From Random Forests to Flood Forecasts: A Research to Operations Success Story

Russ S. Schumacher*, Aaron J. Hill

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

Mark Klein, James A. Nelson

NOAA/NWS/NCEP/Weather Prediction Center, College Park, Maryland

Michael J. Erickson

Cooperative Institute for Research in Environmental Sciences, University of Colorado, College Park, MD and NOAA/NWS/NCEP/Weather Prediction Center, College Park, Maryland

Sarah M. Trojniak

Systems Research Group, Inc., Colorado Springs, Colorado, and NOAA/NWS/NCEP/Weather Prediction Center, College Park, Maryland

Gregory R. Herman†

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

Corresponding author: Russ Schumacher, russ.schumacher@colostate.edu

Current affiliation: Amazon.com, Seattle, Washington

Early Online Release: This preliminary version has been accepted for publication in *Bulletin of the American Meteorological Society*, may be fully cited, and has been assigned DOI 10.1175/BAMS-D-20-0186.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

Excessive rainfall is difficult to forecast, and there is a need for tools to aid Weather Prediction Center (WPC) forecasters when generating Excessive Rainfall Outlooks (EROs), which are issued for the contiguous United States at lead times of 1–3 days. To address this need, a probabilistic forecast system for excessive rainfall, known as the Colorado State University-Machine Learning Probabilities (CSU-MLP) system, was developed based on ensemble reforecasts, precipitation observations, and machine learning algorithms, specifically random forests. The CSU-MLP forecasts were designed to emulate the EROs, with the goal being a tool that forecasters can use as a "first guess" in the ERO forecast process. Resulting from close collaboration between CSU and WPC and evaluation at the Flash Flood and Intense Rainfall experiment, iterative improvements were made to the forecast system and it was transitioned into operational use at WPC. Quantitative evaluation shows that the CSU-MLP forecasts are skillful and reliable, and they are now being used as a part of the WPC forecast process. This project represents an example of a successful research-to-operations transition, and highlights the potential for machine learning and other post-processing techniques to improve operational predictions.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

Capsule summary. Collaboration between university researchers and an operational forecast center led to the development and implementation of a new machine-learning based forecast system for excessive rainfall.

1. Introduction

Excessive rainfall, and the flash flooding it often causes, remains one of the most difficult forecast challenges in meteorology (Fritsch and Carbone 2004; Sukovich et al. 2014; Novak et al. 2014). Unlike many other weather hazards, the definition of "excessive rainfall" varies from location to location, depending on the local and regional climatology. Likewise, the flooding that results from a given amount of precipitation depends strongly on the characteristics of the underlying land surface, among other factors. One important operational forecast product that highlights the potential for excessive rainfall and flash flooding is the Excessive Rainfall Outlook (ERO; NOAA Weather Prediction Center 2021; Erickson et al. 2021; Burke et al. 2021) issued by the NOAA Weather Prediction Center (WPC). The ERO is issued for forecast days 1, 2, and 3. Day-1 forecasts are issued initially at 0900 UTC with regular updates at 1600 and 0100 UTC and unscheduled updates as needed. Day 2 and 3 forecasts are issued initially at 0900 UTC with an update for each at 2100 UTC (for the 0900 UTC outlooks, this nominally covers the 3–27, 27–39, and 39–63-h forecast periods).

One difficulty in generating the daily ERO is that numerical weather prediction (NWP) models output quantitative precipitation forecasts (QPFs), but do not provide direct information about whether the forecast precipitation is excessive for a given location. Furthermore, NWP models have substantial biases for heavy precipitation (e.g., Herman and Schumacher 2016), limiting a forecaster's ability to use the QPFs directly. Many methods have been designed to post-process QPFs from NWP (e.g., Hamill and Whitaker 2006; Gagne et al. 2014; Hamill et al. 2015; Scheuerer

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

and Hamill 2015; Hamill and Scheuerer 2018; Whan and Schmeits 2018; Loken et al. 2019), but these approaches have not focused specifically on rainfall that is excessive with respect to the local climatology. To address these challenges, the authors have developed, evaluated, and transitioned into operations a forecast system based on NWP model reforecasts, historical observations of excessive rainfall, and machine learning algorithms. The goal of this project has been to provide guidance that WPC forecasters can use as a "first guess" when developing their ERO. In this manuscript, we describe the research-to-operations process for this forecast system, known as the Colorado State University-Machine Learning Probabilities (CSU-MLP) system.

2. Background

The CSU-MLP system was first developed by coauthor G. R. Herman during 2015–2018 as a part of his PhD dissertation research at CSU. Details of the methods are provided in Herman and Schumacher (2018c,a), but a brief overview is given here. Probabilistic forecasts are generated using random forests (RFs; Breiman 2001), which are a decision-tree-based machine learning method. The training data for the RF models include approximately ten years of forecasts from NOAA's Second-Generation Global Ensemble Forecast System Reforecast (GEFS/R; Hamill et al. 2013) as predictors, along with occurrences of excessive rainfall from the Climatology-Calibrated Precipitation Analysis (CCPA; Hou et al. 2014) and local reports of flash flooding representing the predictands. Specific definitions of excessive rainfall used in training are discussed further below. The GEFS/R inputs include the QPF along with several atmospheric variables, including precipitable water, mean sea level pressure, CAPE, CIN, 2-m temperature and relative humidity, 10-m wind components, and vertical wind shear. The ensemble median of these variables is used and the GEFS/R information is included from not only the grid point of interest, but over a 3-grid-point radius surrounding that point, every 3 or 6 hours through a 24-hour forecast period.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0186.1.

The CONUS is divided into eight regions (Fig. 1), and RF models are trained for each of these regions.

Real-time probabilistic forecasts are then made using these trained regional RF models and ensemble NWP output. The probabilities for each region are merged together into a single grid of probabilities across the CONUS, with smoothing applied across the regional boundaries to avoid sharp gradients in probabilities, and are displayed graphically to appear similar to the ERO (e.g., Fig. 2a,b). The dynamical model output driving the CSU-MLP forecasts originally came from the GEFS/R, which was routinely run in real time until September 2020; however, the GEFS/R was not an operational product and was subject to occasional supercomputer outages. The authors began using the operational GEFS output to drive the CSU-MLP for operational implementation. The forecasts were now being generated using a real-time dynamical ensemble that was different from the ensemble it was trained on, yet the subjective and objective evaluations showed that the quality of the forecasts were not substantially affected by this change. This enabled the potential for transition to operations at WPC.

Three versions of the CSU-MLP system (Table 1) will be discussed in this manuscript, representing changes that were made in response to feedback gathered during the Flash Flood and Intense Rainfall (FFaIR) experiments in 2017–2019 (e.g., Albright and Perfater 2018; Erickson et al. 2019; Trojniak and Albright 2019) and from WPC staff. The CSU-MLP was first formally evaluated in 2017; that version is as described in Herman and Schumacher (2018c,a) and is denoted herein as the 2017 version. In the 2017 version, forecasts were made for days 2 and 3. Two thresholds for excessive rainfall were used in training the RF models: exceedances of the 1-year and 10-year average recurrence interval (ARI) for 24-h precipitation accumulation using the NCEP Stage IV precipitation analysis (ST4; Nelson et al. 2016). ARIs were defined using NOAA's Atlas 14 where available, and older ARI estimates in other areas (Herman and Schumacher 2018c,b; NOAA Office

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

of Water Prediction 2021). The 2017 version was well received, but suffered from some obvious biases that will be discussed in more detail below. FFaIR participants also noted that the model too infrequently produced high probabilities of exceeding the 10-year ARI for that model to be useful operationally.

To address these concerns, several changes were made between the summers of 2017 and 2018 (summarized in Table 1). First, reports of flash flooding were also incorporated into the model training, so that either an ARI exceedance *or* a flash flood report would be counted as an "event". This was one way to increase the CSU-MLP probabilities, and also better reflect the risk of flash flooding (in addition to just heavy rainfall.) Second, only a single ARI threshold was used (rather than the two separate ARIs used in 2017), and the CCPA was used instead of or in addition to the Stage IV precipitation analysis in some regions (Table 1). The 2018 version resolved many of the issues that arose in the 2017 version, but also introduced new biases, especially with the probabilities being routinely too *high* in many of the regions where they were previously too low. Several additional improvements were made in the 2019 version of the CSU-MLP to address these limitations, including moving from 1- to 2-y ARIs in some regions, and a day-1 model was introduced in 2019. These improvements led to significantly better forecasts in the 2019 version, and as a result we will not present results from the 2018 version in this study in the interest of brevity. Finally, some additional minor improvements, namely using only the CCPA and further adjusting regional ARI thresholds, were made for the 2020 version of the CSU-MLP, which has now been transitioned into operational use at WPC. Quantitative evaluation of the 2017, 2019, and 2020 versions in comparison to the operational ERO will be presented below.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0186.1.

3. Data and methods

A major complication in forecasting excessive rainfall, and in evaluating those forecasts, is how to define "excessive," as no widely accepted definition exists. Flash flood reporting is inconsistent (e.g., Gourley et al. 2013), and thus using solely flash flood reports for forecast verification is inadequate. Therefore, proxies based on precipitation observations are also needed. The current definition of the ERO is based on rainfall exceeding flash flood guidance (FFG) within 40 km of a point (NOAA Weather Prediction Center 2021). However, the way FFG is computed can vary substantially between different NWS River Forecast Centers (e.g., Clark et al. 2014), sometimes resulting in discontinuities across RFC boundaries. This limits the usefulness of rainfall exceeding FFG as a sole definition of excessive rainfall. Frequency-based thresholds such as ARIs are another potential method for defining excessive rainfall, and this is the approach used in the CSU-MLP. But the specific ARI and accumulation period that best corresponds to flash flooding varies from place to place and from dataset to dataset (Herman and Schumacher 2018b; Gourley and Vergara 2021; Schumacher and Herman 2021). To address these complications, WPC developed the Unified Flood Verification System (UFVS; Erickson et al. 2019, 2021), which includes flash flood reports from NWS local storm reports, exceedances of FFG or the 5-y ARI, and reports of flooding from USGS stream gauges. A 40-km radius is applied around these point-based occurrences to match the neighborhood specified by the ERO. Gridded UFVS data are used as an independent source of observations for the forecast evaluation in this study. During the period of study, occurrences of excessive rainfall were most frequent in the eastern United States (Fig. 1), corresponding with a historically wet time period in many areas (e.g., NOAA 2020a).

CSU-MLP forecasts and EROs are evaluated against the UFVS using several metrics. As in Erickson et al. (2019), Hill et al. (2020), and Erickson et al. (2021), forecast reliability is assessed

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

in a spatial sense using the fractional coverage of observed events within probability contours; in other words, when a probability area of N% is forecast, how much of that area experiences excessive rainfall? Similar to Erickson et al. (2021), probabilistic forecast skill is evaluated using the Brier Skill Score (BSS) and the area under the Receiver Operating Characteristic (ROC) curve. In calculating the BSS, first the Brier score is calculated, using the midpoint of the probability ranges defined in the ERO (5–10% for marginal, 10–20% for slight, 20–50% for moderate, and 50–100% for high). CSU-MLP forecasts are also discretized to these probability categories prior to calculating the Brier score. Then, the daily frequency of UFVS occurrences over the four-year period 1 October 2016–30 September 2020, smoothed in time and space using a Gaussian filter, is used as the "climatological forecast" to calculate the BSS. The area under the ROC curve is also calculated using discretized forecast probabilities. Each version of the CSU-MLP was run over a common period for the purposes of evaluation and comparison with the ERO. Depending on the version and the date, these were either the real-time forecasts or were run retrospectively. For day-2 and day-3 forecasts, the evaluation is conducted for daily forecasts from 19 June 2018 through 15 October 2020; for day-1 forecasts a shorter evaluation period from 15 March 2019 through 15 October 2020 is used. The evaluation uses CSU-MLP forecasts driven by 0000 UTC initializations of the operational GEFS, and WPC EROs issued at 0900 UTC. Since late 2019, when the CSU-MLP was first implemented operationally at WPC for the day-2 and 3 forecasts, the ERO and CSU-MLP may not be completely independent, as WPC forecasters may have used the CSU-MLP as guidance. However, quantifying these influences is beyond the scope of this study.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

4. Results

a. Example forecast

Fig. 2 provides an illustrative example comparing the ERO and the CSU-MLP forecast for a day (12–13 April 2020) with widespread excessive rainfall and flooding in the southeastern US. A typical visualization of the CSU-MLP day-2 forecast alongside the analogous ERO is shown in Fig. 2a,b, and Fig. 2c,d shows the UFVS observations of excessive rainfall for the corresponding time period, along with quantitative evaluations for these forecasts. In this case, a broad moderate risk (20–50% probability) area was indicated in both the ERO and CSU-MLP forecast (Fig. 2) ahead of a strong synoptic-scale trough (not shown). Both forecasts skillfully highlighted the areas where excessive rainfall occurred. Although the ERO does not currently have a risk category between "moderate" and "high", the continuous probabilities produced by the CSU-MLP can be used to show areas with such intermediate probabilities. In this example, the 35% probability area in the CSU-MLP forecast (Fig. 2b) corresponded very closely to the observed excessive rainfall and flooding in northern Alabama and Georgia (Fig. 2d). With this example as context, we will proceed to the results of quantitative forecast evaluation.

b. Quantitative evaluation

The frequency of ≥ 10% probabilities (a slight risk or greater) in day-2 forecasts are displayed in Fig. 3 to illustrate the evolution of CSU-MLP versions, in comparison with the WPC ERO. The 2017 version had its highest frequency of 10% probabilities over New Mexico (Fig. 3a), a bias that was routinely noticed by participants in the 2017 FFaIR (Erickson et al. 2019); it was also apparent at the 20% probability threshold (not shown). This was found to be associated with the Stage IV precipitation analysis used to train the model: it often has single grid points

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

that exceed ARI thresholds in New Mexico, but these localized points rarely are associated with flash flooding (Herman and Schumacher 2018b). On the other hand, FFaIR participants noted that the probabilities seemed too low over portions of the midwest and southeastern US; these impressions are also supported when comparing Fig. 3a with Fig. 3d. The 2017 version also forecasts probabilities ≥10% in the northern Great Plains much more frequently than WPC.

Based on this objective and subjective feedback, several changes were made for the 2018 version of the model to address noted issues. The changes reduced the local maximum in New Mexico but also increased the probabilities across most of the CONUS such that higher probabilities were issued much too frequently (not shown). In the 2019 version of the model, these biases were addressed by changing the ARI thresholds in some regions (Table 1), such that the southwestern US and the northern Plains no longer stood out as having abnormally frequent issuance of high probabilities. Another area of concern arose in the 2019 model: the Midwestern US into the mid-Atlantic, which had probabilities exceeding 10% far more frequently than the WPC ERO (Fig. 3b). The Midwest region of the CSU-MLP system was adjusted for the 2020 version (Table 1), reducing this bias (Fig. 3c). The results shown in Fig. 3, and similar analyses at other probability thresholds (not shown), demonstrate that all versions of the CSU-MLP issue probabilities exceeding a given threshold more frequently than the ERO. How do these frequencies compare with the frequency of observed excessive rainfall, or in other words, how reliable are the probabilistic forecasts?

One way to address these questions is to evaluate the average fractional coverage of observed excessive rainfall within forecast probability contours. An example of this analysis is shown in Fig. 4, in which the desired result is for the bar representing fractional coverage to fall between the green and red lines defining the probability range for that threshold. Considering the Slight (10–20% probability) category, if the bar appears between the red and green lines, it indicates that, on average, when forecast probabilities in that range are issued, 10–20% of the area within

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

that contour has excessive rainfall within 40 km of a point. If the bar extends above the red line, it indicates that the forecast contours are (on average) too small; if the bar stays below the green line, the forecast contours are too large or issued too frequently. This method for assessing spatial reliability only evaluates forecasts where at least a 5% probability was issued, and therefore does not include information about missed forecasts (when excessive rainfall occurred but no probability ≥ 5% was forecast).

At all thresholds below High Risk (≥ 50%), the ERO probability contours (yellow bars in Fig. 4) were found to be too small in aggregate when using the UFVS as "truth," consistent with the findings of (Erickson et al. 2021). In contrast, the CSU-MLP forecasts generally fell within the specified probability ranges. The 2020 CSU-MLP versions had greater spatial coverage of observations within each probability range than the 2019 versions, consistent with the changes made to the model training between those versions and bringing the 2020 version closer to the characteristics of the ERO (Fig. 4). On days 2 and 3, despite its substantial biases in specific areas, the original 2017 version's overall reliability was closest to the ERO for most thresholds (Fig. 4).

As noted above, no widely accepted definition of excessive rainfall exists, so these estimates of forecast reliability should be taken with some caution. Recall that the current operational definition of the ERO is based only on precipitation accumulations exceeding FFG, whereas the UFVS includes more types of observations and proxies (and this broader definition is being considered for future implementation). Erickson et al. (2021) show that when considering FFG exceedances alone, the EROs are reliable across all probability categories, whereas considering the full UFVS causes the fractional coverage to fall above the probabilistic definition for some categories, similar to the results shown here. Thus, the finding that the CSU-MLP probabilities appear more reliable (Fig. 4) should not necessarily lead to the conclusion that they would be "better" if issued operationally. Nonetheless, these findings do demonstrate the promising result that the CSU-MLP is capable of producing well-calibrated probabilistic forecasts.

In terms of forecast skill, measured by the BSS with respect to daily "climatology" forecasts, the WPC ERO had greater skill than the CSU-MLP on day 1, whereas the 2020 version of the CSU-MLP had the greatest skill on days 2-3 (Fig. 5a). CSU-MLP skill improved with each updated version on days 1-2, whereas on day 3 the 2017 version slightly outperformed the 2019 version but still underperformed the 2020 version. These results mirror the findings of Hill et al. (2020) for severe weather forecasts, where operational outlooks performed better than machine-learning-based forecasts on day 1, but the opposite was true on days 2-3.

The area under the ROC curve, which measures the ability of different probability thresholds to discriminate between events and non-events (with higher values being better), shows generally higher values for the CSU-MLP than for the ERO (Fig. 5b). This largely stems from the larger number of observed excessive rainfall occurrences that fell outside of 5% probability contours in the ERO compared to the CSU-MLP (not shown; see also Williamson et al. 2021). Among CSU-MLP versions, the 2019 version had the highest ROC area on all forecast days; this result demonstrates the importance of considering multiple skill and reliability metrics as the ROC area does not incorporate any information about forecast bias (e.g., Developmental Testbed Center 2020) and the 2019 version was shown to have a high bias in some regions (Fig. 3). Nonetheless, when considering skill metrics in conjunction with reliability, the results of this study suggest that the skill of the ERO may increase if broader areas of probability were issued, especially on days 2 and 3.

Both the ERO and CSU-MLP day-2 forecasts were most skillful as defined by BSS in the south-central and southeastern United States where excessive rainfall was relatively frequent during the period of study, with low or negative skill in much of the western US (Fig. 6; day 1 and 3

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

forecasts exhibit similar spatial patterns, not shown). The CSU-MLP had negative skill (with large magnitudes) in many areas in the interior western US (Fig. 6a). These are areas with zero or one occurrence of excessive rainfall during the study period that had a few instances of probability forecasts exceeding 5%; this small sample size results in a negative BSS with large magnitude. This issue is muted in the ERO (Fig. 6b), where no excessive rainfall probabilities were issued in the study period (e.g., Fig. 3d). In contrast, in some of the western-US areas with relatively frequent excessive rainfall (e.g., the higher terrain of Washington, Oregon, and California, along with southern California), the CSU-MLP showed much greater skill. Two areas where the CSU-MLP outperformed the ERO were the southern US inland from the Gulf of Mexico coast (extending from central Texas eastward into Georgia) and the mid-Atlantic coast into the northeast (Fig. 6c). On the other hand, the ERO strongly outperformed the CSU-MLP in pockets of the northern Great Plains and Midwest (Fig. 6c).

When aggregated over the CSU-MLP regions (shown in Fig. 1), updates to the CSU-MLP version improved skill on day 2 across most regions (Fig. 7). However, three regions were notable exceptions: the SW, ROCK, and NGP, in which the original 2017 version was found to have the most skill. These results represent a paradox, as these were the regions identified as having unrealistic high biases in both subjective and objective evaluation (Fig. 3a). In other words, it seems that those high probability biases actually *benefited* that version in terms of forecast skill when using the UFVS as “truth”. It is possible that some of the same issues with defining excessive rainfall in these regions that affected the training of machine-learning models also appear in the UFVS. These are important challenges to address in future work.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0186.1.

c. Subjective evaluation

In addition to the evaluation discussed above, CSU-MLP forecasts have been evaluated subjectively by participants in the FFaIR experiments from 2017–2020. The FFaIR final reports (Perfater and Albright 2017; Albright and Perfater 2018; Trojniak and Albright 2019; Trojniak et al. 2020) include quantitative subjective evaluations, but the method (and the CSU-MLP products being evaluated) varied from year to year, so it is difficult to draw meaningful information about forecast improvements from those results. Thus, we focus here on summarizing the written comments from those reports.

When the CSU-MLP was first evaluated in 2017, "participants overwhelmingly agreed that the CSU-MLP First Guess Field is an excellent step in providing an initial starting point for WPC ERO forecasts, which has been a long requested tool." Furthermore, it was "recommended that the CSU developers work to reduce recurring biases and continue to refine the tool and reintroduce it into the testbed next year for further evaluation" (Perfater and Albright 2017). As discussed previously, these biases were especially apparent in the southwest US during the North American monsoon season. The models were updated to reduce these biases, but those changes introduced additional biases in the 2018 version. Nonetheless, the outcome of the 2018 FFaIR experiment was that, "The WPC-HMT recommends the Day 2 and Day 3 ERO CSU-MLP First Guess Field for operations as it showed great potential and was scored well by participants. It is recommended that the CSU developers work to refine some of the high probabilities in the High Plains and low probabilities in the Southeast..." (Albright and Perfater 2018).

In 2019, the focus shifted from the day 2–3 forecasts to day 1, and a day-1 version (analogous to the day 2–3 versions, with additional updates to address the biases discussed above) was introduced. These forecasts were also well received, with the 2019 FFaIR report concluding that "participants

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

felt that the guidance provided a great ‘starting spot’ for creating the experimental FFaIR ERO. However, there are a few regions across the CONUS that the products do not appear to be well-calibrated for, such as the Northern Rockies and Northern Plains. . .Further refinement of how the flooding risk is determined should be done in these regions. These refinement will likely help with the calibration of the marginal risk, which was generally too large spatially. . .” (Trojniak and Albright 2019). With further refinements to the CSU-MLP in the 2020 version evaluated during FFaIR, it was found that “Objectively and subjectively the CSU ML First Guess Day 1 GEFS ERO performed well and was often comparable to both the operational and FFaIR ERO. The GEFS Day 1 ERO should be transitioned to operations, however with the upgrade to the GEFS system it must be demonstrated that the retrained ML product on the new GEFS climatology is comparable to the one evaluated in FFaIR” (Trojniak et al. 2020). The last sentence refers to a major upgrade that was made to the GEFS in 2020; the implications of this upgrade will be discussed in the concluding remarks.

d. Use in operations

The CSU-MLP system was transitioned to WPC operations in March 2019. At this time, forecasters could view guidance from the 2017 and 2018 versions of the modeling system to assist with the day-2 and day-3 outlooks. The 0000 and 1200 UTC operational GEFS drove both versions, and the 2017 variant included output from the 0000 UTC GEFS/R, when available. The 2019 model was introduced in August 2019 after positive evaluations from that year’s FFaIR, along with objective verification showing notable improvements over the prior versions.

WPC forecasters began incorporating the CSU-MLP into their operational workflow during the latter part of the 2019 warm season. At the time, it was most used for the day-3 ERO issued at 0900 UTC, since this is the first product that is completely developed from scratch. While

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0186.1.

subsequent to the 0900 UTC day-3 ERO, forecasters incorporate some level of continuity from previous forecasts, the alternative solutions provided by the CSU-MLP highlighted areas where forecasters may consider changes in continuity.

In the spring of 2020, CSU made available to WPC the fourth iteration of the modeling system. At that time, through operational evaluation and verification of the available versions, forecasters found the 2018 version to have a significant high bias across much of the CONUS, thus this version was discontinued in favor of implementing the updated 2020 model. The 2020 model also introduced the first day-1 forecasts into operations.

In the short time that the CSU-MLP has been operational, WPC has increasingly utilized the guidance into the production of the ERO. Prior to its availability, WPC forecasters examined an abundance of numerical model and observational data to create the product, which was often challenging given tight product deadlines. Consequently, the CSU-MLP has resulted in notable time savings to forecasters. Forecaster feedback has been positive, and while there is still a perception of a high bias, objective evidence that the CSU-MLP is well-calibrated with respect to fractional coverage has resulted in a general trend of forecasters to increase the areal extent of their risk areas. In addition, there have been several cases where the CSU-MLP has correctly highlighted a high impact event where WPC had relatively low risk potential. Fig. 8 shows an example from July 2020 when a flash flood emergency occurred in the Philadelphia, PA area. CSU-MLP forecasts highlighted the potential for this event three days in advance, with probabilities reaching a moderate risk on day 1. In addition to the flash flooding in the city of Philadelphia, there were multiple reports of flash flooding from the Washington, DC area through the Philadelphia suburbs on the afternoon of 6 July 2020 (not shown).

5. Summary, conclusions, and recommendations

This manuscript outlines the development of a forecast system based on ensemble reforecasts, observations of excessive rainfall, and machine learning algorithms. The CSU-MLP produces skillful, reliable probabilistic forecasts of excessive rainfall on days 1–3, and they are being used as “first guesses” by operational forecasters when generating their nationwide excessive rainfall outlooks. Thorough evaluation shows that the most recent version of the CSU-MLP system is less skillful than the operational ERO on day 1, but more skillful on days 2 and 3, though these conclusions are sensitive to uncertain definitions of “excessive rainfall.” This project has demonstrated that forecast tools based on machine learning have a beneficial place in the operational forecaster’s toolbox, and that close collaboration between forecasters and researchers yields important improvements that benefit all those involved.

This project has had many successes, though there remain numerous challenges and opportunities for future work. The lack of broadly accepted definitions and datasets documenting “excessive rainfall” and “flash flooding” hinder the development of post-processing tools, as considerable uncertainty surrounds what should be considered “truth” for the purpose of training and evaluating such tools. As noted above, the GEFS underwent a major upgrade in September 2020 (NOAA 2020b) that includes a new model dynamical core, the Finite Volume Cubed-Sphere (FV3; Lin et al. 2017). The CSU-MLP continues to run using the upgraded GEFS (even though it was trained on the earlier version), and evaluation thus far does not show substantial differences in the forecast output (in fact, the last three weeks of forecasts included in the formal evaluation in this study were driven by the new FV3-GEFS). Nonetheless, future development will require retraining the CSU-MLP models using the FV3-GEFS; fortunately a comprehensive reforecast dataset has been generated that will enable this future work. Research is also ongoing to incorporate convection-allowing model (CAM) output into the CSU-MLP. Day-1 models trained on the 4-km NSSL-WRF CAM have been evaluated in the FFaIR experiment since 2018, and although these have not yet been recommended for transition to operations, they also show promising results (e.g., Hill and Schumacher 2021). Work is underway to incorporate CAM output beyond the NSSL-WRF as well. As archives of these forecasts continue to grow, it will allow for the development of further post-processing products. The relative importance of various atmospheric predictors within the machine learning system have been analyzed in aggregate (Herman and Schumacher 2018c,a), but better quantification and visualization of how the machine-learning system is performing its post-processing calculations in different weather situations may build more trust among forecasters in these tools (e.g., McGovern et al. 2019). In total, further efforts to integrate ever-growing datasets, new research tools, and meteorological expertise should lead to continued advances in probabilistic forecasting of high-impact weather.

Acknowledgments. The authors are grateful to Ben Albright, Jimmy Correia, and Sarah Perfater for their efforts in organizing the FFaIR experiment and synthesizing the feedback from participants each year. The authors also thank all of the WPC forecasters and FFaIR participants who have given their perspectives on the CSU-MLP forecasts and offered constructive suggestions to improve the forecasts. This research and operational transition was supported by NOAA Joint Technology Transfer Initiative grants NA16OAR4590238 and NA18OAR4590378. Thanks go to JTTI program manager Chandra Kondragunta for constructive feedback and continued support of this project. The authors also thank two anonymous reviewers and editor Jeff Waldstreicher for their suggestions that helped to strengthen the manuscript.

Data availability statement. UFVS data are available from WPC at `https://ftp.wpc.ncep.noaa.gov/ERO_verif/`. The CSU-MLP forecasts used in this study are available at `http:`

//dx.doi.org/10.25675/10217/222367. Graphical real-time CSU-MLP forecast output is available at http://schumacher.atmos.colostate.edu/hilla/csu_mlp/.

References

Albright, B., and S. Perfater, 2018: 2018 Flash Flood and Intense Rainfall experiment: Findings and results. Available online at https://www.wpc.ncep.noaa.gov/hmt/2018_FFaIR_final_report.pdf.

Breiman, L., 2001: Random forests. *Machine Learning*, **45 (1)**, 5–32, doi:10.1023/A:1010933404324.

Burke, P., A. Lamers, G. Carbin, M. J. Erickson, M. Klein, and M. Chenard, 2021: The modern Excessive Rainfall Outlook at the Weather Prediction Center. *Wea. Forecasting*, submitted.

Clark, R. A., J. J. Gourley, Z. L. Flamig, Y. Hong, and E. Clark, 2014: CONUS-wide evaluation of National Weather Service flash flood guidance products. *Wea. Forecasting*, **29 (2)**, 377–392, doi:10.1175/WAF-D-12-00124.1.

Developmental Testbed Center, 2020: Receiver operating characteristic. Available online at https://dtcenter.github.io/MET/latest/Users_Guide/appendixC.html#receiver–operating–characteristic, accessed 15 December 2020.

Erickson, M. J., B. Albright, and J. A. Nelson, 2021: Verifying and redefining the Weather Prediction Center's Excessive Rainfall Outlook forecast product. *Wea. Forecasting*, **36 (1)**, 325–340, doi:10.1175/WAF-D-20-0020.1.

Erickson, M. J., J. S. Kastman, B. Albright, S. Perfater, J. A. Nelson, R. S. Schumacher, and G. R. Herman, 2019: Verification results from the 2017 HMT–WPC Flash Flood and Intense Rainfall experiment. *J. Appl. Meteor. Climatol.*, **58 (12)**, 2591–2604, doi:10.1175/JAMC-D-19-0097.1.

Fritsch, J. M., and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85 (7)**, 955–966, doi:10.1175/BAMS-85-7-955.

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29 (4)**, 1024–1043, doi:10.1175/WAF-D-13-00108.1.

Gourley, J. J., and H. Vergara, 2021: Comments on "Flash flood verification: Pondering precipitation proxies". *J. Hydrometeor.*, **22 (3)**, 739–747, doi:10.1175/JHM-D-20-0215.1.

Gourley, J. J., and Coauthors, 2013: A unified flash flood database across the United States. *Bull. Amer. Meteor. Soc.*, **94 (6)**, 799–805, doi:10.1175/BAMS-D-12-00198.1.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94 (10)**, 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

Hamill, T. M., and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146 (12)**, 4079–4098, doi:10.1175/MWR-D-18-0147.1.

Hamill, T. M., M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143 (8)**, 3300–3309, doi:10.1175/MWR-D-15-0004.1.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134 (11)**, 3209–3229, doi: 10.1175/MWR3237.1.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31 (6)**, 1853–1879, doi:10.1175/WAF-D-16-0093.1.

Herman, G. R., and R. S. Schumacher, 2018a: "Dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea. Rev.*, **146 (6)**, 1785–1812, doi:10.1175/MWR-D-17-0307.1.

Herman, G. R., and R. S. Schumacher, 2018b: Flash flood verification: Pondering precipitation proxies. *J. Hydrometeor.*, **19 (11)**, 1753–1776, doi:10.1175/JHM-D-18-0092.1.

Herman, G. R., and R. S. Schumacher, 2018c: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146 (5)**, 1571–1600, doi:10.1175/MWR-D-17-0250.1.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148 (5)**, 2135–2161, doi:10.1175/MWR-D-19-0344.1.

Hill, A. J., and R. S. Schumacher, 2021: Forecasting excessive rainfall with random forests and a deterministic convection-allowing model. *Wea. Forecasting*, submitted.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15 (6)**, 2542–2557, doi:10.1175/JHM-D-11-0140.1.

Lin, S.-J., W. Putman, and L. Harris, 2017: FV3: The GFDL finite-volume cubed-sphere dynamical core. available online at: https://www.gfdl.noaa.gov/wp–content/uploads/2020/02/FV3–Technical–Description.pdf.

Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Postprocessing next-day ensemble probabilistic precipitation forecasts using random forests. *Wea. Forecasting*, **34 (6)**, 2017–2044, doi:10.1175/WAF-D-19-0109.1.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100 (11)**, 2175–2199, doi:10.1175/BAMS-D-18-0195.1.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31 (2)**, 371–394, doi:10.1175/WAF-D-14-00112.1.

NOAA, 2020a: National temperature and precipitation maps. Available online at https://www.ncdc.noaa.gov/temp–and–precip/us–maps/12/201 912?products[]=statewidepcpnrank#us–maps–select, accessed 15 December 2020.

NOAA, 2020b: NOAA upgrades Global Ensemble Forecast System. Available online at https://www.noaa.gov/media–release/noaa–upgrades–global–ensemble–forecast–system, accessed 15 December 2020.

NOAA Office of Water Prediction, 2021: Current NWS Precipitation Frequency Documents. available online at https://www.weather.gov/owp/hdsc_currentpf, accessed 6 April 2021.

NOAA Weather Prediction Center, 2021: Excessive Rainfall Outlooks. available online at https://www.wpc.ncep.noaa.gov/html/fam2.shtml#excessrain, accessed 21 January 2021.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29 (3)**, 489–504, doi:10.1175/WAF-D-13-00066.1.

Perfater, S., and B. Albright, 2017: 2017 Flash Flood and Intense Rainfall experiment: Findings and results. Available online at https://www.wpc.ncep.noaa.gov/hmt/2017_FFaIR_final_report.pdf.

Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143 (11)**, 4578–4596, doi:10.1175/MWR-D-15-0061.1.

Schumacher, R. S., and G. R. Herman, 2021: Reply to "Comments on 'Flash flood verification: Pondering precipitation proxies"'. *J. Hydrometeor.*, **22 (3)**, 749–752, doi:10.1175/JHM-D-20-0275.1.

Sukovich, E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds, and D. R. Novak, 2014: Extreme quantitative precipitation forecast performance at the Weather Prediction Center from 2001 to 2011. *Wea. Forecasting*, **29 (4)**, 894–911, doi:10.1175/WAF-D-13-00061.1.

Trojniak, S., and B. Albright, 2019: 2019 Flash Flood and Intense Rainfall experiment: Findings and results. Available online at https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2019_FFaIR.pdf.

Trojniak, S., J. Correia, Jr, and B. Albright, 2020: 2020 Flash Flood and Intense Rainfall experiment: Findings and results. Available online at https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2020_FFaIR_Experiment_Nov13.pdf.

Whan, K., and M. Schmeits, 2018: Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Mon. Wea. Rev.*, **146 (11)**, 3651–3673, doi:10.1175/MWR-D-17-0290.1.

Williamson, M., K. Ash, M. J. Erickson, and E. Mullens, 2021: How much do flash flooding events outside of an excessive rainfall outlook (ERO) matter? 35th Conference on Hydrology, Amer. Meteor. Soc., Available online at: ttps://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/381 012.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

LIST OF TABLES

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

Table 1. Definitions of "excessive rainfall" used in training of the day 2–3 versions of the CSU-MLP. Columns of the table indicate the precipitation dataset(s) (ST4=Stage IV; FFR=flash flood reports) and the 24-h ARI threshold used in training (1-yr or 2-yr). In the 2019 version, the day 2 and 3 models were inadvertently trained differently in the NGP and SE regions; these are shown by the entries before (day 2) and after (day 3) the slash in that year and regions. The 2020 version of the day-1 CSU-MLP uses the same definitions as the 2020 day 2–3 versions.

	Version			
Region	2017	2018	2019	2020
MDWST	ST4, 1-yr	FFR+CCPA+ST4, 1-yr	FFR+CCPA+ST4, 1-yr	FFR+CCPA, 2-yr
NE	ST4, 1-yr	FFR+CCPA+ST4, 1-yr	FFR+CCPA, 1-yr	FFR+CCPA, 1-yr
NGP	ST4, 1-yr	FFR+CCPA, 1-yr	FFR+CCPA, 2-yr / 1-yr	FFR+CCPA, 1-yr
PCST	ST4, 1-yr	FFR+CCPA, 1-yr	FFR+CCPA+ST4, 1-yr	FFR+CCPA, 2-yr
ROCK	ST4, 1-yr	FFR+CCPA, 1-yr	FFR+CCPA, 2-yr	FFR+CCPA, 2-yr
SE	ST4, 1-yr	FFR+CCPA+ST4, 1-yr	FFR+CCPA, 1-yr / FFR+CCPA+ST4, 2-yr	FFR+CCPA, 1-yr
SGP	ST4, 1-yr	FFR+CCPA+ST4, 1-yr	FFR+CCPA, 1-yr	FFR+CCPA, 1-yr
SW	ST4, 1-yr	FFR+CCPA, 1-yr	FFR+CCPA, 2-yr	FFR+CCPA, 2-yr

LIST OF FIGURES

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0186.1.

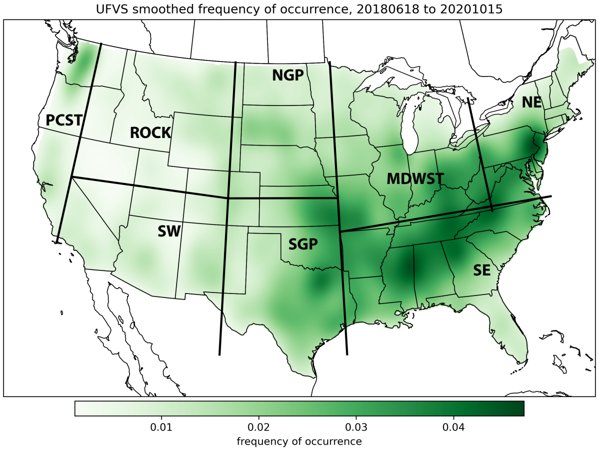

FIG. 1. Frequency of occurrence (in fraction of days) of excessive rainfall in the UFVS between 18 June 2018–15 October 2020. Values have been smoothed with a 9-sigma Gaussian filter. The regions used to train the CSU-MLP models are also shown: PCST=Pacific Coast; SW=Southwest; ROCK=Rocky Mountains; NGP=Northern Great Plains; SGP=Southern Great Plains; MDWST=Midwest; NE=Northeast; SE=Southeast.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

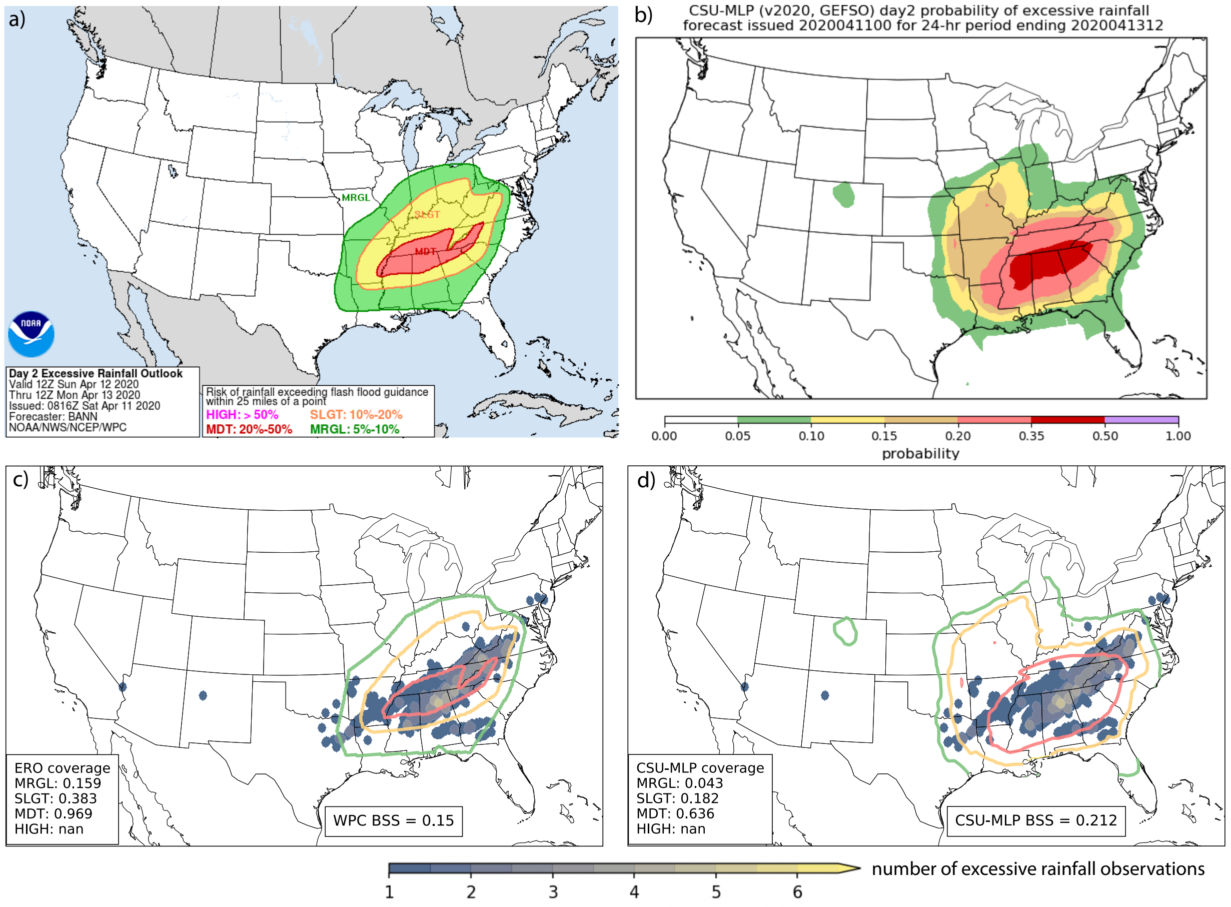
a)
MRGL
SLGT
MDT
Day 2 Excessive Rainfall Outlook
Valid 12Z Sun Apr 12 2020
Thru 12Z Mon Apr 13 2020
Issued: 0816Z Sat Apr 11 2020
Forecaster: BANN
NOAA/NWS/NCEP/WPC
Risk of rainfall exceeding flash flood guidance within 25 miles of a point
HIGH: > 50%
SLGT: 10%-20%
MDT: 20%-50%
MRGL: 5%-10%
b)
CSU-MLP (v2020, GEFSO) day2 probability of excessive rainfall
forecast issued 2020041100 for 24-hr period ending 2020041312
0.00
0.05
0.10
0.15
0.20
0.35
0.50
1.00
probability
c)
ERO coverage
MRGL: 0.159
SLGT: 0.383
MDT: 0.969
HIGH: nan
WPC BSS = 0.15
d)
CSU-MLP coverage
MRGL: 0.043
SLGT: 0.182
MDT: 0.636
HIGH: nan
CSU-MLP BSS = 0.212
1
2
3
4
5
6
number of excessive rainfall observations

FIG. 2. (a) WPC day-2 ERO issued at 0816 UTC 11 April 2020 and valid for the 24-h period from 1200 UTC 12 April to 1200 UTC 13 April 2020. (b) CSU-MLP day-2 forecast based on the 0000 UTC 11 April 2020 GEFS initialization and valid for the same time period. In addition to the probability bins defined by WPC, intermediate contours at 15 and 35% are also shown. (c) As in (a), but with observations of excessive rainfall within 40-km neighborhoods from the UFVS shown by shaded circles. The proportions of the forecast probability areas covered by observed excessive rainfall are shown in the lower left, and the Brier Skill Score for this forecast is shown at the bottom. (d) As in (c) but for the CSU-MLP forecast.

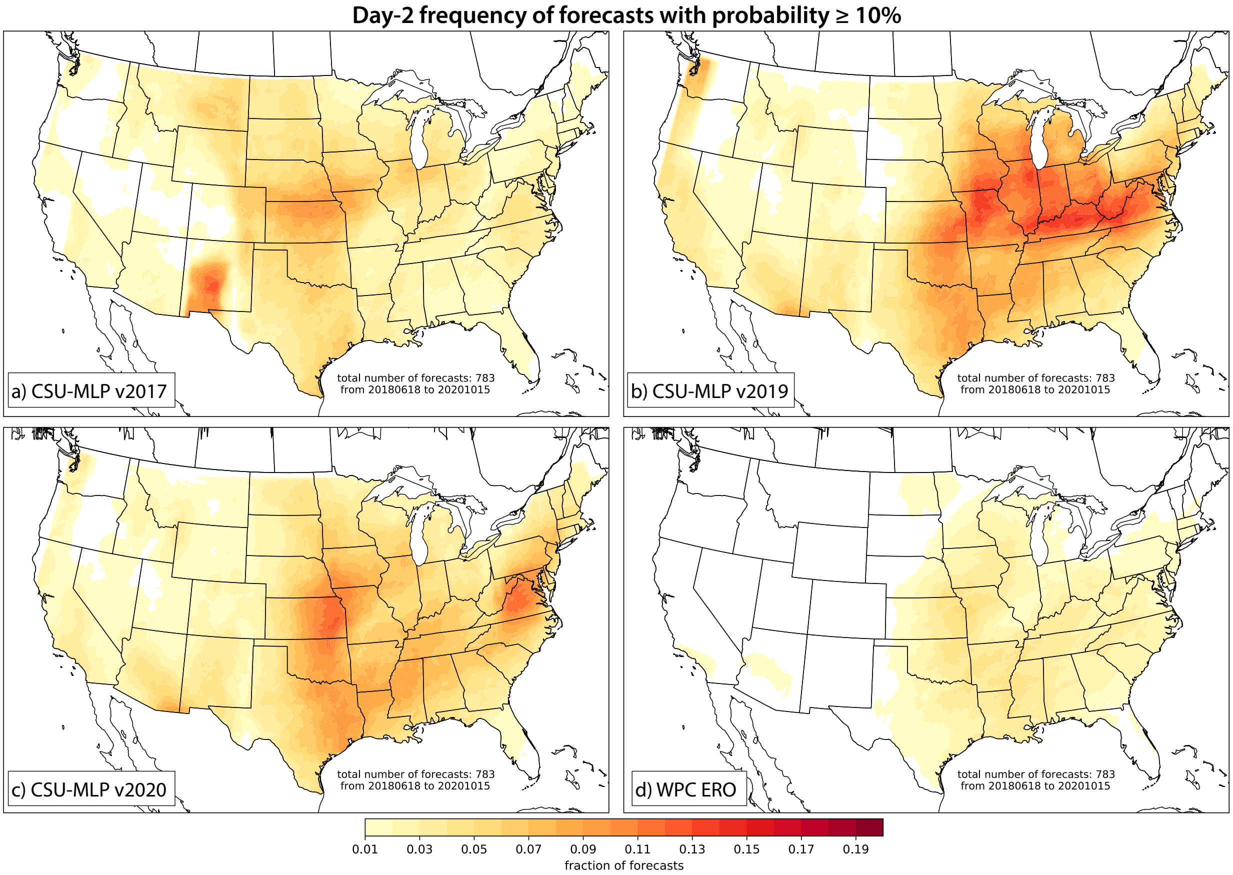

FIG. 3. Frequency of forecasts with probability ≥10% for day-2 forecasts from (a) CSU-MLP 2017 version; (b) CSU-MLP 2019 version; (c) CSU-MLP 2020 version; and (d) WPC ERO.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

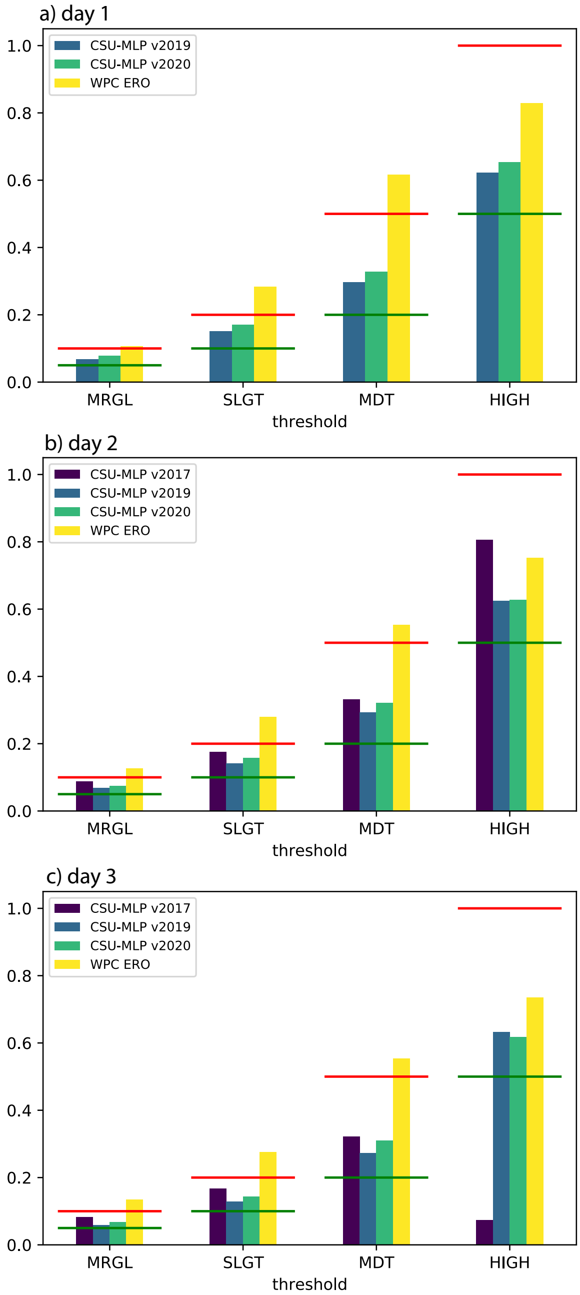

FIG. 4. Fractional coverage of observations for each ERO probability category over CONUS, for (a) day 1; (b) day 2; and (c) day 3 forecasts. Green horizontal lines indicate the lower bound of each probability category; red horizontal lines indicate the upper bound.

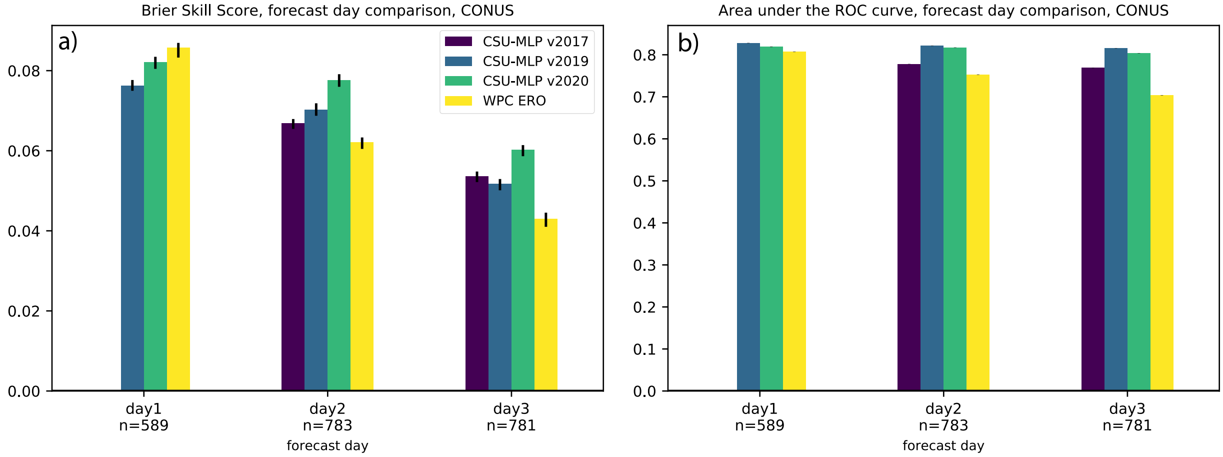

FIG. 5. (a) BSS and (b) area under the ROC curve comparing CSU-MLP versions and the ERO for forecast days 1-3. Uncertainty (shown by the black lines) is calculated using 100 bootstrap samples of the probabilistic forecast contingency table; the 2.5 to 97.5 percentile values are shown.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-20-0186.1.

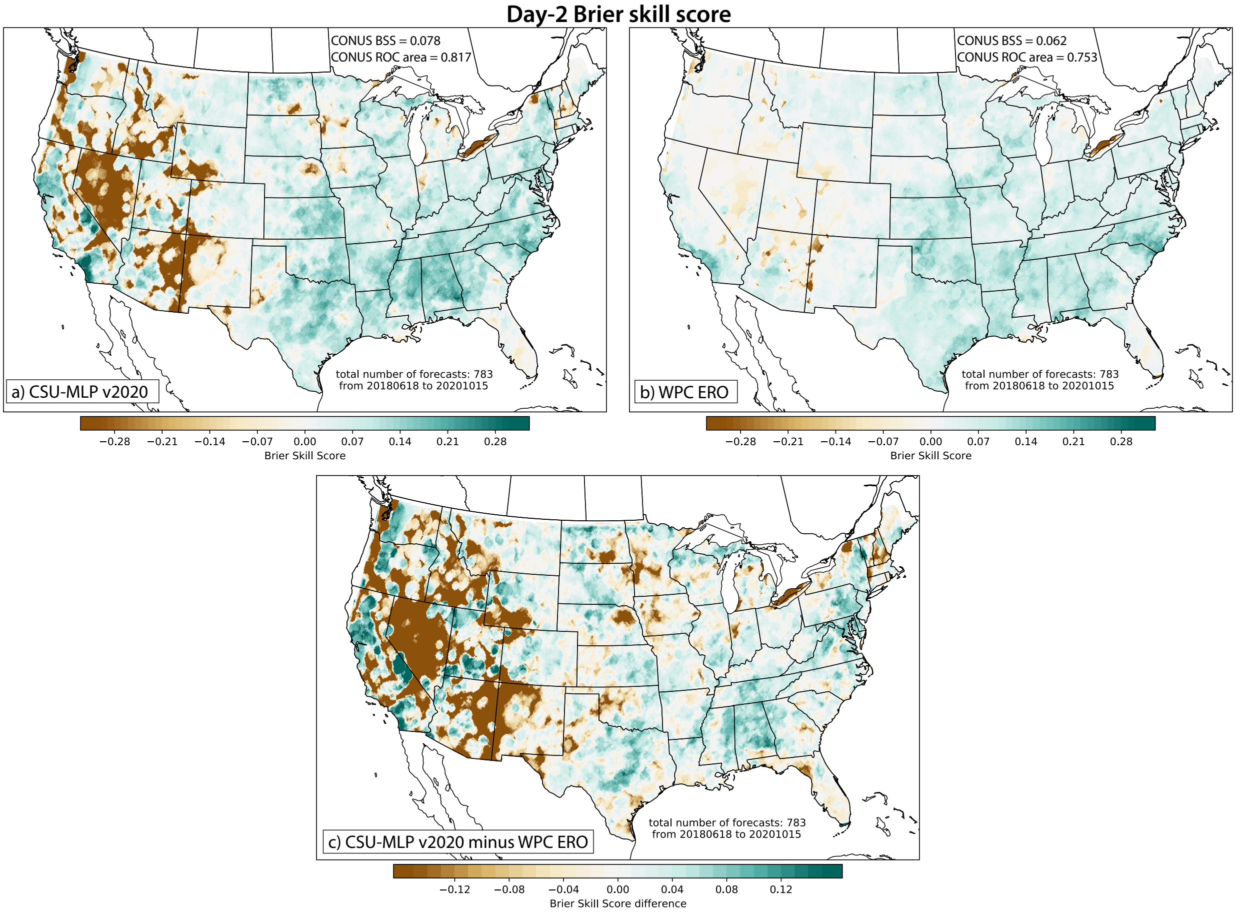

FIG. 6. (a,b) Map of BSS for day-2 forecasts from (a) CSU-MLP 2020 version and (b) ERO. (c) Map of BSS difference, CSU-MLP 2020 version minus ERO; green shading indicates where the CSU-MLP had greater skill, brown shading where the ERO had greater skill.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-20-0186.1.

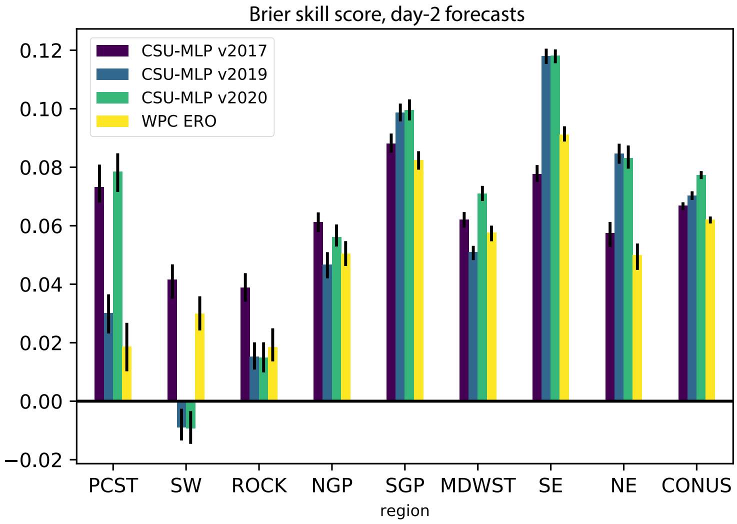

FIG. 7. BSS comparison of ERO and CSU-MLP versions for day-2 forecasts, aggregated by region (see Fig. 1 for region definitions). Uncertainty (shown by the black lines) is calculated using 100 bootstrap samples of the probabilistic forecast contingency table; the 2.5 to 97.5 percentile values are shown.

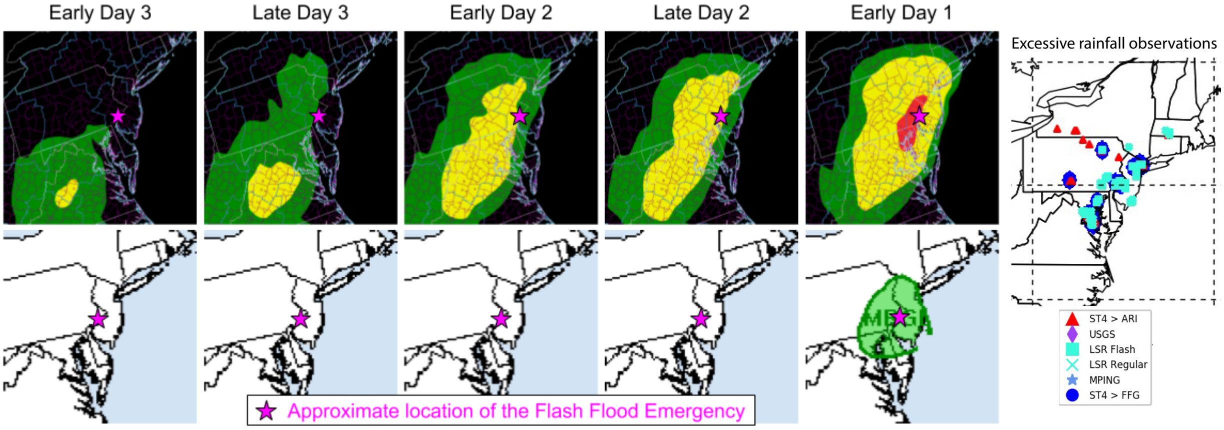

FIG. 8. CSU-MLP version 2020 forecasts from day 3 to day 1 (top row) and corresponding WPC EROs (bottom row), valid for the 24-h period ending 1200 UTC 7 July 2020. Early forecasts represent the 0900 UTC WPC ERO and the CSU-MLP driven by the 0000 UTC GEFS; late forecasts are the 2100 UTC WPC ERO and the CSU-MLP driven by the 1200 UTC GEFS. The color scale for both the CSU-MLP and ERO is as in Fig. 2a. The purple star indicates the location of a flash flood emergency issued for the Philadelphia area on the afternoon of 6 July 2020. The rightmost panel shows the locations of excessive rainfall from the different data sources used in the UFVS for this 24-h period.