Central Region Technical Attachment
Number 21-03
December 2021

Ensemble Streamflow Probabilistic Categorical Verification

Scott D. Dummer, Development and Operations Hydrologist
National Weather Service, Missouri Basin River Forecast Center, Pleasant Hill, Missouri

A. Juliann Meyer, Senior Hydrologist
National Weather Service, Missouri Basin River Forecast Center, Pleasant Hill, Missouri

1. Introduction

There are several categorical verification measures available for probabilistic forecasts. Some of these as mentioned by Wilks (1995), Jolliffe, et al. (2004), and Reconda et al. (2006) are Brier score, ranked probability score, rank histogram, distributions-oriented measures, resolution reliability, discrimination, sharpness, and relative value. As further discussed by Reconda et al. (2006) there are eight measures that can be considered to answer the question, What makes an ensemble forecast “good”? These measures look at accuracy, bias, association, skill, reliability, resolution, sharpness, and spread. The goal for this project was to use a measure or set of measures that would be easy to understand by the many users of probabilistic Ensemble Streamflow Prediction (ESP) outlooks. For this, we chose a simple categorical verification that measures the reliability of the ESP forecasts over the exceedance forecast probability range (i.e., 95, 90, 75, 50, 25, 10, and 5 exceedance probabilities). This analysis attempts to answer the question how often are the probabilistic forecasts being exceeded by the observational record, (i.e., how often is the 50% exceedance forecast verified). Forecast error and bias are also considered over the entire probability range.

This paper describes the categorical verification of the probabilistic ESP outlooks by measuring their accuracy, forecast bias, and reliability. The paper also covers the website that makes the results available to users and other interested parties. The Missouri Basin River Forecast Center (MBRFC) provides monthly 90-day ESP outlooks for most of its river forecast locations. The ESP system utilizes historical precipitation and temperature data as input along with the current river, snow, and soil moisture conditions in the hydrologic river model to produce long-range probabilistic outlooks. These long-range outlooks provide information on future river stage, flow, and volume possibilities. The categorical verification of exceedance probabilistic outlooks, or forecasts will assist the MBRFC to determine where to focus development efforts to improve the river model accuracy, and quality of the probabilistic outlooks. Another benefit of the MBRFC probabilistic verification services is that the National Weather Service (NWS) Weather Forecast Offices (WFOs) can use this information as part of their hydrologic Impact Decision Support Services (IDSS) they provide to emergency management, water managers, and other consumers.

2. Background – ESP

The ESP system produces long-range probabilistic forecasts of discharge, stage, and volume within the Community Hydrologic Prediction Service (CHPS) model and Flood Early Warning System (FEWS) framework. This system assumes that historical temperature and precipitation data are equally likely to occur in the future. In other words, ESP assumes that temperature and precipitation patterns that occurred in the past are representative of events that may occur in the future. These historical precipitation and temperature data are used as forcings to create many equally likely scenarios of future hydrologic conditions each starting with the current hydrologic conditions. This system is designed to run a large river system with many years of historical data where each year is a particular scenario. These scenarios are then summarized to provide probabilistic stage, flow, and volume forecast information on the order from weeks to seasons. Graphical depiction of this information is available on the NWS's Hydrology Web Services (https://water.weather.gov/ahps/).

ESP outlooks were implemented at MBRFC over a twenty-year period. This included collecting historical precipitation and temperature data for computation of mean areal precipitation (MAP) and mean areal temperature (MAT) datasets going back to 1948. These data sets are used for both calibrating the model and as model forcings for ESP. Historical time-series were collected for over 800 modeled locations and hundreds of diversion and return sites for calibration and model evaluation. In addition, rules of operation for 82 reservoirs were modeled. The Missouri Basin is split into 25 forecast groups. So, implementation of ESP was accomplished by configuring and calibrating each of these forecast groups individually to enable ESP computations and generation of outlooks. The first ESP outlook of any kind was issued in January 2001 for a nine-month valid period for selected locations in the Upper Missouri and Big Horn forecast groups; the Yellowstone forecast group was implemented in March of that year. These longer period outlooks were issued to capture the spring-snowmelt peak events. The first ESP 90-day outlook was issued for 21 forecast locations in the Siouxs forecast group (Floyd, Little Sioux, Big Sioux and their tributaries). The ESP outlook for the last of the twenty-five forecast groups, the lower Missouri River Mainstem, was issued in October 2018. The current implementation of the Hydrologic Ensemble Forecasting System (HEFS) at the Missouri Basin RFC will utilize the modeling framework done for implementing ESP outlooks.

3. Analysis Procedure

River verification requires a continuous observational record that is typically only available with an automated recorded river gage. The Missouri Basin RFC currently provides forecast services for 432 river gage locations. Of the 432 locations, 392 have an automated gage with a continual observational record. Unfortunately, 40 forecast locations in the Missouri Basin do not have a gage or are manually read, so they lack the continuous observations to perform the desired verification. As such these locations are not included as part of the river verification analysis.

For forecast point locations where adequate data are available, a simple categorical verification that measures the reliability of the ESP forecasts over the exceedance forecast probability range the 90-day valid period was done. Exceedance probability information is provided for 95%, 90%, 75%, 50%, 25%, 10%, and 5% categories. If the outlook indicated a stage value should be exceeded 50% of the time, was it actually exceeded 50% of the time? By looking at the distribution of the categorized data, one can provide the user with the forecast probability accuracy and probability bias of the outlook for a particular forecast point location. These outlooks are issued generally once a month. The analysis determines the observed exceedance percentage (OEP) for each location and forecast exceedance probability (FEP). The analysis for each location and FEP category consists of the following steps:

(1) Find the maximum observed value for each 90-day forecast period
(2) Count the number of times the maximum observed value was greater than the FEP value for each category
(3) Count the total number of forecast issued for each FEP category
(4) Calculate the category forecast error (FE), that is the difference between FEP category and OFP
(5) Calculate the net mean absolute error for all categories (netMAE95to5). This is the sum of the absolute value of each forecast error divided by the number of categories.
(6) Calculate the net mean error for all categories (netME90to10). This is the sum of the sum of the forecast errors divided by the number of categories.

Forecast quality is the average MAE95to5 of all the locations for that WFO or forecast group (FGRP). Forecast bias is the average of the ME95to5 of all the locations for that WFO or FGRP.

For the overall forecast quality, the distribution of the MAE values was reviewed and a forecast quality distribution was determined. The forecast quality categories were divided into five groups, define as:

➢ very good	>= 0% and < 5%
➢ good	>= 5% and < 10%
➢ fair	>= 10% and < 20%
➢ poor	>= 20% and < 30%
➢ very poor	>= 30%

4. Example Calculation

For this example, the forecast location, Big Creek near Blairstown, Missouri (BLRM7) is used. Blairstown is within WFO – Kansas City/Pleasant Hill (WFO-EAX) hydrologic service area and is within the MBRFC Osage forecast group. The Missouri Basin RFC began issuing ESP outlooks for this location in August 2013. Between August 2013 and February 2020, there are 80 outlooks available that have a full 90-day set of observations available.

Figure 1 shows one ESP outlook for BLRM7 that was issued on October 26, 2019 along with the observed data for the 90-day outlook window. This plot shows 1) the ESP probabilistic outlook as a whisker-box showing the 95%, 75%, 50%, 25%, and 5% exceedance values, and 2) shows the observations for the valid period.

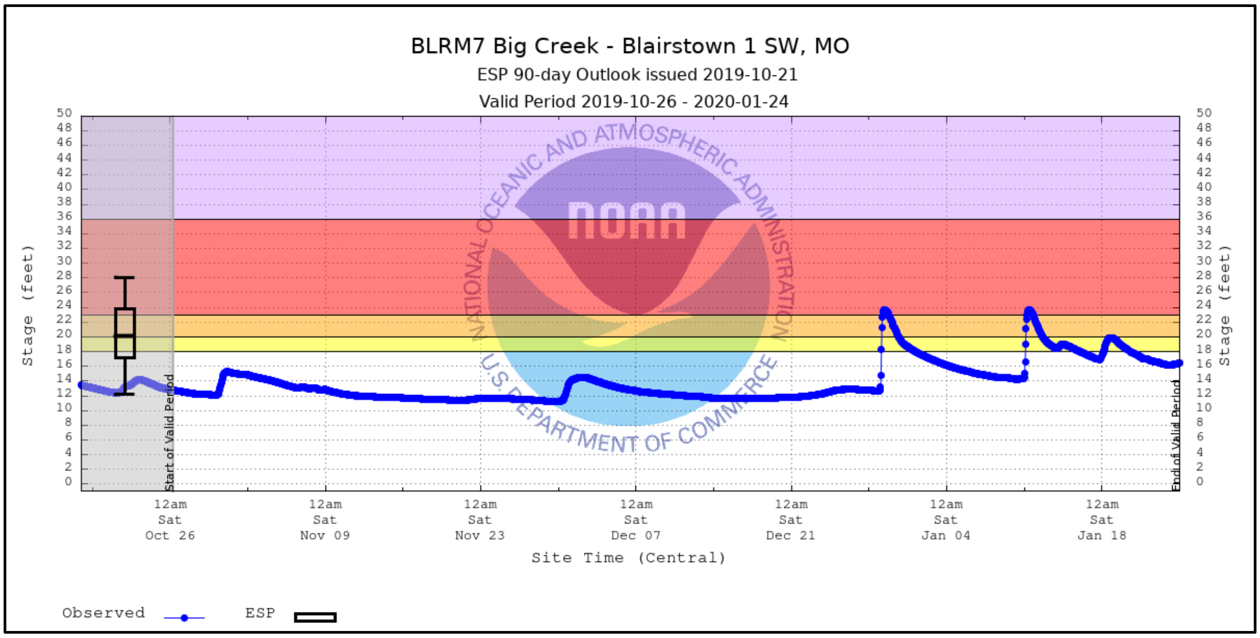

Figure 1. Example ESP outlook with observations for the 90-day valid period for BLRM7

Step 1 is to find the maximum stage that occurred. A review of the observations indicates the maximum stage that occurred of 23.68 ft. on December 29, 2019. The same step was done for the other 79 outlooks; a summary can be found in Appendix A.

The results for step 2 – 6 are shown in Table 1.

Table 1. Summary of Calculations (Steps 2-6) for BLRM7

	Exceedance Probabilities (FEP)						
	95%	**90%**	**75%**	**50%**	**25%**	**10%**	**5%**
# times exceeded	75	65	56	45	17	1	0
Total	80	80	80	80	80	80	80
OEP	95%	94%	81%	56%	21%	1%	0%
Fcst Error = FEP-OEP	0	-4	-6	-6	4	9	5

5. Group Calculations

In addition to calculating the individual forecast error and bias for each forecast point location, a consolidated forecast error and bias was calculated by WFO hydrologic service area and by RFC forecast group.

Forecast error, also known as forecast quality, is based on the magnitude of the mean absolute error (MAE). MAE is the average error of the absolute value between the difference of forecast and observed values for each FEP category, that is, netMAE95to5. The closer the MAE value is to zero the better the forecast.

Forecast bias is based on the magnitude of the mean error (ME). ME is the average error between the difference of forecast and observed value for each FEP category, that is netME95to5. A positive value indicates an over-forecasting bias, a negative value indicates an under-forecasting bias, while a zero indicates equal chance of over- or under-forecasting, i.e., no forecast bias.

6. Example Results for Group Calculations

For these examples, WFO Springfield, Missouri (SGF) and forecast group Osage are used. Figure 2 (WFO-SGF) and Figure 3 (Osage) show the forecast quality by the categories previously mentioned in section 2.

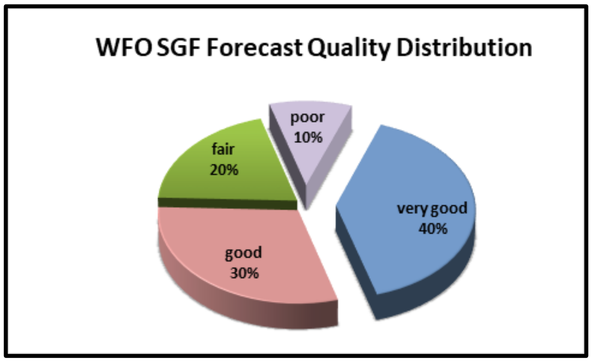

Figure 2. Forecast Quality for WFO SGF

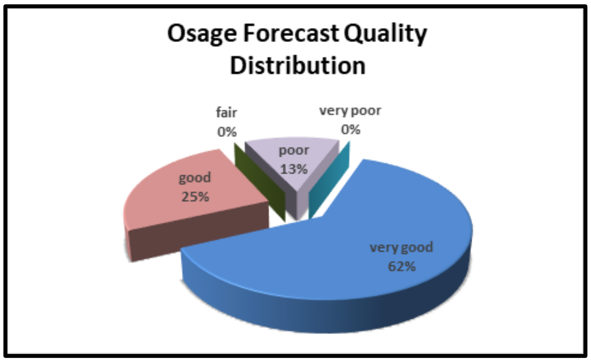

Figure 3. Forecast Quality for Osage FGroup

The forecast bias for WFO SGF and Osage forecast group are shown in Figures 4 and 5, respectively. Figure 4 shows that, in general, for the forecast point location in their area, the ESP outlook is more likely to under-forecast then over-forecast - 60% and 40% respectively.

However, in general, for the Osage forecast group the ESP outlooks are more likely to over-forecast than under-forecast.

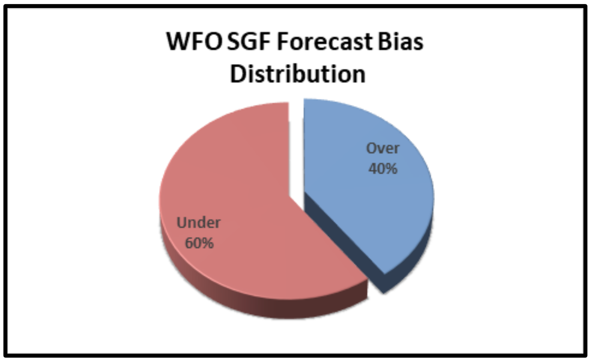

Figure 4. Forecast Bias for WFO SGF

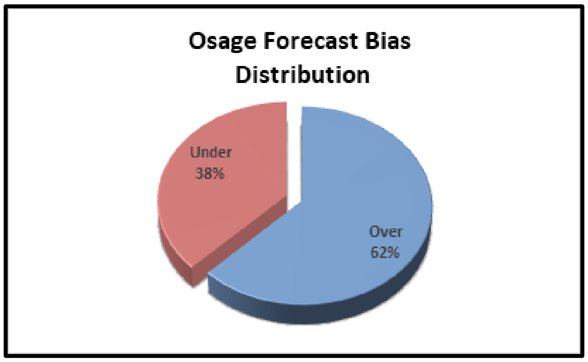

Figure 5. Forecast Bias for Osage FGroup

Figures 6 through 9 show the forecast error and bias for the individual forecast point locations that make up the WFO SGF area and the Osage forecast group.

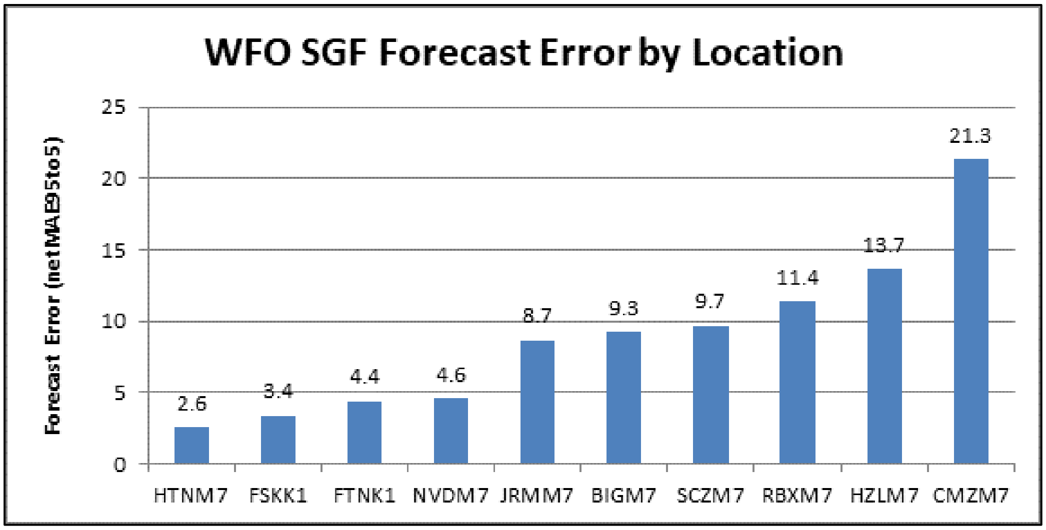

Figure 6. Forecast error by location for WFO SGF area

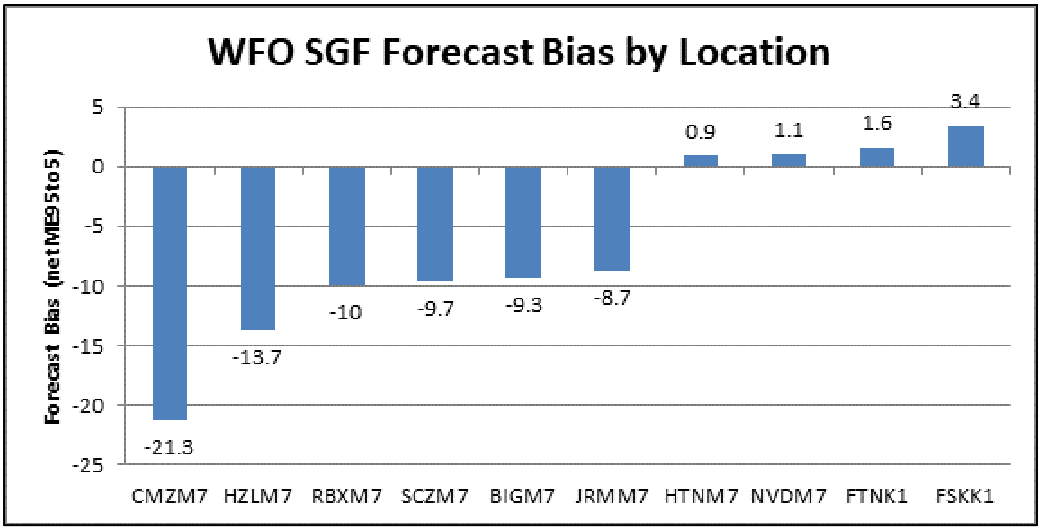

Figure 7. Forecast bias by location for WFO SGF area

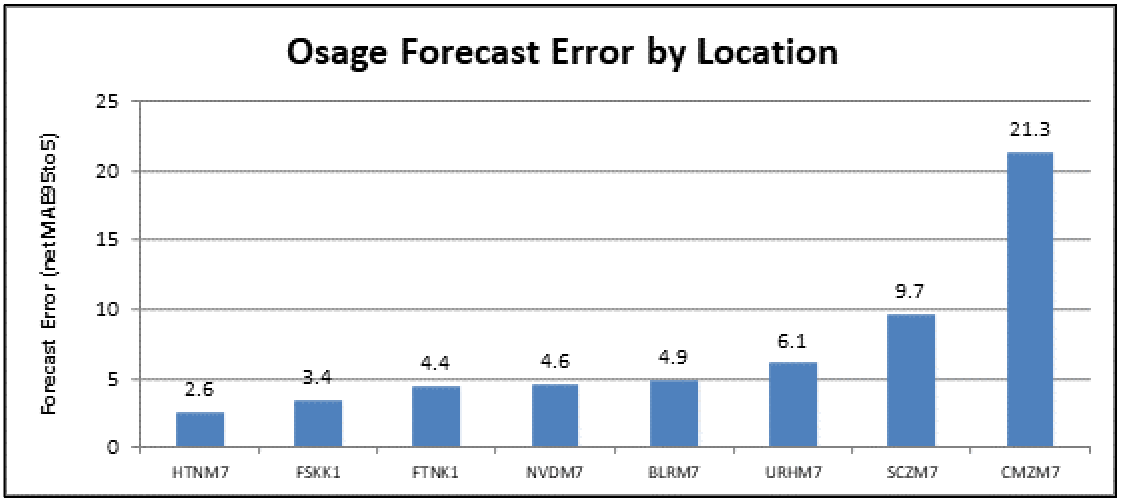

Figure 8. Forecast error by location for Osage forecast group

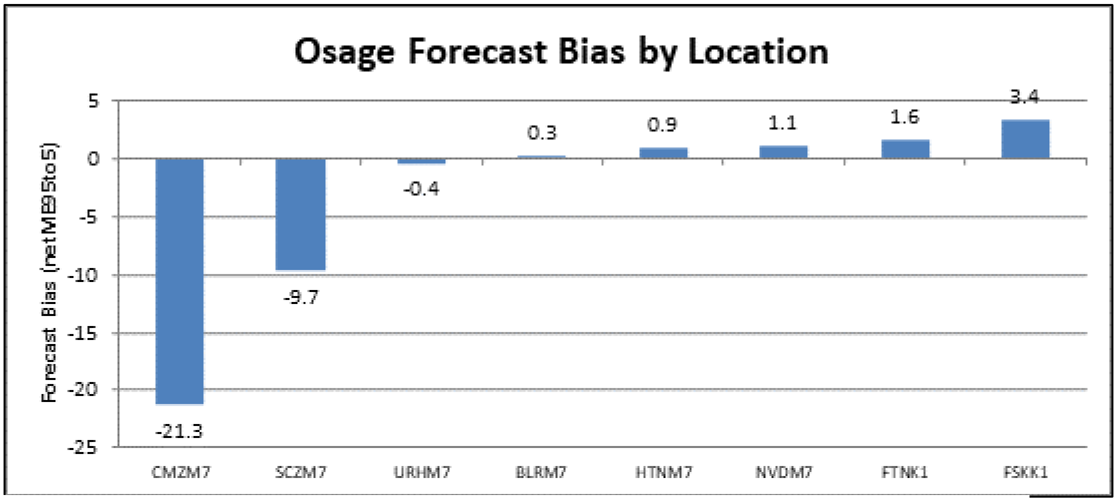

Figure 9. Forecast bias by location for Osage forecast group

7. Exceedance Probability Distribution Plots

For every forecast location, where adequate observations were available, the plot was generated of observed exceedance probability (OEP) versus the forecast exceedance probability category – an example is shown in Figure 10. In Figure 10 it is shown that forecast quality for Big Creek near Blairstown, Missouri was good (10% exceedance) to very good (95, 75, 50, 25, 5% exceedances) with a slight under-forecasting bias for the 95 through 50% exceedances, and a slight over-forecasting bias for the 25 through 5% exceedances.

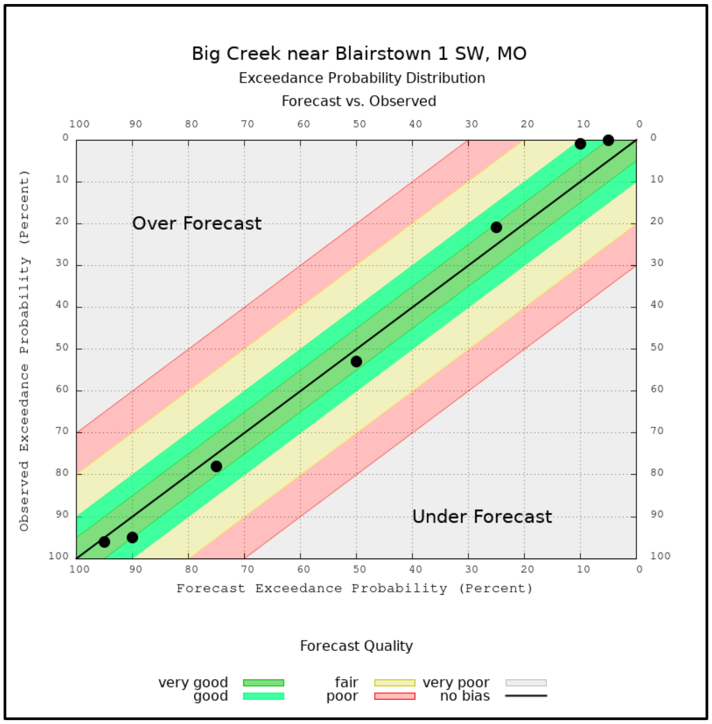

Figure 10. Exceedance probability distribution plot for BLRM7

8. ESP Verification webpages

The information described in the previous sections has been made available to the Missouri Basin RFC's users as part of their public website. It can be found at https://www.weather.gov/mbrfc/espvfy_main. These webpages include all the various figures for the forecast points by WFO and forecast group as shown in the examples (Figure 2 – 10) shown in this paper. In addition, links to the data analysis summary in both simple text and comma-delimited format are provided. This information can be utilized by the RFC's users to aid in their decision-support process.

9. Summary

Forecasting stage, flow and volumetric exceedance probabilities is critical for emergency managers, water managers, business owners, and the general public in order to properly assess risk and prepare in advance of flooding. Assessing and understanding the accuracy and bias of NWS probabilistic forecasts is critical for emergency management, water management, and decision makers. With this understanding the appropriate level of confidence in current NWS probabilistic forecasting capabilities, their limitations, and reasonable expectations can be set. Furthermore, well-informed, preemptive and mitigating measures can be taken due to these probabilistic Impact Decision Support Services+.

10. Acknowledgements

The authors wish to thank Jim Terrell (MBRFC Hydrologist) for his work in coding the spreadsheet macros, Peggy Lee (former MBRFC Hydrologist and NWC Development and Techniques Hydrologist), and Lisa Holts (MBRFC Senior Hydrologist) for their thoughtful review of the verification procedures, calculations and analysis. In addition, the authors thank Gregg Schalk (MBRFC Senior Hydrologist) for reviewing the section 2 (Background – ESP), and Kevin Low (MBRFC Service Coordination Hydrologist), Lisa Holts, and Briona Saltzman (MBRFC Hydro-meteorologist Analysis and Support Forecaster) for taking the time to review and provide feedback on this paper.

References

Jolliffe, I. T., & Stephenson, D. B., "Forecast Verification - A Practitioner's Guide in Atmospheric Science", Wiley, 2004

Regonda, S., Demarggne, J., & Seo,D.-J, "Hydrologic Forecast Verification", NWS, November 2006 https://www.weather.gov/media/owp/oh/hrl/docs/STEWksp_Training_Hydro_Verification_30Nov06.pdf

Wilks, D. L., "Statistical Methods in the Atmospheric Sciences", Academic Press, 1995

Appendix A

Table A-1. Summary of ESP Outlooks and maximum observation for Aug. 2013 – Oct. 2019 for BLRM7

Outlook Issued	Exceedance Probabilities							maximum observation
	95%	90%	75%	50%	25%	10%	5%	
2013-08-21	6.9	8	13.6	21.1	25.2	28.7	30	9.8
2013-09-24	6.6	9.9	12.8	17.3	21.5	26.9	28.3	9.8
2013-10-22	6.5	6.6	11	14.9	18.1	23.1	25	9.8
2013-11-26	13.8	15.3	16.1	18.2	22.3	25	26.9	16.75
2013-12-24	15.2	16.4	17.7	21.4	24.7	26.5	27.3	16.75
2014-01-21	15.7	17.2	18.9	22.3	25.1	27.3	28.4	27.44
2014-02-18	16.1	17.4	20.6	23.6	25.4	28.2	29.5	27.44
2014-03-04	17.8	19.1	22.7	24.7	26.4	28.5	29.6	27.44
2014-03-25	15.1	17.1	21.3	24.3	27.1	28.2	29.5	27.44
2014-04-22	9.5	14.6	21.8	24.7	27.3	28.5	29.5	24.19
2014-05-20	7.5	13	15.7	22	25.9	28.4	29.5	24.19
2014-06-25	9.4	12.5	16.6	21.5	26.5	29.3	30.9	23.76
2014-08-26	7.3	10.6	15.2	22.3	25.4	28.8	30.1	23.96
2014-09-23	6.7	12.6	14.4	19	23.6	27.2	28.4	23.96
2014-10-23	10.1	10.1	17.1	20.6	24.8	27.1	28	14.87
2014-11-25	13.1	14.5	16	17.9	21.5	24.7	26.5	14.87
2014-12-23	15	16.1	17.5	20.8	24.5	26.4	27.3	14.76
2015-01-20	15.5	17.1	18.6	22.3	25.1	27.3	28.3	21.59
2015-02-17	16.7	17.9	21.6	24.2	25.6	28.3	29.5	22.44
2015-03-03	16.4	17.6	21.8	23.9	25.7	28.2	29.5	23.27
2015-03-24	14.8	17.1	22	24.6	27.3	28.2	29.6	23.27
2015-04-21	14.1	16.4	22.4	25.2	27.5	28.6	29.5	23.64
2015-05-26	13.3	15.1	19.1	23.7	26.5	28.4	29.1	23.64
2015-06-23	12.3	14.3	18.8	23.1	27.2	29.4	31	23.64
2015-07-21	16.1	16.3	20.6	24.2	27.6	29.6	31.2	15.31
2015-08-25	7.3	12.9	17.3	24.3	27.1	29.1	30.8	15.31
2015-09-22	7.5	14.6	16.6	21.6	26.4	28	28.9	24.6
2015-10-20	6	6	13.3	16.8	21	24.8	25.8	24.6
2015-11-23	17.1	17.1	17.6	20.1	22.8	25.4	27.7	24.6
2015-12-21	16.3	17.8	20	22.7	24.8	26.7	27.3	22.63
2016-01-26	17.1	18.1	21.1	23.9	25.9	28.2	29.1	24.29
2016-02-16	17.5	18.6	22.8	24.7	26.6	28.7	29.6	24.29
2016-03-01	16.2	19	22.8	24.6	26.2	28.5	29.6	24.29
2016-03-22	15.2	17.1	20	23.5	26.3	28	29.4	24.29
2016-04-26	15.9	15.9	19.7	23.4	26.2	28	29.4	24.29
2016-05-03	7.5	11.8	18.7	23.6	26.4	28.4	29.5	24.29
2016-05-24	13.5	15.7	17.8	23.3	26.4	28.6	30.7	24.12
2016-06-21	8.7	11.6	16.3	20.2	26	30.9	36.4	23.59
2016-07-26	13.6	14.9	18.5	23.3	27.3	31.8	37.3	23.59
2016-08-23	3.8	10.6	17.1	23.9	27	31	36.3	23.59

Outlook Issued	Exceedance Probabilities							maximum observation
	95%	90%	75%	50%	25%	10%	5%	
2016-09-20	11.6	15	17.3	21.5	26.4	28.5	30	19.69
2016-10-25	4.2	4.2	16.3	18.6	23.4	26.1	27.8	12.92
2016-11-21	15.6	16.7	17.7	20.6	23	25.1	28	12.92
2016-12-19	15.9	16.4	17.6	20	23.4	26.3	27.1	12.92
2017-01-24	17.5	18.4	21.1	23.8	25.8	27.8	29.4	20.39
2017-02-13	17.5	18.3	21.8	24.4	26.3	29.5	32.2	26.68
2017-02-27	13.7	16.8	20.3	22.6	25.3	28.6	31.9	26.68
2017-03-21	14.4	16.6	18.7	22.5	25.9	28	31.5	26.68
2017-04-25	23.7	23.7	23.7	25.1	27.8	29	31.9	26.68
2017-05-23	12.2	14.2	17.1	21.9	25.4	28.2	30.4	27.04
2017-06-20	12.3	14.7	17.3	21.5	26.2	29.1	35.7	27.04
2017-07-25	12.2	14.2	17.2	20.7	25.9	31.3	36.6	27.04
2017-08-22	15.7	16.1	19.8	25.4	27.7	31.6	37.2	22.96
2017-09-26	10.5	15.4	16.9	22.4	25.2	27.8	28.8	22.96
2017-10-24	11.3	12.9	16.9	20.4	24.9	26.8	27.9	15
2017-11-21	12.1	15.6	16.3	19	22.9	25.6	26.9	21.67
2017-12-26	9.8	14.4	16.6	20.3	24.7	26	26.9	23.41
2018-01-23	16.1	16.9	19.1	22.3	24.8	27.1	28.5	23.41
2018-02-12	17.1	18.4	21.9	24	26.4	28.7	32.1	23.41
2018-03-19	15.7	18.6	22.4	24.4	26.8	28.9	32.1	23.41
2018-04-23	14.1	15.9	21.8	24.4	27.2	29.4	32.3	17.54
2018-05-21	8	10.6	16.1	22.7	26	28.3	32.1	17.54
2018-06-25	8	9.2	14.5	17	24.6	30.9	34.8	17.54
2018-07-23	13.5	14.2	15.5	19.1	26	33	35.5	18.85
2018-08-20	10.4	12.4	15.3	21.6	26.5	30.5	33.7	18.85
2018-09-24	10.2	11	14.8	17.5	22.9	27.2	28.5	21.85
2018-10-22	8	11.7	17.1	19.7	24.3	27.2	28.4	21.85
2018-11-19	13	13.9	16.9	20.6	23.5	26.9	29.5	21.85
2018-12-24	14	14.4	18.2	21.4	25	27.6	30	22.36
2019-01-21	16.1	20.6	21.9	24.6	26.6	27.8	30.8	22.36
2019-02-18	17.1	21.7	22.9	25	27.1	29.2	32.5	26.63
2019-03-04	23.7	23.7	23.7	25.4	27.2	29.4	32.6	26.63
2019-03-25	15.6	22.5	23.5	25.6	27.2	29.4	32.5	26.63
2019-04-22	10.9	14.4	21.4	24.6	27.3	29.5	31.9	28.13
2019-05-20	25.3	25.3	25.3	25.3	26.8	29.4	32.9	28.13
2019-06-24	11.7	12	13.5	21.2	26.3	29.7	36.5	28.13
2019-07-24	12.3	12.3	13.3	17.3	22.3	30.1	34.2	24.28
2019-08-19	18.6	18.6	20.3	23.6	26.3	30.3	34	24.28
2019-09-23	18.7	18.7	20.3	22.8	26.6	28.7	30.2	23.26
2019-10-21	12.2	14.4	17.2	20.1	23.8	26.7	28.1	23.68