Probabilistic Verification of Calibrated CDFs

John Wagner
Dana Strom
Daniel Eipper
Erin Thead
Gavin Harrison
Gordana Sindic-Rancic
Keenan Stone
Michael Levine
Samantha Walley
Tamarah Curtis

Introduction

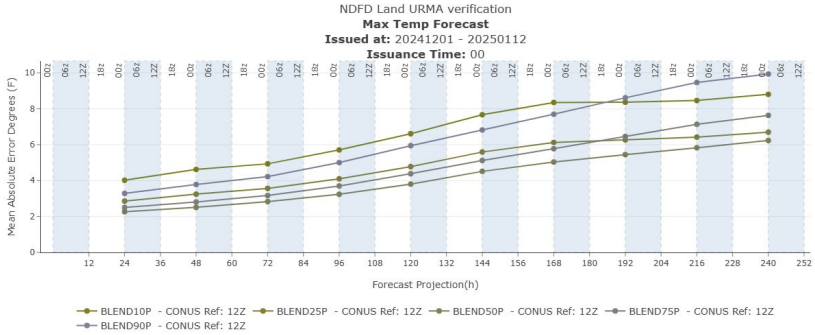
Cumulative Distribution Functions (CDFs) are used to express the expected error of calibrated probabilistic forecast systems. Probabilistic verification scores will be introduced to measure the accuracy, reliability, and quality of CDFs. We will be introducing the following scores to the Verification Viewer to verify calibrated CDFs from the National Blend of Models and the Weather Prediction Center:

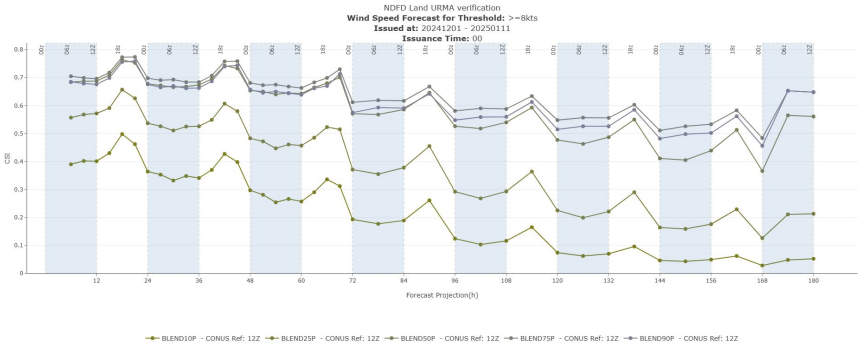
- Continuous Ranked Probability Score (CRPS)
- Spread-Skill Score
- Probability Integral Transform (PIT) Histograms
- Receiver Operating Characteristic (ROC) Curves

Probabilistic verification will be computed in our new python-based verification system. Scores will be computed on Amazon Web Services (AWS) in MDL's ParallelCluster, using data primarily from the NOAA Open Data Dissemination (NODD) Program.

Deterministic Scores

We are focusing here on probabilistic scores, but it is still important to look at the deterministic verification scores from calibrated CDFs. The verification viewer currently displays deterministic scores for 5 percentiles (10/25/50/75/90%) from the National Blend of Models for MaxT/MinT/ws/wg.





Types of Probabilistic Verification

1. CDF of Ranked Ensemble Members (Relative Frequency)
    - Typically used to tune an ensemble prediction system
    - Look at scores like CRPS and Spread-Skill Score (Ensemble Spread vs RMSE)
    - Can use MET to verify ensemble members, as MET will build a CDF
2. Full Cumulative Distribution Function
    - Save and share the moments of the distribution
    - Can compute true scores
3. Select percentiles from the calibrated CDF (highlighted in this presentation)
    - Often easier to share and understand
    - Computationally efficient verification, especially for estimated CRPS
    - Compare verification for more sources, provided that the same percentiles are available
    - Will have to make assumptions about data between and outside of the selected percentiles

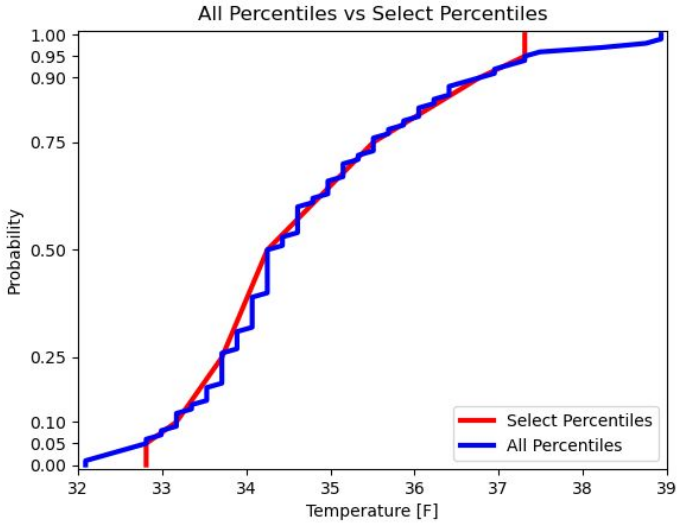
Continuous Ranked Probability Score

CRPS is a measure of accuracy and reliability of a continuous forecast. It is a negatively oriented score with a perfect score of CRPS=0. It is the squared difference between the forecast probability and the observation, which collapses to the MAE for a deterministic case.

$$CRPS = \int_{-\infty}^{\infty} (P(f) - P(o))^2 dx$$

Continuous Ranked Probability Score

A true CRPS is computed from a full CDF, shared either by the ensemble members or the moments of the distribution. An estimated CRPS is computed using select percentiles from the CDF. In doing so, we assume a linear relationship between the selected percentiles and no knowledge of the distribution outside of the lowest and highest percentiles.



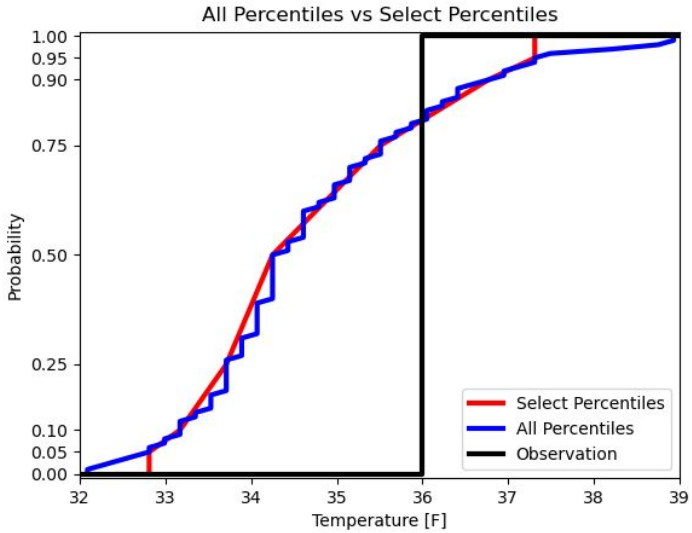
Blend V5.0 MaxT KDCA
Issued: 20250111 00z
Valid: 20250113 00z

$CRPS_{all}$ = 0.91
$CRPS_{sel}$ = 0.98

Continuous Ranked Probability Score

A Heaviside function is used display the deterministic observation as a non-exceedance probability.

$$H(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$



Blend V5.0 MaxT KDCA
Issued: 20250111 00z
Valid: 20250113 00z

$CRPS_{all} = 0.91$
$CRPS_{sel} = 0.98$

Continuous Ranked Probability Score

Why use CRPS?

- It is a proper score
- It can be estimated efficiently given a select number of percentiles
- It represents both the accuracy and reliability of the CDF, though accuracy will be the dominant part of the score for most elements

Continuous Ranked Probability Score

Hans Hersbach and Dave Unger have both decomposed the CRPS in literature. Unger's decomposition helps us understand why the CRPS will always be less than the MAE of the median.

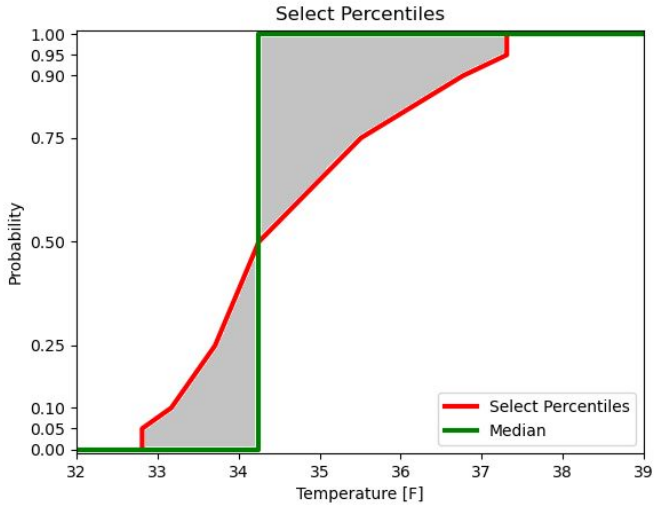
$$CRPS = S_0 + A - W$$

Where
S_0 is the uncertainty about the median
A is the absolute error of the median
W is the forecast likelihood of departure of the observation from the median

Continuous Ranked Probability Score

S_0 is a measure of the degree of uncertainty which is expressed in the forecast, and is independent of the verifying observation. For sharper CDFs, S_0 will be smaller.

$$S_0 = \int_{-\infty}^{\infty} Q^2(x)dx$$



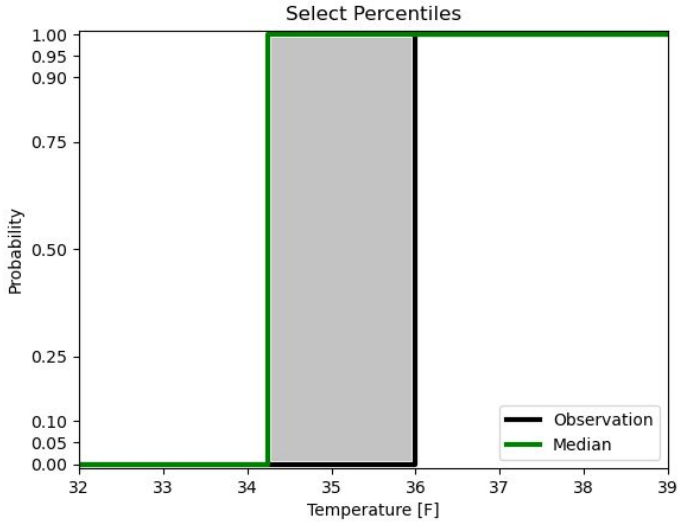
Blend V5.0 MaxT KDCA
Issued: 20250111 00z
Valid: 20250113 00z

S_0 = 0.313
CRPS = 0.901

Continuous Ranked Probability Score

A is the absolute difference of the observation from the median of the forecast distribution. This term measures the accuracy of the forecast independently of the uncertainty estimate.

$$A = |T - K|, \text{ where } T \text{ is the obs and } K \text{ is the median}$$



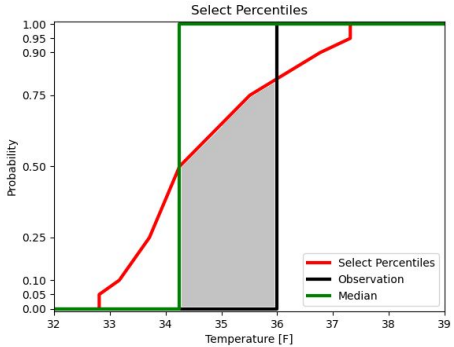
Blend V5.0 MaxT KDCA
Issued: 20250111 00z
Valid: 20250113 00z

A = 1.750
CRPS = 0.901

Continuous Ranked Probability Score

W is a measure of the forecasted likelihood of the departure of the verifying observation from the median. When departures are considered likely (the forecast probability that the observation will fall between the observation and median is small) W is close to A. When the forecast distribution in the range between the observation and median indicates that the verifying observation is expected to occur much closer to the median than it actually occurred, W << A. A larger value for W represents a better forecast.

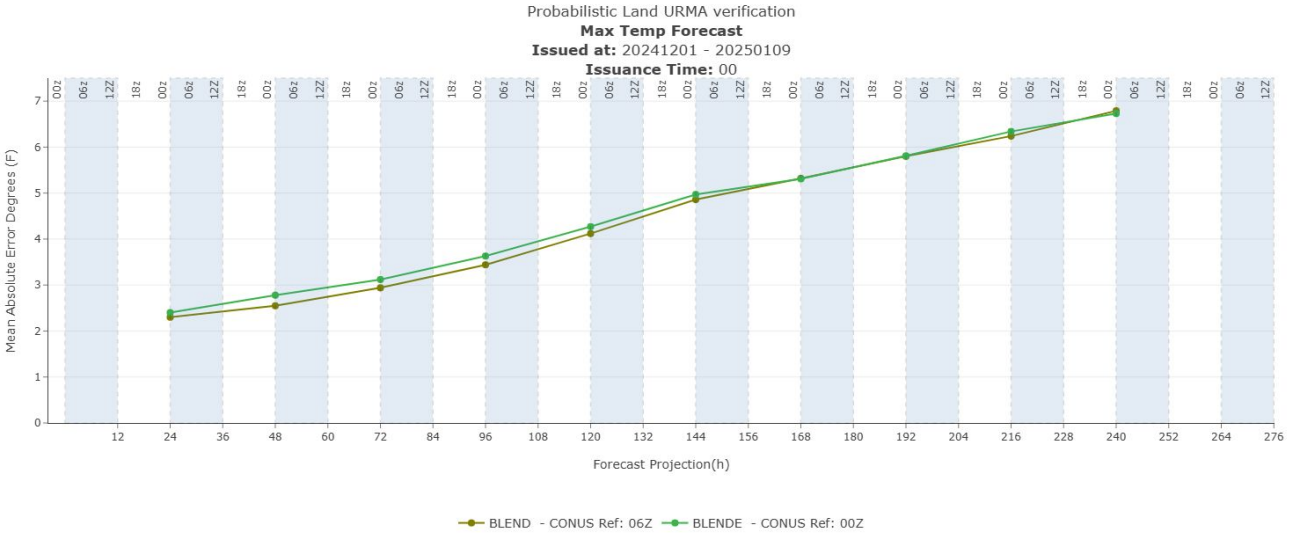
$$W = 2\left|\int_{T}^{K} Q(x)dx\right|, \text{ where } T \text{ is the ob and } K \text{ is the median}$$



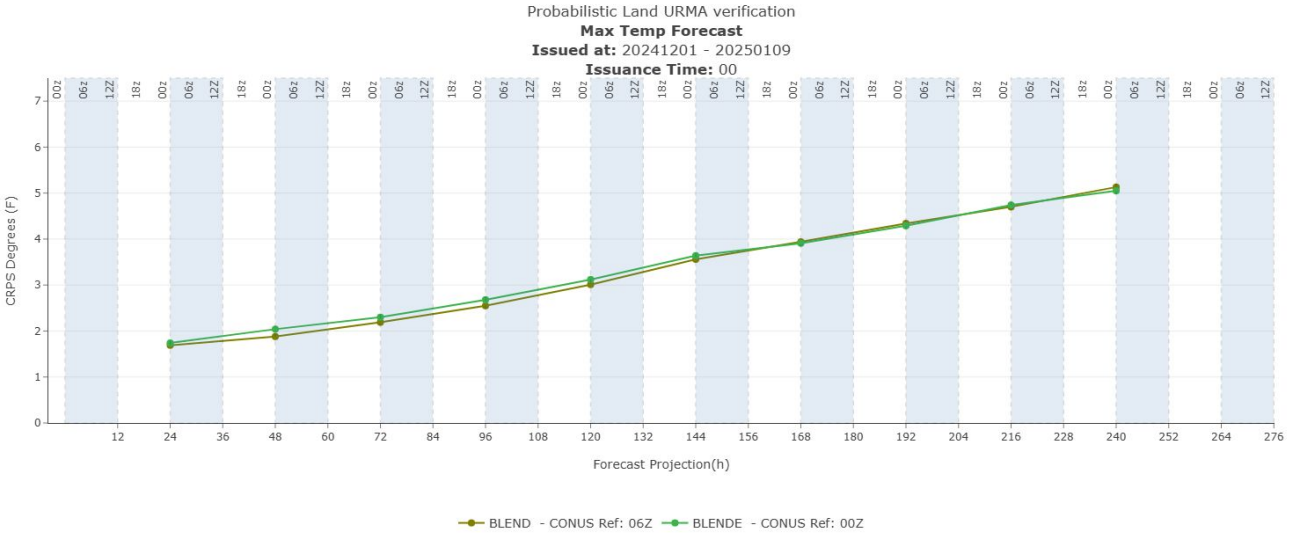
Blend V5.0 MaxT KDCA
Issued: 20250111 00z
Valid: 20250113 00z

W = 1.162
CRPS = 0.901

MAE vs CRPS


Probabilistic Land URMA verification
Max Temp Forecast
Issued at: 20241201 - 20250109
Issuance Time: 00
Mean Absolute Error Degrees (F)
Forecast Projection(h)
BLEND - CONUS Ref: 06Z
BLENDE - CONUS Ref: 00Z

MAE vs CRPS


Probabilistic Land URMA verification
Max Temp Forecast
Issued at: 20241201 - 20250109
Issuance Time: 00
CRPS Degrees (F)
Forecast Projection(h)
BLEND - CONUS Ref: 06Z
BLENDE - CONUS Ref: 00Z

Spread-Skill Score

This version of the spread-skill score is an **experimental** score that relates the accuracy and reliability of a probabilistic forecast. For a forecast with a **near-normal error distribution**, the ratio of the 50% spread to the MAE of the median should be near 1 for a reliable system, as they should both capture half of the expected error. Scores less than 1 indicate under-dispersion while scores greater than 1 indicate over-dispersion.

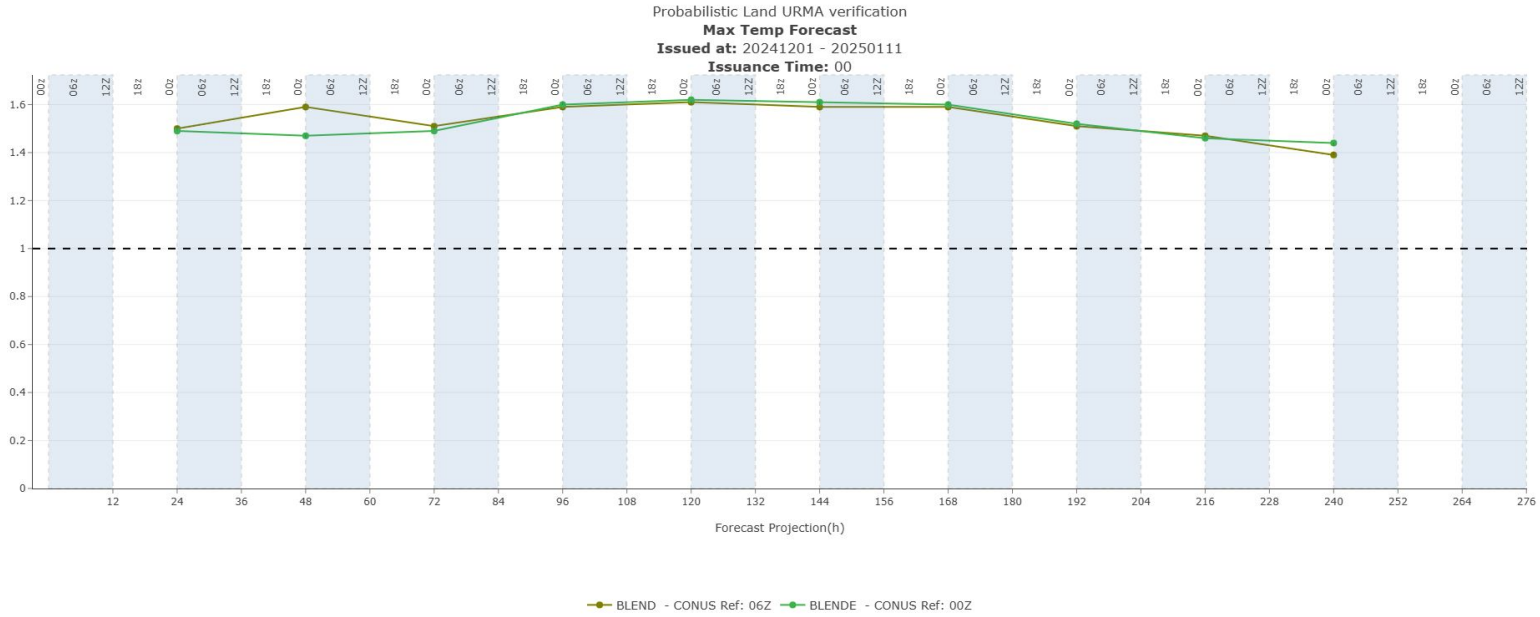
$$SSS = \frac{\frac{1}{N}\sum_{i=1}^{N}(P_{75i} - P_{25i})}{\frac{1}{N}\sum_{i=1}^{N}|P_{50i} - o_i|}$$

Spread-Skill Score

Why use a spread-skill score?

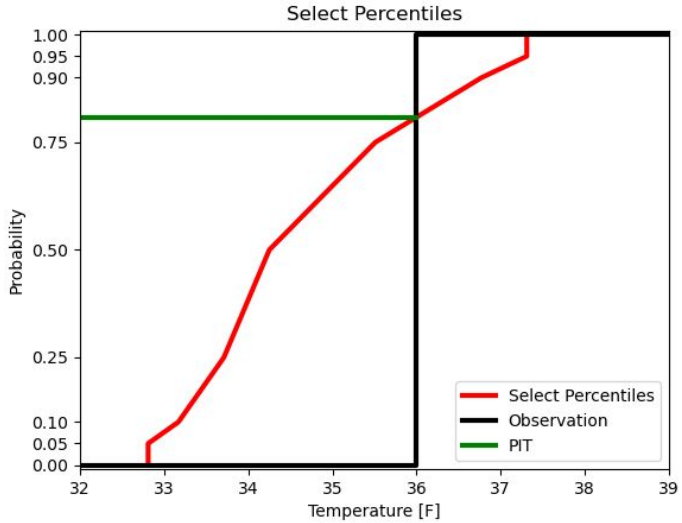
- A probabilistic forecast is reliable when the spread of the forecast represents the expected error of the forecast.
- Looking at the ratio of the average spread to the average skill is a quick way to assess how reliable a set of forecasts are and to compare the reliability of different forecast systems

Spread-Skill Score



Probability Integral Transform Histogram

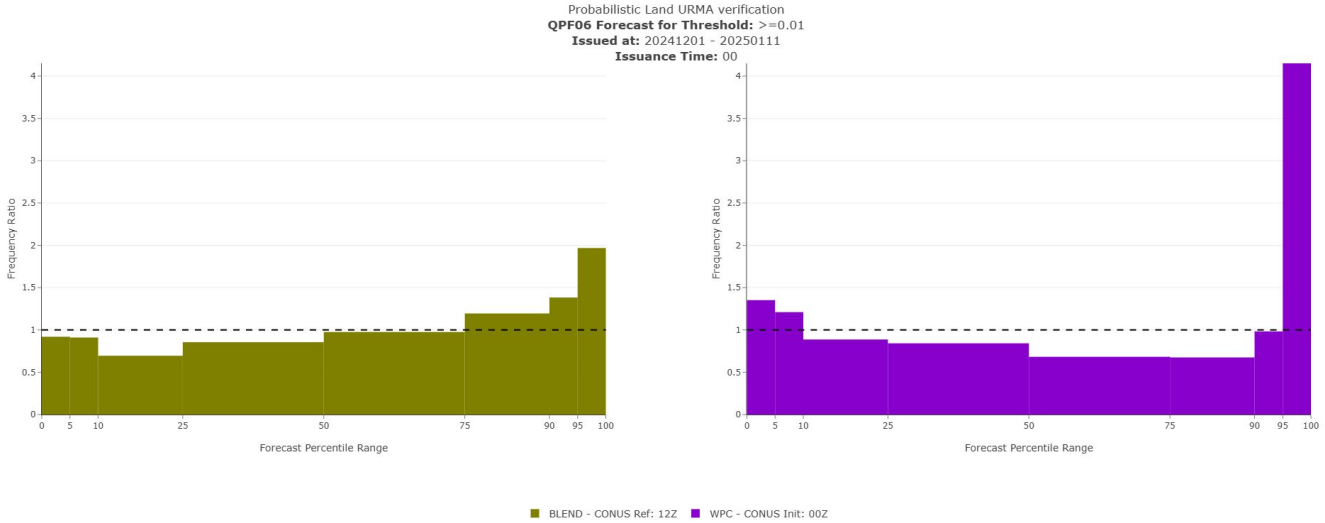
The Probability Integral Transform (PIT) is the percentile location on a CDF of the observation. A histogram can be made by placing each PIT into a bin, based on the selected percentiles. The count of each bin is divided by the expected number of hits for that bin.



Blend V5.0 MaxT KDCA
Issued: 20250111 00z
Valid: 20250113 00z

Probability Integral Transform Histogram

A PIT Histogram can be used to assess the reliability of a probabilistic forecast system. A reliable system will have a flat PIT Histogram with a height of 1. A PIT Histogram with a U-shape [O-shape] shows under-dispersion [over-dispersion]. A PIT Histogram skewed to the left [right] shows over-forecasting [under-forecasting].



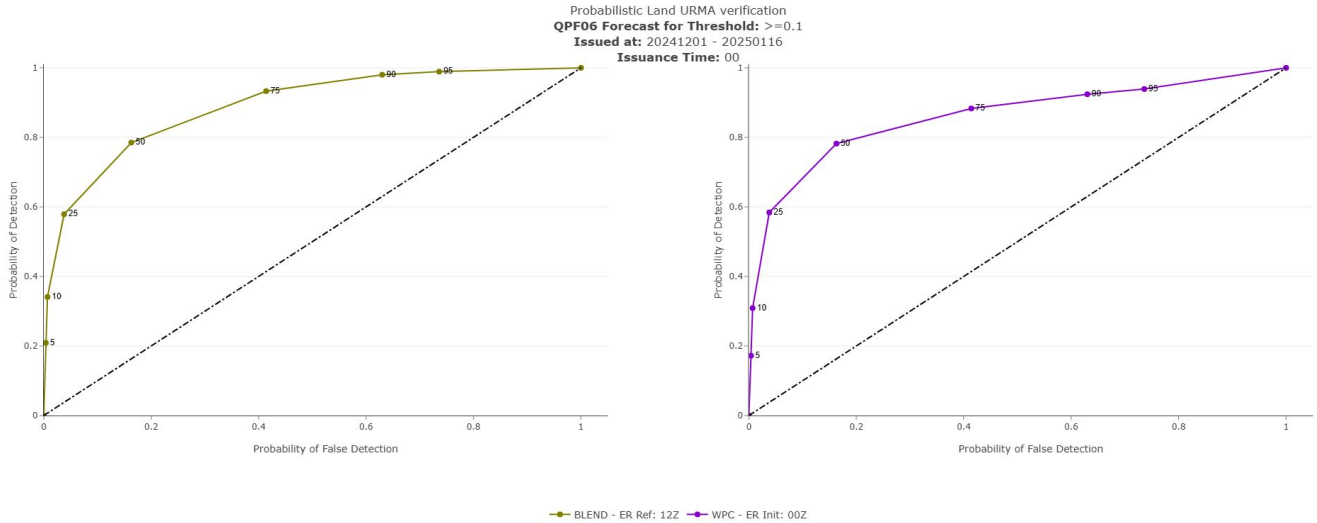
Probability Integral Transform Histogram

Why use PIT Histograms instead of Ranked Histograms?

- For calibrated CDFs, ranked histograms are generated using evenly spaced bins (e.g every 10%). A perfect ranked histogram will be flat at 1/# of bins (e.g. .10)
- PIT Histograms are more flexible. The percentiles used to split the bins do not need to be regularly spaced, but they do need to match for each source.
- Since the bins are normalized based on the expected number of cases, the perfect score (1) will always be the same regardless of the number of percentiles sampled.

Receiver Operating Characteristic Curve

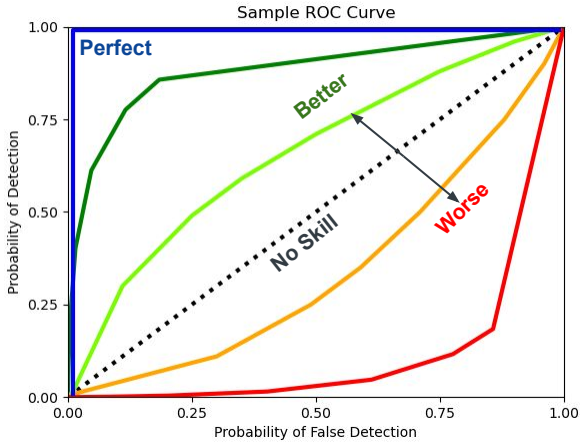
The ROC Curve is a plot of the Probability of False Detection (False Alarm Rate) and the Probability of Detection for each selected percentile (5, 10, 25, 50, 75, 90, 95). The ROC Curve displays the quality of the CDF. A perfect score would consist of a probability of detection of 1 and a probability of false detection of 0, resulting in all percentiles plotting at (0,1), with an area under the curve of 1.



Receiver Operating Characteristic Curve

Why use ROC Curves?

- Only score that we are adding that measures the overall quality of a forecast.
- Easy visual display to distinguish between skillful and unskillful forecasts, with skillful forecasts having points above the diagonal no skill line.
- ROC curves probably shouldn't be used to distinguish between two or more skillful models as they can be hard to visually compare. ROC Area Under the Curve (AUC; future enhancement) should be used to assess which model is of higher quality.

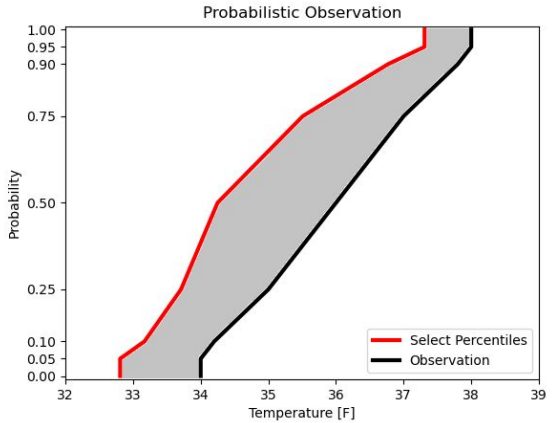


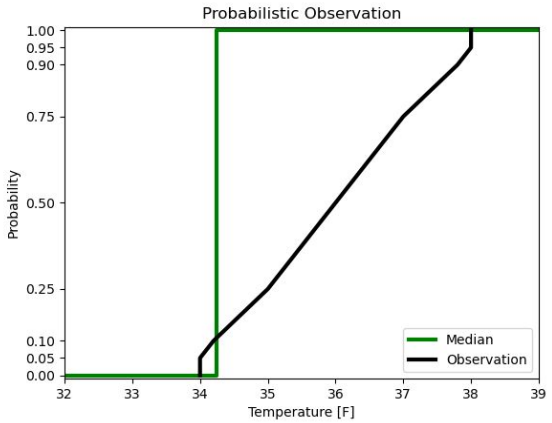
Probabilistic Observations

- This is a future enhancement that we would like to explore
- Like probabilistic forecasts, observations do not need to be deterministic values, expressed only as Heaviside functions.
- Uncertainty in station values and gridded analyses should be accounted for.
  - For station observations, including a measure of expected instrumentation error can be included.
  - For gridded analyses, a probabilistic URMA could include a measure of how representative the single gridpoint value is of the area it represents, including changes in terrain and elevation within the gridpoint.
- A probabilistic forecast should represent the expected error of the median forecast. This spread should include any uncertainty in the observed values used in the training sample.

Probabilistic Observations

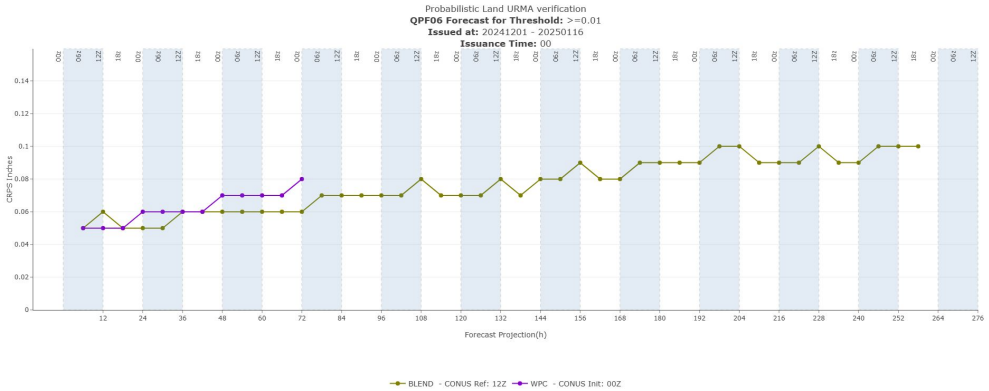
- Verification scores of how well the median forecast fits into the expected range of observations can be computed and could be used to justify large errors when there is more uncertainty in the observations.
  - CRPS can still be computed as the squared difference between two curves
  - PIT Histograms can be generated to show how well median forecasts fit into gridpoints that are not easily represented by a single value





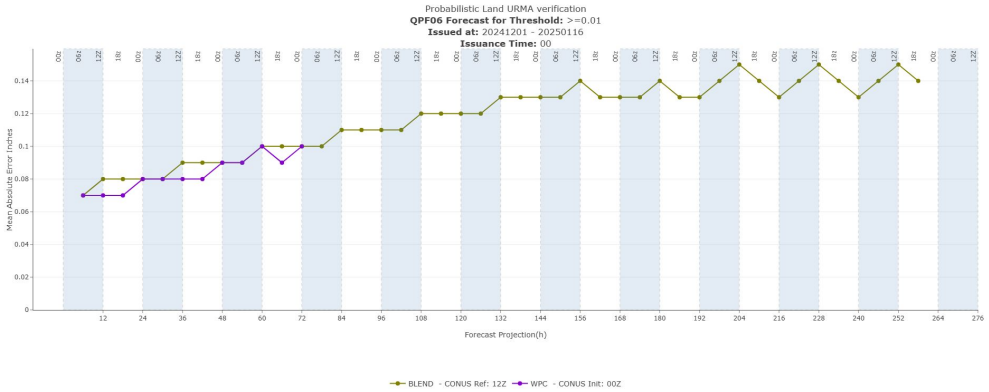
Key Takeaways

- The Verif team is currently finalizing scores and viewer enhancements. Once we have analyzed a sufficient amount of data, we will make these updates available on the verification viewer.
- Always look at multiple scores to get the whole story on probabilistic verification
  - Accuracy - Look at **CRPS** to get the accuracy of the CDF and the MAE/Bias/RMSE to get the accuracy of the median forecast
  - Reliability - Look at CRPS/SSS/PIT Histograms to measure the reliability of the CDF and the CSI/POD/FAR/GSS/PC/FB to measure the reliability of the ensemble mean for each threshold
  - Quality - Look at the ROC curves to determine whether CDFs are skillful or unskillful
  - Continue to look at deterministic scores from the percentiles



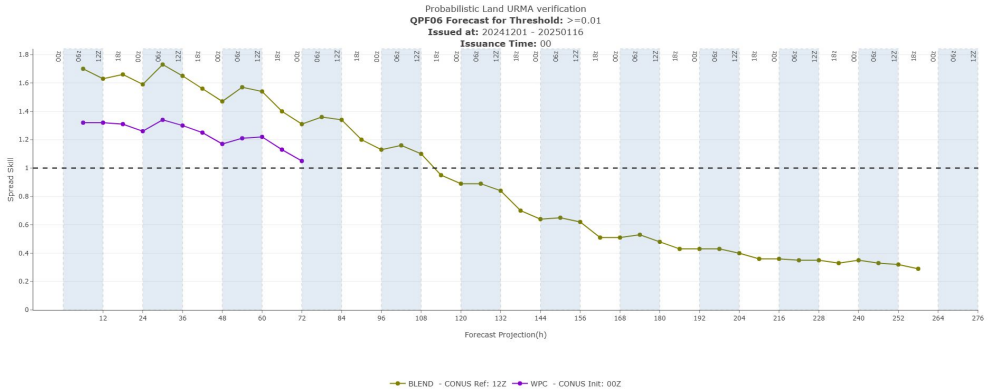
Key Takeaways

- The Verif team is currently finalizing scores and viewer enhancements. Once we have analyzed a sufficient amount of data, we will make these updates available on the verification viewer.
- Always look at multiple scores to get the whole story on probabilistic verification
  - Accuracy - Look at CRPS to get the accuracy of the CDF and the **MAE**/Bias/RMSE to get the accuracy of the median forecast
  - Reliability - Look at CRPS/SSS/PIT Histograms to measure the reliability of the CDF and the CSI/POD/FAR/GSS/PC/FB to measure the reliability of the ensemble mean for each threshold
  - Quality - Look at the ROC curves to determine whether CDFs are skillful or unskillful
  - Continue to look at deterministic scores from the percentiles



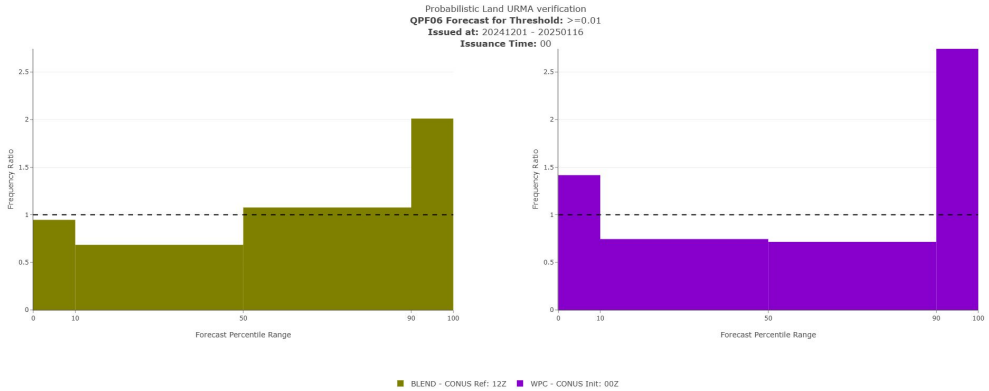
Key Takeaways

- The Verif team is currently finalizing scores and viewer enhancements. Once we have analyzed a sufficient amount of data, we will make these updates available on the verification viewer.
- Always look at multiple scores to get the whole story on probabilistic verification
  - Accuracy - Look at CRPS to get the accuracy of the CDF and the MAE/Bias/RMSE to get the accuracy of the median forecast
  - Reliability - Look at CRPS/**SSS**/PIT Histograms to measure the reliability of the CDF and the CSI/POD/FAR/GSS/PC/FB to measure the reliability of the ensemble mean for each threshold
  - Quality - Look at the ROC curves to determine whether CDFs are skillful or unskillful
  - Continue to look at deterministic scores from the percentiles



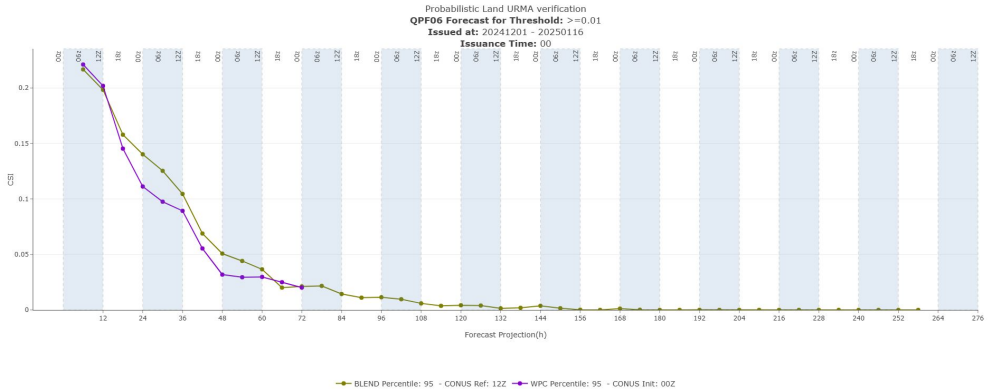
Key Takeaways

- The Verif team is currently finalizing scores and viewer enhancements. Once we have analyzed a sufficient amount of data, we will make these updates available on the verification viewer.
- Always look at multiple scores to get the whole story on probabilistic verification
  - Accuracy - Look at CRPS to get the accuracy of the CDF and the MAE/Bias/RMSE to get the accuracy of the median forecast
  - Reliability - Look at CRPS/SSS/**PIT Histograms** to measure the reliability of the CDF and the CSI/POD/FAR/GSS/PC/FB to measure the reliability of the ensemble mean for each threshold
  - Quality - Look at the ROC curves to determine whether CDFs are skillful or unskillful
  - Continue to look at deterministic scores from the percentiles



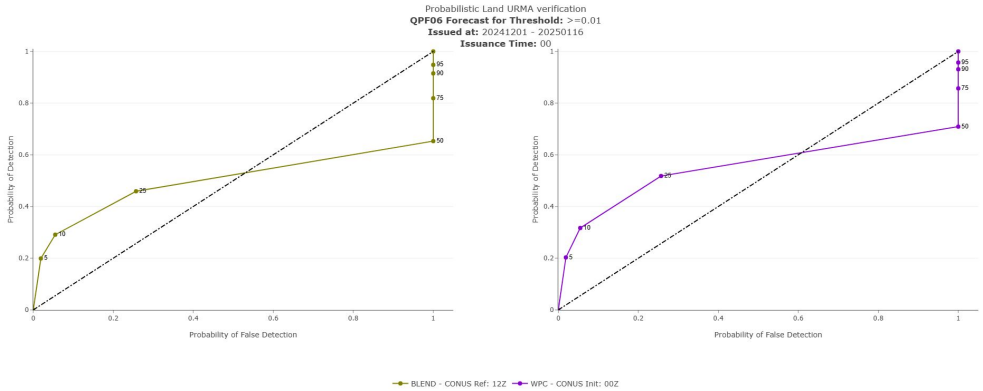
Key Takeaways

- The Verif team is currently finalizing scores and viewer enhancements. Once we have analyzed a sufficient amount of data, we will make these updates available on the verification viewer.
- Always look at multiple scores to get the whole story on probabilistic verification
  - Accuracy - Look at CRPS to get the accuracy of the CDF and the MAE/Bias/RMSE to get the accuracy of the median forecast
  - Reliability - Look at CRPS/SSS/PIT Histograms to measure the reliability of the CDF and the **CSI**/POD/FAR/GSS/PC/FB to measure the reliability of the ensemble mean for each threshold
  - Quality - Look at the ROC curves to determine whether CDFs are skillful or unskillful
  - Continue to look at deterministic scores from the percentiles



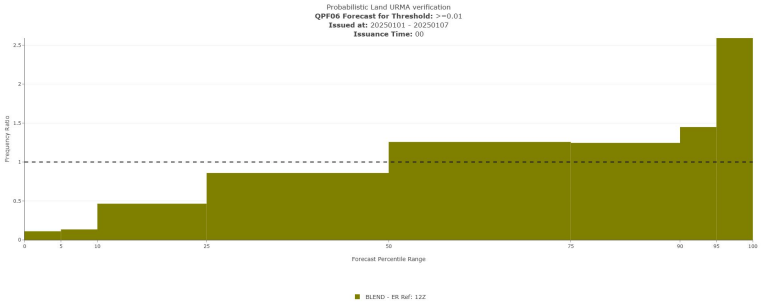
Key Takeaways

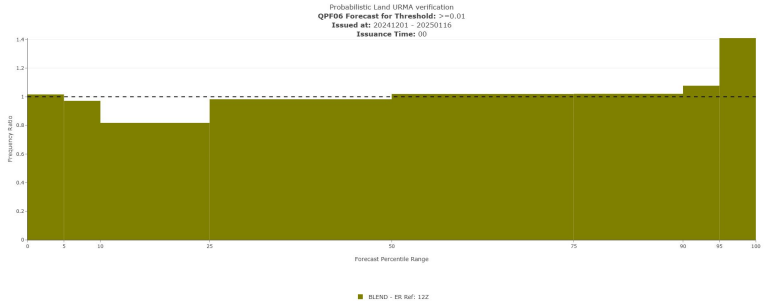
- The Verif team is currently finalizing scores and viewer enhancements. Once we have analyzed a sufficient amount of data, we will make these updates available on the verification viewer.
- Always look at multiple scores to get the whole story on probabilistic verification
  - Accuracy - Look at CRPS to get the accuracy of the CDF and the MAE/Bias/RMSE to get the accuracy of the median forecast
  - Reliability - Look at CRPS/SSS/PIT Histograms to measure the reliability of the CDF and the CSI/POD/FAR/GSS/PC/FB to measure the reliability of the ensemble mean for each threshold
  - Quality - Look at the **ROC curves** to determine whether CDFs are skillful or unskillful
  - Continue to look at deterministic scores from the percentiles



A Note About Sufficient Cases

- MDL's verification viewer allows you to slice and dice data in many different ways
  - Breaking up data by discontinuous dates
  - Focusing on high impact events
  - Looking at verification for a single day
- **It is important not to make decisions based on small samples of data.**
- Looking at a limited number of cases can lead to misleading results, especially when looking at scores that focus on reliability.
- When possible, a sufficient amount of data should be selected. What constitutes a sufficient amount of data can vary based on element, threshold, and area of interest.





PIT Histograms - Eastern Region - Blend 00z QPF06

7 days 7 weeks

References and Links

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Unger, D. A., 1985: A method to estimate the continuous ranked probability score, Preprints *Ninth Conference on probability and Statistics in Atmospheric Sciences*, Virginia Beach, Amer. Meteor. Soc., 206-213.

EMC Verification Page - https://www.emc.ncep.noaa.gov/users/verification/

MET User's Guide - https://met.readthedocs.io/en/main_v11.0/Users_Guide/index.html

MDL Verification Viewer - https://veritas.mdl.nws.noaa.gov/ndfd-stats/comparative/verification.php

NODD - https://www.noaa.gov/information-technology/open-data-dissemination

What Are Accuracy, Reliability, and Quality?

From “What is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting” by Allan Murphy:

Accuracy is the average correspondence between individual pairs of forecasts and observations.

Reliability is the correspondence between the conditional mean observation and conditioning forecast, averaged over all forecasts.

Quality is the degree of correspondence between forecasts and observations. Forecasts of high quality exhibit a close correspondence with the observations.