRRR at the 99.9% significance level
	Not statistically relevant

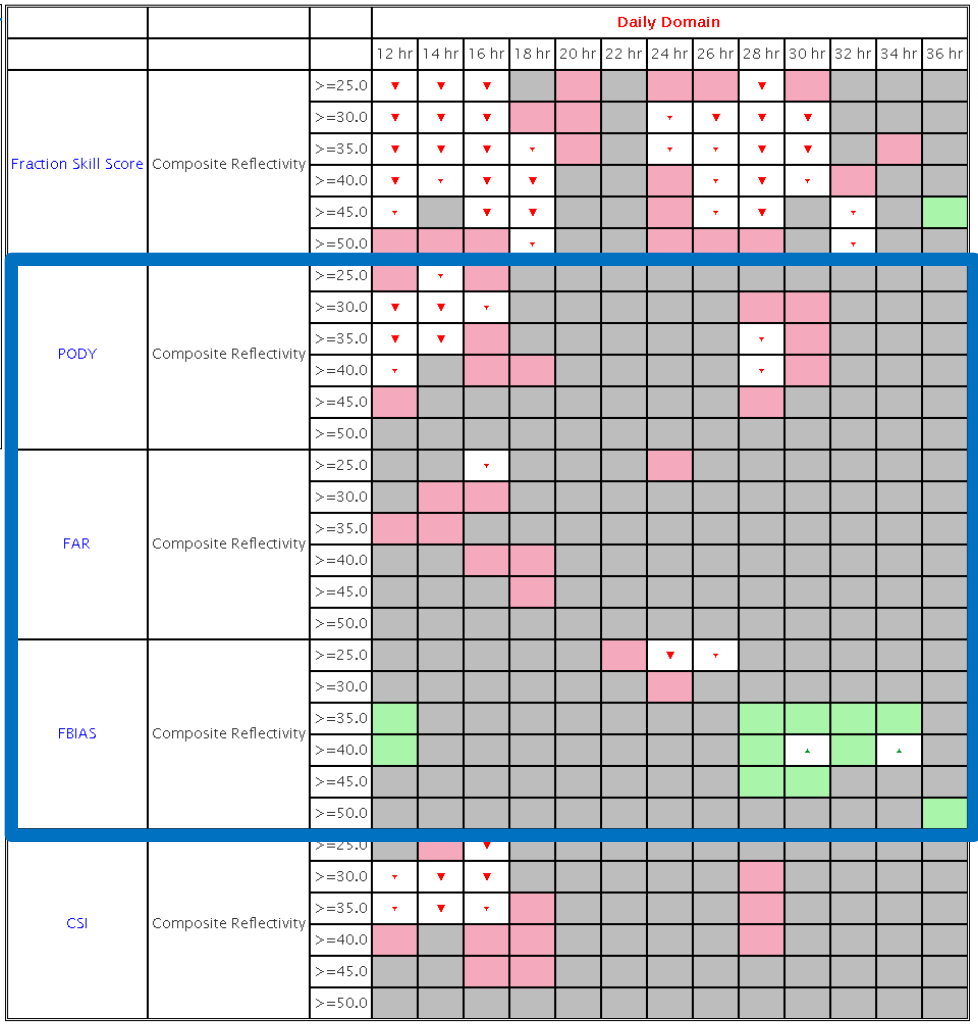
Very little difference – HRRR better when there is

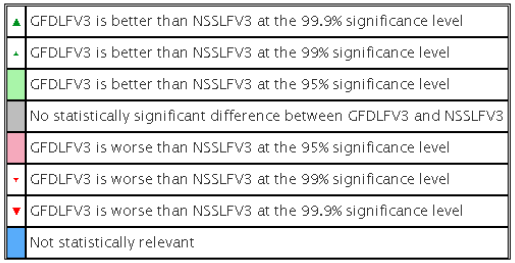
Summarizing the Scorecard?





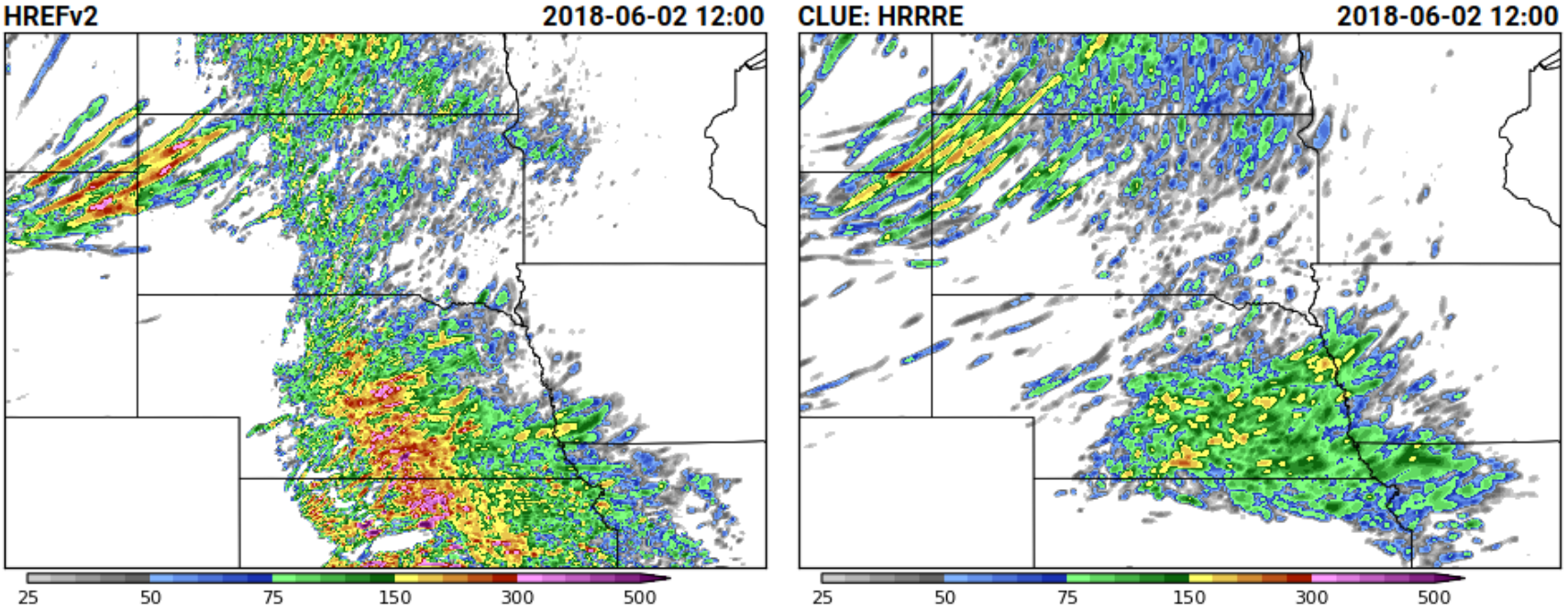






Question:
Do more statistics help?

Example Product – Updraft Helicity



Surrogate Severe Based on Updraft Helicity
Evaluated Using Practically Perfect Prog

Surrogate Severe

METViewer CAM Scorecard
for GFDLFV3 and NSSLFV3

2018-04-30 00:00:00 - 2018-05-22 00:00:00

			Daily Domain	CONUS
			Daily	
CSI	NBR 50	>=0.02		▲
		>=0.05		▲
		>=0.10		▲
		>=0.15		▲
		>=0.30		
		>=0.45		
		>=0.60		
	NBR 75	>=0.02		▲
		>=0.05		▲
		>=0.10		▲
		>=0.15		▲
		>=0.30		
		>=0.45		
		>=0.60		
	NBR 100	>=0.02		▲
		>=0.05		▲
		>=0.10		▲
		>=0.15		▲
		>=0.30		▵
		>=0.45		
		>=0.60		
	NBR 125	>=0.02		▲
		>=0.05		▲
		>=0.10		▲
		>=0.15		▲
		>=0.30		▵
		>=0.45		
		>=0.60		

▲	GFDLFV3 is better than NSSLFV3 at the 99.9% significance level
▵	GFDLFV3 is better than NSSLFV3 at the 99% significance level
	GFDLFV3 is better than NSSLFV3 at the 95% significance level
	No statistically significant difference between GFDLFV3 and NSSLFV3
	GFDLFV3 is worse than NSSLFV3 at the 95% significance level
▾	GFDLFV3 is worse than NSSLFV3 at the 99% significance level
▼	GFDLFV3 is worse than NSSLFV3 at the 99.9% significance level
	Not statistically relevant

METViewer CAM Scorecard
for NSSLFV3 and HRRR

2018-04-30 00:00:00 - 2018-05-22 00:00:00

Daily & CONUS Domains

			Daily Domain	CONUS
			Daily	
CSI	NBR 50	>=0.02		▼
		>=0.05		▼
		>=0.10	▾	▼
		>=0.15	▾	▼
		>=0.30		▼
		>=0.45	▾	▾
		>=0.60		
	NBR 75	>=0.02		▼
		>=0.05		▼
		>=0.10		▼
		>=0.15	▾	▼
		>=0.30	▾	▼
		>=0.45	▾	▾
		>=0.60		
	NBR 100	>=0.02		▼
		>=0.05		▼
		>=0.10		▼
		>=0.15		▼
		>=0.30	▾	▼
		>=0.45		▾
		>=0.60		
	NBR 125	>=0.02		▼
		>=0.05		▼
		>=0.10		▼
		>=0.15		▼
		>=0.30		
		>=0.45		
		>=0.60		

Prob

UH exceeding

▲	NSSLFV3 is better than HRRR at the 99.9% significance level
▵	NSSLFV3 is better than HRRR at the 99% significance level
	NSSLFV3 is better than HRRR at the 95% significance level
	No statistically significant difference between NSSLFV3 and HRRR
	NSSLFV3 is worse than HRRR at the 95% significance level
▾	NSSLFV3 is worse than HRRR at the 99% significance level
▼	NSSLFV3 is worse than HRRR at the 99.9% significance level
	Not statistically relevant

METViewer CAM Scorecard
for HREFv2 and HRRRE

2018-04-30 00:00:00 - 2018-05-22 00:00:00

			Daily Domain	CONUS
			Daily	
CSI	NBR 50	>=0.02		
		>=0.05		
		>=0.10		
		>=0.15		
		>=0.30		
		>=0.45		
		>=0.60		
	NBR 75	>=0.02		
		>=0.05		
		>=0.10		
		>=0.15		
		>=0.30		
		>=0.45		
		>=0.60	▵	▵
	NBR 100	>=0.02		
		>=0.05		
		>=0.10		
		>=0.15		▵
		>=0.30		▵
		>=0.45		
		>=0.60		
	NBR 125	>=0.02		▵
		>=0.05		▲
		>=0.10		▲
		>=0.15		▲
		>=0.30		
		>=0.45		
		>=0.60	▲	▲

▲	HREFv2 is better than HRRRE at the 99.9% significance level
▵	HREFv2 is better than HRRRE at the 99% significance level
	HREFv2 is better than HRRRE at the 95% significance level
	No statistically significant difference between HREFv2 and HRRRE
	HREFv2 is worse than HRRRE at the 95% significance level
▾	HREFv2 is worse than HRRRE at the 99% significance level
▼	HREFv2 is worse than HRRRE at the 99.9% significance level
	Not statistically relevant

Immediate Future Work

- Complete HWT SFE 2018 evaluation
- Enhance MET+ to compute additional Severe Weather specific fields
- Work with community to formulate CAM Severe scorecard (version1)
- Extend CAM scorecard to other fields beyond “Severe” as specified by UFS CAM Working Group
- Participate in HWT SFE 2019

Questions?

- Emails: jensen@ucar.edu; kalb@ucar.edu
- MET Help: met_help@ucar.edu
- MET Info: https://dtcenter.org/met/users/

- HWT 2018 Experiment Page: https://hwt.nssl.noaa.gov/sfe/2018/

  Select “Objective Verification” page

This totality of MET+ work is funded by the NGGPS program, USWRP R2O grants, and DTC partners (NOAA, Air Force and NSF)