NBM v4.0

Summer 2022 Probabilistic Review

Caleb Steele // WR-STID // 10-Nov-2022

NATIONAL OCEANIC & ATMOSPHERIC ADMINISTRATION | NATIONAL WEATHER SERVICE
WESTERN REGION HEADQUARTERS

STID | SCIENCE & TECHNOLOGY INFUSION DIVISION

The Details

12Z Cycle NBM v4.0 QMD (19Z NBM)

PMaxT*, PMinT*, PQPF24

~5 day lead

Precipitation day = 12Z Day - 12Z Day + 1

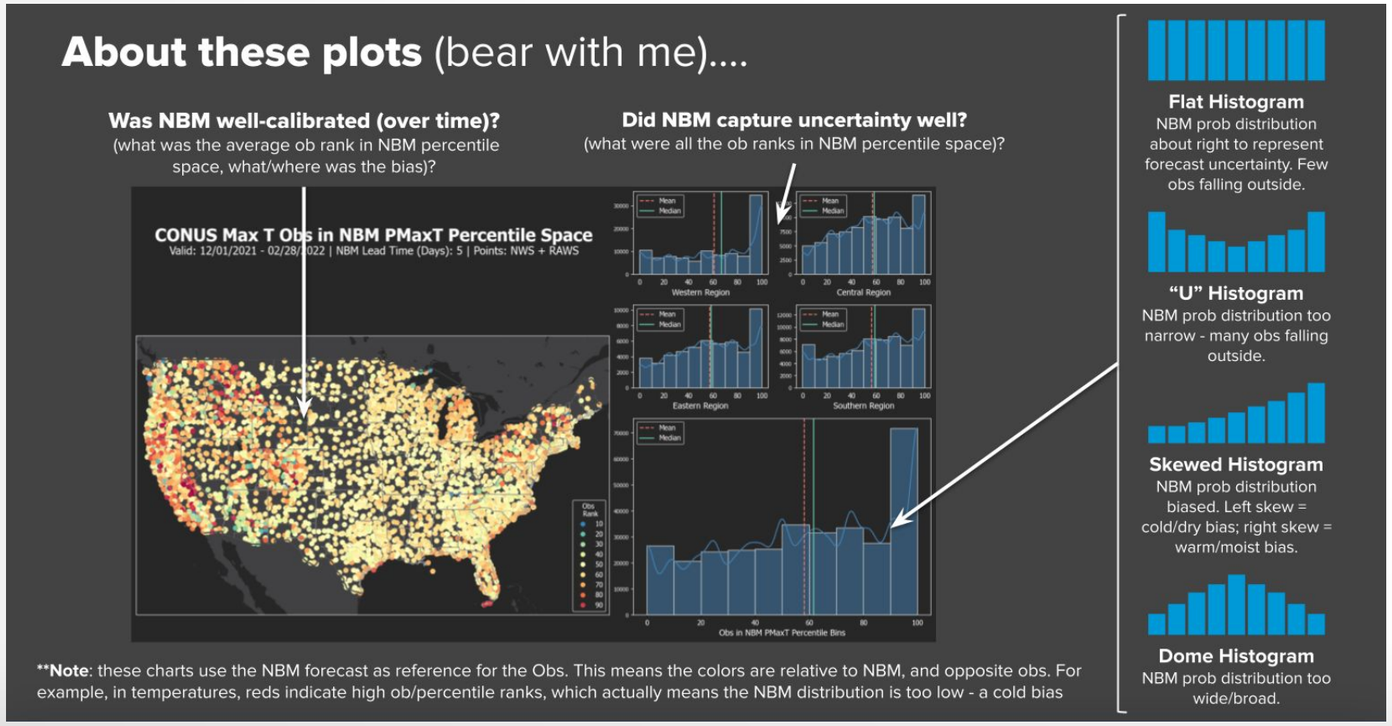
Processed data and figures on Google Drive

Preliminary analysis with unofficial and non-QC'd obs

**NWS 10-201 defines the Max/Min Temp grids as daytime and nighttime respectively (not 24-hr values), so max temp ob search time was 12Z-06Z / min 00Z-18Z*

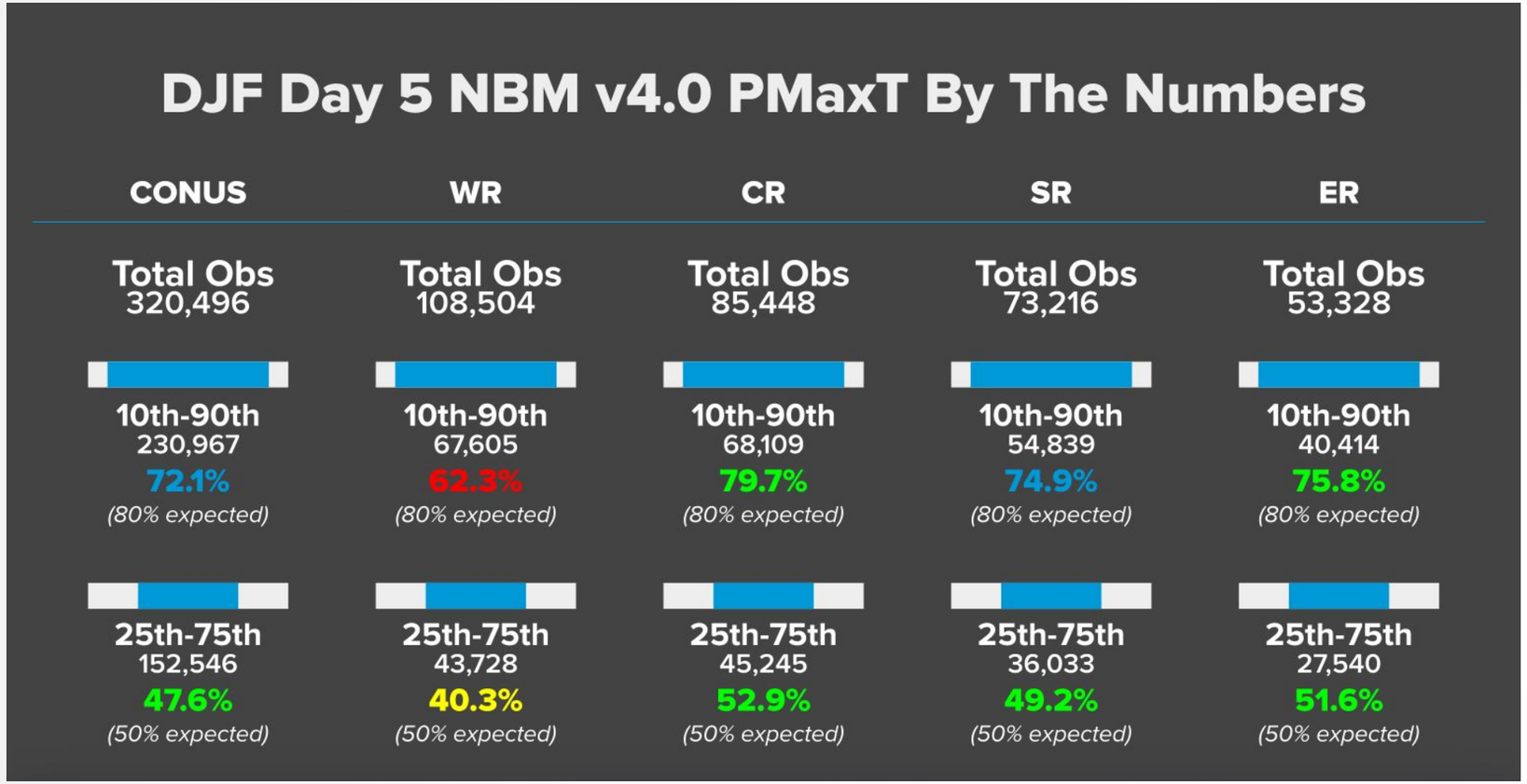
How Probabilistic Data is Verified - Method 1 (Ob Ranking)

1. Request and process observations from synoptic api into dataframe
2. Fetch NBM files from AWS
3. Extract NBM data for stations and add to dataframe
4. Loop over stations and interpolate Ob and/or deterministic value in NBM percentiles
5. Compile stats & visualizations using matplotlib & seaborn



- Rank the obs in the context of the probabilistic distribution. In this case, 1st, 5th, 10th, 20th……80th, 90th, 95th, 99th percentiles extracted and a univariate spline function calculated
- Repeat the process over a long period of time (3 months).
- The mean rank should end up around the 50th.
- The shape of a histogram will tell you information about bias and dispersion.

How Probabilistic Data is Verified - Method 1 (Ob Ranking)



Another way to look at the distribution is the interquartile (25th - 75th) and interdecile (10th-90th) ranges. Here, even skewed/biased distributions can show good stats, especially if the problems are just on the tails.

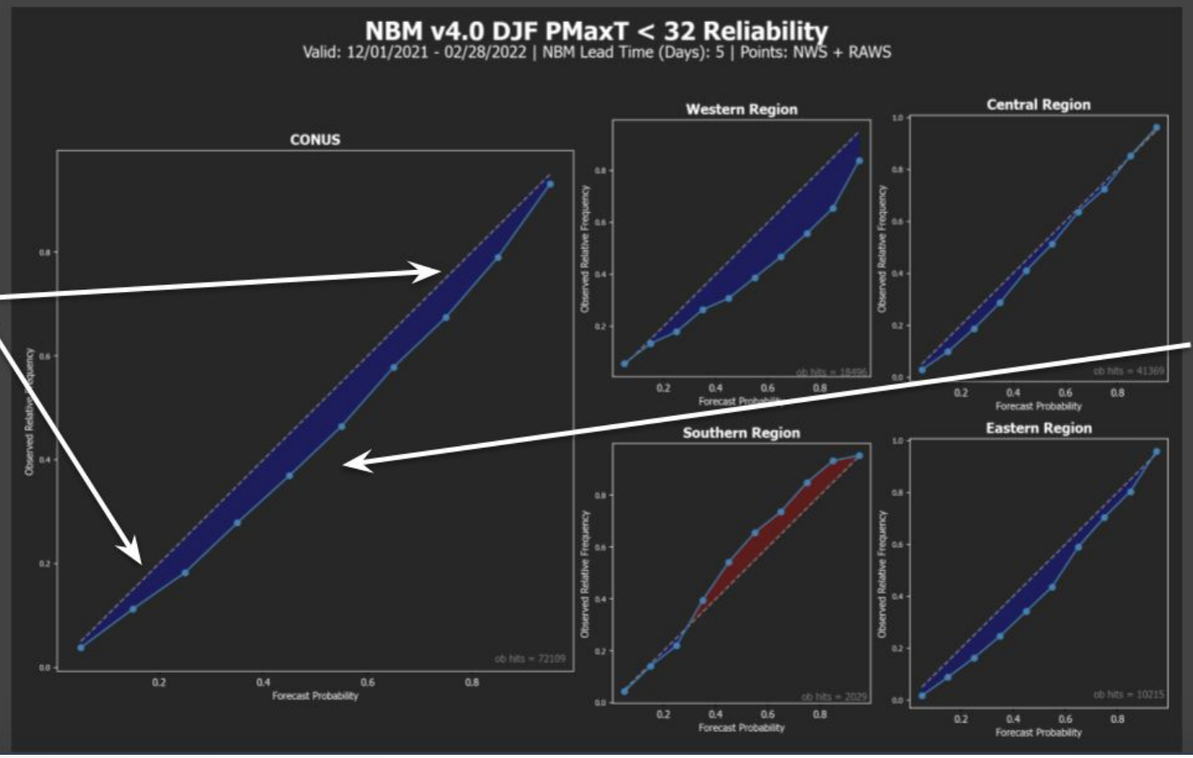
How Probabilistic Data is Verified - Method 2 (Reliability)

About these plots (continued)....

Were NBM forecasts reliable?

(did a 20% chance really only happen 20% of the time)?

The dashed diagonal grey line is perfect calibration (20% forecast happens 20% of the time, etc.)



The light blue line with points is NBM reliability - x,y pairs of forecast bins centered at the 5s (ie the point at 0.05/5% is 0-10% and so on) with verified frequency. Which side of the line it is on can mean different things depending on if you are verifying a probability of exceedance or a probability of non-exceedance. Shading has been added to help with interpretation (blue shading = cold bias, red shading = warm bias, brown shading = dry bias, teal shading = wet bias)

1. Bin probabilities of non/exceedance.
2. Pair with ob hits to get relative frequency
3. Make sure sample size is large (ob hits)
4. Relative frequency should closely match forecast frequency (diagonal line)
5. Exceedance or nonexceedance choice determines which bias is on which side of the line

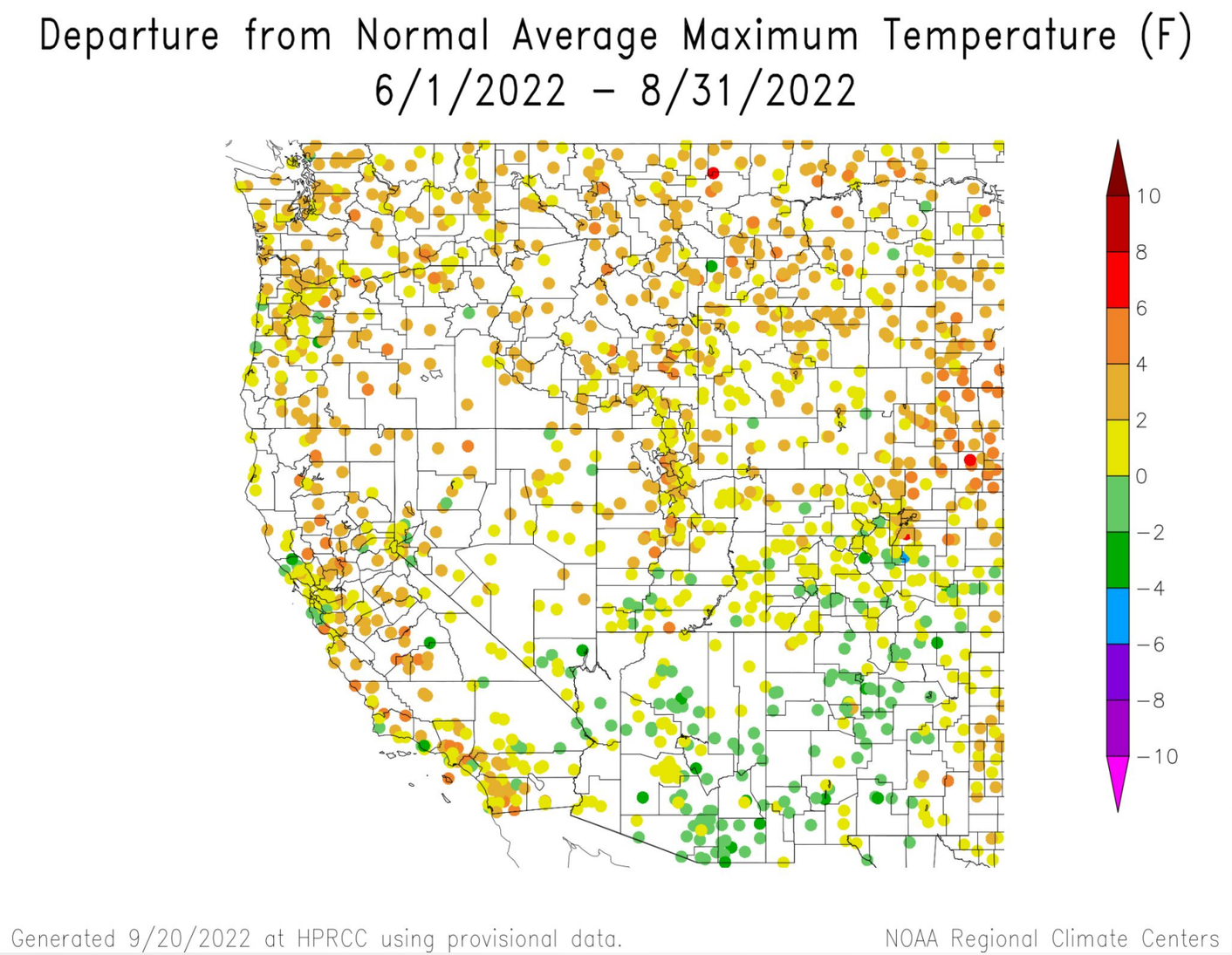
NBM 4.0 Day 5 PMaxT

1,307 Obs (NWS + RAWS) with obs spanning every day | **92** days = **120,244** sample size

Missing: NONE!

Confidence in analysis: **HIGH**

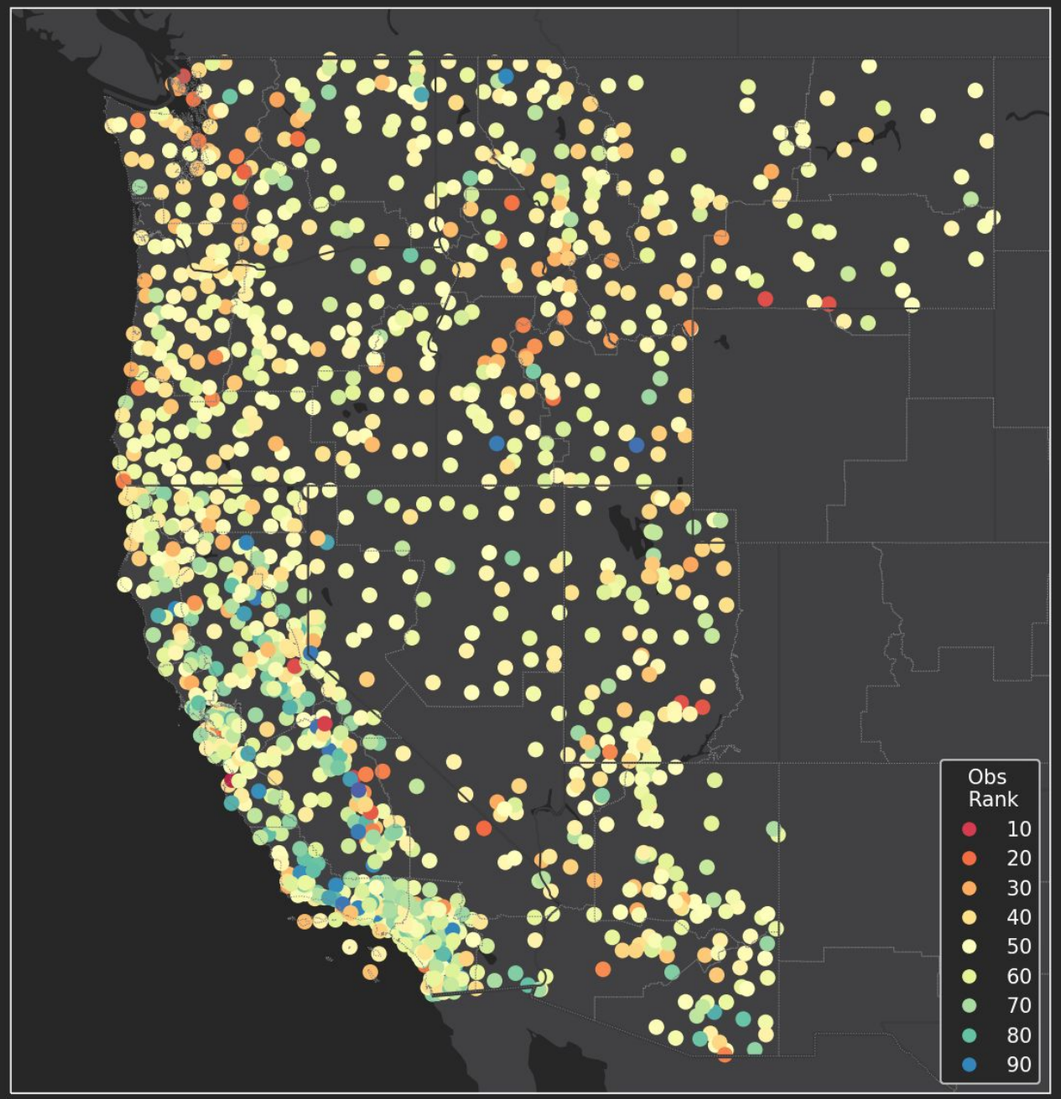
Jun-Aug MaxT Summary

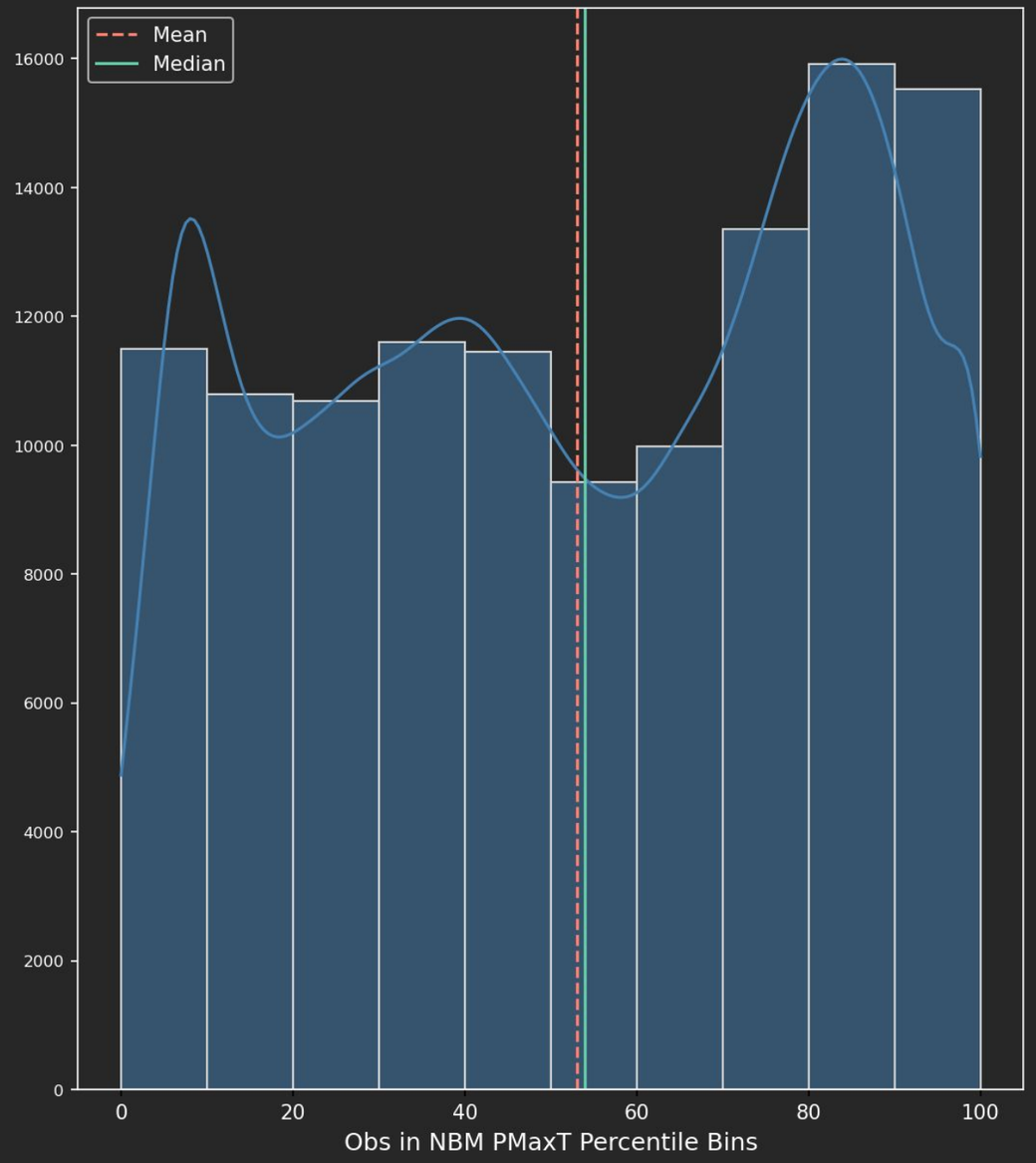


- On the cooler side in active monsoon areas
- Near to above normal everywhere else

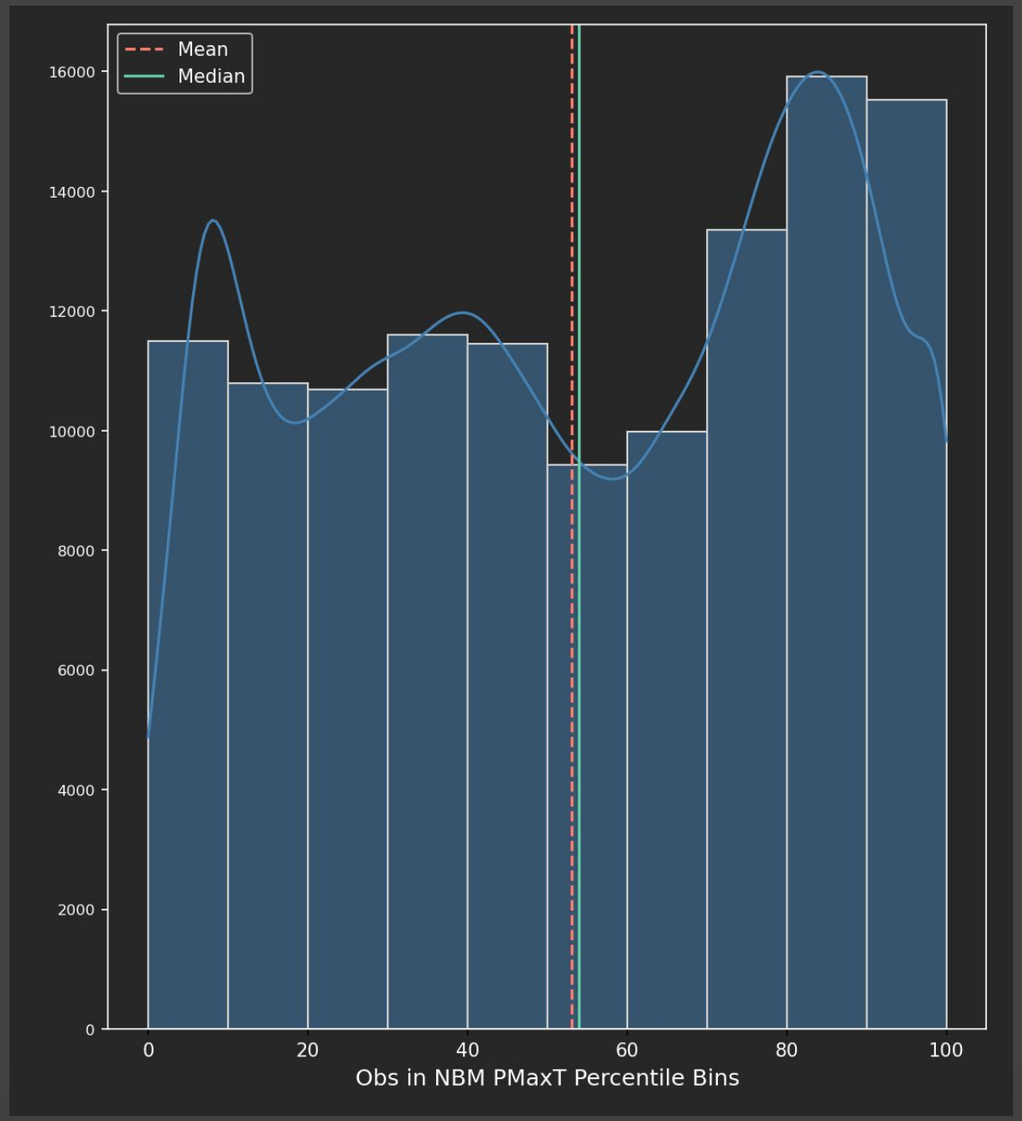
WR Max T Obs in NBM PMaxT Percentile Space

Valid: 06/01/2022 - 08/31/2022 | NBM Lead Time (Days): 5 | Points: NWS+RAWS





Summer 2022 Day 5 NBM v4.0 PMaxT By The Numbers



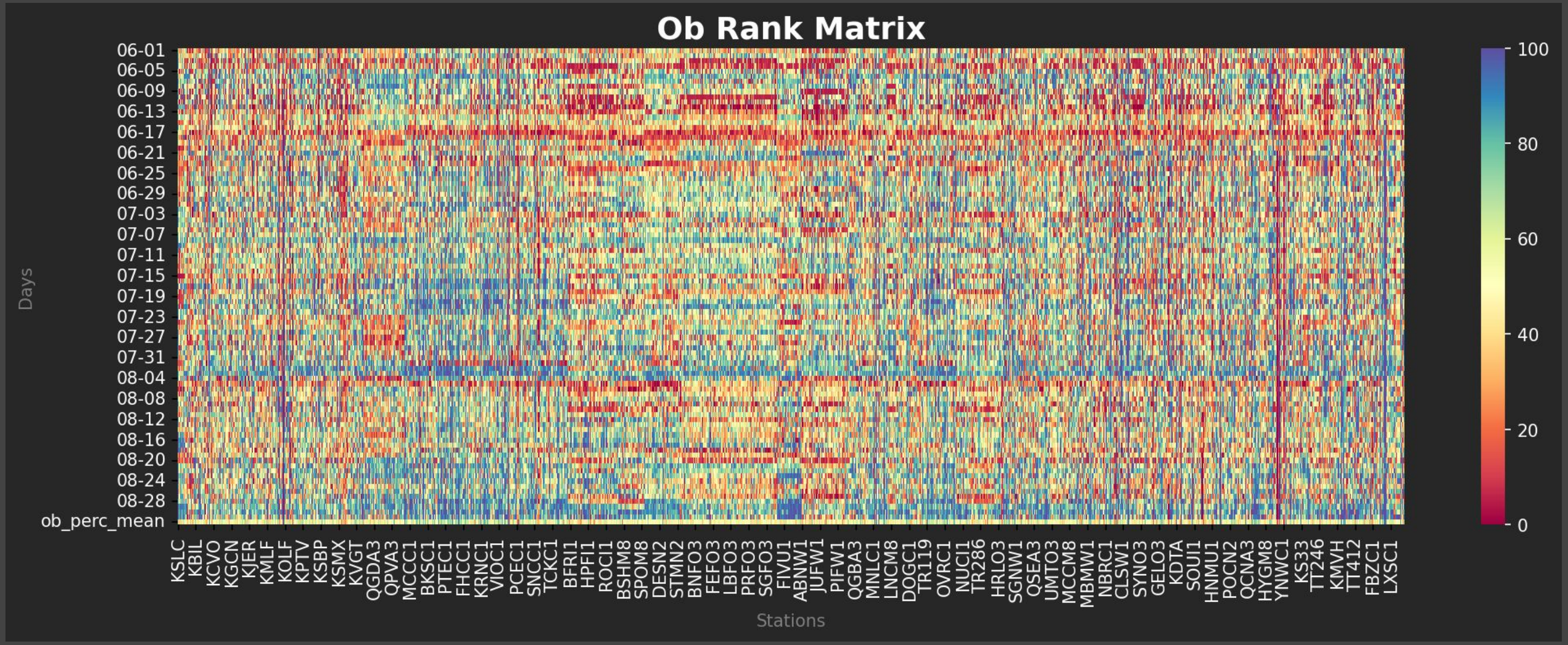
WR

Total Obs
120,244

10th-90th
94,672
78%
(80% expected)

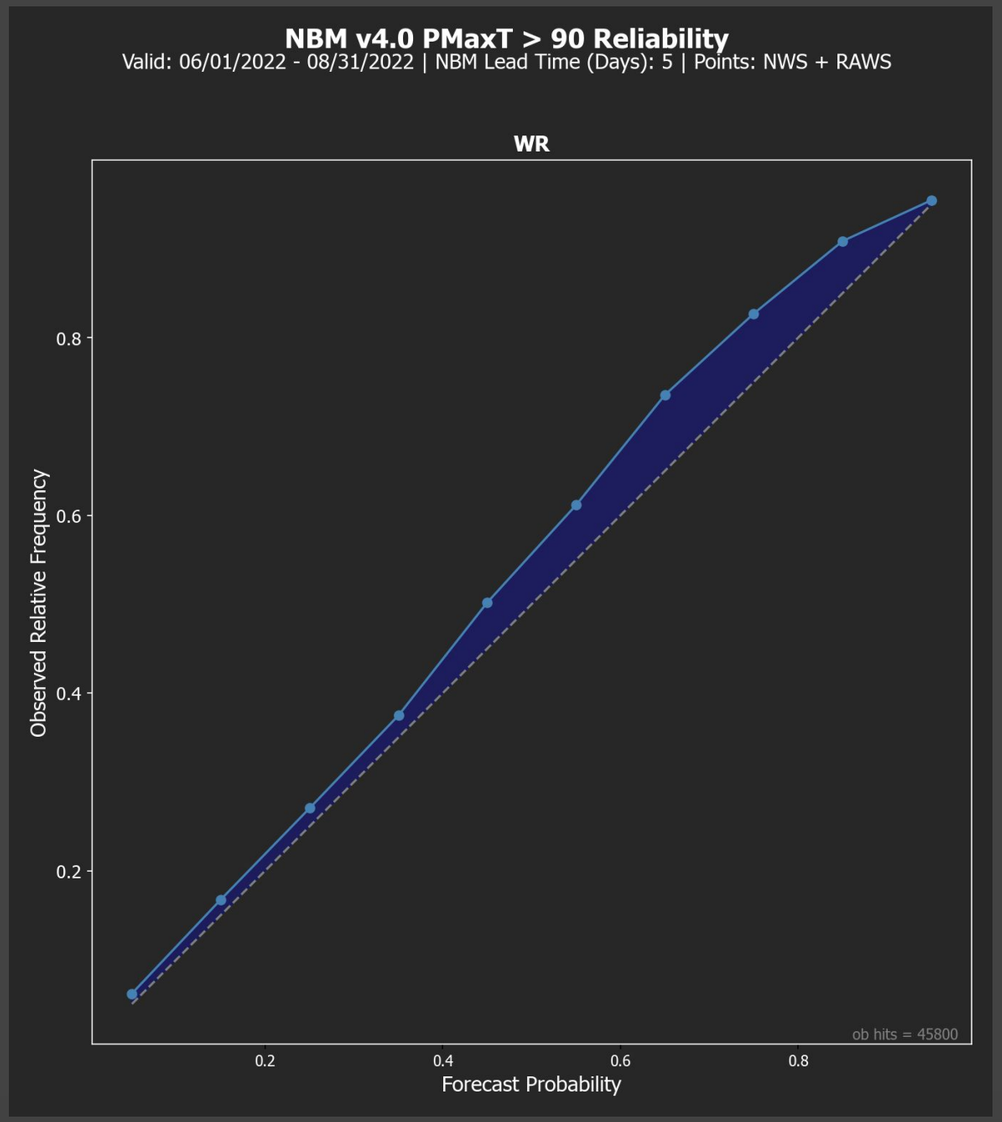
25th-75th
55,355
46%
(50% expected)

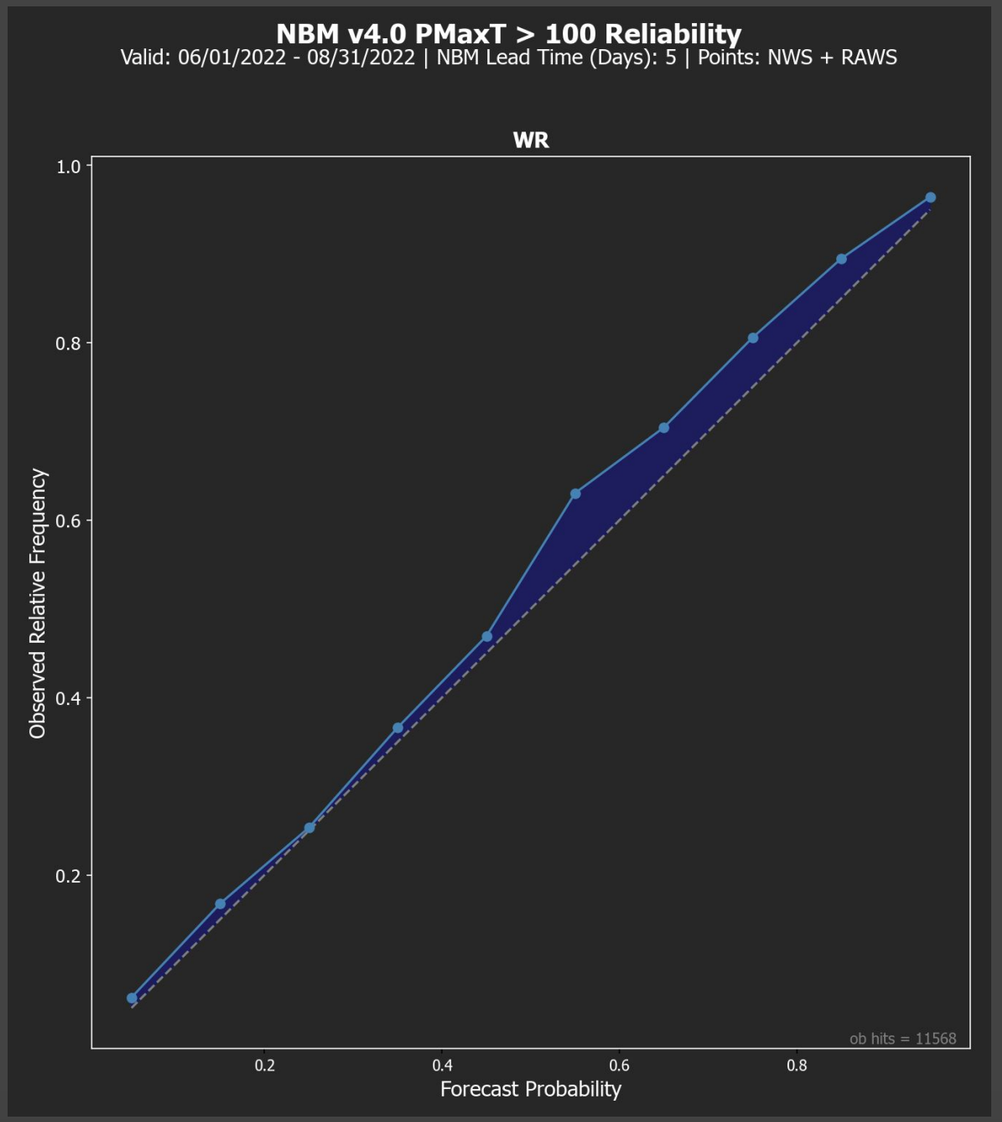
- On the previous slide, a lot of yellow on the map, which is good, with no spatial groupings of colors indicating any systemic bias (though a few ob sites to look at).
- On the histogram, there is not much of a “U” shape, which is good, and indicative that NBM is representing the spread well, with an appropriate number of obs falling outside the overall distribution. (We want that histogram to be flat)
- The histogram does show a *left skew*, indicating an overall **cold bias**. The mean and median are also displaced upward from 50, confirming the bias in the NBM distribution, though this is **very slight** (less than 5%).
- The amount of obs falling inside the interquartile range (25th-75th) of the NBM distribution was only 4% off our expected 50%. Observations falling within in the interdecile range is even better, only missing the mark by 2% - which indicates the bias was slight enough to still be contained within the bounds of the probabilistic distribution.



This plot once again looks noisy, which is good/desired. Rows of color are usually weather events, and columns of color are either persistent problems with the forecast, or with the ob. We could use the rows to pick out weather events to look at, but we'll already know what the map will look like, so instead, we'll try to start from impacts to identify cases for closer looks.

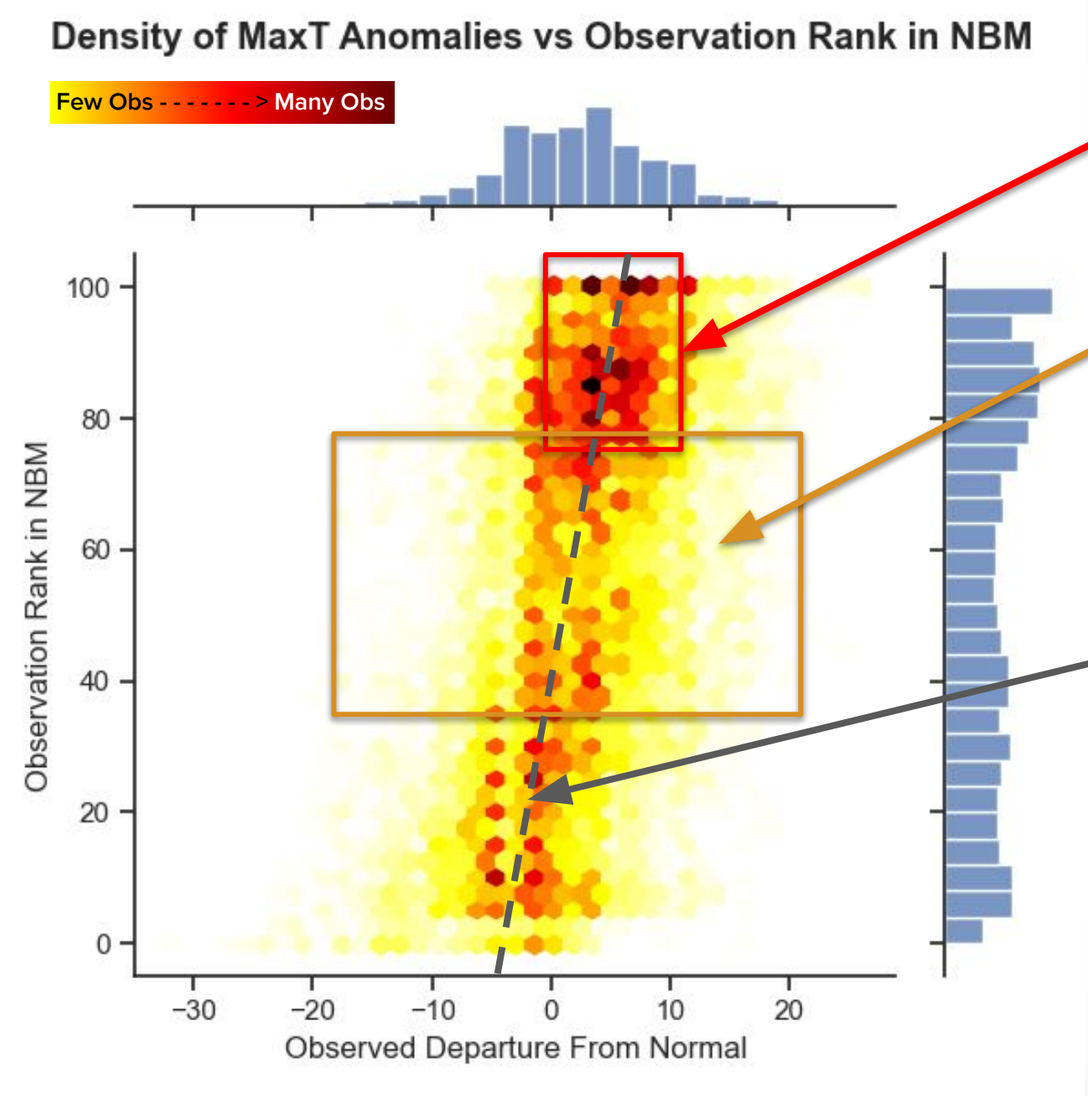
Day 5 NBM PMaxT Threshold Reliability





The reliability for both PMaxT > 90 and >100 (both shown off the to the left) for Western Region was quite good, with the NBM line closely following the diagonal. We do notice a slight **cool bias**, with observed frequencies slightly higher than forecast frequency, but this bias is **very slight**, only about a 5% bias at the most on these curves.

Did the NBM Do Well With Extremes in MaxT?



Highest concentration of obs above the 75th percentile. Also generally with temperatures near/slightly above normal (-2 to +10 departure from normal).

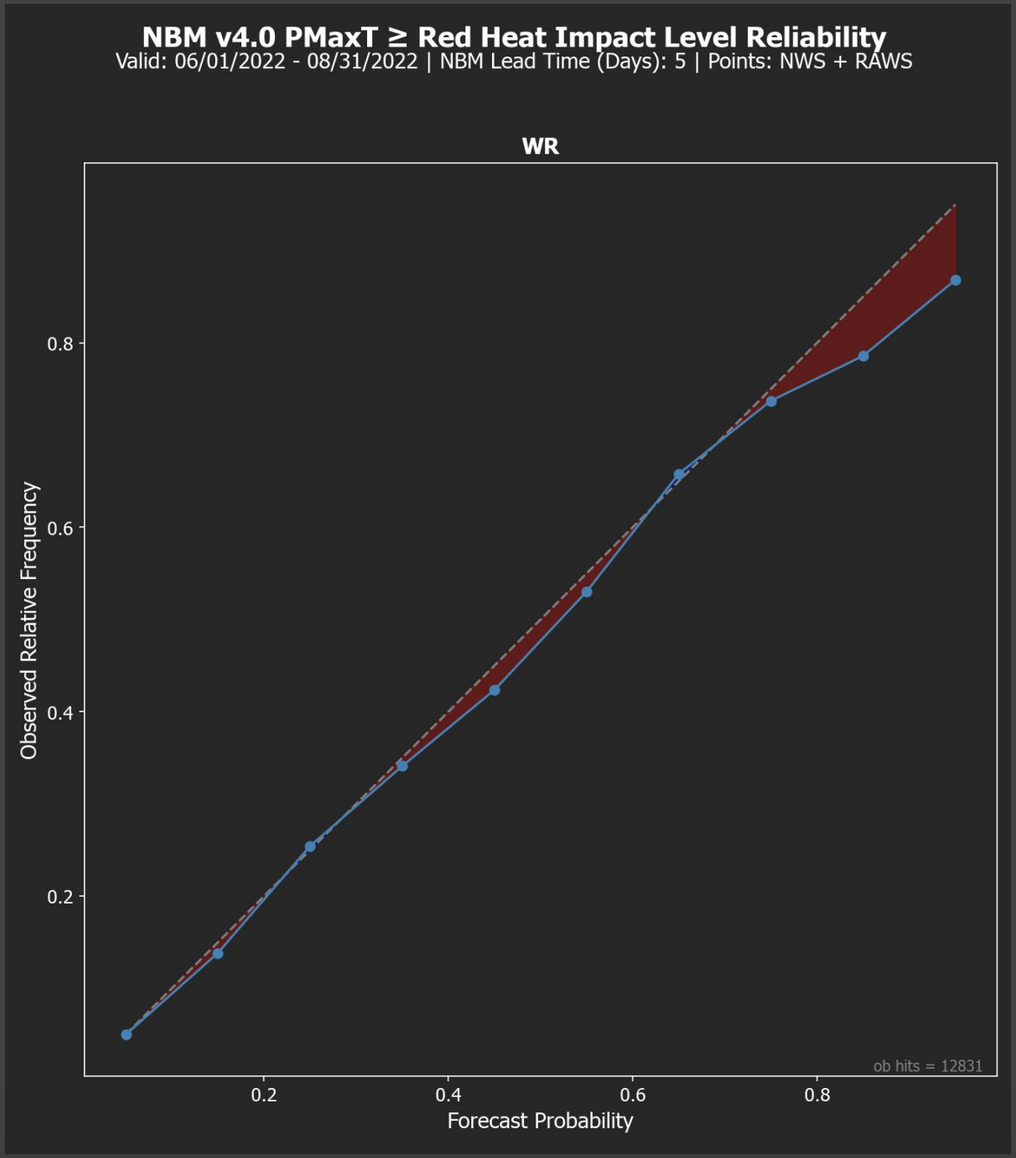
While the vast majority of these obs remained near climatological values (center of x axis), if we look at the interquartile range of the ob ranks (between 25 and 75 on the y axis), at least some extreme obs (+/- 20 degrees from normal) were still contained in the NBM interquartile range.

With proper spread, we expect the grey dashed line [rough approximation of a linear regression] to be something like a vertical indicating good spread. The positive tilt indicates that 1) warm observed max temperature anomalies were far more common than cold ones (notice the histogram at the top is centered to the right of 0), and 2) with those positive temperature anomalies came high NBM ob ranks (that is, a cool bias, with more extreme obs tending to fall above the 80th percentile or even beyond the NBM distribution).

In a bulk sense, the NBM did *okay* but there is room for improvement. In general, positive observed temperature anomalies fell above the 80th percentile. In other words, when it was warmer than normal, the NBM tended to be too cool. Keep in mind any URMA vs station biases (probabilistic data in 4.0 is tuned to URMA, not METAR).

Limited to stations that start with "K" AND had climo data available on acis: ***n=146 stations / 13,524 obs***

Did the NBM Do Well With Extremes in MaxT?

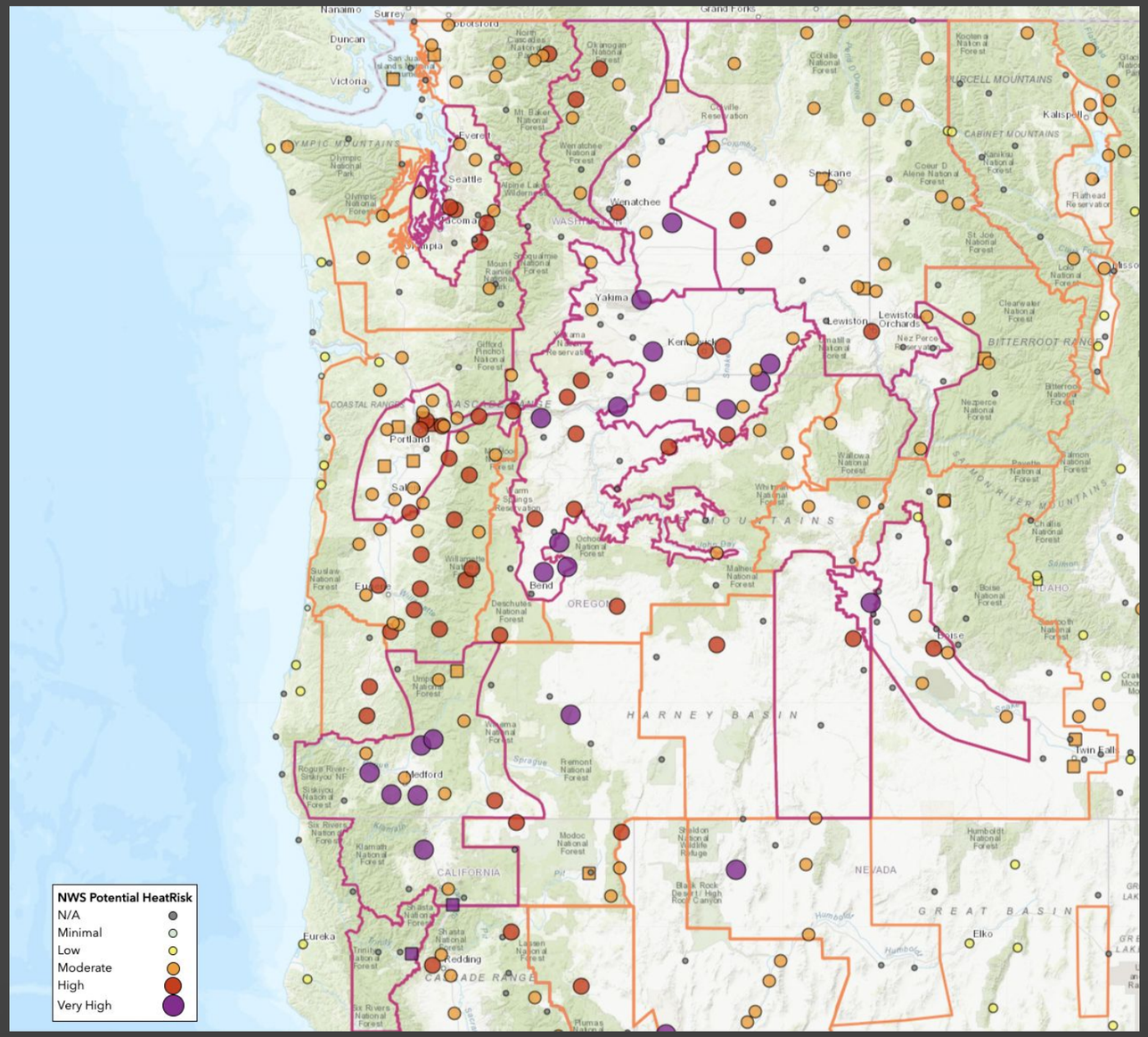


Another way we can measure performance in extremes, is asking "**how well did the NBM do when it mattered most?**"

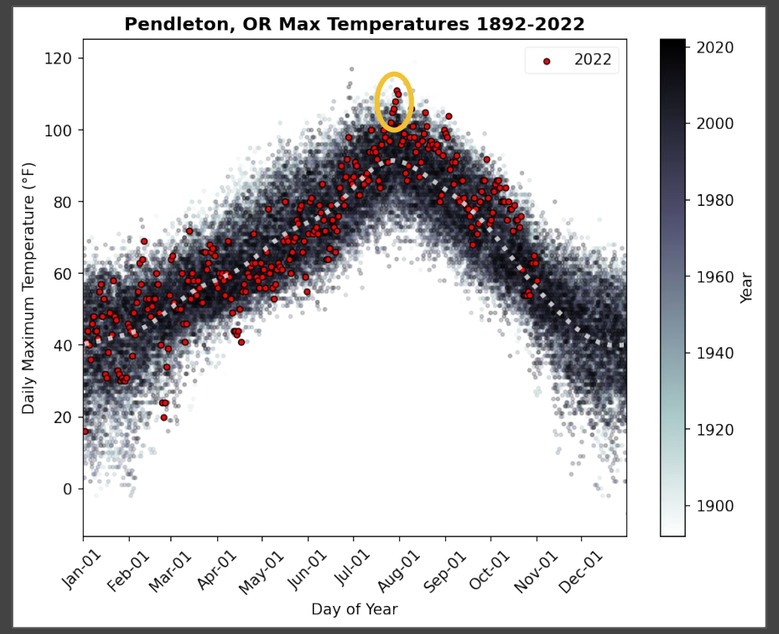
HeatRisk is a powerful impact-based messaging tool, so why not measure NBM reliability at HeatRisk thresholds (in this case, Heat Impact Level, since we are only looking at one out of the three components of daily HeatRisk)? That is just what is done here. Instead of a static threshold for everywhere, all the time, the threshold measured varied by space and time based on gridded Red Max T Heat Impact Levels. Once again, Heat Impact Level is a single component of HeatRisk. HeatRisk is a 24 hour value that incorporates the day's MaxT along with the shouldering MinTs.

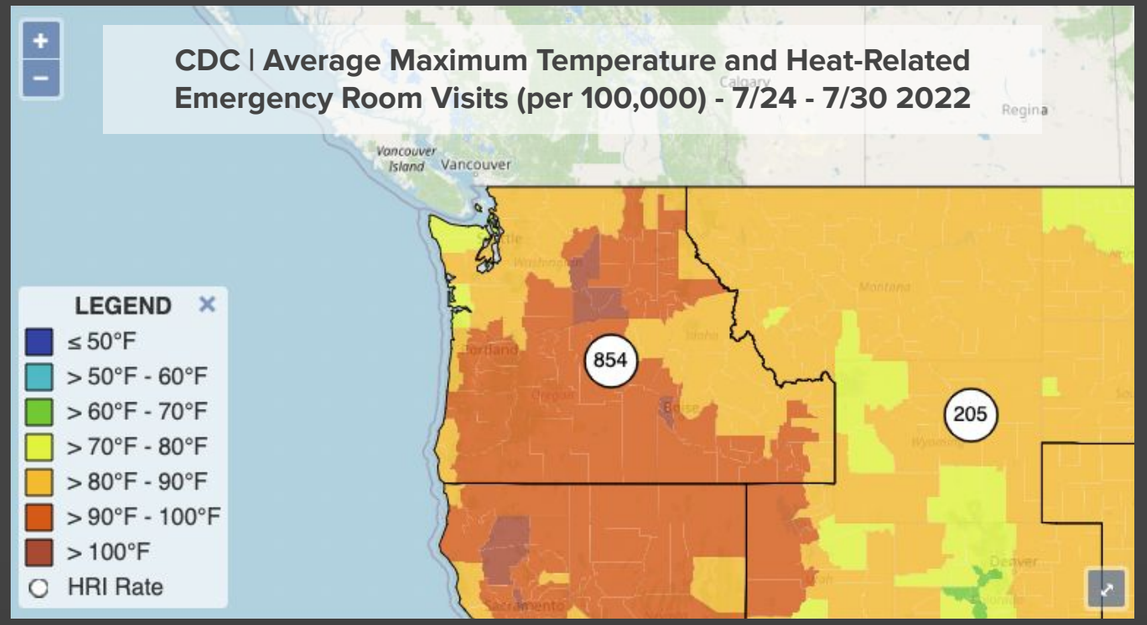
As is turns out, we see **near perfect calibration!** There are a few areas at higher probabilities where a very slight (<5%) warm bias starts showing, but you would be hard pressed to find a better reliability curve on live data. But this is a different story than the previous slide, right? Well, verification is complicated. One number does not tell the story. With a threshold probability, there is no sensitivity to how much you crossed the threshold by, just that you crossed it (threshold verification is binary: a yes or a no). If we go back to the original question, **the NBM does very well for probabilistic Heat Impact Level Red Max T thresholds**.

Late July Heatwave | 30-July-2022



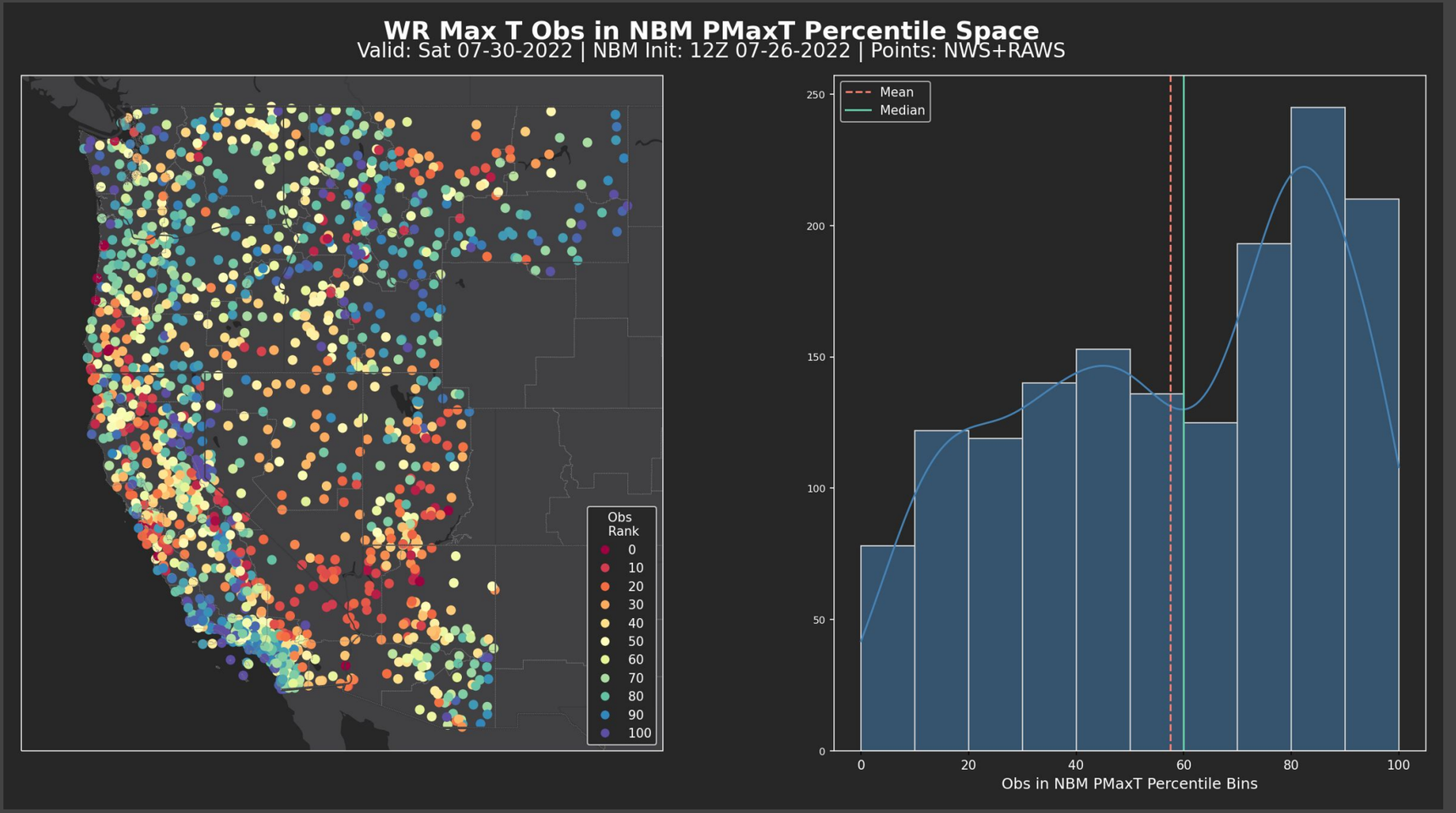
A heatwave ramped up during the last week of July for northwestern portions of the region with 126 stations setting new daily records across Washington and Oregon. Observed HeatRisk had widespread areas of Magenta/Very High and there was an accompanying uptick in heat-related emergency room visits.





Late July Heatwave | 30-July-2022

So how did the NBM do? In this first plot below, we can see some yellows, but also a lot of greens and blues in the Pacific Northwest, which would suggest obs falling above the 50th percentile, but not terribly so. This isn't bad in and of itself.

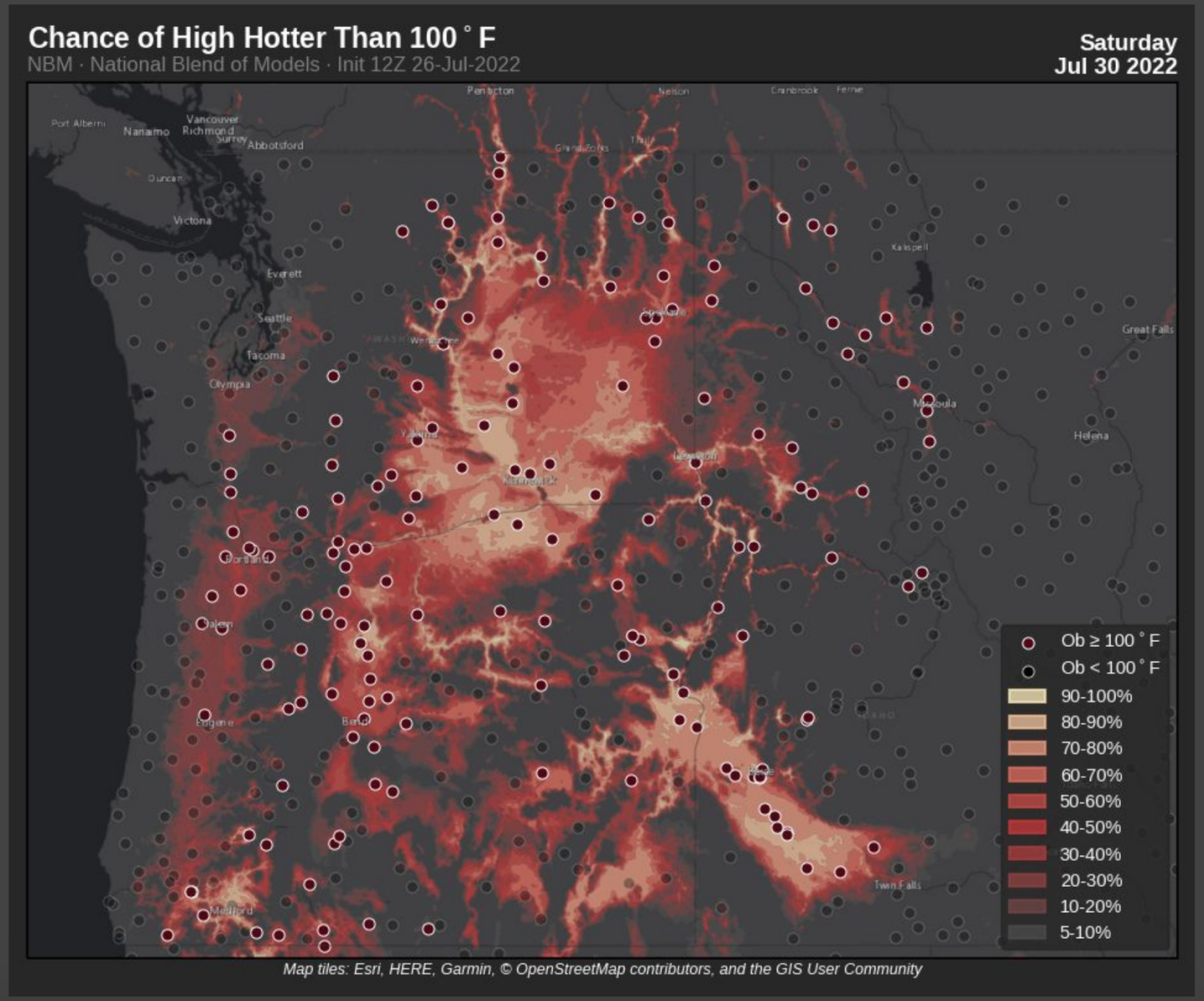


Make this plot in Google Colab!

Late July Heatwave | 30-July-2022

This second plot shows a threshold of 100 degrees, with the NBM probability shaded, and the ob hits and misses in dots. We can see the NBM actually did a pretty good job of the footprint of temperatures hotter than 100. And while probabilistic forecasts cannot be verified with a single time, perhaps the coverage of the reds and yellows is a little low - especially across Oregon and Idaho (though a lot of these are valley locations).

Make this plot in Google Colab!



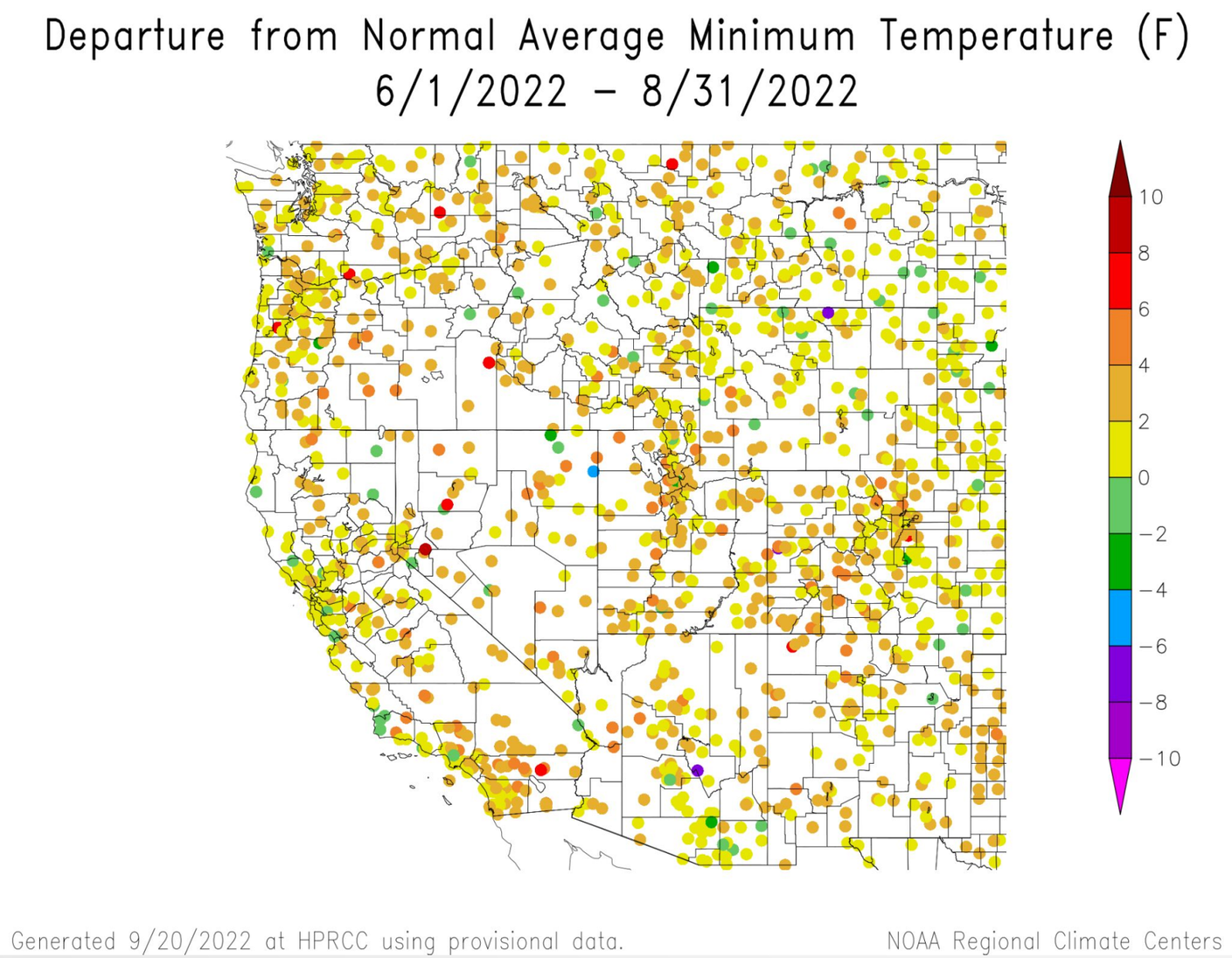
NBM 4.0 Day 5 PMinT

1,298 Obs (NWS + RAWS) with obs spanning every day | **91** days = **118,118** sample size

Missing: 7/12 (grib files incomplete)

Confidence in analysis: **HIGH**

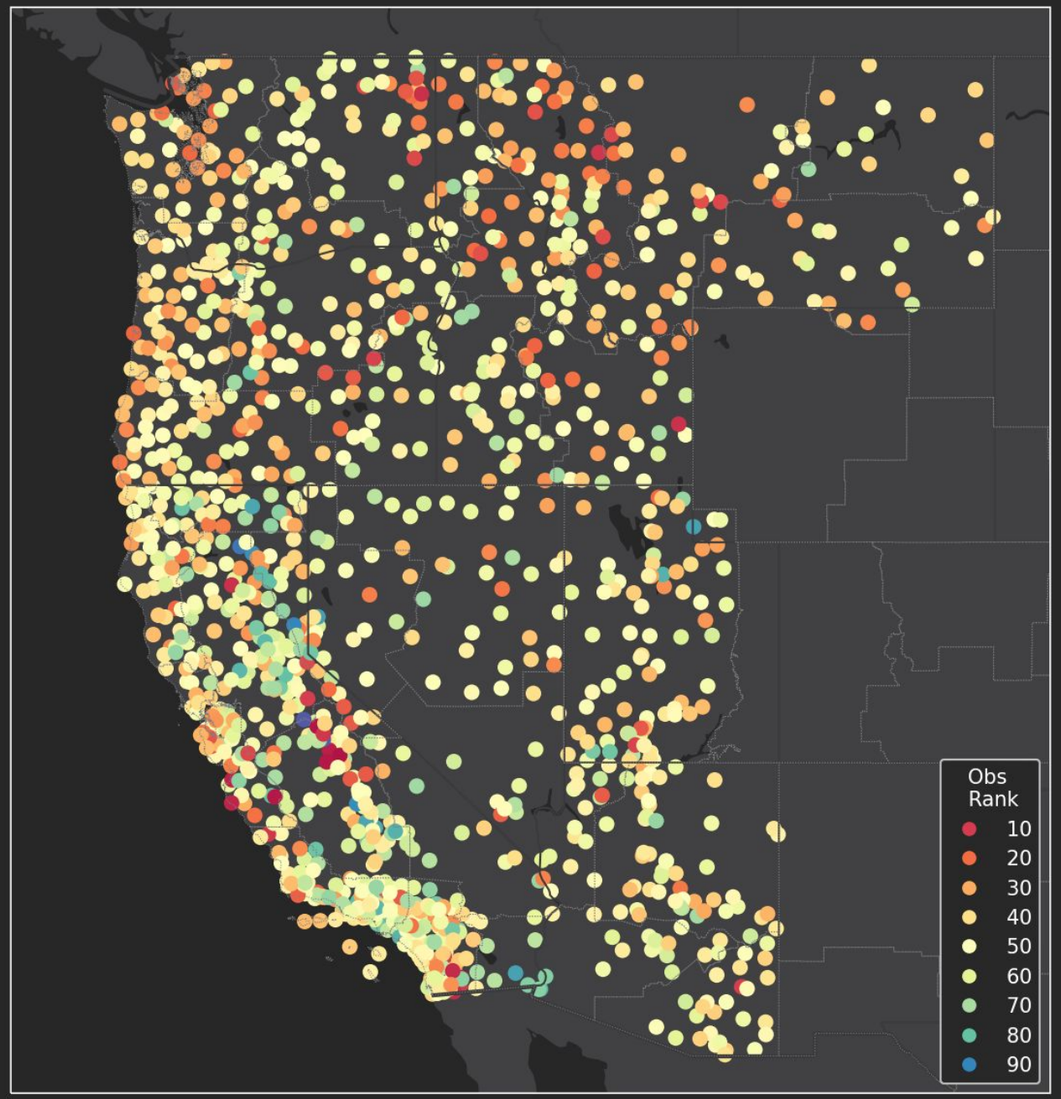
Jun-Aug MaxT Summary

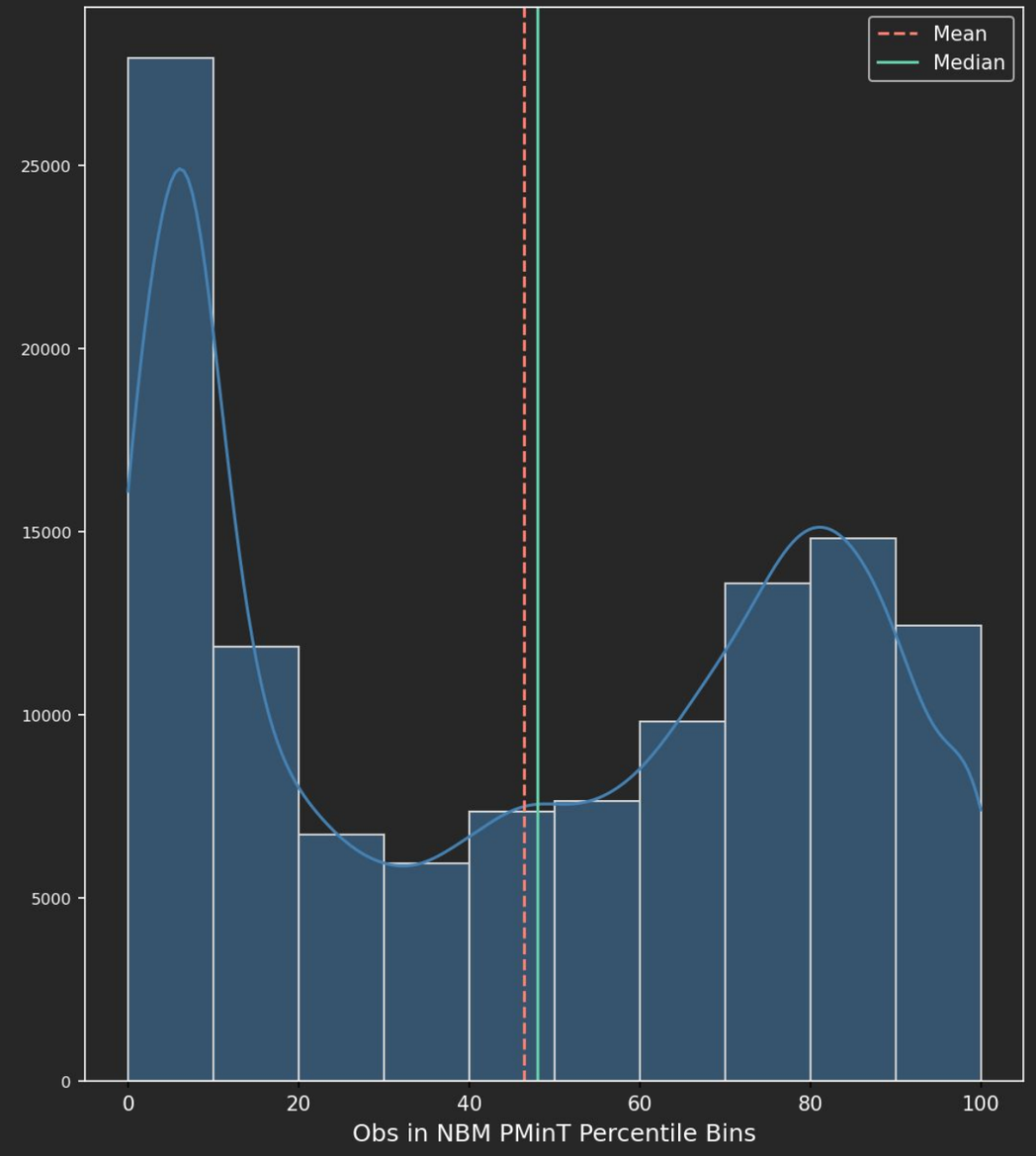


- Widespread near to above normal lows region-wide

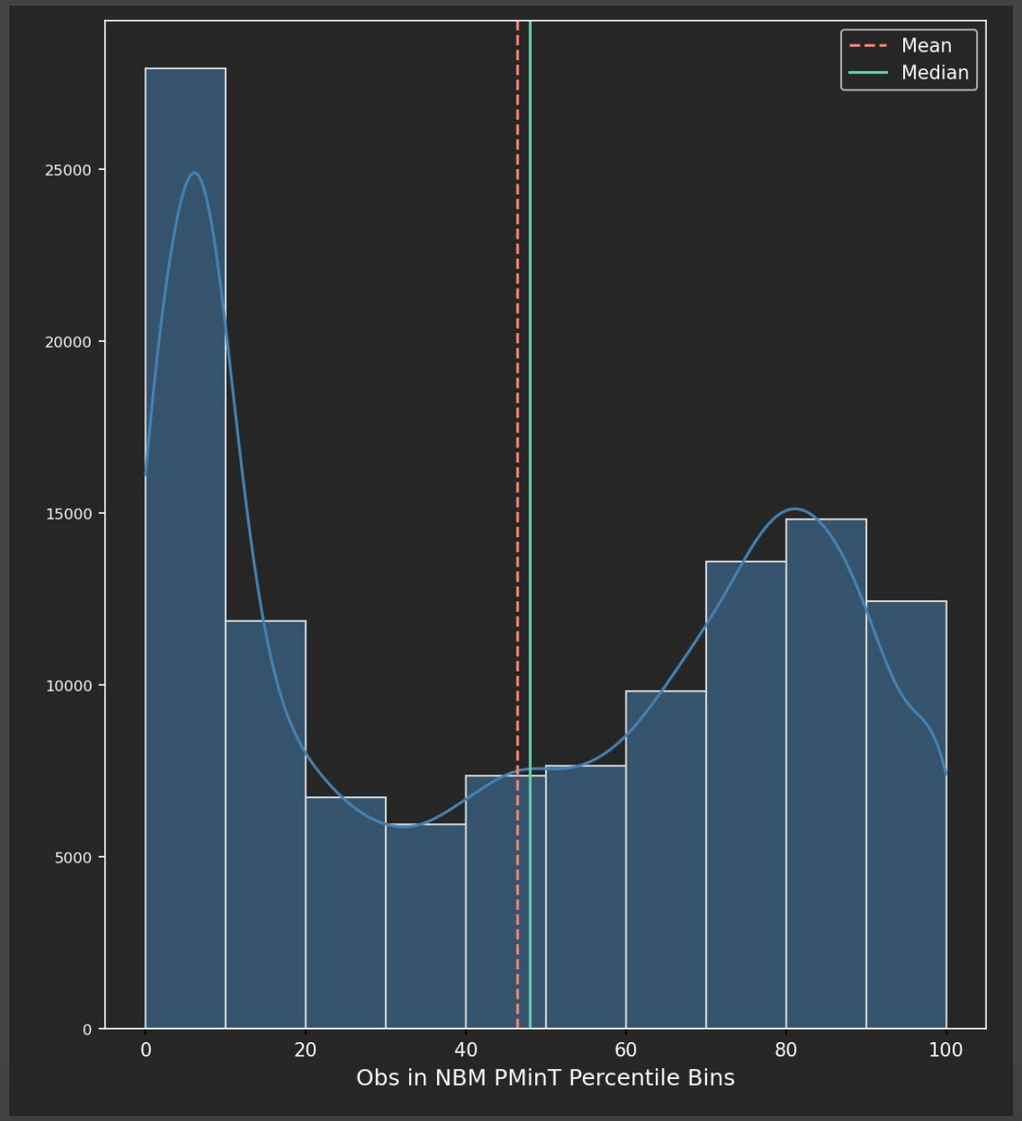
WR Min T Obs in NBM PMinT Percentile Space

Valid: 06/01/2022 - 08/31/2022 | NBM Lead Time (Days): 5 | Points: NWS+RAWS





Summer 2022 Day 5 NBM v4.0 PMinT By The Numbers



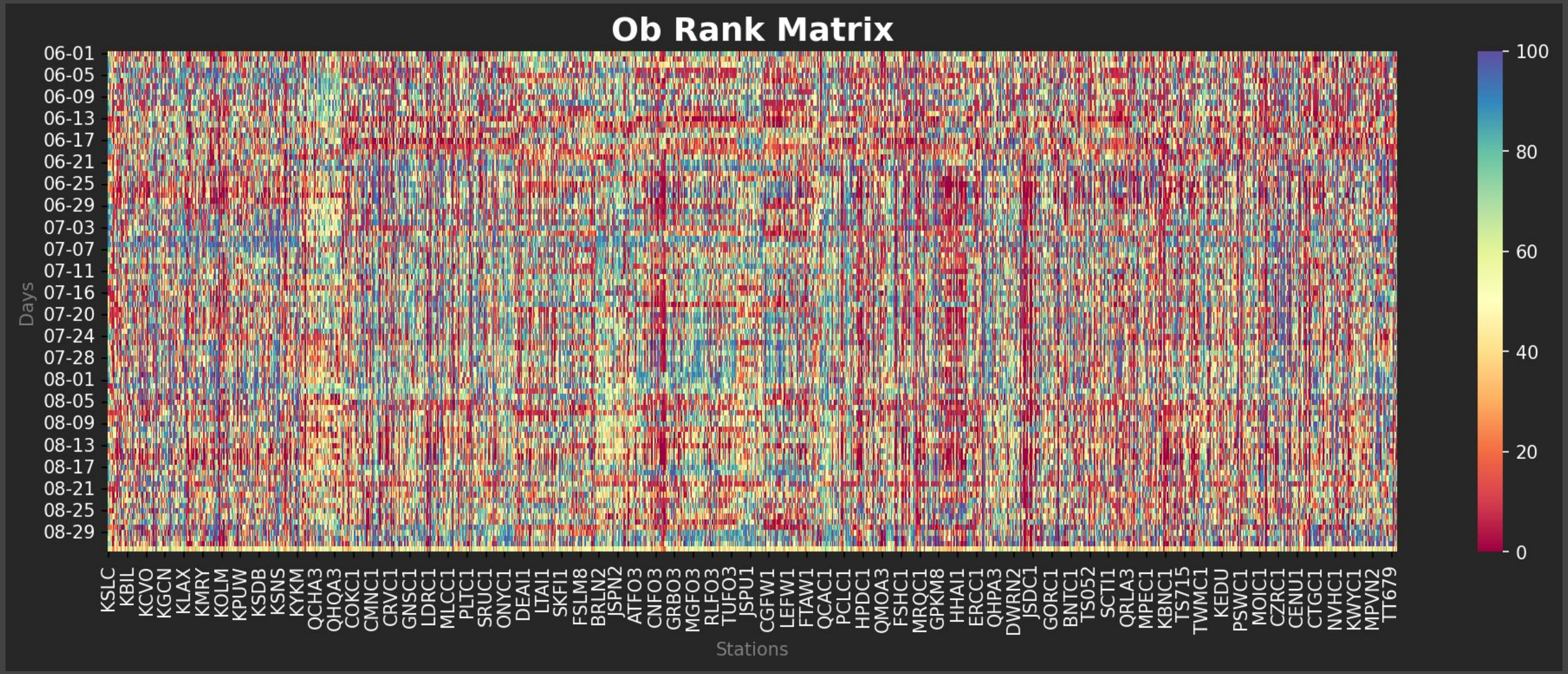
WR

Total Obs
118,118

10th-90th
78,924
67%
(80% expected)

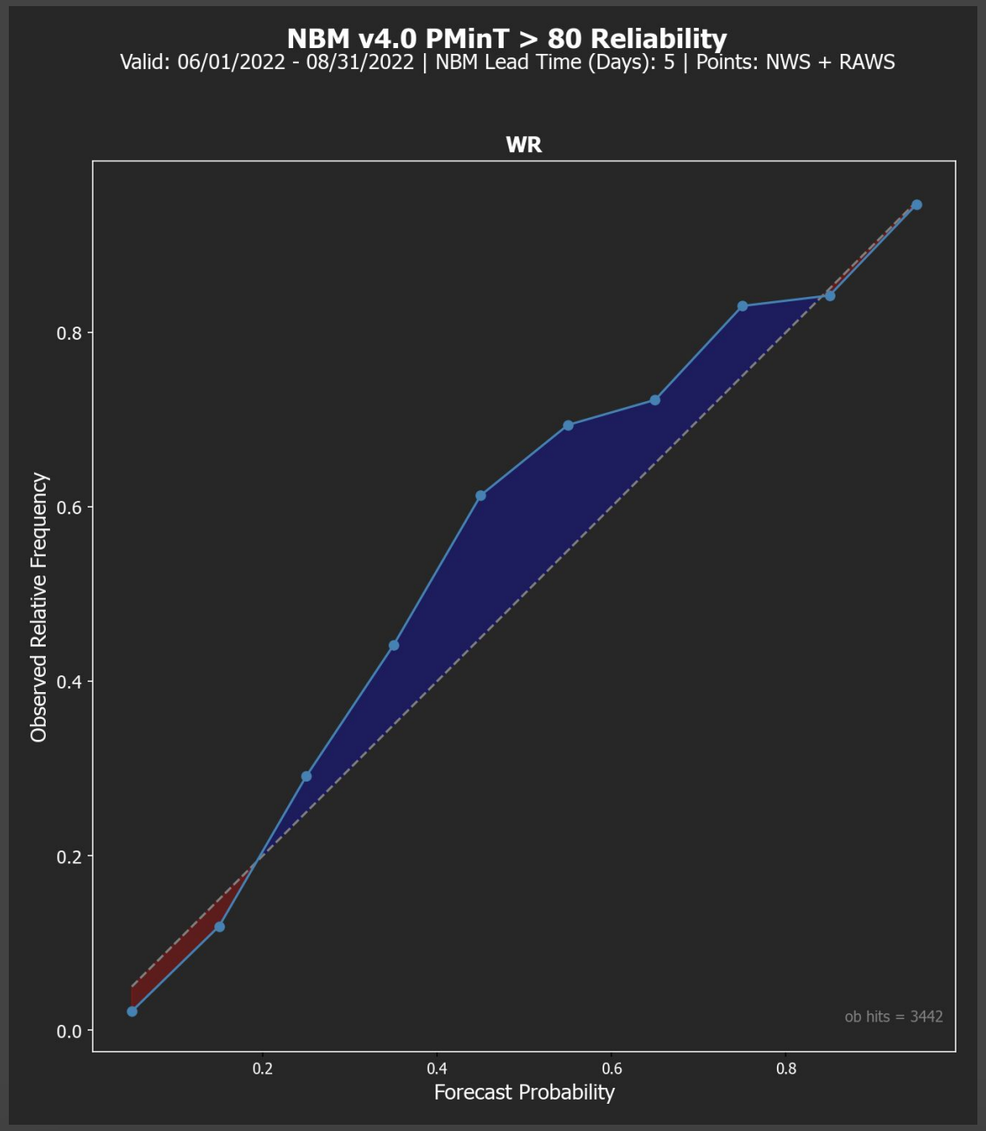
25th-75th
41,355
35%
(50% expected)

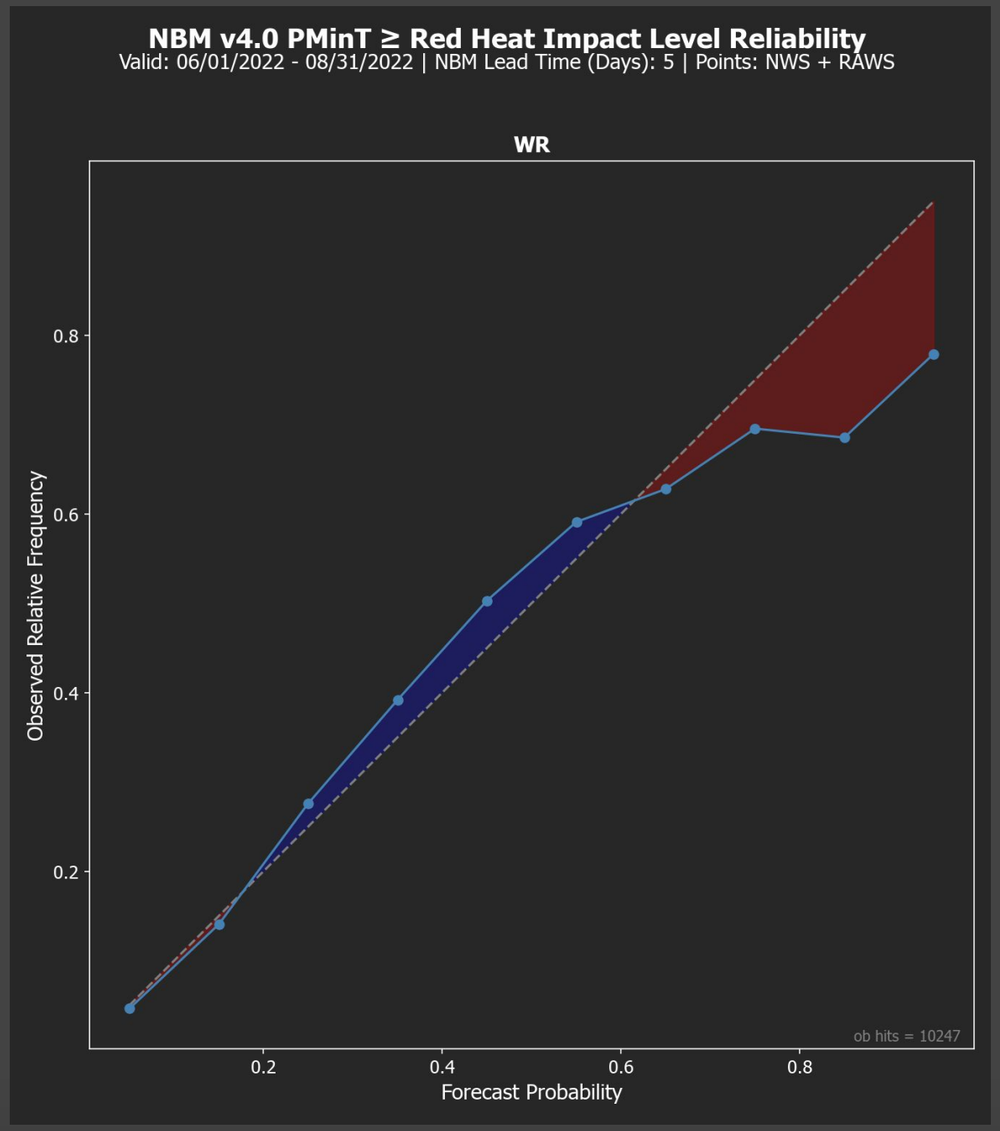
- On the previous slide, a lot of yellow on the map, which is good, but the histogram shows this is an average of extremes.
- On the histogram, we see both a half “U” to the left, as well as a left skew, with a secondary mode just above the 80th percentile. The next plot, the ob rank matrix, should let us know if there were persistent problem stations, but from the map, the central Sierra looks like an issue (which suggests a thermal belt issue?).
- As mentioned before, the histogram is bimodal, which makes it a bit more difficult to interpret. The left skew on the right side indicates a cold bias, but the obs falling outside of the bottom of the distribution also indicates a warm bias; so it is a very mixed bag.
- The amount of obs falling inside the interquartile range (25th-75th) of the NBM distribution was 15% off our expected 50%. Observations falling within in the interdecile range is not much better, still missing the mark by 13% - which again illustrates issues with obs falling either to the extremes or completely outside the NBM distribution (an uncontained bias).



This plot once again looks noisy, and the bottom row has a lot of yellow, but in this case, that comes from an average of extremes - with a lot of blue and red up above the bottom row, with some exceptions. We can also see quite a few columns of color - especially in California and Oregon RAWS stations. (the underlying data is here)

Day 5 NBM PMinT Threshold Reliability



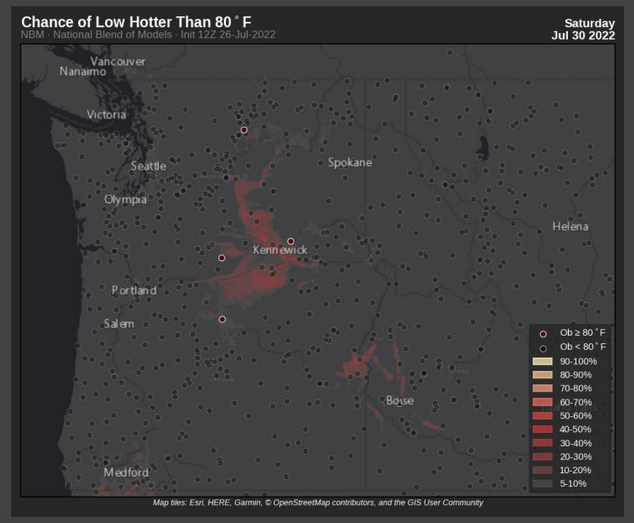


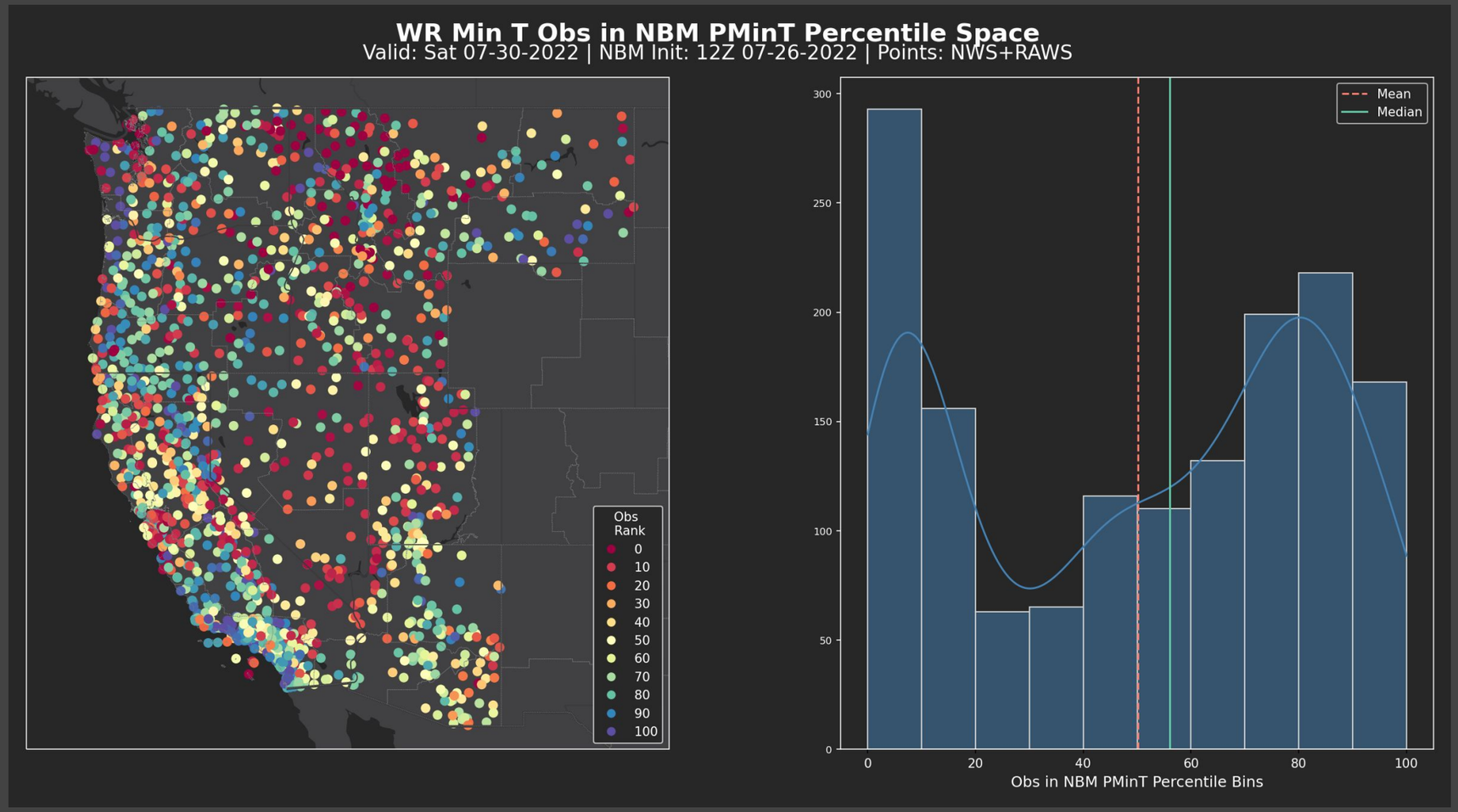
The reliability for PMinT is mixed, as we might have expected given the bimodal histogram of ob ranks. For a fixed threshold like 80 degrees, we see a bit of a cool bias. For a variable threshold, like Red HIL MinT, we see encouraging reliability up through the mid range of probabilities, then a stark warm bias, or over-forecast bias of Red HIL MinT.

HIL = Heat Impact Level, a component of HeatRisk. HeatRisk is a 24 hour value that incorporates the day's MaxT along with the shouldering MinTs.

Late July Heatwave | 30-July-2022

Let’s look at the same event we looked at for PMaxT, that late-July heat wave across the Northwest. Here, we see a much murkier picture, with a “U” shaped histogram, and a mix of both high and low ob ranks. Especially for our impact region, not sure if this is telling us anything just yet. Unfortunately, the only warm low threshold the NBM precomputes is 80 degrees, which probably isn’t all that useful for this case (or this region). But I’ll show it below if you don’t believe me.





Make this plot in Google Colab!

NBM 4.0 Day 5 PQPF

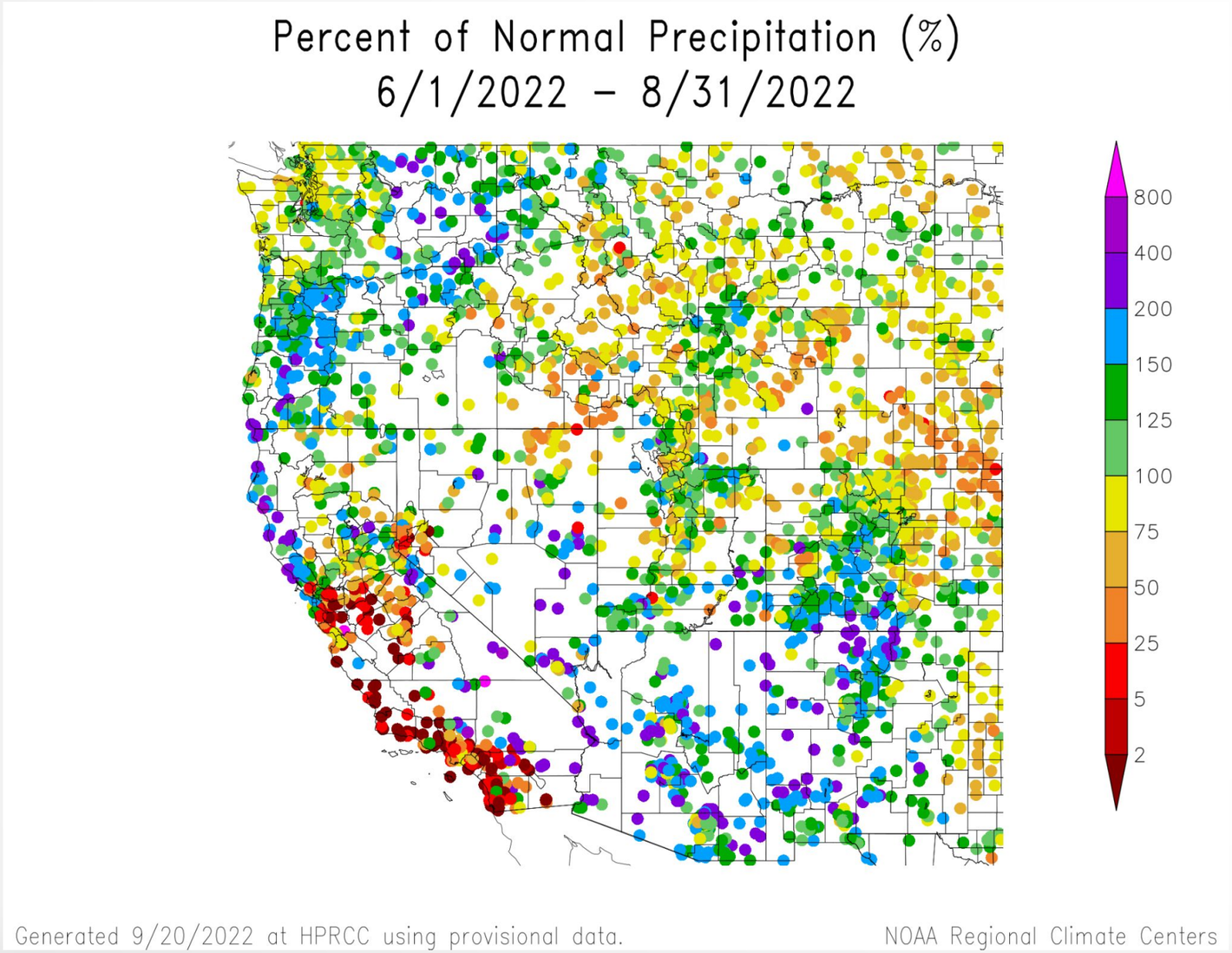
21,165 Obs (NWS + RAWS) => 0.01” spanning **91** days

Missing: 7/12 (grib files incomplete)

Confidence in analysis: **MEDIUM**

Low sample size should limit our ability to generalize these results

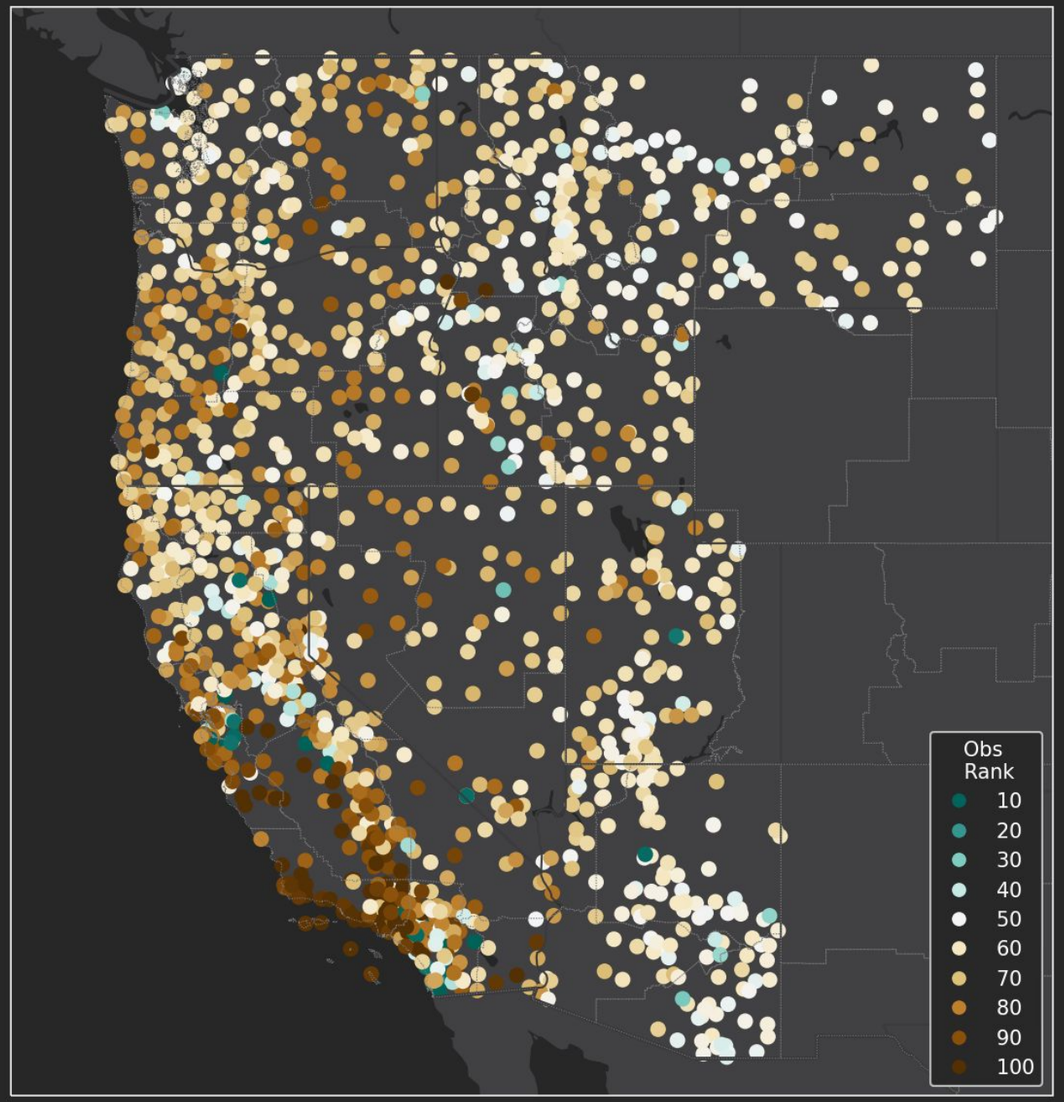
Jun-Aug Precipitation Summary

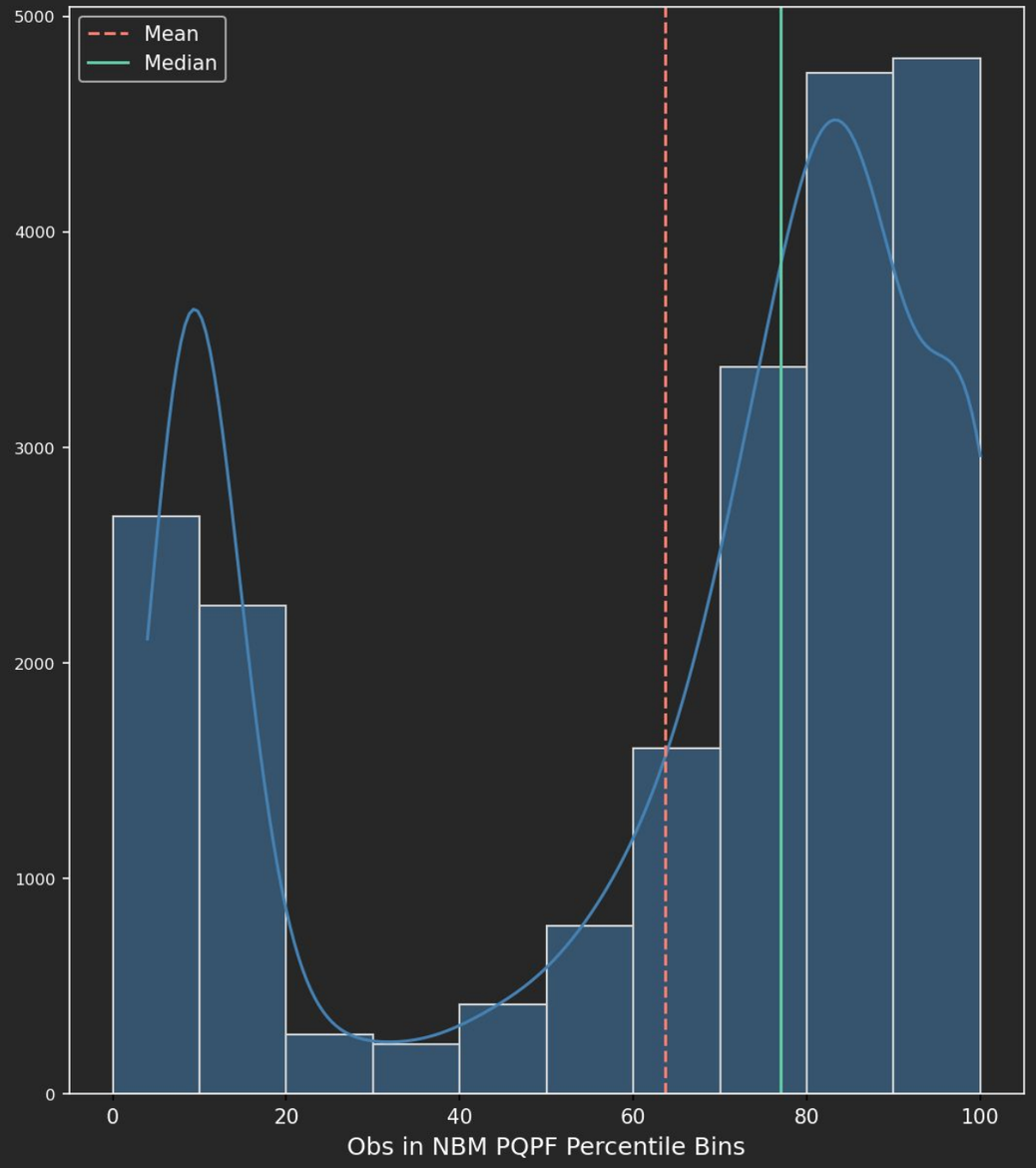


- Active/wet monsoon
- Extremely dry southern and central California

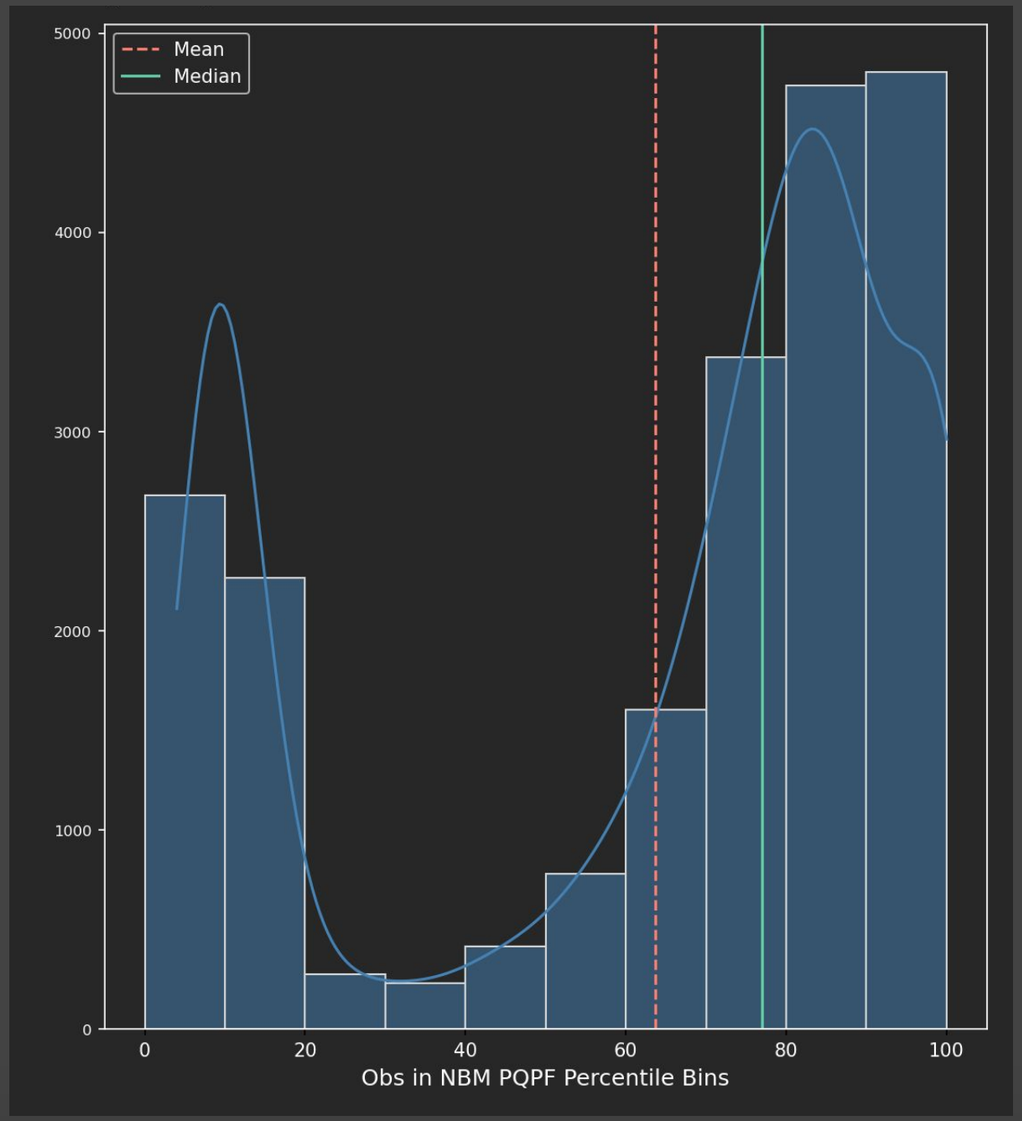
WR Precip Obs in NBM PQPF Percentile Space

Valid: 06/01/2022 - 08/31/2022 | NBM Lead Time (Days): 5 | Points: NWS+RAWS





Summer 2022 Day 5 NBM v4.0 PQPF By The Numbers



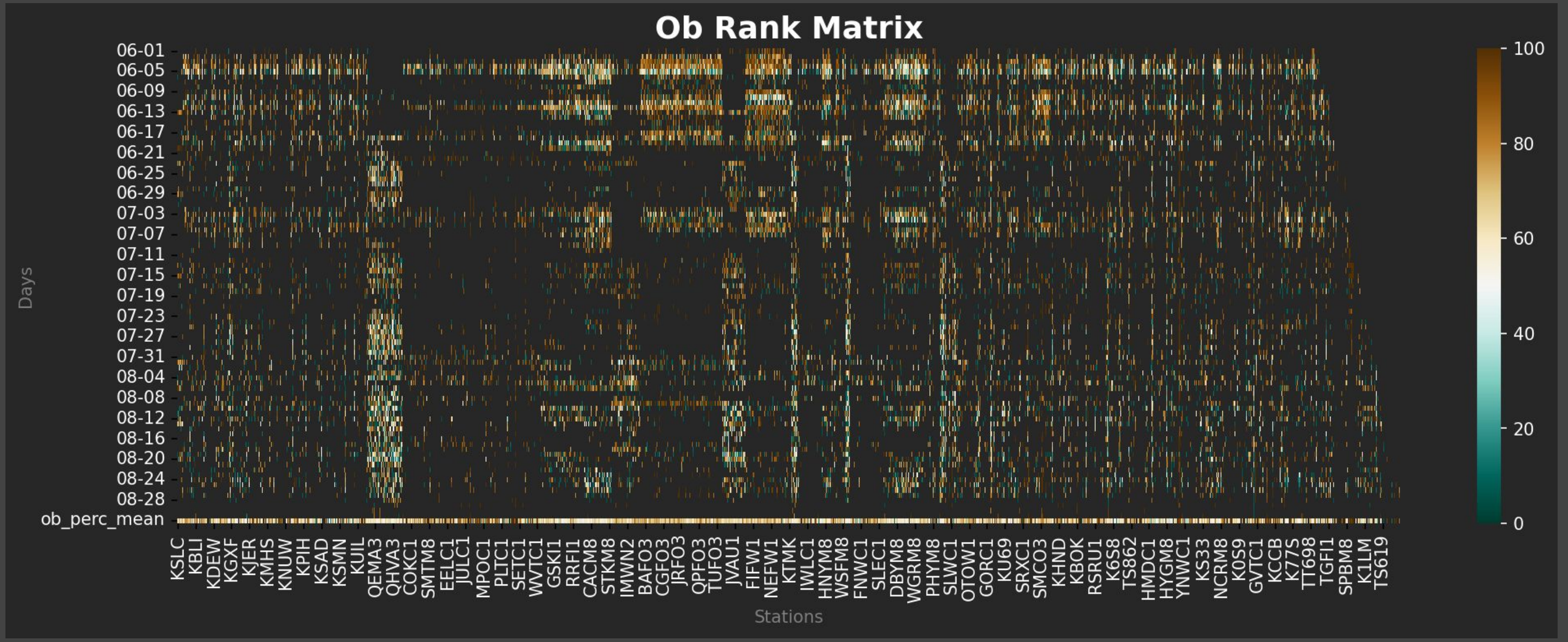
WR

Total Obs
118,118

10th-90th
78,924
66%
(80% expected)

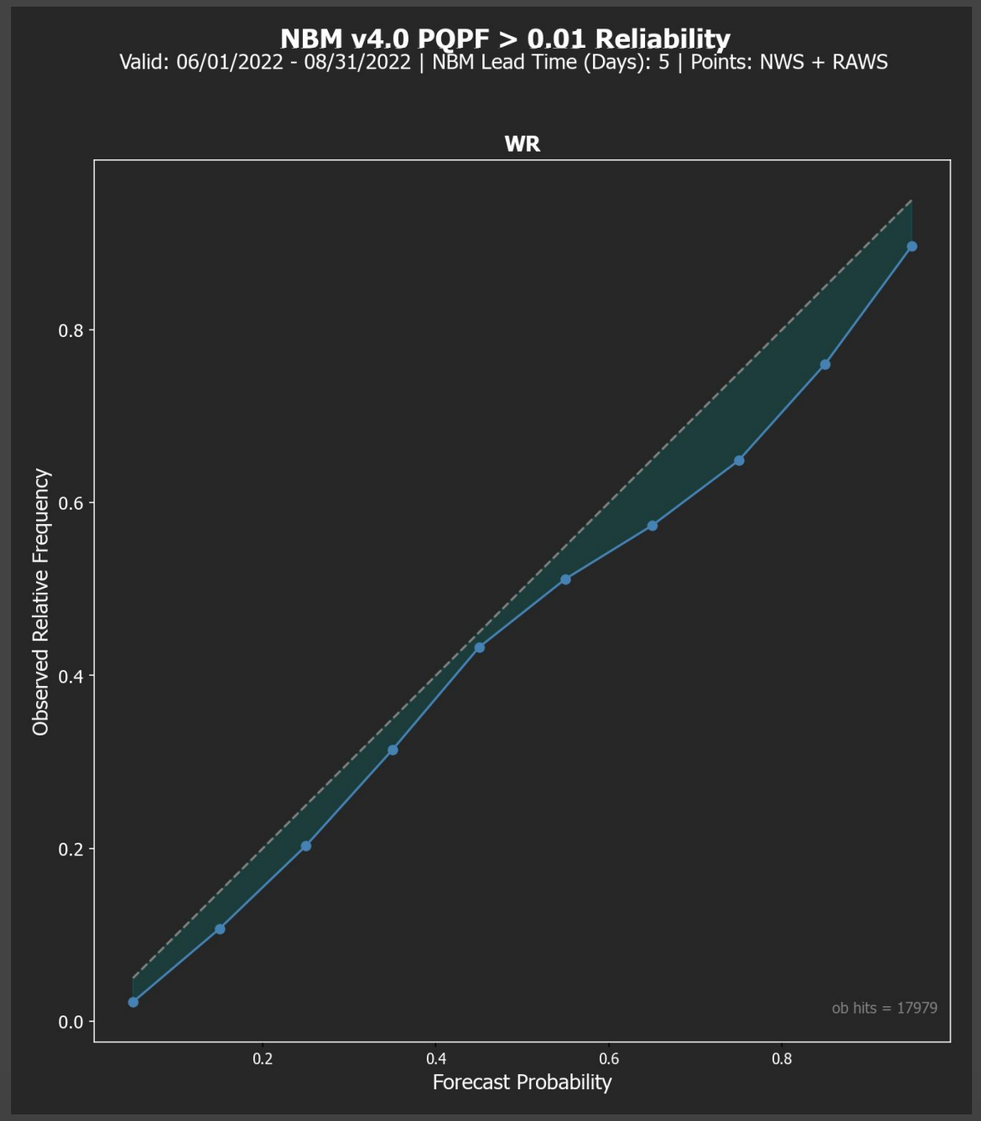
25th-75th
41,355
23%
(50% expected)

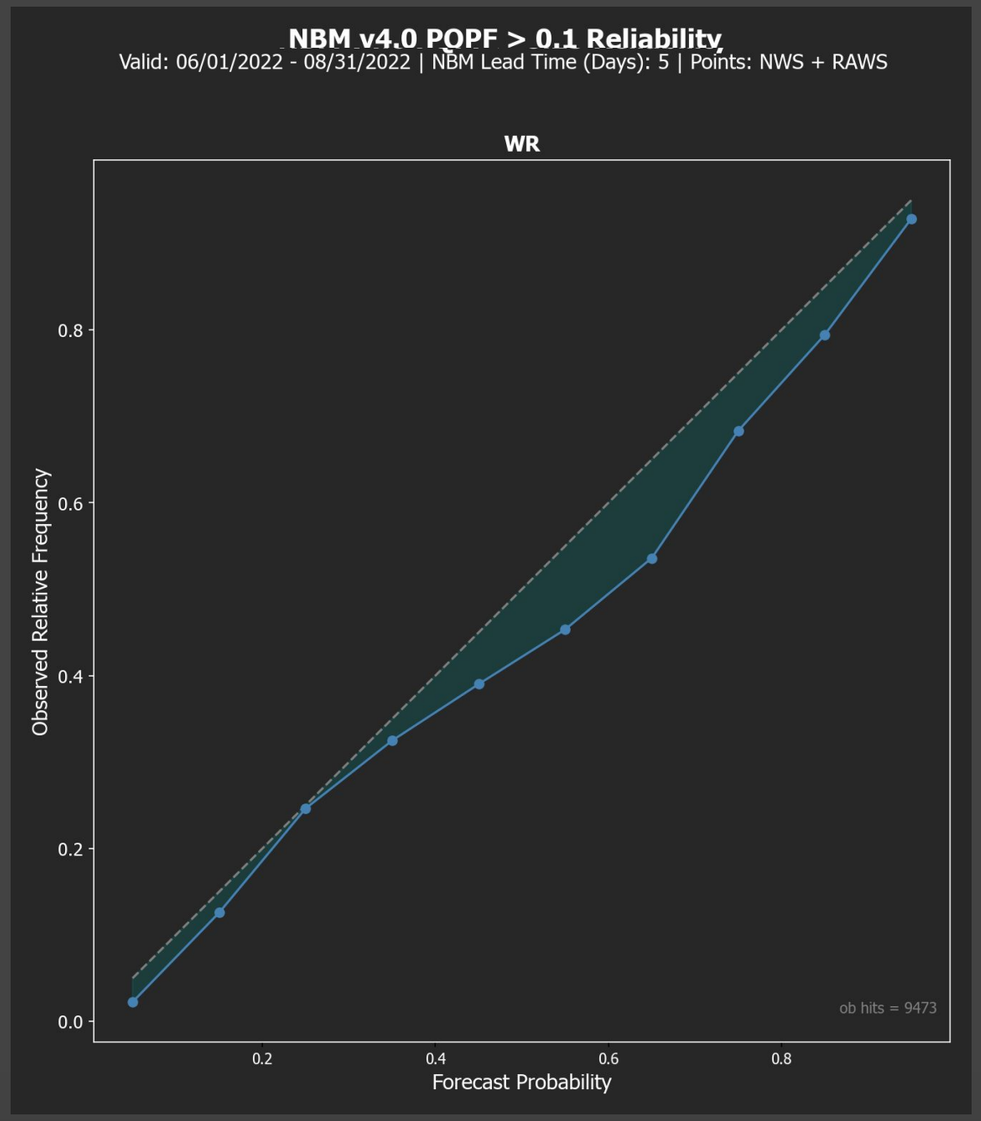
- While we probably shouldn't expect a flat histogram here, since we usually only have outliers in one direction (wet - since you can't have negative precipitation), this histogram looks a lot worse than previous seasons' analysis.
- It is bimodal, with peaks around both the 10th and 90th percentiles.
- In previous seasons, even though we could note a bias with the skewed distributions and shifted mean, interquartile and interdecile obs still "contained" the bias. Here, we have obs falling outside both ends, but an overall dry bias seems to be most dominant.
- Not all of this is surprising as summer is dominated by convective precipitation - for which the predictability horizon is very narrow.



In this plot, we can see a few widespread precipitation events in (early) June, with monsoon regions stealing the show, with consistent activity shown in Arizona RAWS (the underlying data is here)

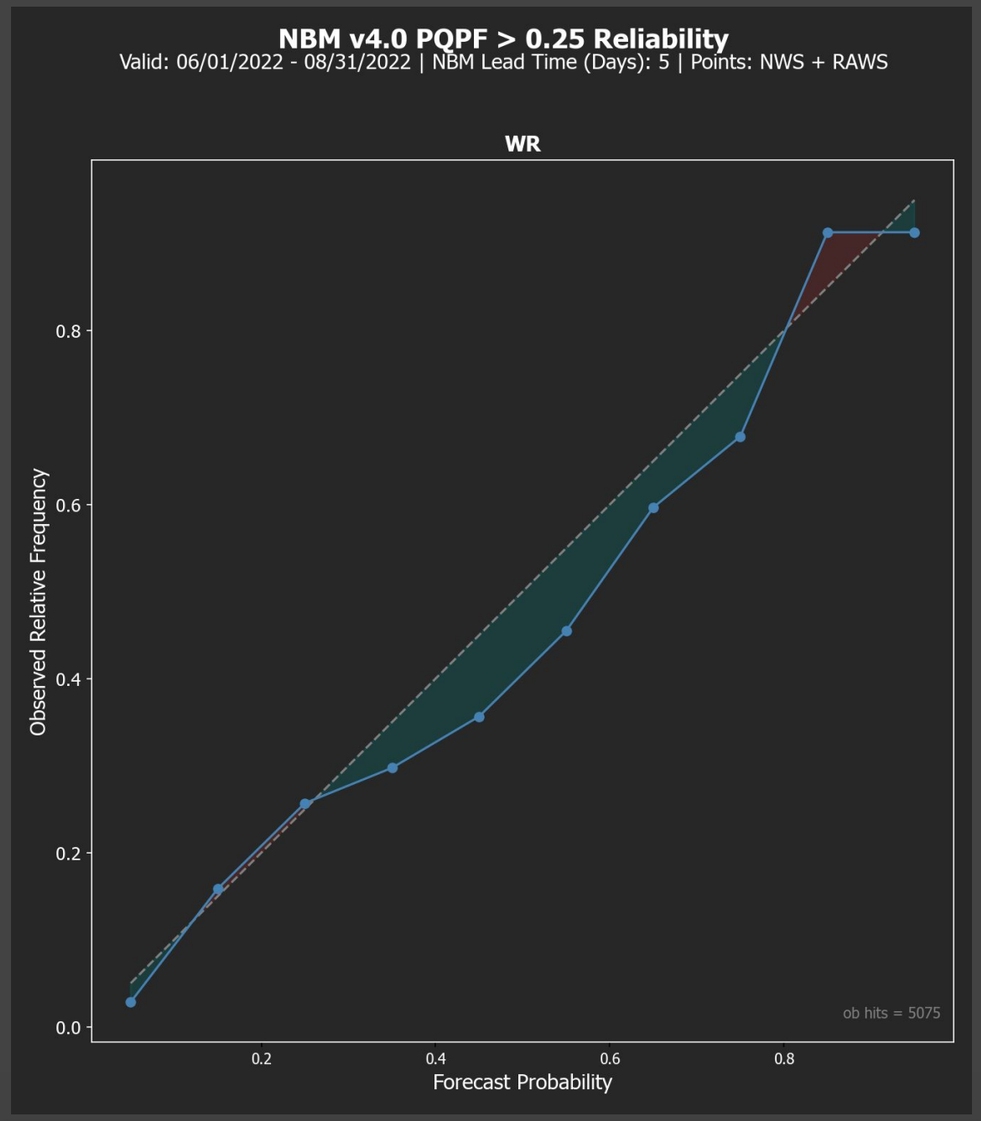
Day 5 NBM PQPF Threshold Reliability

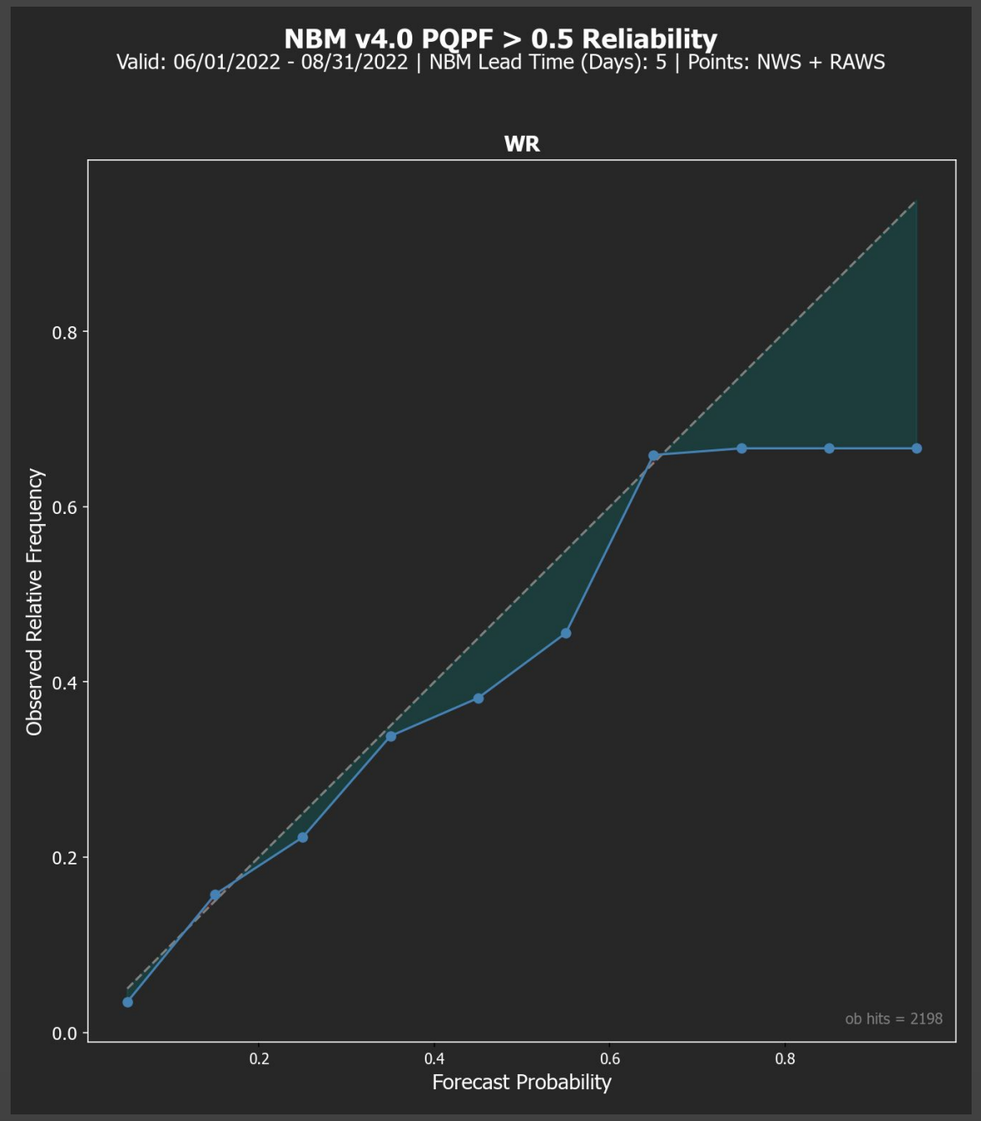




For low thresholds (0.01” left, 0.10” right), we see the line hugging the diagonal, displaying very good calibration. Amounts were observed a little less frequent than forecast (line below diagonal), so we we a ***slight*** (5-10%) wet bias at these thresholds.

Day 5 NBM PQPF Threshold Reliability

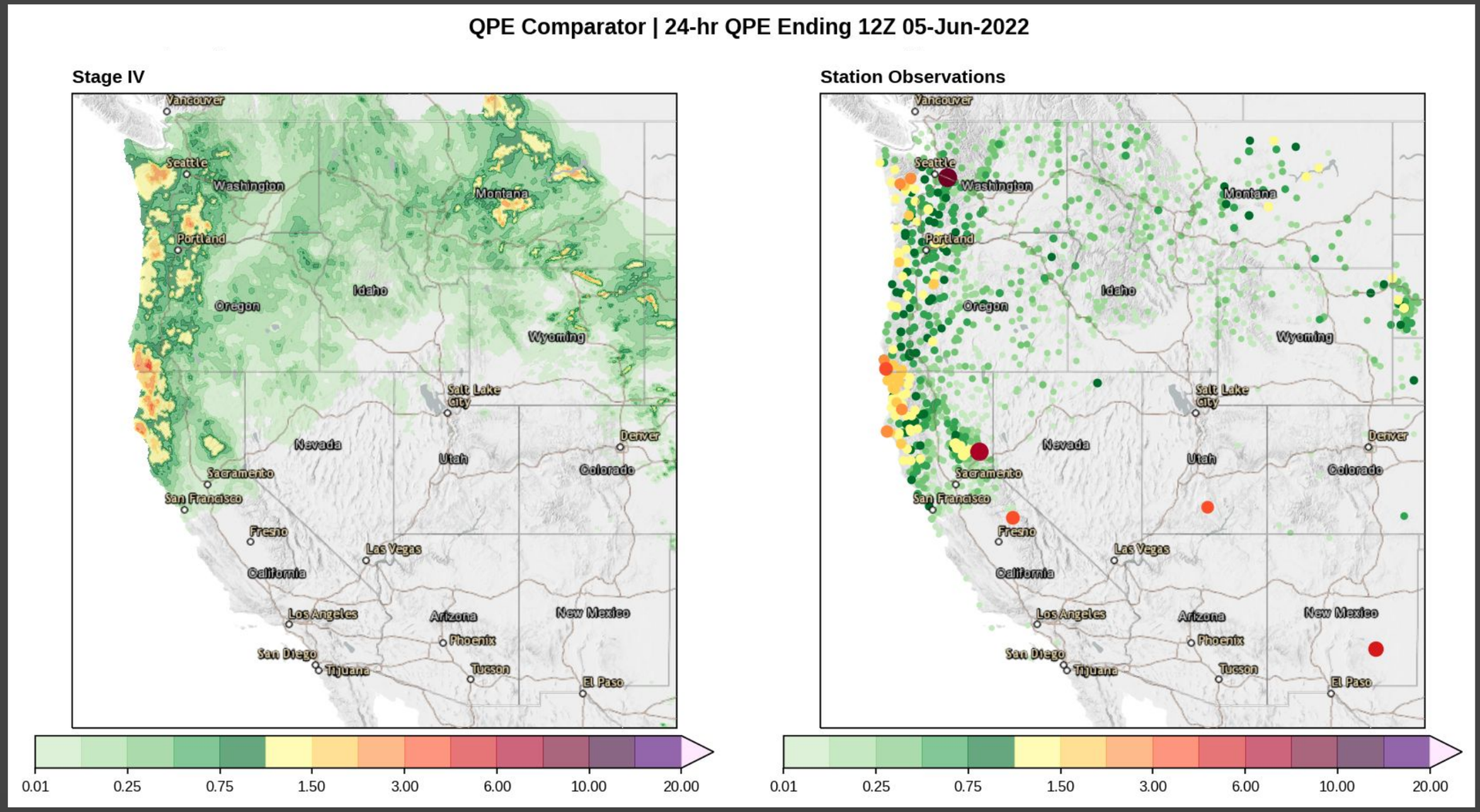




As we step up in thresholds, notice our sample size just about halving each time. We maintain good reliability with slight wet bias through the 0.25” threshold, but our sample size becomes too small to accurately resolve our curve by the time we get up to 0.5 inches, so that’s where we’ll stop.

Notice these tell a little bit different story than our bulk distribution (ob ranking) just showed? Verification is indeed complicated and one number/approach is unlikely to tell the whole story!

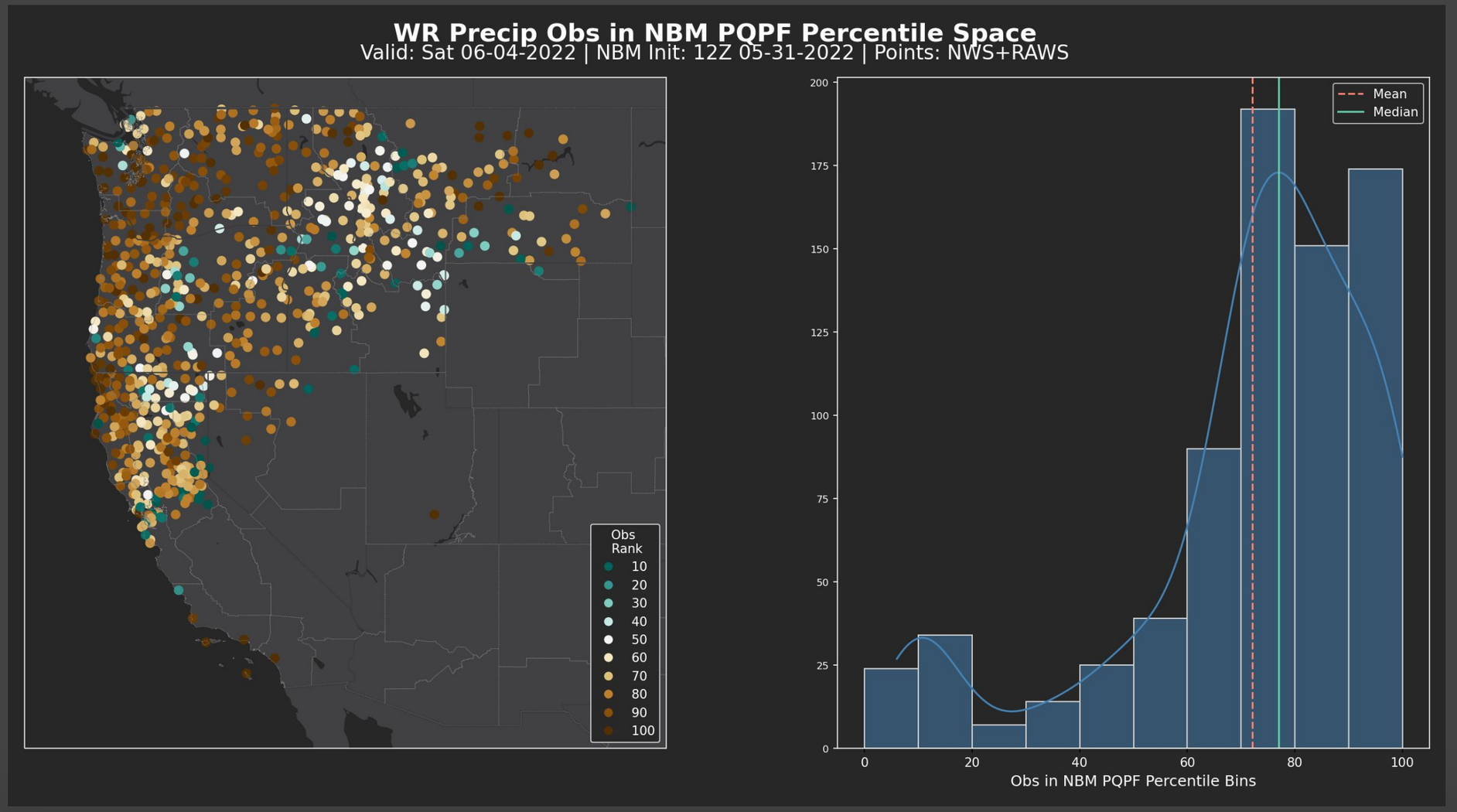
Case | 4-5 Jun-2022



This was a multi-day event with widespread precipitation for the northern half of the region, and especially along the coast from San Francisco northward. Because NBM v4.0 only has 24-hr PQPF, we'll have to break this into two periods, but we'll also take a peak at some 48-hr NBM v4.1 PQPF.

Observations for the first 24-hr period (12Z 4th - 12Z 5th) are on the right with gridded Stage IV (RFC) QPE on the left.

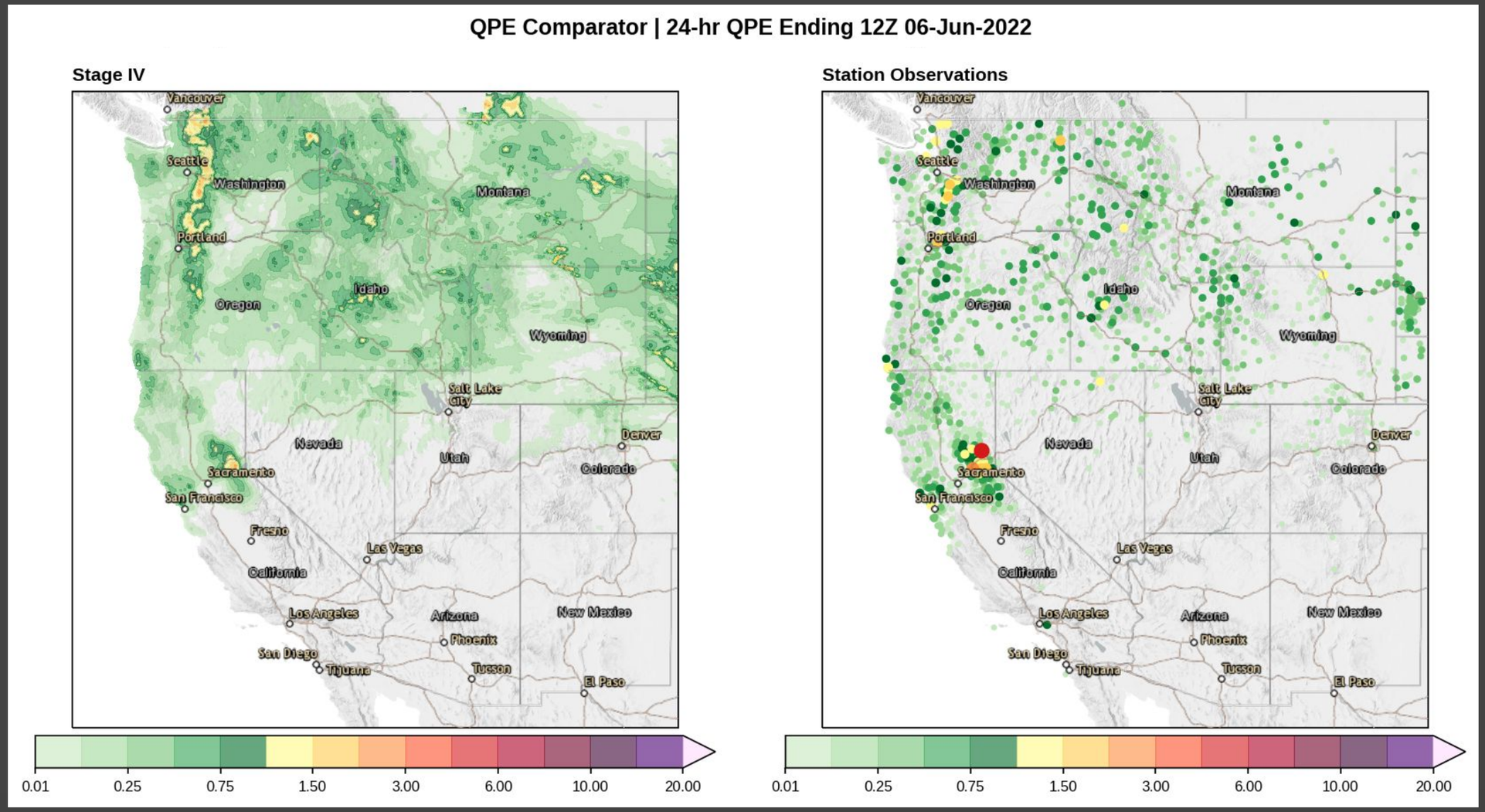
Case | 4-5 Jun-2022



If we rank those obs in the NBM PQPF distribution, we can see overall, obs ranked in the upper percentiles of the NBM, indicating the NBM distribution was probably a little dry here, especially where the heavy rain fell. The mean ob rank is displaced around 12% up from the 50th percentile, confirming this tendency to be dry. Note our QPF analysis is usually from 12Z-12Z. So while I correct NBM plots to say “Valid 06-04-2022”, the actual valid time would extend to 12Z 06-05-200

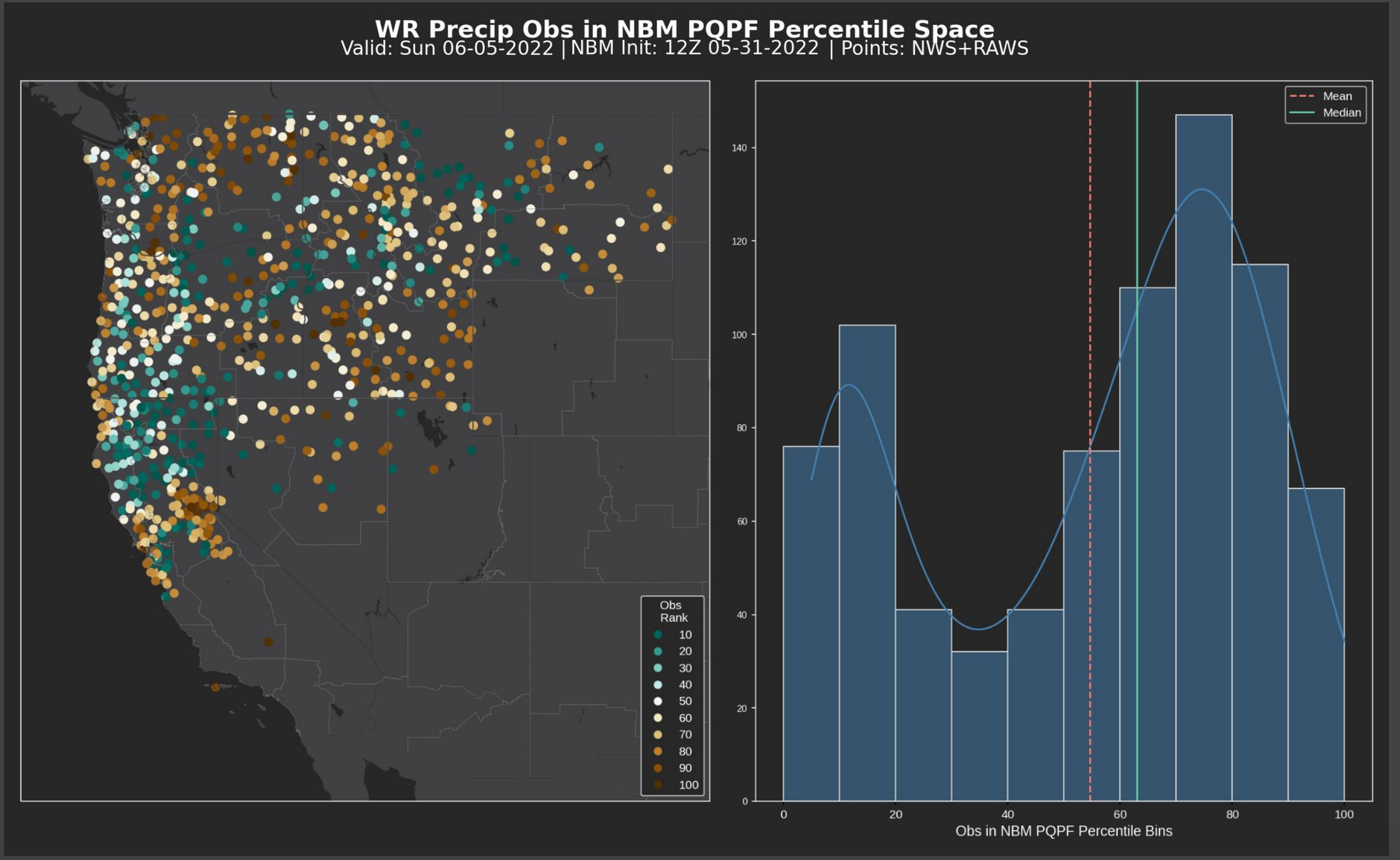
However, notice this wasn’t equally dry, nor was the “bias” the same sign everywhere.

Case | 4-5 Jun-2022



For the second period, we spread the precipitation further inland across the region, with generally lighter amounts.

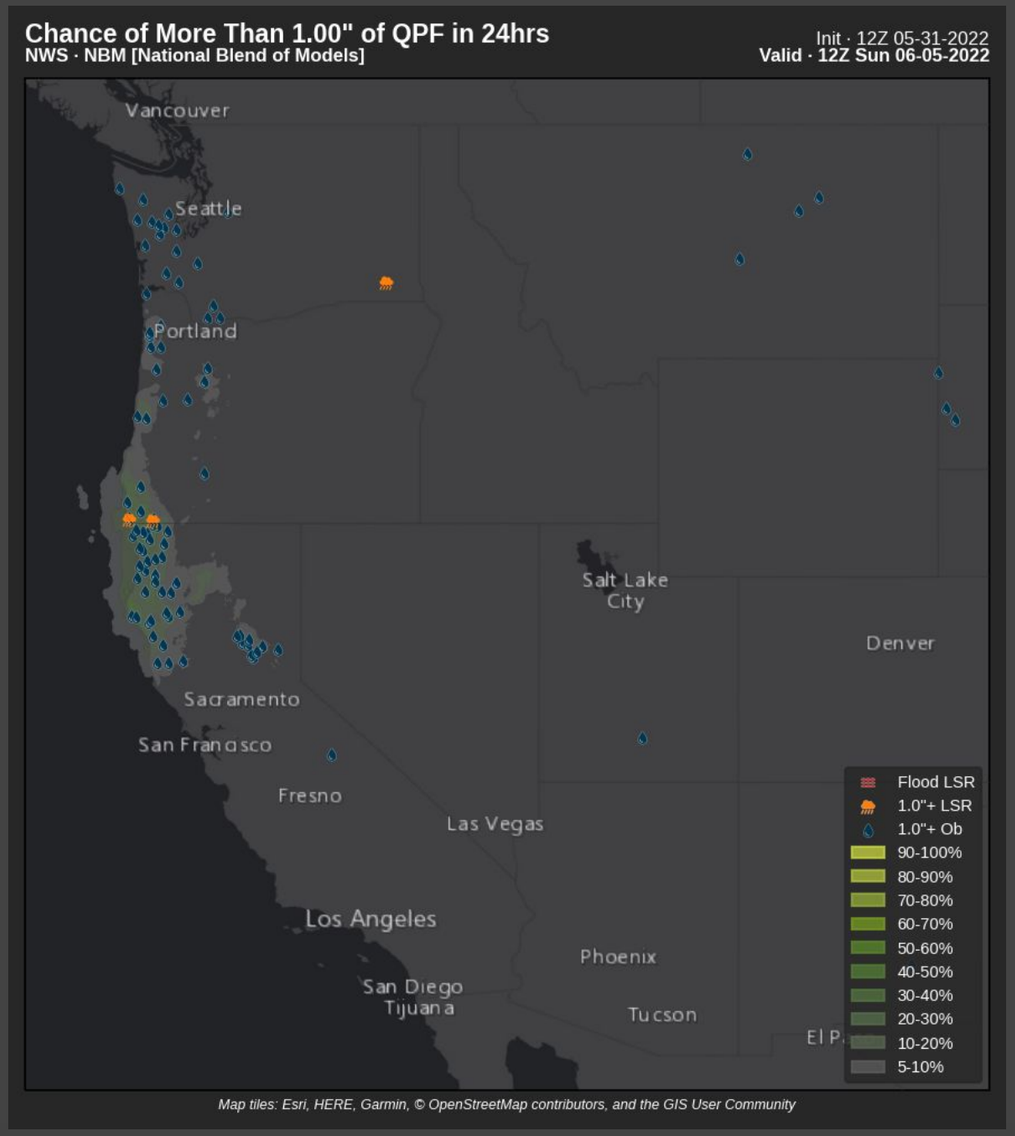
Case | 4-5 Jun-2022

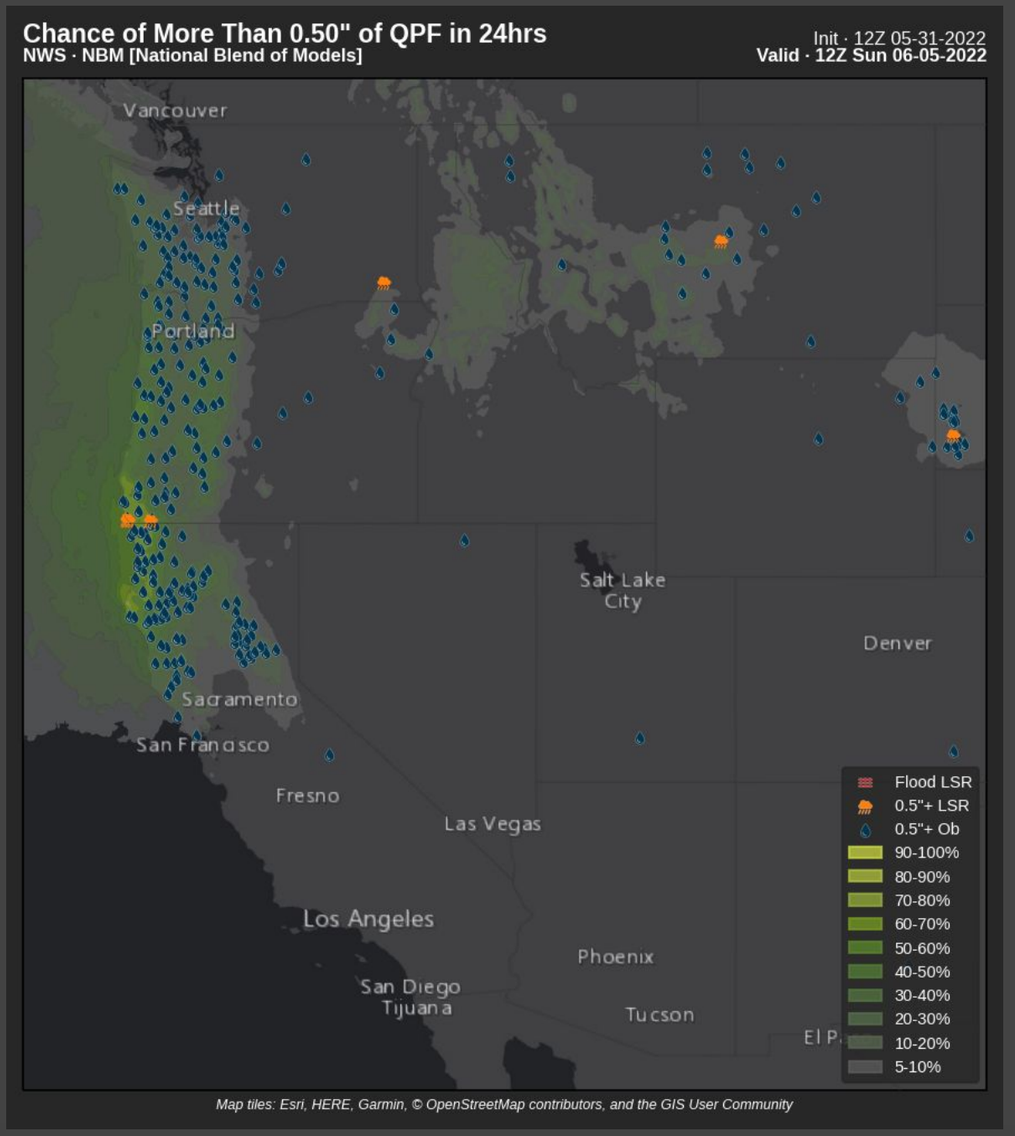


For this second period, let's keep the same cycle (so, effectively now looking at a day 6 forecast). We have much more of a mixed bag for NBM PQPF performance. The histogram looks to be more bimodal, but a much lower/unclear bias (although obs above the 50th outnumber obs below, so still a bit of a dry bias). Also notice from the previous day, the percentile ranks did not stay the same - which would make substituting something like a single percentile level in for a forecast for an event would not yield good results.

Case | 4-5 Jun-2022

In previous seasonal analysis, and even in this one, we can see seemingly different stories developing between the bulk ob ranking and threshold verification, and that even in a messy distribution, we can still get good threshold probabilities. Looking at the 1” and 0.5” thresholds for this first period, we can see a pretty small footprint for the 1” threshold, but generally correct in northern California while missing out on the Oregon and Washington coasts. We improve that a lot by dropping down to 0.5”.

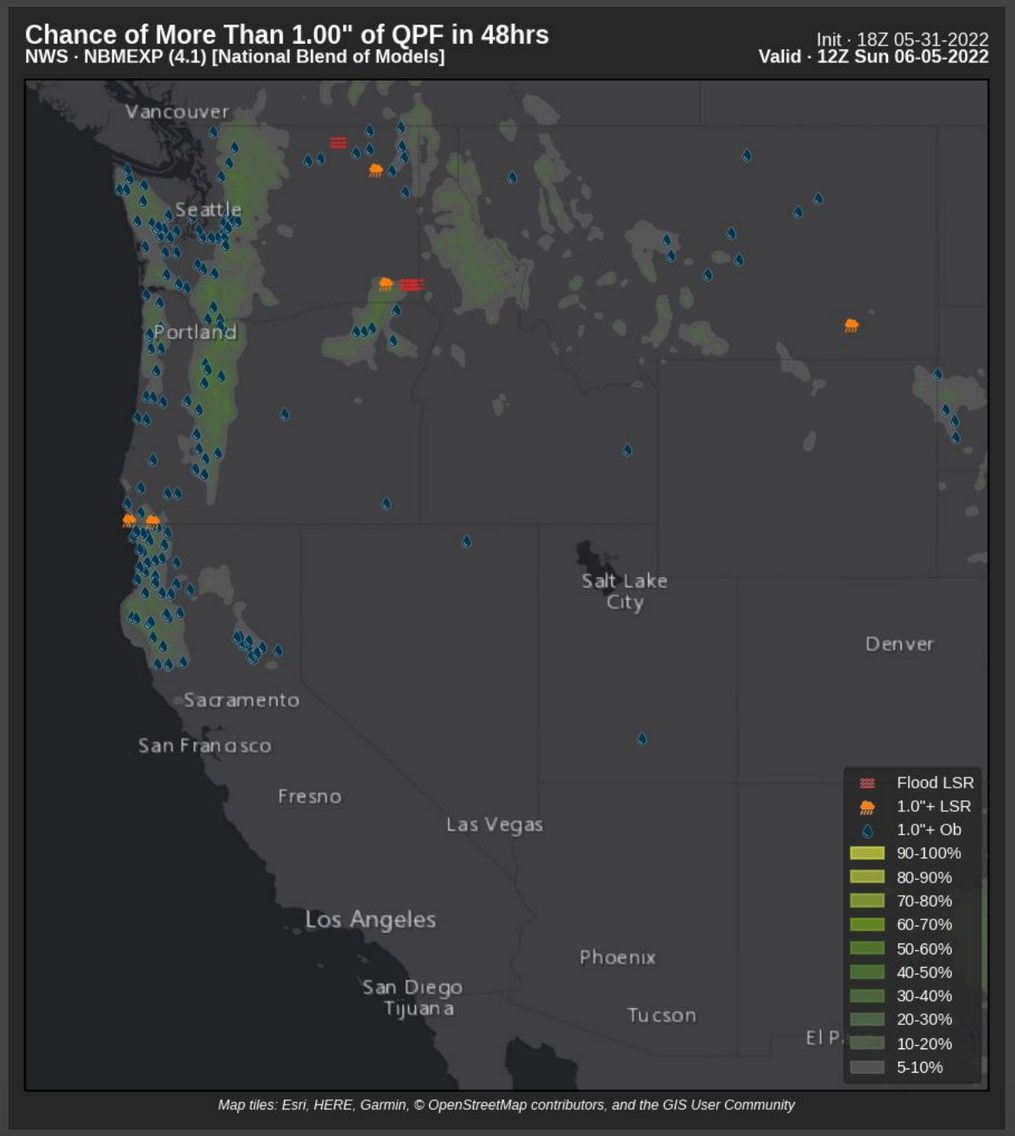


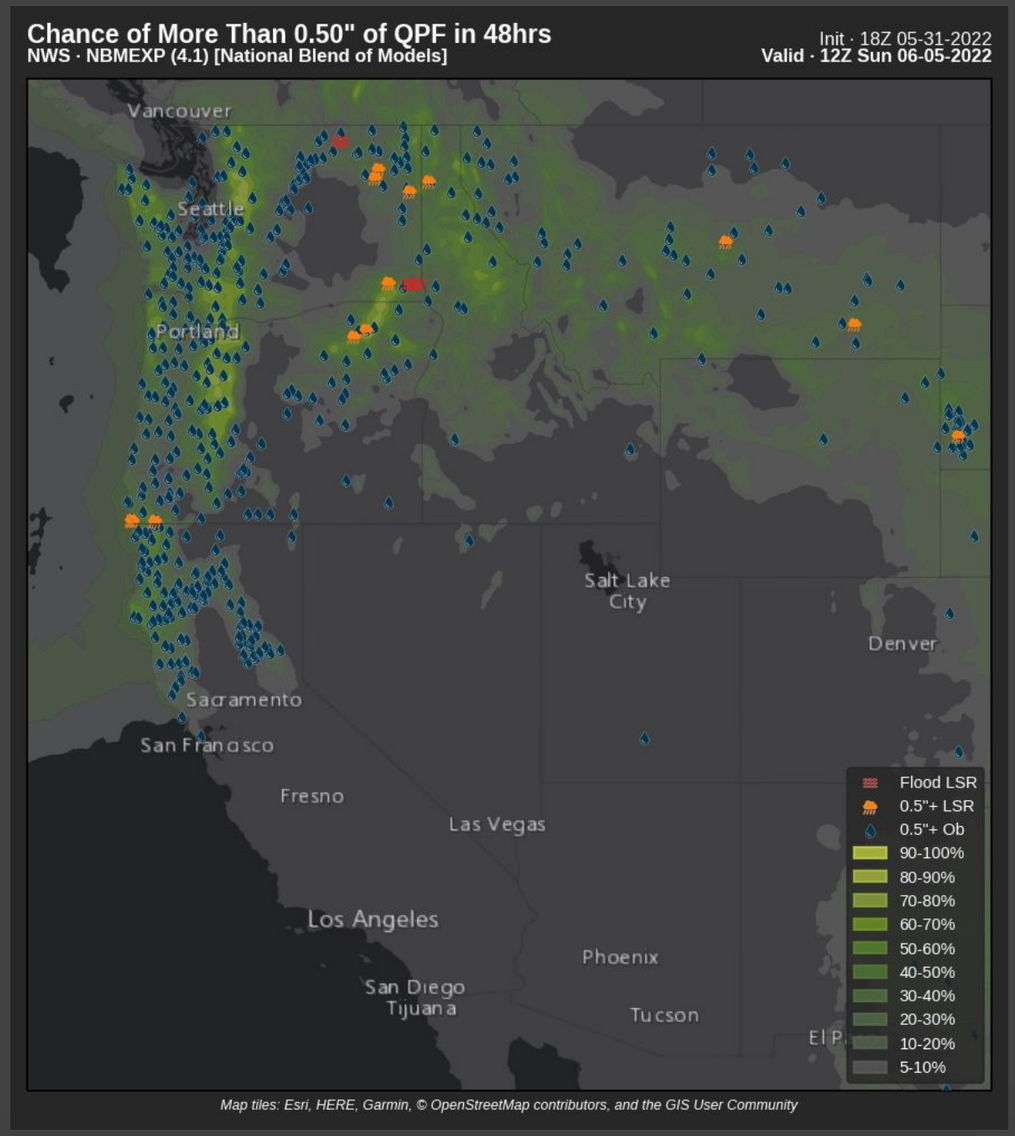


Make these plots in Google Colab!

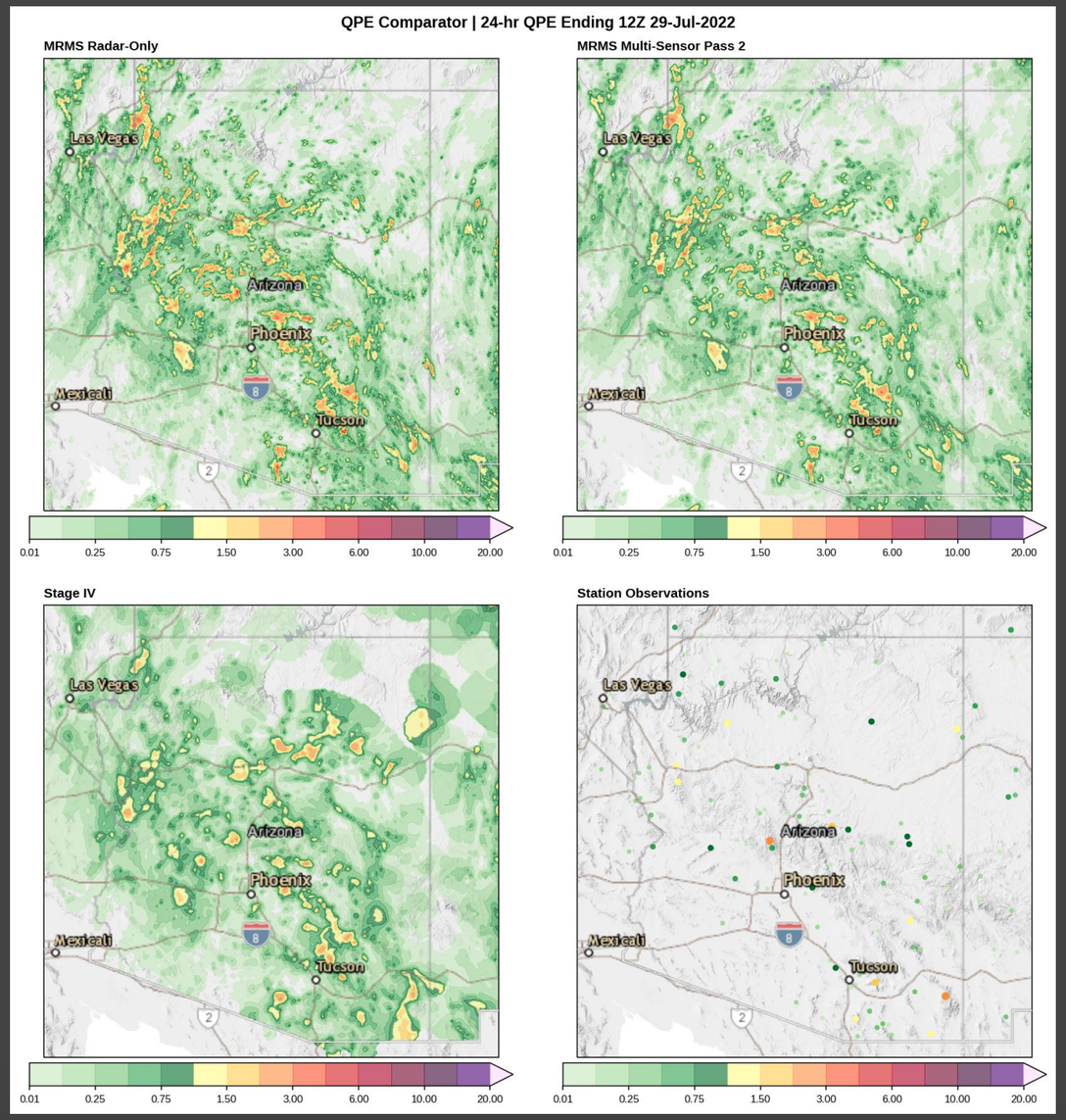
Case | 4-5 Jun-2022

This was a two-day event, so did the new longer-period PQPF in v4.1 yield better results? Generally yes, although it is a mixed bag. 4.1 correctly identified probabilities of 1” or more along the Cascades, but actually had lower 48 probabilities than the 24 hour 4.0 probabilities in northern and central California. For half an inch, the coverage is much better. In the Pacific Northwest, and especially around Seattle, 4.1 reduced probabilities for the valleys, where the rain did in fact occur. But again, verifying probabilistic forecasts with a single event is tricky business and is not real robust, but still useful for subjective verification.

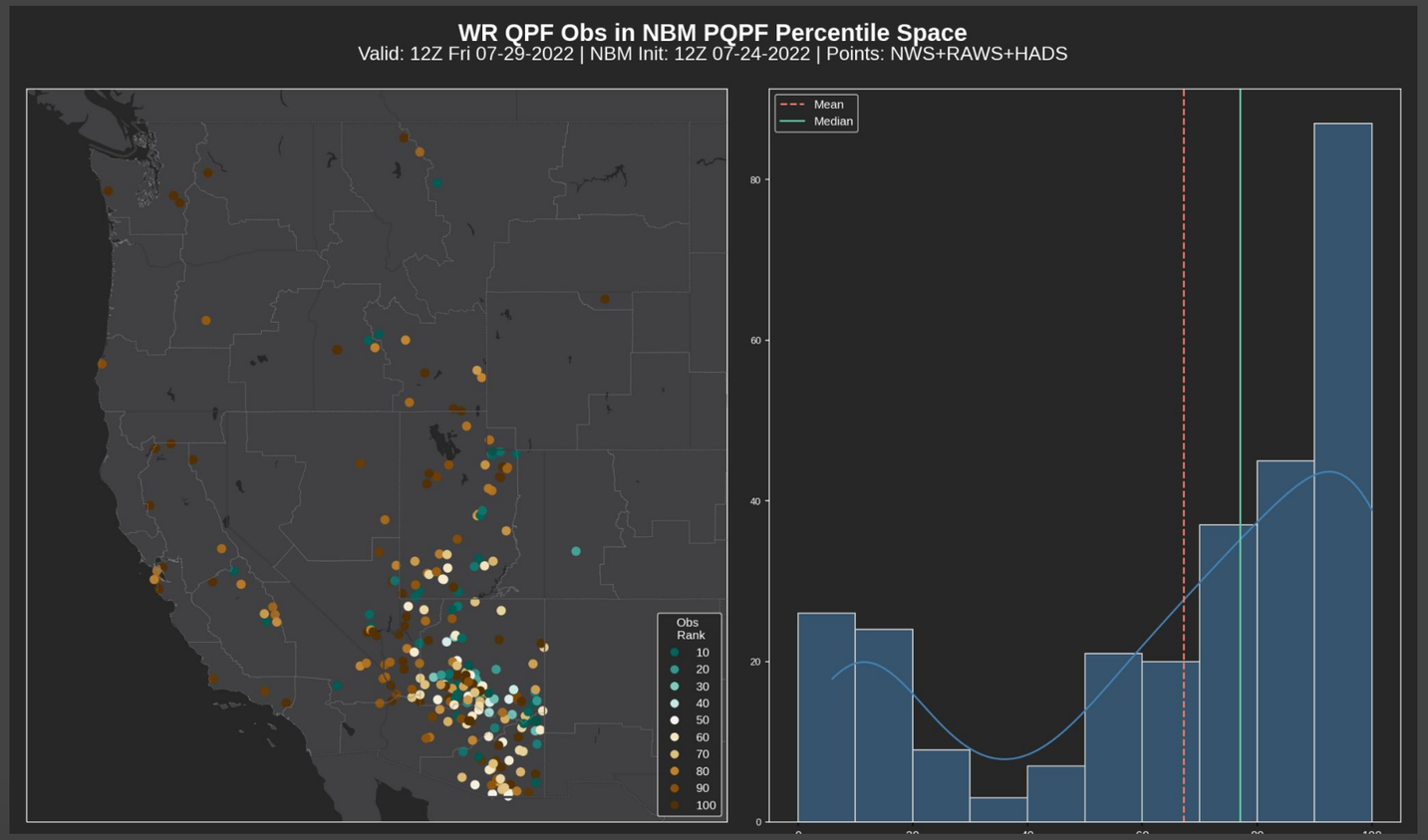




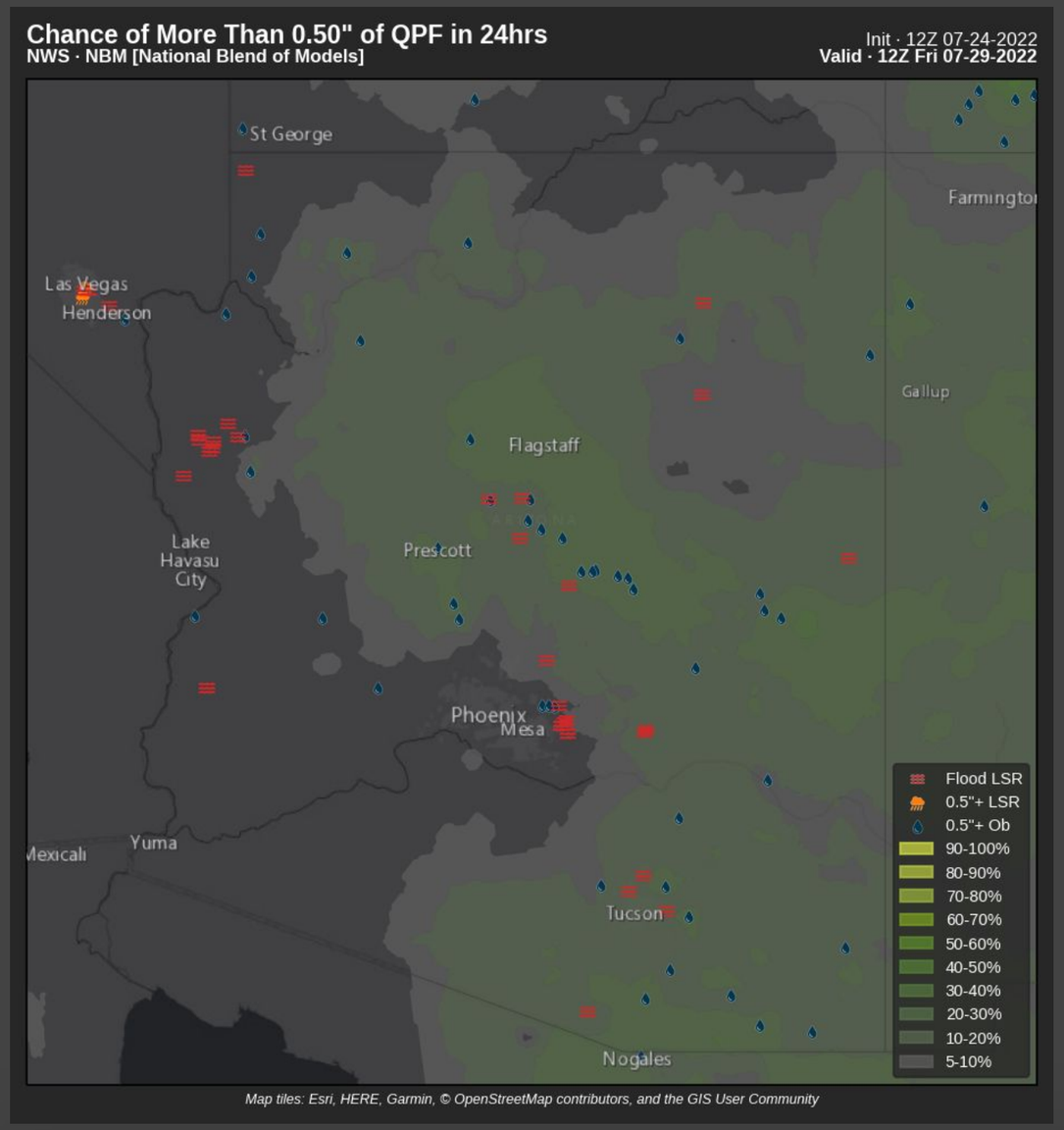
Monsoonage | 28-July-2022



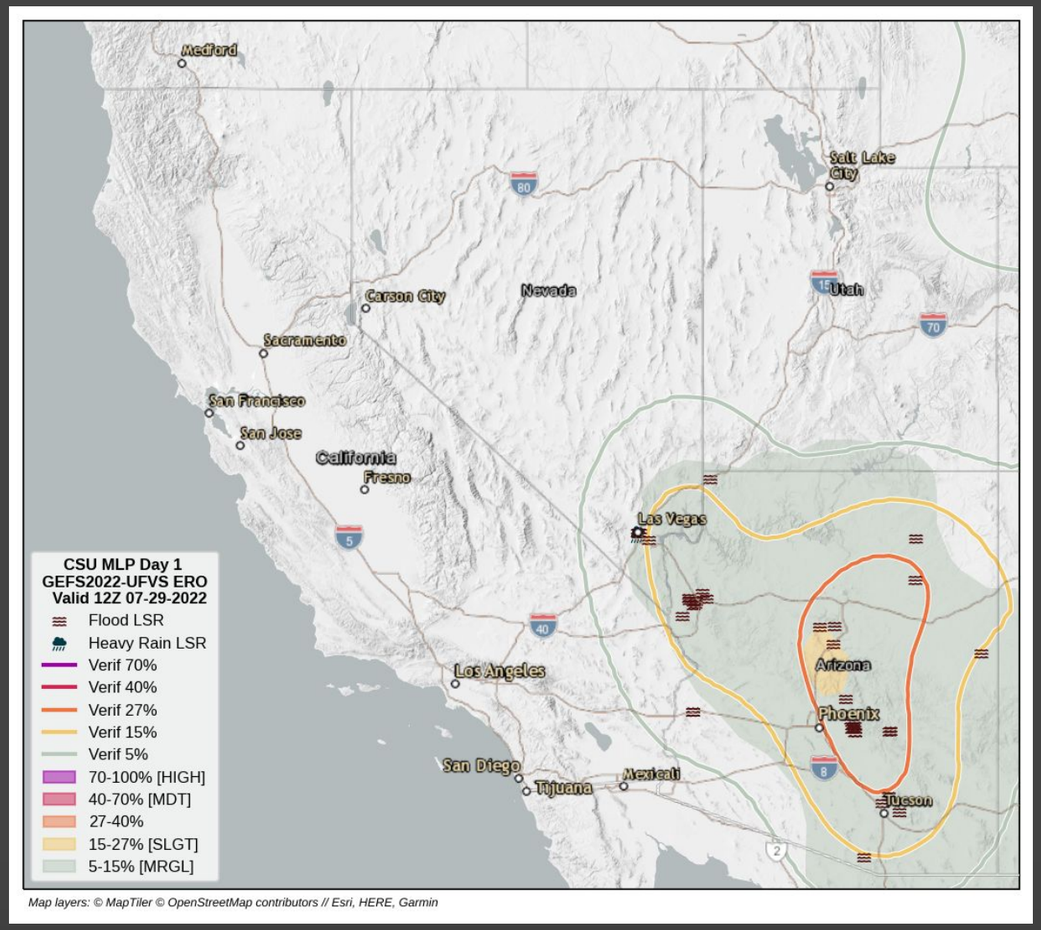
So, how did NBM do for monsoon activity? About as well as you would expect for global models (non-CAMS) at this forecast range. While convection is very much unresolved by global models, the monsoon circulations are, and some signal usually shows up. And indeed we see a very mixed bag, although our histogram even looks better than the bulk.



Monsoonage | 28-July-2022

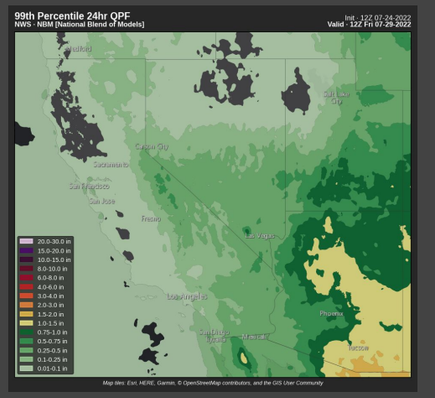


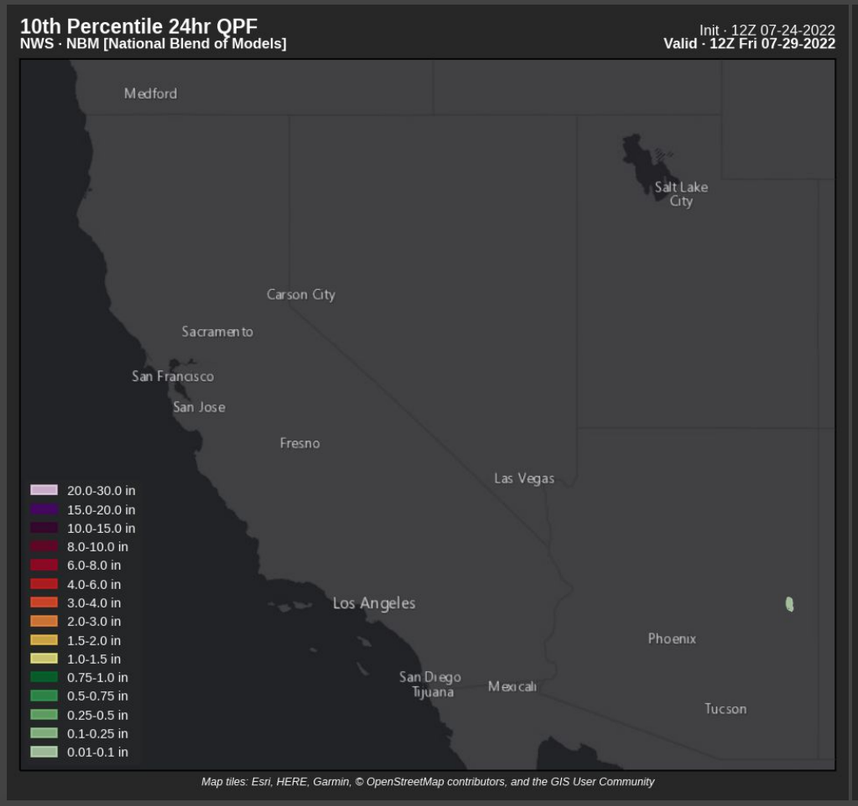
While we did have some signals in threshold probabilities, and areas around Tucson and up along the Mogollon Rim were highlighted, the probabilities were overall a bit low, and the footprint too small. Just to illustrate how difficult exact monsoon impacts are for global NWP, we can look at something that is leading the way in post-processing of NWP using Artificial Intelligence to predict impacts, not just amounts - the Colorado State University Machine Learning Probabilities. While it is a topic for another presentation, even this sophisticated processing often struggled for monsoon activity, even at day 1.

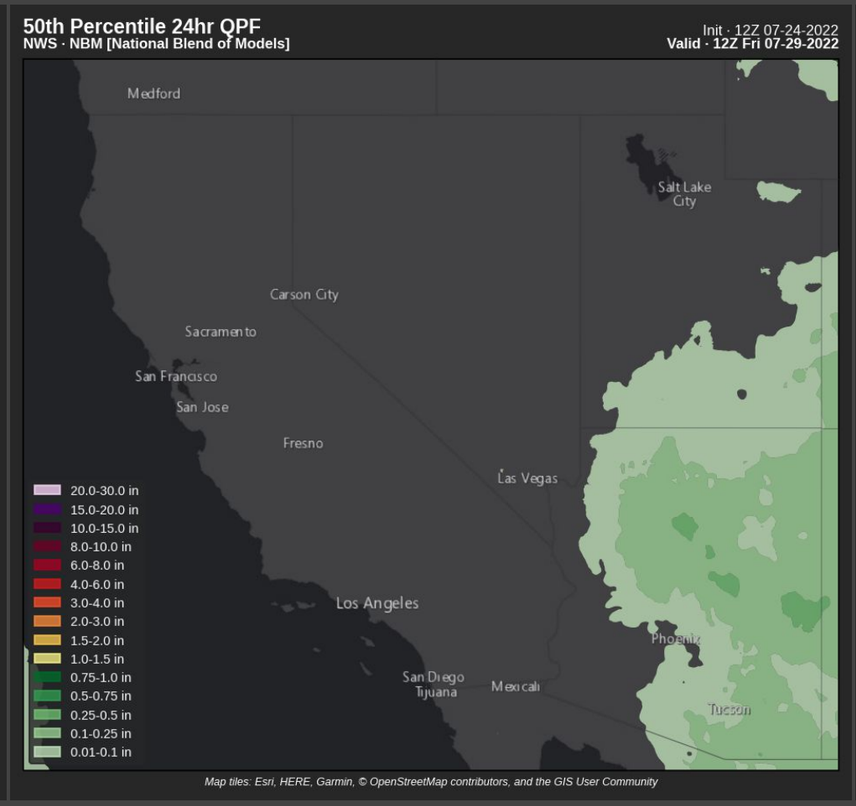


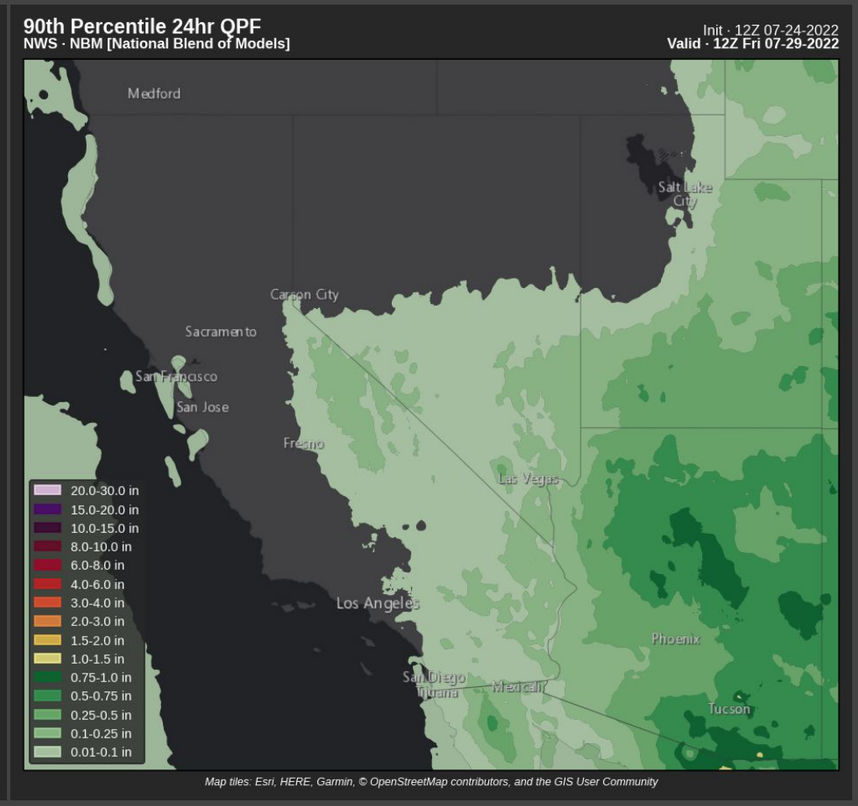
Monsoonage | 28-July-2022

We can dive into the percentiles directly to try to look for a signal. From the analysis plots, we saw widespread splotches of over an inch of rain during this period across Arizona. Using the typical three percentiles of 10th, 50th, and 90th, we don't see any indication of that amount. We do however at least see those amounts when we crank it all the way up to the 99th percentile (upper right). Remember, this is day 5 and there are no CAMS at this range.



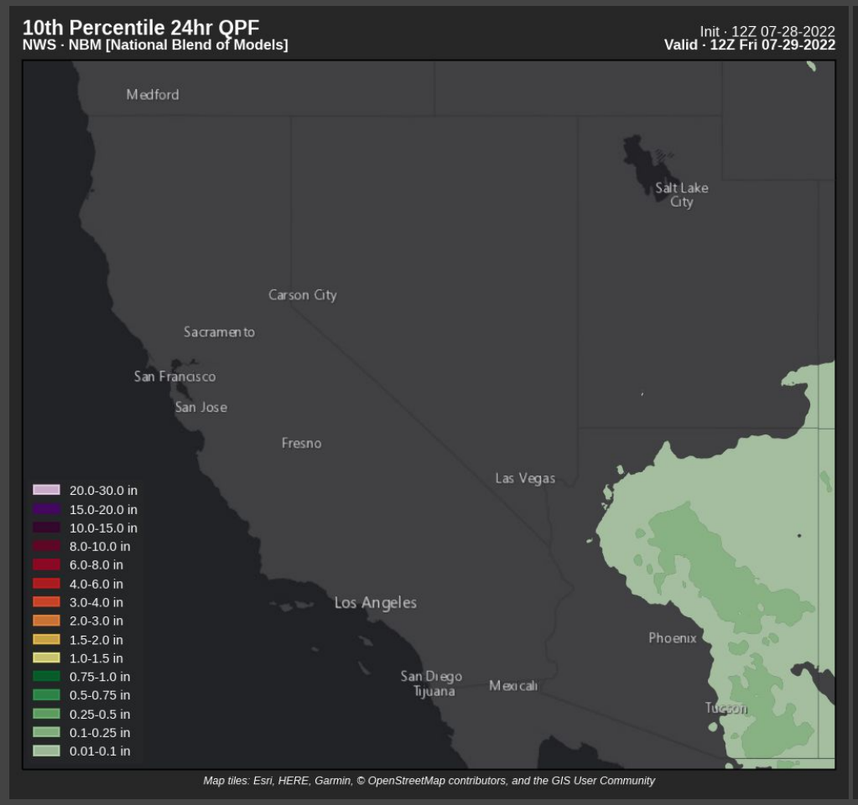


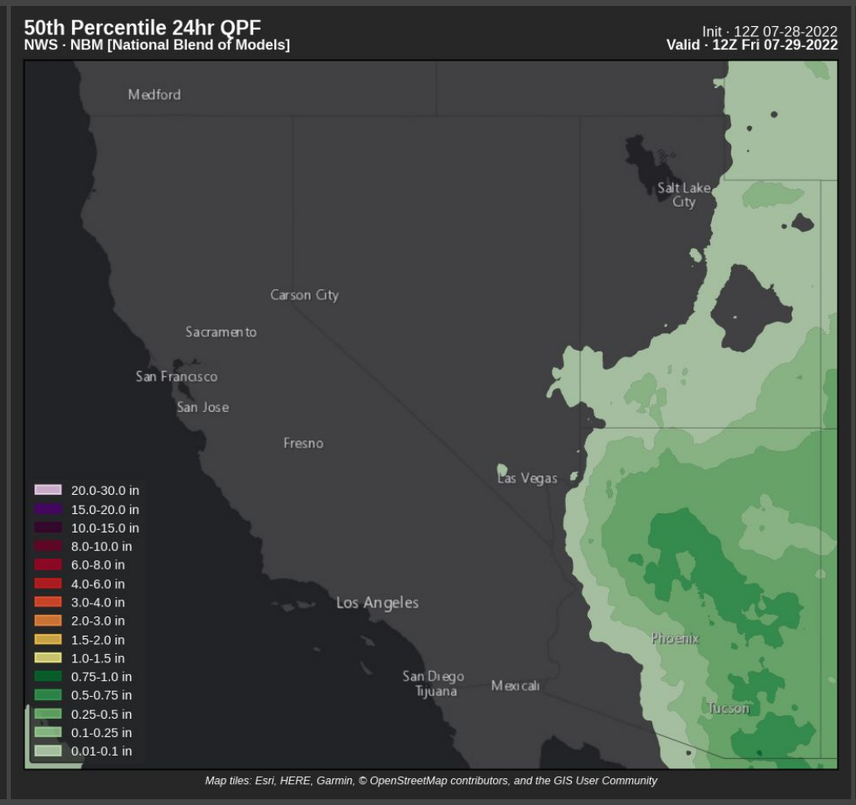


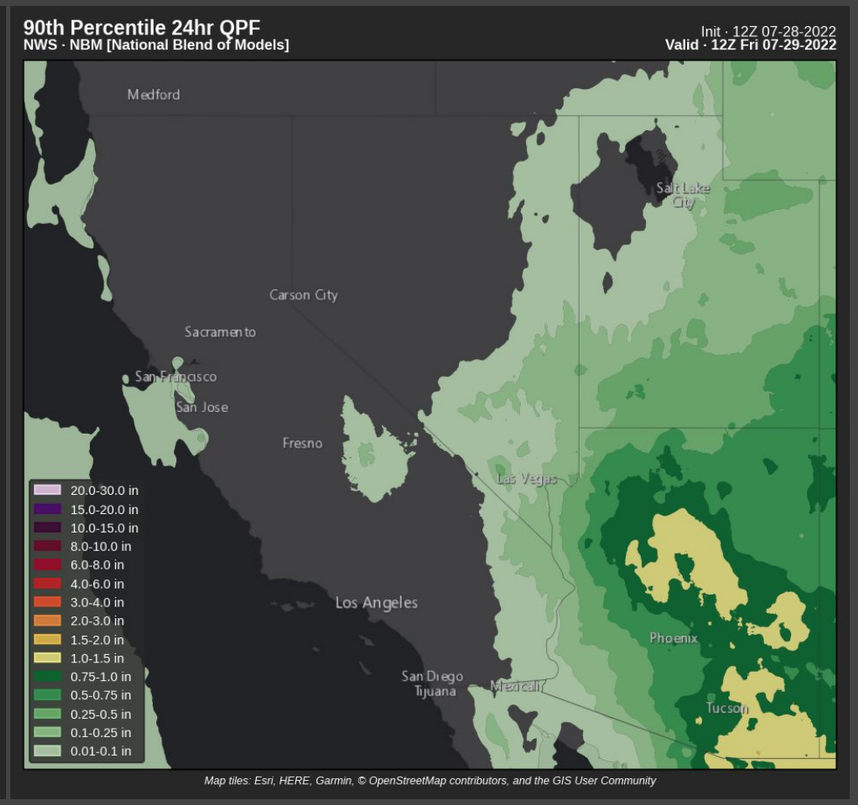


Monsoonage | 28-July-2022

Let's take off our "Day 5" hat just to look if we get any better of a signal with a 24-hr forecast, with the NBM now ingesting CAMS. We indeed see higher amounts and a better footprint, so it has improved by this lead time, even if the footprint still ended up a bit on the low side. Unfortunately, this falls in a gap of our NBM4.1 LDM data archive, so we can't compare to 4.1 for this event.







Takeaways - Day 5 NBM v4.0 Summer 2022

All elements besides PMaxT displayed a little worse performance than the Spring analysis. PMaxT showed encouraging performance, especially for both fixed (100F) and normalized (Heat Impact Level) thresholds. However, PMinT struggled in bulk, and PQPF struggled due in large part to 1) fewer samples and 2) convective nature of precipitation in the summer lowering predictability.

PMaxT

- Slight cold bias in bulk and for thresholds.
- Impressive reliability “when it mattered” for things like Red Heat Impact Level, which make a good case for the feasibility of a probabilistic HeatRisk.

PMinT

- Underdispersive (too many obs falling outside) and bimodal (mixed bias).
- Lack of useful precomputed thresholds to further investigate.
- Encouraging reliability for Red Heat Impact Level, but with some issues at high probabilities.

PQPF24

- Poor predictability and low sample size of convective precip results in rather poor performance in bulk with severely underdispersive distribution.
- Despite poor performance in bulk, threshold probabilities for lower amounts (<0.5”) still show considerable reliability.
- Even looking at a synoptic event, there remained inconsistent bias, which challenges a meteorologist to manually determine a direction to bias correct toward or to "pick a winner" in percentile space.