What's In A Number?
Sensitivity of Object-Based Verification Results to Configuration Options Using MODE

Jeffrey D. Duda[1,2]

David D. Turner[2]

[1]*Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder*

[2]*NOAA Global Systems Laboratory, Boulder, CO*

(Which number? ***I***, the object pair interest value)

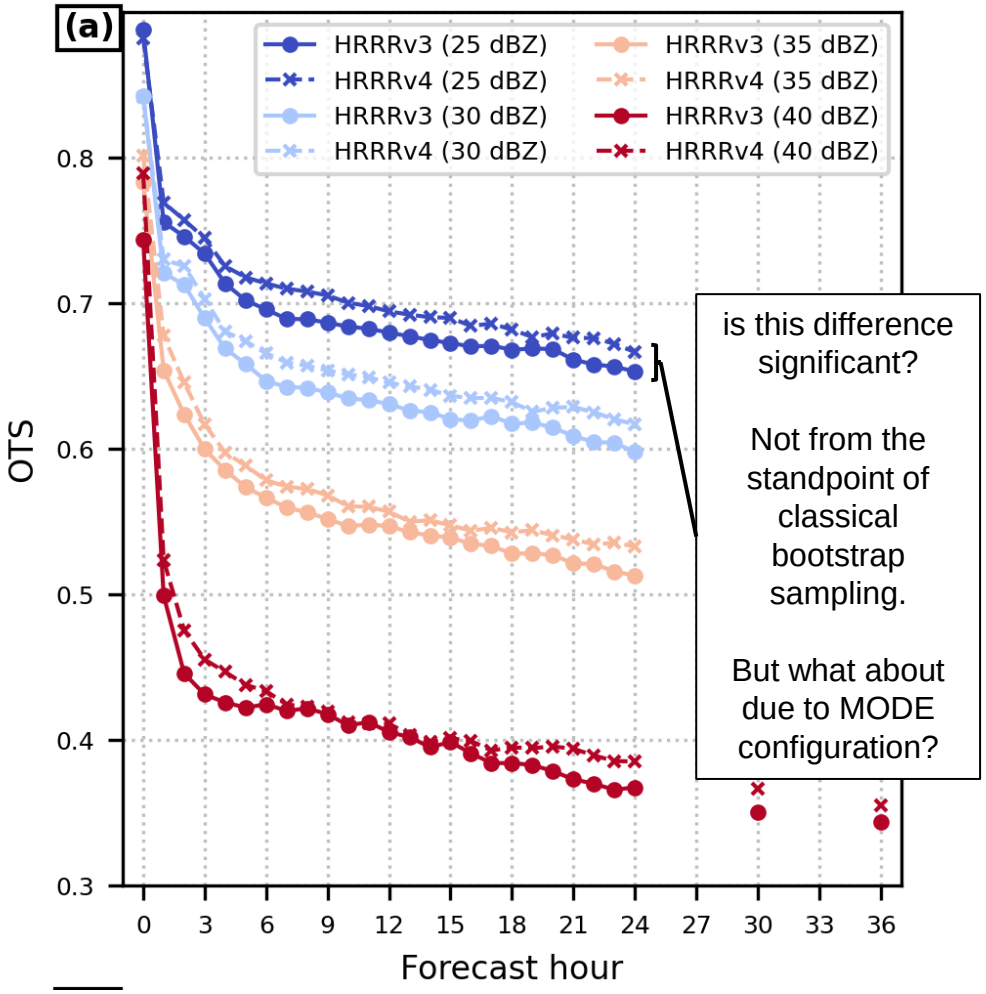
32nd Conf. on WAF/28th Conf. on NWP/20th Conf. on Mesoscale Processes – 20 July 2023 – presentation 16.3

Motivation: results from Duda and Turner (2021, 2023)

HRRRv3 vs. HRRRv4 in object-based space

- Field: composite reflectivity
- Sample: every 3-h forecasts from Aug/Sep 2019 and Apr-Sep 2020
- Both HRRRv3 (operational) and HRRRv4 (parallel, code frozen) running at the same time
  - 1300+ cases, O[100,000] objects per forecast hour
- Some distinction between v3 and v4 in OTS (and other metrics), but classical bootstrap testing revealed no statistically significant differences

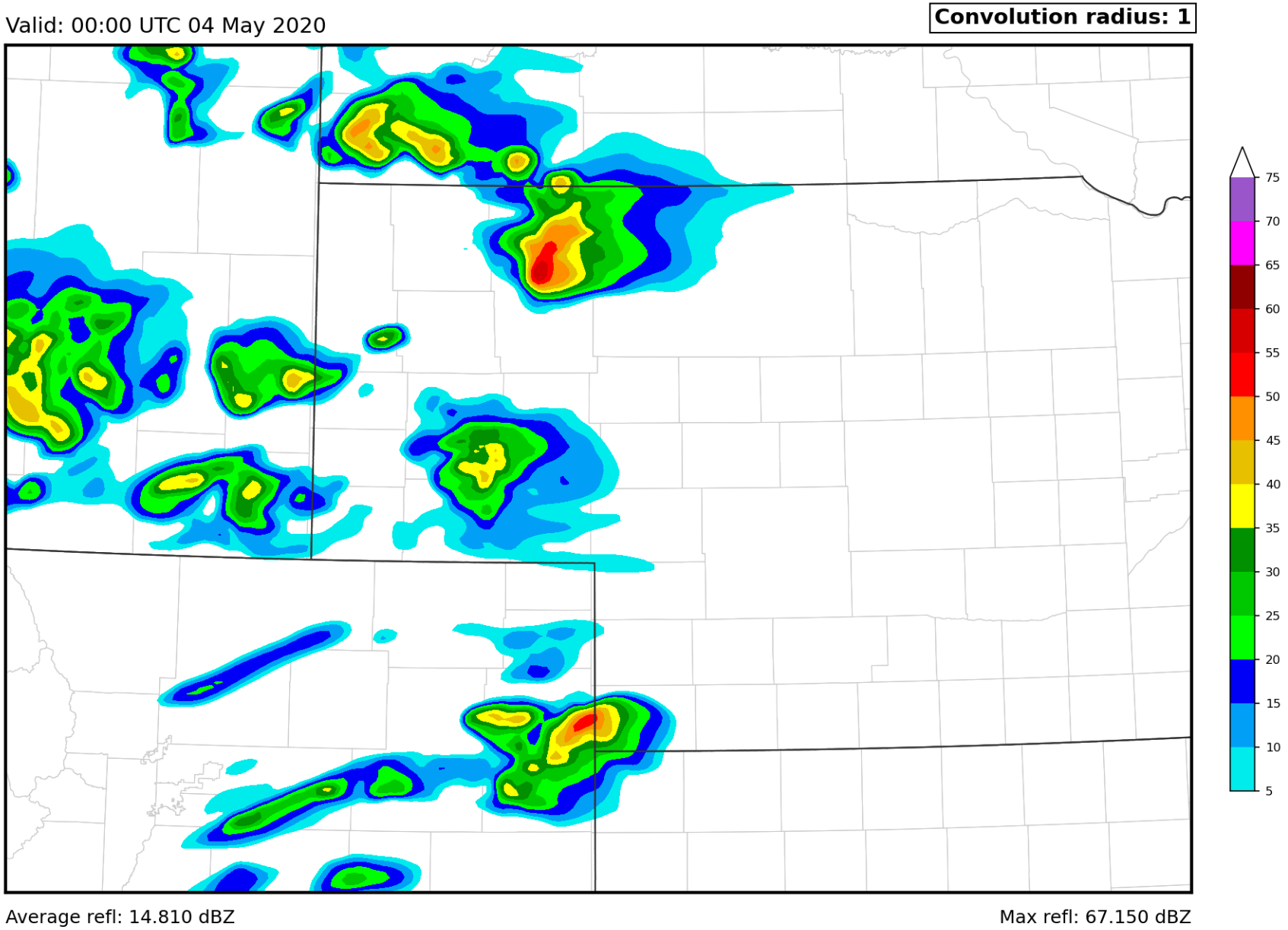
PROBLEM: Would we get the same or even similar results if we tried a different configuration of MODE?



MODE process Step 1: Field convolution

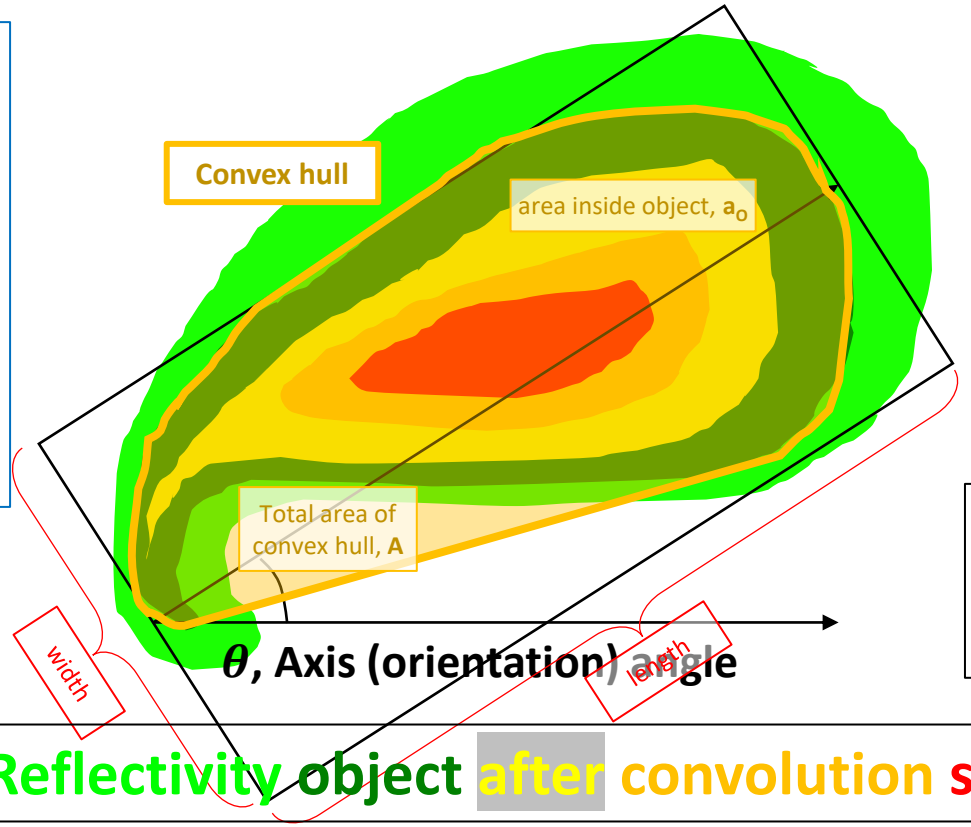
Really just a circular average filter…

<u>Convolution radius of 1 grid square used; minimal smoothing</u>



Step 2 – identify objects and calculate attributes

Classified object from contiguous points exceeding convolved field threshold (25 dBZ used throughout, WLOG)



Object complexity = $(A-a_o)/A$

Not shown: quantiles of field within object (95th percentile used throughout; near-max reflectivity)

Reflectivity object after convolution step

Aspect ratio = width/length

Also not shown: (radius of) curvature
Computed from higher-order moments of field

Step 3 – calculate attribute comparisons for object pairs

index	attribute name	description	Range
0	centroid distance	distance between object centroids	[0,∞) [km]
1	boundary distance	Closest that object boundaries come to each other	[0,∞) [km]
2	convex hull distance	same as boundary distance, except using convex hull boundary	[0,∞) [km]
3	angle difference	Difference in orientation angle of objects	[0°,90°]
4	aspect ratio difference	Difference between object aspect ratios	[0.0,1.0]
5	area ratio	Ratio of object areas*	[0.0,1.0]
6	consumption ratio	Fraction of smaller object encompassed by larger object	[0.0,1.0]
7	curvature ratio	ratio of object curvatures	[0.0,1.0]
8	complexity ratio	ratio of object complexities	[0.0,1.0]
9	intensity percentile ratio (pXX ratio)	ratio of pXX value of object* (95^{th} percentile used for composite reflectivity)	[0.0,1.0]

*MODE forces this ratio to be $\in [0,1]$, and will invert the ratio to enforce this range, but computations herein did not use this formulation.

Step 4 – compute object pair interest

- Key output: object pair <u>total interest</u> (for F-O object pair p)

$$I^p = \frac{\sum_{k=1}^{10} C_k * w_k * i_k^p}{\sum_{k=1}^{10} C_k * w_k}$$

(summing index k represents the 10 attribute comparisons)

C	**Confidence** (robustness of attribute comparison value) Disregarded in this investigation, and most values are 1.0 anyway
i	**single-attribute interest** Maps actual attribute difference to normalized range [0.0,1.0]
w	**attribute weight** – importance of a given attribute

- *Problem: optimal values to use for these settings not established, and likely dependent on individual user's application, so what values do we pick?*
  - Short and straightforward solution: tuning
    - Problem: values would need to be tuned for every variable in the verification, including field and cases
- Present objective: identify robustness of interest values (and metrics derived from it) to modifications of attribute weight sets (***w***) and interest maps (***i***)

Experiment setup

- Attribute weights: $W_n = [w_{n,0}, w_{n,1}, w_{n,2}, \cdots, w_{n,9}]\ \forall n \in \#\ of\ realizations$
  - Sample randomly from uniform distribution in range [0.1,5.0] (strictly 2 sig figs)
  - 200 random realizations (for most tests)
- Preliminary tests revealed two guiding principles:
  1. Maximum weight value not very important
  2. More extreme behavior when attribute weights set to 0.0
- Main test: "support" experiments
  - Nomenclature: *support_N*, for N in [2,10] indicating the count of attributes allowed to be nonzero (the particular attribute weights set to 0.0 allowed to vary randomly)
  - Additional experiment: *support_fixed*, a subset of *support_6*, fixes the attributes to be 0.0 to the same attributes used to verify HRRRv3/v4 in Duda and Turner (2021; 2023)
- Cases from June-August 2020 every 3 hours
  - O(10000) objects and 600 cases for each forecast hour (halved after f18)

Example of support experiments

Number of distinct sets possible for support_N = ${}_{10}C_N * 50^N$

Examples:

N=2 → 112500

N=5 → 7.875×10^{10}

N=10 → ~9.8×10^{16}

Name/ attribute	centroid distance	boundary distance	convex hull distance	angle difference	aspect ratio difference	area ratio	consumption ratio			
support_1	**0.**	**0.**	**0.**	**0.**	2.7	**0.**	**0.**			
support_2	**0.**	**0.**	4.9	**0.**	**0.**	**0.**	**0.**			
support_3	5.5	3.1	**0.**	**0.**	**0.**	4.6	**0.**	**0.**	**0.**	**0.**
support_4	**0.**	1.0	**0.**	**0.**	**0.**	**0.**	1.1	2.7	2.5	**0.**
support_5	**0.**	**0.**	**0.**	1.9	3.3	1.4	**0.**	**0.**	0.7	4.8
support_6	0.2	0.8	4.1	2.5	**0.**	**0.**	4.3	3.5	**0.**	**0.**
support_7	**0.**	2.0	5.0	3.9	1.8	**0.**	0.6	1.6	3.2	**0.**
support_8	1.5	4.6	1.1	**0.**	3.6	2.6	2.7	**0.**	4.9	0.1
support_9	3.9	**0.**	2.7	3.6	4.1	1.2	2.3	4.4	3.5	1.0
support_10	3.3	3.9	1.1	0.6	4.7	2.2	5.0	1.1	1.2	3.5
support_fixed	4.3	3.7	**0.**	**0.**	**0.**	0.4	1.8	**0.**	1.3	0.6
HRRR*	5.0	4.0	**0.**	**0.**	**0.**	4.0	2.0	**0.**	0.5	3.5

*Values used in Duda and Turner (2021), WAF and Duda and Turner (2023), *in review*

But wait! Wouldn't it help to have a baseline?
Well…yes!

- What would serve as a baseline, though?
  - Equal weighting
- Baseline: *equal-cN* experiments
  - 10 attributes – 10th row of Pascal's triangle gives the number of possible combinations for each comparable N from the *support* experiments:

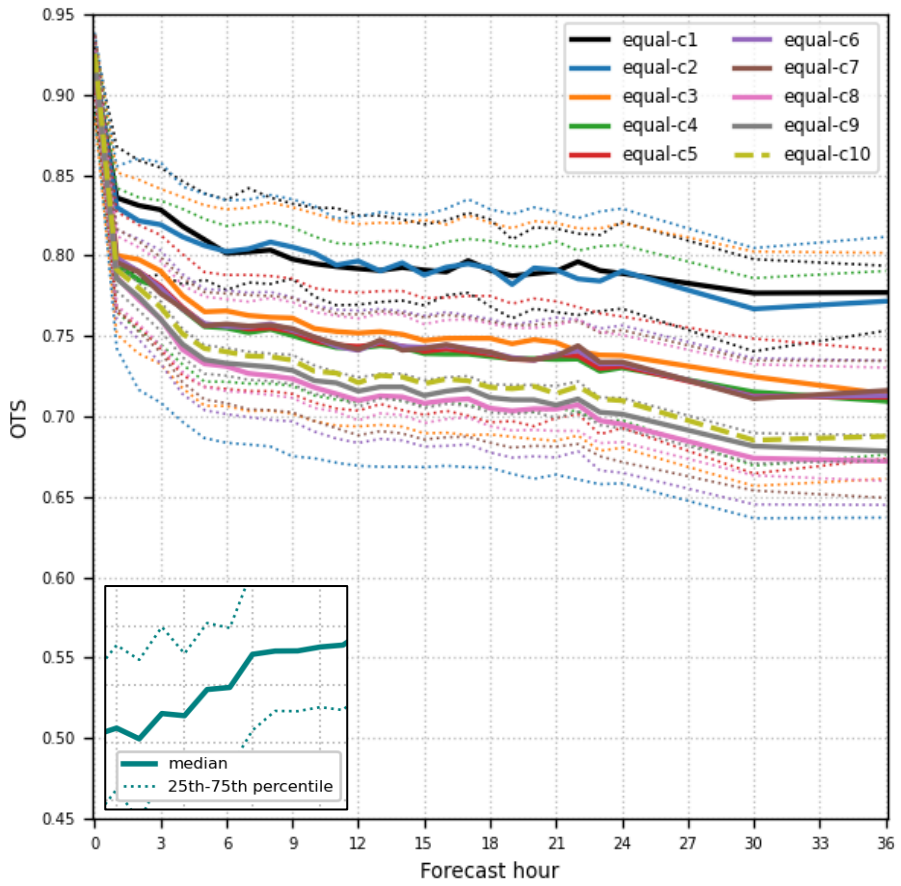
N	0	1	2	3	4	5	6	7	8	9	10
$_{10}C_N$	1	10	45	120	210	252	210	120	45	10	1

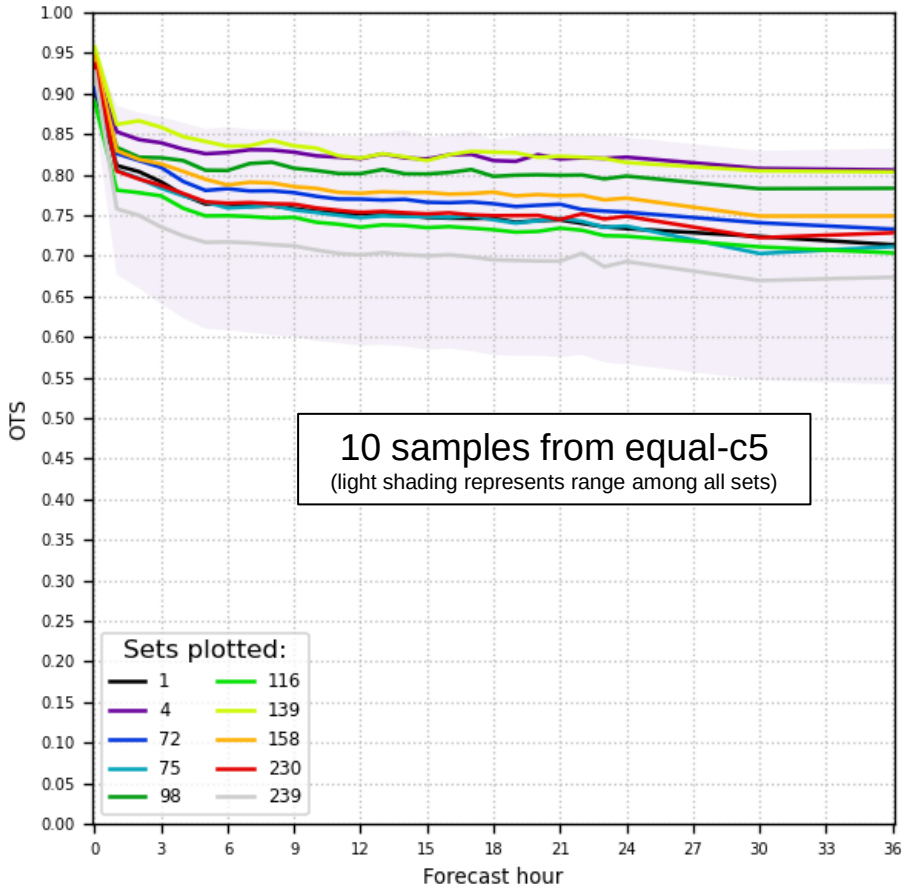
  - All combinations used for each N
  - WLOG, each weight value was set to 1.0
- On second thought, shouldn't there be a baseline for the interest maps, too?
  - Answer: yes, and "equal interest maps" were created, too.

Baseline via OTS using ”equal interest maps” and equal weighting

-Median OTS tends to decrease with increasing cN, as does range/variability

$$OTS = \frac{1}{A_f + A_o} \sum_p I^p \left(a_f^p + a_o^p \right)$$

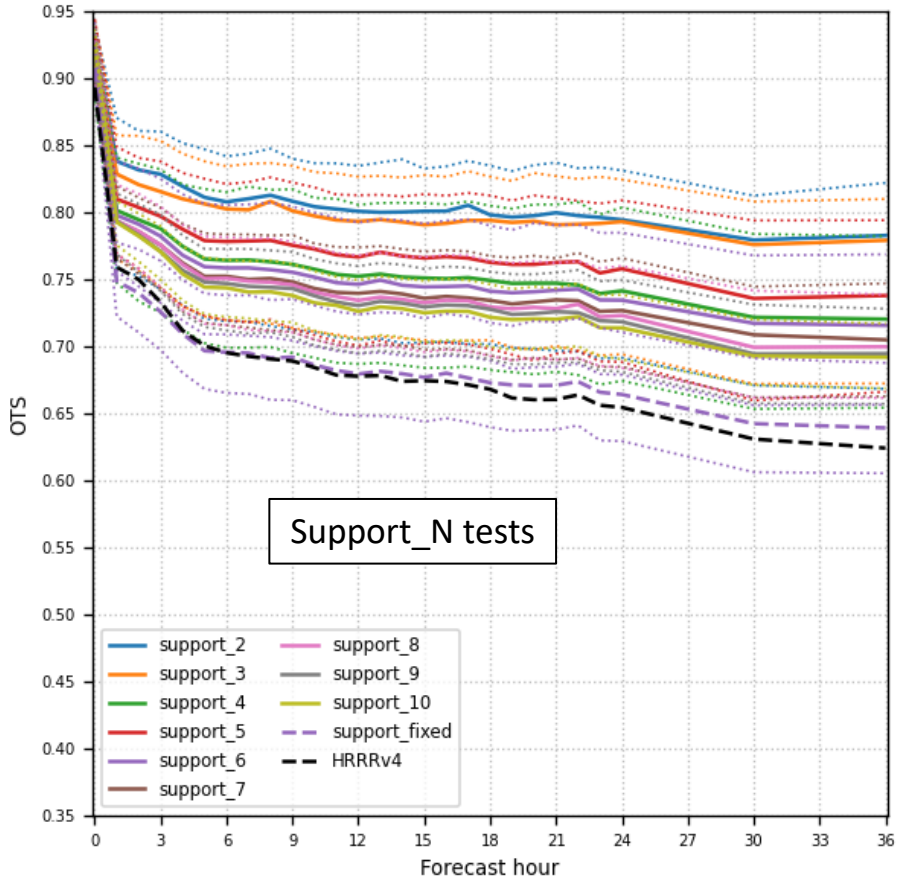


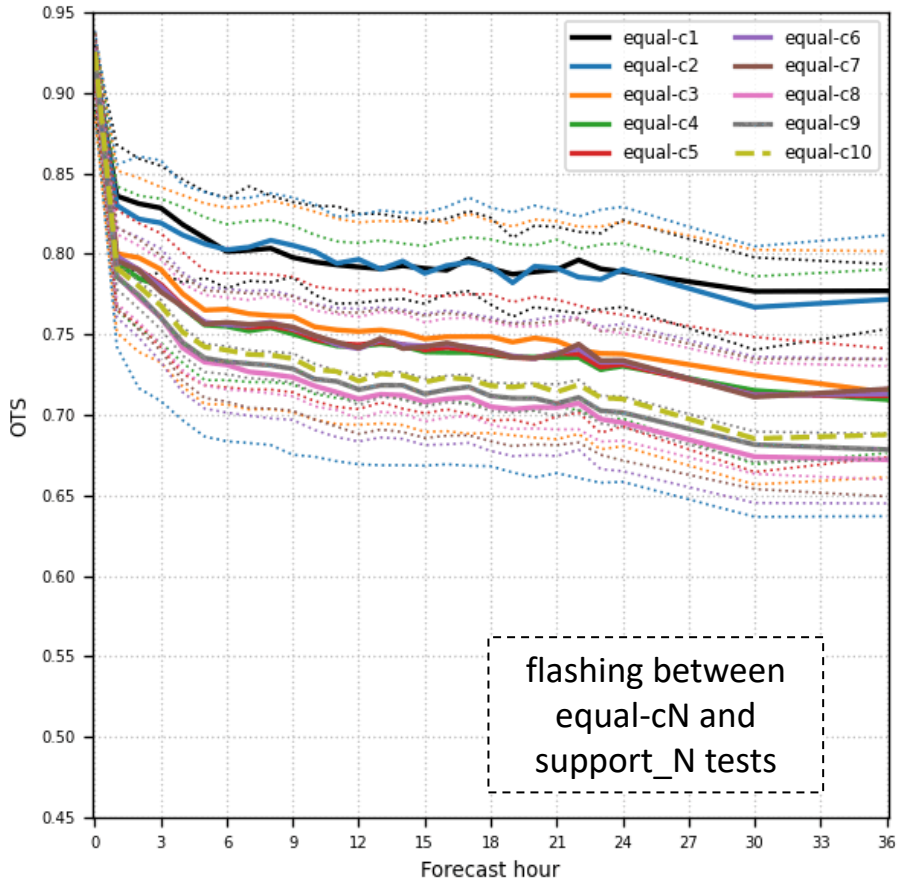


- A_f, A_o – total area of F/O objects
- a_f^p, a_o^p – area of p^{th} pair F/O object
- I^p – interest of p^{th} object pair
- OTS ϵ [0,1]:
  0.0: no meaningful correspondence
  1.0: effectively perfect

This is substantial variability considering weight value is fixed, only nonzero attributes impacted!

How does incorporating the attribute value variability modify the OTS diversity?



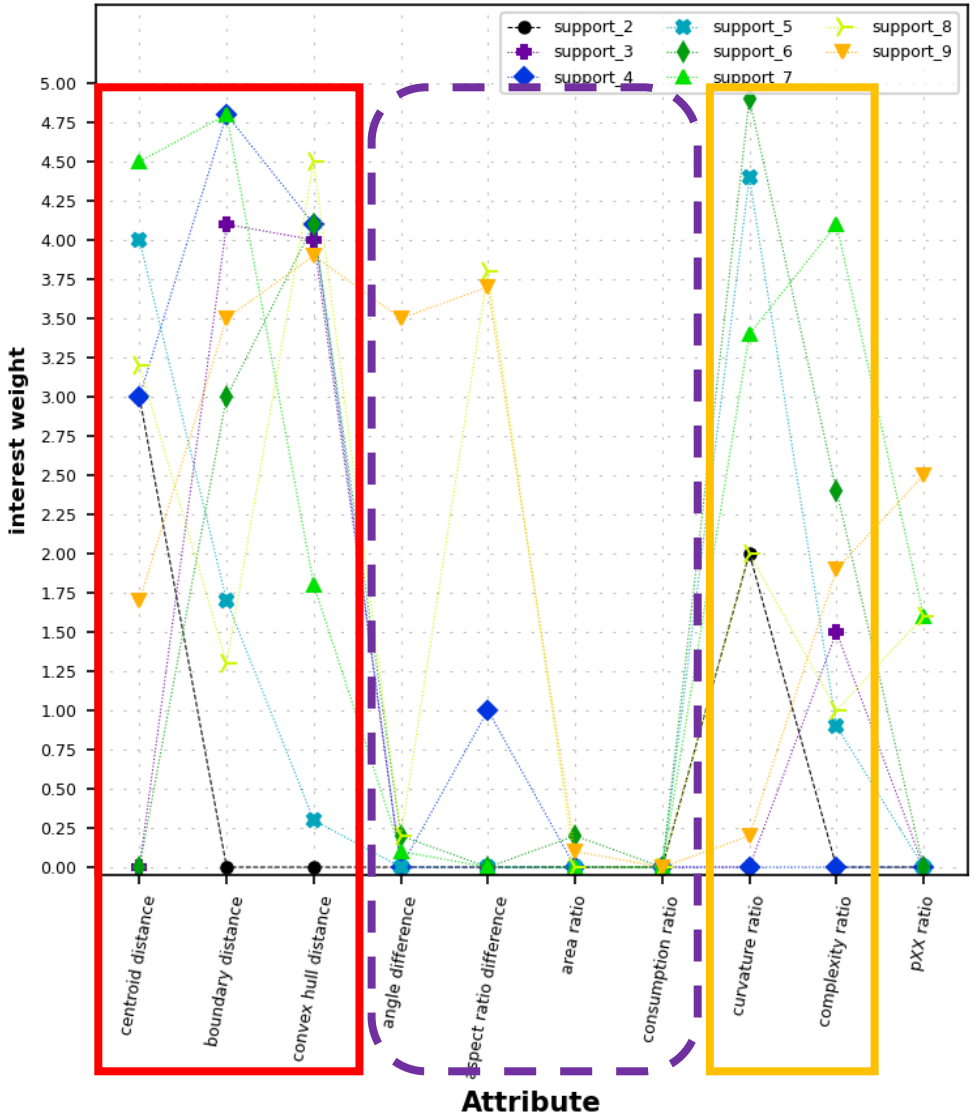


It almost doesn't seem to matter if you use equal weighting or random weighting!

Which attribute weight vector gave the best score?

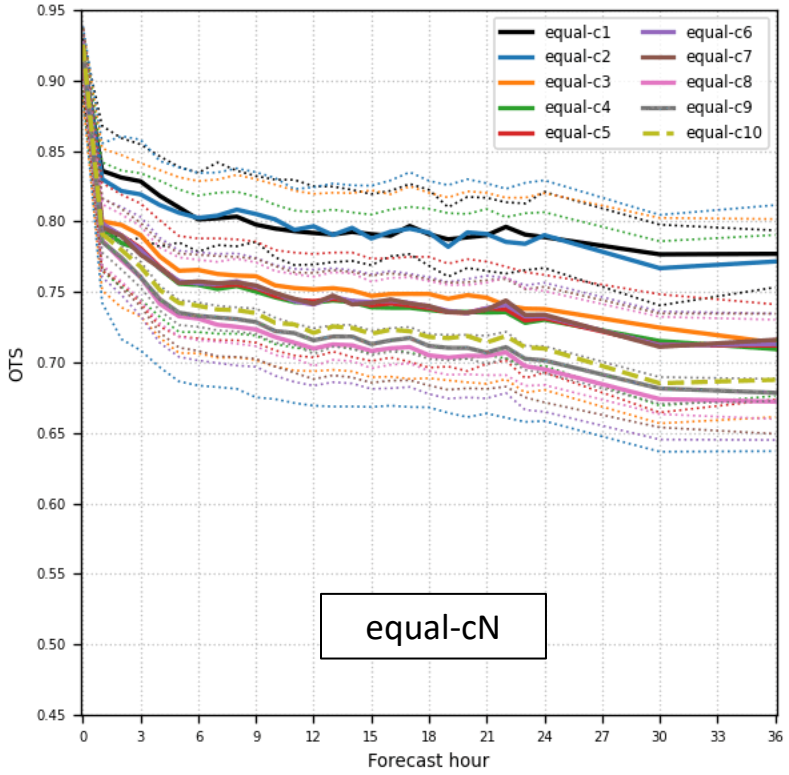
- Large values of distance-related attribute weights
- Large value of either complexity ratio or curvature ratio (or both)
- Following attributes resulted in best OTS when reduced or zeroed out:
  - angle diff
  - aspect ratio diff*
  - area ratio*
  - consumption ratio
- Complicating factors:
  - Which attributes were overall better forecast?
  - Which vector resulted in second best OTS? (sensitivity)

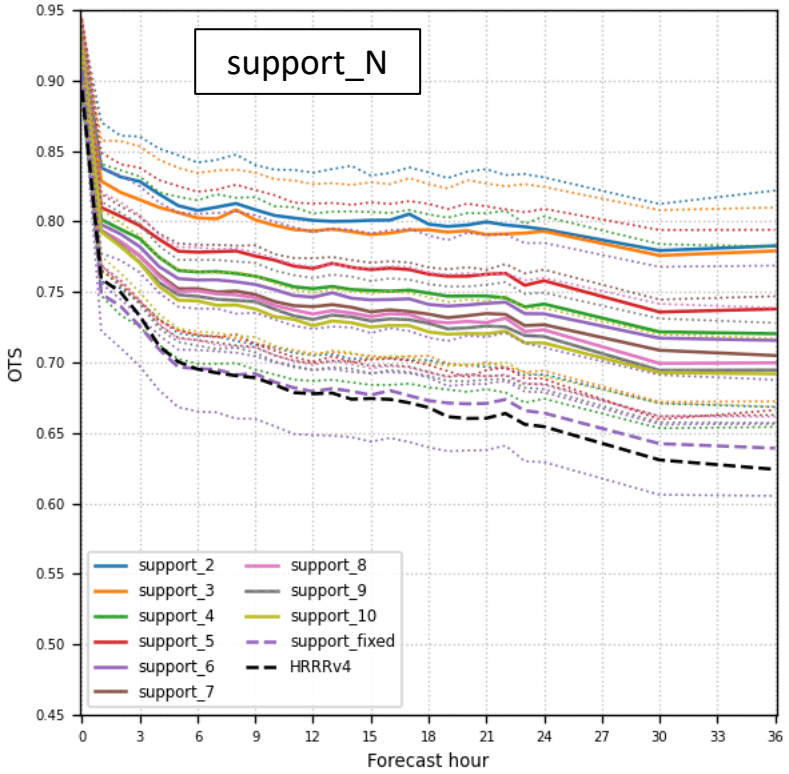
*High weights for these attributes tended to result in the lowest OTSs.



Duda and Turner (2021) used a ”stricter” interest map for verifying HRRRv3. What difference does the interest map make?

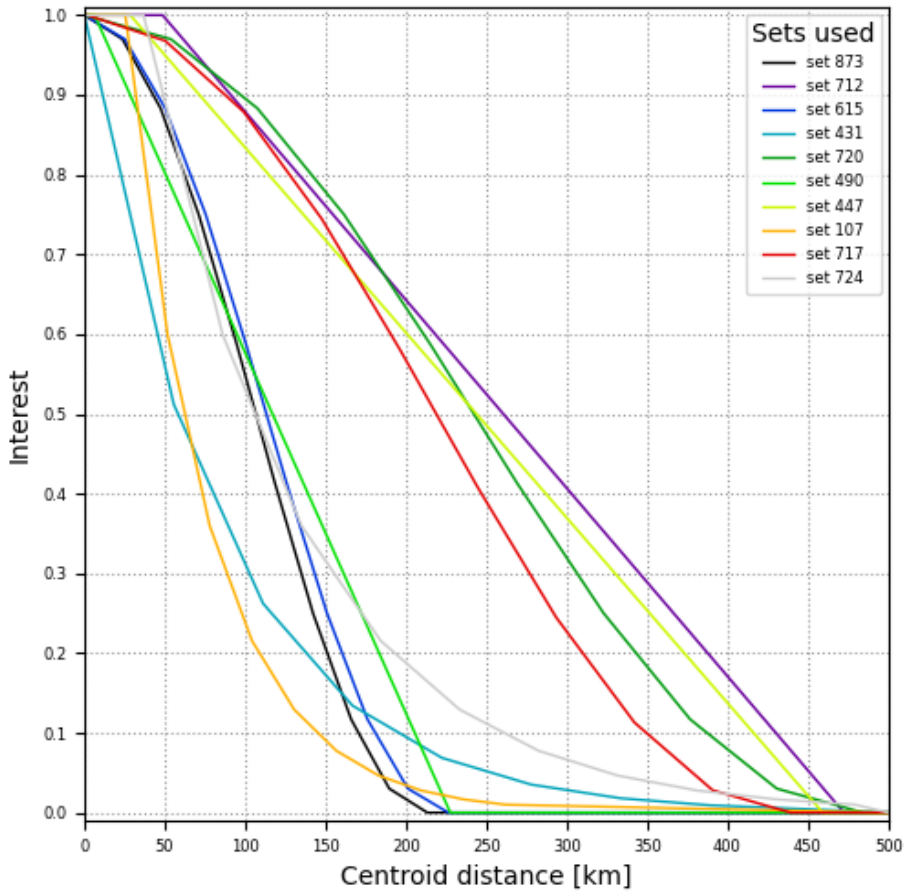
The impact of a stricter set of interest maps manifests as overall lower OTSs in most experiments, but the impact is larger for higher numbers of nonzero attribute interest weights (both equal-CN and support_N)





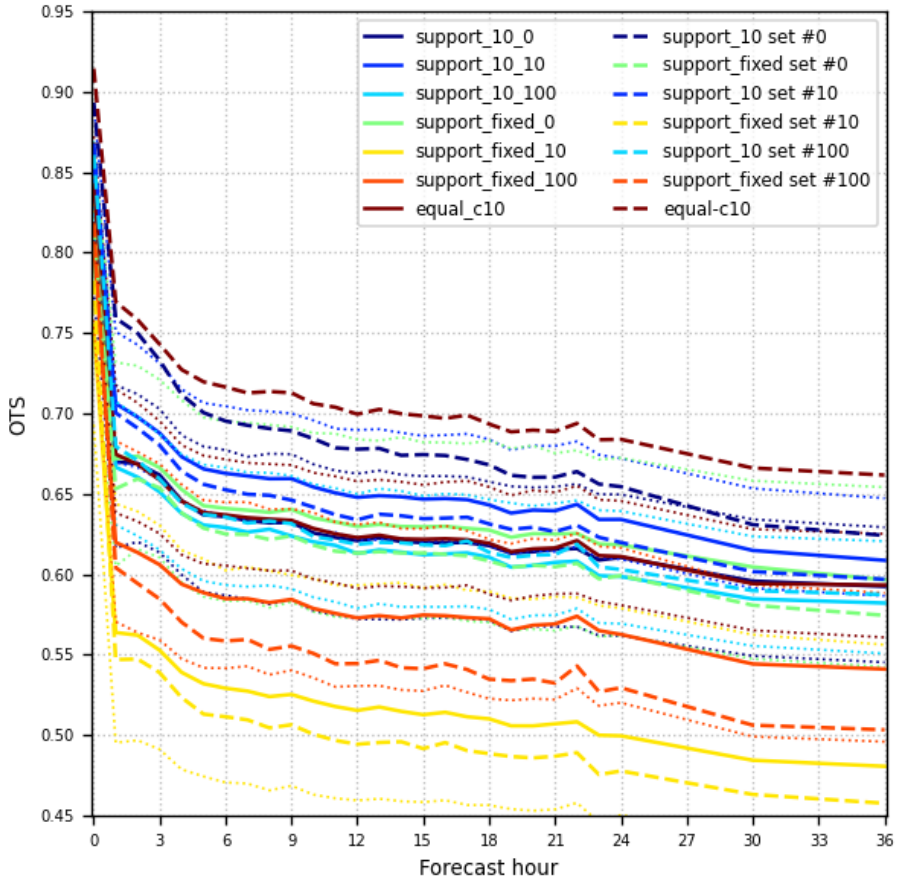
What about testing all sorts of interest maps?

- How does one settle on such an arbitrary choice?
  - One solution: test a variety of maps
- Three functional forms tested:
  - linear
  - exponential decay
  - cosine-squared
- Optional flat spot at 1.0 for “good enough (for perfect)” region of attribute comparison value
- Interest forced to reach 0.0 within some reasonable range (i.e., “not good enough”)



Apply this to all attribute comparisons

- Variability between many test-base pairs larger than variability between interest map varieties
  - But, some central tendency noted
- OTS values overall lower than those using fixed equal or strict interest maps



General findings

- Magnitude of attribute weight means less than whether attribute weight is nonzero or not
- Using varied weights not much different than using equal weighting
- More nonzero weights tend to shrink the uncertainty
- Ultimately, however, the particular attribute weight and interest map configuration will be different for each MODE user, so nearly all possible configuration settings are scientifically legitimate
- So…which set of weights should you use? ¯_(ツ)_/¯ (Choose your own adventure)

So what?

- This testing reveals substantial variability in object-based metrics that depend on the object pair interest value (including, OTS, MMI, generalized metrics presented in Duda and Turner 2023)
- Renders the object-based verification process less certain/less robust to arbitrary decisions made by verification suite designers
  - Especially applies to metrics that can be used as direct comparisons to observations/forecast accuracy
- It offers a different means of statistical significance testing, though
- HRRRv3 vs HRRRv4 differences well within range of values obtained from using reasonable and random sets of attribute weights and maps, so changing the MODE configuration could result in substantially different results

BUT

If the <u>same set</u> of interest weights and interest maps are used, then comparisons between different forecast systems are likely <u>still meaningful</u>