Evaluating Machine Learning–Based Probabilistic Lightning Forecasts Using the HRRR: A Comparison to Three Forecast Baselines

RYAN A. SOBASH[a] AND DAVID A. AHIJEVYCH[a]

[a] *National Science Foundation, National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 24 July 2024, in final form 4 February 2025, accepted 19 February 2025)

ABSTRACT: Probabilistic forecasts of lightning for the CONUS were generated by postprocessing the HRRR with neural networks (NNs). These NN probability forecasts (NNPFs) were produced for HRRR forecasts in 2021–22 using NNs trained with 0000 UTC HRRR forecasts from 2019 to 2020 paired with ≥1 observed cloud-to-ground (CG1) flash as the target variable. CG1-NNPF skill was evaluated against three baselines: smoothed HRRR-based surrogate lightning forecasts, SPC probabilistic thunderstorm outlooks, and calibrated ensemble thunderstorm guidance from the High-Resolution Ensemble Forecast (HREF) system. The hourly maximum updraft speed (UP) diagnostic was the most skillful HRRR surrogate for predicting CG1, outperforming diagnostics such as hourly maximum lightning threat and updraft helicity. The UP forecasts were compared to the NNPFs by using the most skillful combination of UP threshold and smoothing length scale at each forecast hour. The 4- and 1-h CG1-NNPFs exceeded the skill of the UP forecasts in 2021 for nearly all forecast hours, with reduced skill differences overnight compared to the daytime. The NNPFs exhibited excellent reliability, with slight overforecasting of probabilities overnight. The NNPFs were more skillful than NOAA Storm Prediction Center (SPC) Thunderstorm Outlooks in both 2021 and 2022, especially overnight, while during the daytime, the SPC forecasts had similar or slightly greater skill. Finally, the NNPFs outperformed calibrated thunder guidance from the HREF system evaluated across forecasts in 2022. These findings imply that using NNs for thunderstorm prediction can improve upon existing non–machine learning baselines from deterministic and ensemble systems and may improve operational SPC thunderstorm forecasts.

SIGNIFICANCE STATEMENT: The prediction of where and when thunderstorms will occur remains a significant challenge for forecasters. In this work, we attempt to improve upon existing methods for thunderstorm prediction in the contiguous United States (CONUS) by using machine learning (ML). The ML system compares forecasts from a commonly used atmospheric model to observed lightning occurrence, correcting errors in the model forecasts and outputting probabilities for the occurrence of thunderstorms. The ML thunderstorm forecasts were compared to three different methods of generating thunderstorm forecasts during the years 2021 and 2022. In most situations, the ML thunderstorm forecasts were superior. Thus, using ML may improve thunderstorm forecasts across the CONUS.

KEYWORDS: Lightning; Thunderstorms; Forecast verification/skill; Neural networks

1. Introduction

Lightning is one of the primary hazards associated with convective storms. More accurate forecasts of lightning, and thus thunderstorms, can improve preparedness for both lightning occurrence and other thunderstorm hazards such as hail and tornadoes. Forecasts of lightning are often derived from the output of numerical weather prediction (NWP) models. Early forms of lightning guidance were based on large-scale environmental predictors from coarse NWP guidance, which rely on convective parameterization to represent simulated convective overturning {e.g., the U.S. National Centers for Environmental Prediction (NCEP) short-range ensemble forecast [SREF; Du et al. (2014)] or Global Forecast System (GFS) outputs}. For example, the NOAA Storm Prediction Center (SPC) has historically used SREF-based lightning predictions as a "first-guess" field when creating 1–2-day forecasts of thunderstorms across the contiguous United States (CONUS). SREF thunderstorm guidance, derived using fields such as convective available potential energy within the level from 0° to −20°C and forecast accumulated precipitation, is calibrated based on the historical occurrence of cloud-to-ground (CG) lightning flashes in these conditions to produce reliable, probabilistic forecasts of lightning (Bright et al. 2005).

Convection-allowing model (CAM) output has allowed for direct predictions of thunderstorms as convective-scale circulations are partially resolved in these NWP systems. For instance, the aforementioned SREF-based thunderstorm guidance at the SPC has been adapted to use high-resolution CAM ensemble information from the High-Resolution Ensemble Forecast system (HREF; Roberts et al. 2019, 2020). This HREF-based calibrated thunder guidance (HREF-CT; Harrison et al. 2022, 2023) leverages the explicit predictions of thunderstorms provided by the HREF and is superior to SREF-based CT guidance. Other forms of operational thunderstorm guidance [e.g., the Localized Aviation Model Output Statistics (MOS) Program; Charba et al. 2019] use the High-Resolution Rapid Refresh (HRRR; Dowell et al. 2022), which provide deterministic storm-scale forecasts across CONUS with an hourly update cadence. Novel storm-scale diagnostics are often available within these CAM systems, such as midlevel updraft helicity (UH; Kain et al. 2008; Sobash et al. 2011) and lightning threat diagnostics (McCaul et al. 2009; Kain et al. 2010). The primary

Corresponding author: Ryan A. Sobash, sobash@ucar.edu

DOI: 10.1175/WAF-D-24-0140.1

lightning threat diagnostic (LTG) is a weighted combination of the vertical flux of graupel at the −15°C level and the vertically integrated ice hydrometeor content (McCaul et al. 2009). Blaylock and Horel (2020) used LTG, combined with NOAA *GOES-16* Gridded Lightning Mapper (GLM; Goodman et al. 2013) observations to validate HRRR predictions of lightning during the 2018 and 2019 warm seasons. Lightning forecasts based on LTG were found to be skillful, although predominantly at larger spatial scales; at small scales, forecast skill diminished quickly during the first few hours of the forecast. Other lightning diagnostics and parameterizations have been proposed, based on updraft speed, cloud-top height, and microphysical mixing ratios (e.g., Price and Rind 1992; Yair et al. 2010) although these have not been extensively used in operational NWP systems.

While statistical calibration, such as MOS, has been used to generate thunderstorm guidance, machine learning (ML) techniques offer the ability to fuse prior NWP forecasts and lightning observations and produce robust, reliable, estimates of lightning occurrence and its uncertainty without requiring assumptions regarding the underlying distribution of data. ML algorithms have already been extensively applied to generate probabilistic predictions of other convective hazards for lead times of 1–48 h, such as hail, tornadoes, and convective wind gusts (Hill et al. 2020, 2023; Loken et al. 2022; Sobash et al. 2020; Sobash and Ahijevych 2024; Flora et al. 2021). These studies have largely used National Weather Service (NWS) storm reports (i.e., reports of wind gusts $\geq$ 50 kt, hail $\geq$ 1 in. in diameter, or a tornado; 1 kt $\approx$ 0.51 m s^{-1}) as ground truth, and generated probabilities of the occurrence of a storm report within a specified time and space scale (e.g., 40 km of a point).

Most previous work using ML to predict lightning occurrence has focused on generating thunderstorm nowcasts (i.e., forecasts with lead times of $\leq$1 h). Many of these algorithms [e.g., NOAA Probability of Severe (ProbSevere); Cintineo et al. 2018, 2020] merge remotely sensed satellite and radar observations [e.g., GOES GLM, GOES Advanced Baseline Imager (ABI), and Multi-Radar Multi-Sensor (MRMS) system radar observations] and NWP forecasts, using either basic probabilistic classifiers (e.g., naive Bayes, as in the case of ProbSevere) or more recently, ML techniques such as convolutional neural networks (CNNs) to generate 1-h gridded probabilities of severe weather for individual thunderstorms.

For example, Cintineo et al. (2022) developed an extension of ProbSevere, referred to as ProbSevere LightningCast. This system was built upon a U-net CNN and generated next-hour lightning probabilities with patches of two-dimensional *GOES-16* visible, short-wave infrared, and long-wave infrared channel data as input. *GOES-16* GLM data were used as the target data, or ground truth. Their system produced skillful predictions of lightning, with the most skillful forecasts occurring during the spring and summer when lightning is most prevalent; skill was reduced during the cool-season. Other studies, as summarized in Wang et al. (2023), have used ML, combined with satellite and lightning observations to also generate short-term nowcasts of lightning (e.g., Zhou et al. 2020; Ortland et al. 2023; Song et al. 2023; Leinonen et al. 2022).

Here, we generate ML-based predictions of lightning using the HRRR for 2021–22, but at lead times extending to 48 h, beyond the nowcasting range explored in previous work that evaluated thunderstorm predictions. The present work is an extension of Sobash and Ahijevych (2024), hereafter SA24, which produced 1–48-h predictions of severe thunderstorm hazards using the HRRR. We apply a similar methodology to SA24, using neural networks (NNs) to postprocess HRRR forecast output to generate lightning probabilities, utilizing both CG and IC data from two distinct lightning networks, i.e., *GOES-16* GLM and the Earth Networks Total Lightning Network. We primarily focus on the occurrence of $\geq$1 CG lightning flash (CG1), to determine the skill of the ML forecasts at predicting whether there will be a thunderstorm or not at a given grid point, although forecasts of any [CG or intracloud (IC)] lightning will also be evaluated. While generating forecasts using larger flash thresholds may be useful to anticipate more intense convection (e.g., Harrison et al. 2023), this will be left for future work.

The probabilistic CG1 NN-based forecasts examined here are produced on finer scales (for CG1 within 20 km of a point) than SA24 and for two different time windows (1 and 4 h). While SA24 used a single non–ML UH surrogate forecast for comparison, we compare the CG1 NN-based forecasts to three different non–ML probabilistic baselines: surrogate CG1 forecasts using six HRRR diagnostics, SPC Thunderstorm Outlooks, and HREF-CT guidance. These three forecasts serve as non-ML benchmarks to quantify the added skill of the ML forecasts and how they perform relative to human-generated thunderstorm forecasts.

The paper is organized as follows. Section 2 discusses the training and lightning datasets, NN architecture, and verification strategy. Section 3 provides probabilistic verification statistics for the three baseline forecasts and the NN-based thunderstorm forecasts, including a comparison of the NN forecasts to the SPC Thunderstorm Outlooks. Finally, section 4 provides a summary, discussion, and implications of the findings.

2. Methodology

a. HRRR forecasts and diagnostics

The HRRR training dataset consisted of 342 0000 UTC HRRR forecasts initialized between 0000 UTC 2 October 2019 and 0000 UTC 2 December 2020; this training dataset is identical to that used in SA24. The forecasts extended to 48 h and were part of an experimental HRRRv4 dataset that was generated in real time by the NOAA Global Systems Laboratory concurrently with the operational HRRRv3, prior to the operational HRRRv4 implementation on 1200 UTC 2 December 2020. Operational HRRRv4 forecasts initialized between 0000 UTC 1 January 2021 and 0000 UTC 31 December 2022 were used as a testing dataset to compute verification statistics. The HRRR output fields used for training include all fields used in SA24, plus three additional diagnostics: the HRRR surface hail diagnostic and two lightning diagnostics (Table 1). These three fields were unavailable in HRRRv3, but since we do not use HRRRv3 forecasts, unlike SA24, we have added the diagnostics given their potential relationship to the occurrence of lightning.

TABLE 1. Base predictors used for model training. Diagnostics are identical to those used in SA24, except for the addition of the three starred (*) diagnostics. The mean of the environmental and upper-air fields, and the maximum of the explicit fields, over all native HRRR grid points within each 80-km grid box, was used to upscale the fields onto the 80-km grid. Neighborhood predictors were generated for each of the environmental and explicit fields as described in the text.

Field	Type
Forecast hour, latitude, and longitude	Static
Day of year (encoded with sine and cosine)	Static
Local solar hour (encoded with sine and cosine)	Static
Surface-based, mixed-layer, and most-unstable CAPE	Environment
Surface-based and mixed-layer CIN	Environment
Surface-based lifted condensation level	Environment
0–6 and 0–1 km AGL bulk wind difference	Environment
0–1 and 0–3 km AGL storm-relative helicity	Environment
2-m temperature and dewpoint temperature	Environment
Surface pressure	Environment
Product of most-unstable CAPE and 0–6 km AGL bulk wind difference	Environment
Significant tornado parameter	Environment
1-h precipitation accumulation	Environment
Composite reflectivity	Environment
Height of freezing level	Environment
Hourly max 1 km AGL vertical vorticity	Explicit
Hourly max 2–5, 0–3, and 0–2 km AGL cyclonic UH	Explicit
Hourly min 2–5 km AGL anticyclonic UH	Explicit
Hourly max updraft and downdraft speed below 400 hPa	Explicit
Hourly max 10 m AGL wind speed	Explicit
Hourly max vertically integrated graupel	Explicit
HRRR lightning diagnostic No. 1*, No. 2*, and No. 3	Explicit
HRRR surface hail diagnostic*	Explicit
700–500-hPa lapse rate	Upper air
925-, 850-, 700-, and 500-hPa zonal wind speed	Upper air
925-, 850-, 700-, and 500-hPa meridional wind speed	Upper air
925-, 850-, 700-, and 500-hPa temperature	Upper air
925-, 850-, 700-, and 500-hPa dewpoint temperature	Upper air

The diagnostics were first upscaled from the 3-km native HRRR to an 80-km grid across the CONUS using the mean for the environmental and upper-air fields, and the maximum or minimum for the explicit fields (i.e., the maximum and minimum was used for fields that are strictly positive and negative, respectively) of all 3-km grid points within each 80-km grid box. Additional neighborhood diagnostics were generated for the environmental and explicit fields by either computing an average (for the environmental fields) or a maximum/minimum (for the explicit fields) across 3×3 and 5×5 spatial neighborhoods, within 1 and 2 h of the valid time. This led to an additional six input fields for those diagnostics; the total number of input fields after preprocessing was 240.

While previous work, e.g., Clark et al. (2022), Loken et al. (2020), and Sobash et al. (2020), has generated severe weather probabilities from ML models that are valid within 40 km (≈25 mi) of a point, matching the 80-km grid spacing, we wish to generate lightning probabilities within 20 km (≈12 mi) of a point, given the presumably higher predictability of parent thunderstorms relative to their hazards. Additionally, SPC thunderstorm outlooks are valid on this spatial scale, and other work generating calibrated thunderstorm probabilities has defined the probabilities as an event occurring within 20 km of a point (e.g., Harrison et al. 2022). Thus, we used the preprocessed 80-km HRRR output to train ML models to predict the probability of lightning within 20 km of each 80-km grid point (rather than within 40 km of each 80-km grid point as in SA24). This allows us to use the same preprocessed input that was used in SA24, with the only changes being the addition of the three new diagnostics and the target observations.

b. *Lightning observations*

The NOAA *GOES-16* GLM (Goodman et al. 2013) was the first source of lightning data used to train and test the ML models. The GLM instrument detects total (CG and IC) lightning by assessing changes in brightness every ≈2 ms relative to a continuously updating background image, across an array of grid points with horizontal grid spacing ranging from 8 km at nadir to 14 km along the edge of the *GOES-16* field of view (Bruning et al. 2019). Since the GLM instrument provides observations with good spatial resolution and detection accuracy across the contiguous United States, we do not use *GOES-17* GLM data in this work.

GLM data were retrieved from Amazon Web Services (https://registry.opendata.aws/noaa-goes/) in hourly tar files, each comprising 180 files, each file covering 20 s. We used the *flash*-level data and aggregated the flashes in time windows of 1 and 4 h centered on each hour. The binary data were read and filtered spatially with the help of the Python module

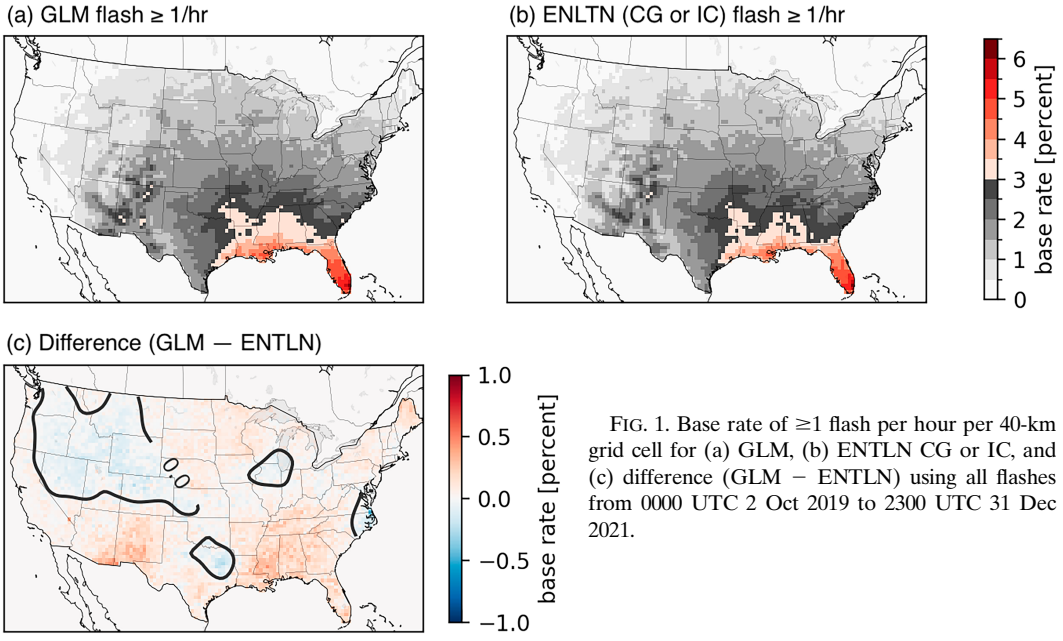

FIG. 1. Base rate of ≥1 flash per hour per 40-km grid cell for (a) GLM, (b) ENTLN CG or IC, and (c) difference (GLM − ENTLN) using all flashes from 0000 UTC 2 Oct 2019 to 2300 UTC 31 Dec 2021.

glmtools (Bruning 2019) and the interpolation to 40-km grid boxes was done with KDTree from scikit-learn (Pedregosa et al. 2011). Corrupt files were discarded. If more than one 20-s file was discarded or missing in the 1-h time window (or more than four in the 4-h time window), the GLM flash count for that time window was considered missing. This arbitrary threshold of one missing 20-s file per hour was a compromise between maximizing usable hourly accumulations and minimizing the effect of missing data, which could artificially lower the flash totals.

The second source of lightning observations was the Earth Networks Total Lightning Network (ENTLN), a ground-based detection system (Thompson et al. 2014; Lapierre et al. 2019). ENTLN comprises surface stations that continuously record lightning pulse waveforms across frequencies from 1 Hz to 12 MHz. Lightning pulse locations are calculated through triangulation, using the time of arrival of each pulse at nearby stations. Pulses are grouped into flashes if they occur within 700 m s and 10 km of each other and are classified as either IC or CG flashes based on the presence of a return stroke. We filtered out individual pulse-level events and flashes with corrupt metadata, which constituted 0.016% of flashes (with errors such as invalid latitude, longitude, time, or intracloud flags).

Flashes were then aggregated in 30-min intervals starting at the top of the hour and were further aggregated into 1- and 4-h windows, centered on the top of each hour. There were several ENTLN outages ranging from a few hours to days, during which no flashes were recorded. To avoid these gaps, periods entirely within these outages were excluded from training and testing. For partial dropouts—where the missing data covered less than half the window—missing 30-min bins were filled by the mean of nonmissing bins. This scaling was only necessary for 1.76% of the windows. If more than half the window overlapped with missing data, the window was marked as missing.

Both GLM and ENTLN flashes were summed within 40-km grid boxes that were half the length and half the width of the larger 80-km grid boxes used to upscale the HRRR. The 80-km grid size started as 93 × 65 in the east–west and north–south dimensions, respectively, and the 40-km grid size started as 185 × 129; the spatial extents of the two grids were identical. After dropping grid boxes outside the CONUS, there were 1308 80-km grid boxes and 5207 40-km grid boxes.

Overall, the base rates using the 40-km, 1-h observations were similar between the GLM and ENTLN both in space (Fig. 1) and across forecast hours (Fig. 2), suggesting that both should provide similar forecasts when used for training NNs to predict the occurrence of lightning. After separating IC and CGs in the ENTLN data, CGs occurred in 70%–80% of the grid boxes where ≥1 flash occurred (Fig. 2). While we focus mostly on the skill of ML lightning forecasts for CG1, generated with the ENTLN dataset, we also provide the skill when using all flashes to fully quantify thunderstorm predictability in the HRRR.

c. *Generation of neural network lightning forecasts*

The HRRR diagnostics and lightning observations were used to train NNs with an architecture nearly identical to the one used in SA24, with the only differences being the number of input fields and output neurons. The input layer had 240 predictors, an increase due to the addition of three new predictors and their associated neighborhood predictors, as described in section 2a. The output layer produced eight independent probabilities, corresponding to the likelihood of ≥1 flash occurring within the specified time window. The flash types considered were 1) GLM, 2) ENTLN CG, 3) ENTLN IC, and 4) ENTLN CG or IC. Each flash type was paired with two distance thresholds (20 and 40 km), resulting in eight forecast probabilities. This

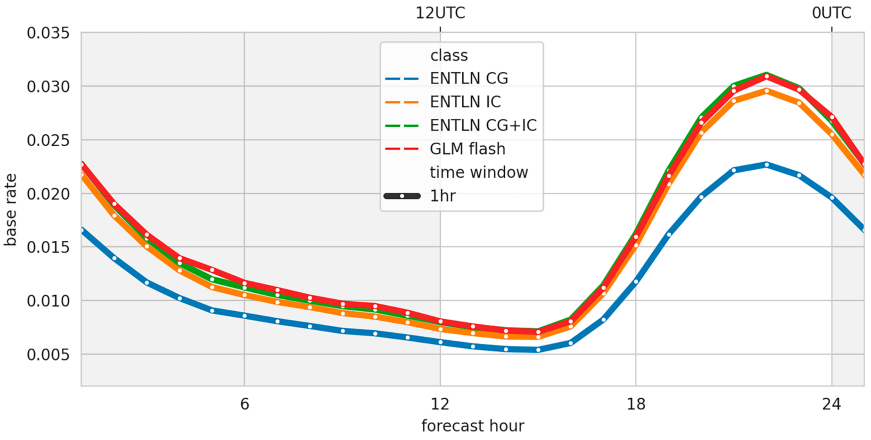

FIG. 2. Diurnal cycle of base rate for ≥1 flash (CG, IC, CG + IC, or GLM-detected) per hour per 40-km grid box using all flashes from 0000 UTC 2 Oct 2019 to 2300 UTC 31 Dec 2021.

study only presents results from the 20-km distance threshold. Separate NNs were trained for 1- and 4-h time windows; no constraints were applied to ensure the 4-h probability was greater than or equal to the 1-h probabilities.

The number of hidden layers and other hyperparameters were nearly identical to those used in SA24 and are provided in Table 2. These hyperparameters and training choices were determined through a five-fold cross-validation process using the HRRR training dataset. Training was repeated 10 times with different seeds to the random number generator to generate 10 different NNs. Predictions from the 10 NNs were averaged to get the final CG1 neural network probability forecasts (NNPFs) for a particular HRRR forecast; these CG1-NNPFs for 2021 and 2022 HRRR forecasts are used for subsequent verification.

d. Generation of surrogate lightning forecasts

To put the skill of the CG1-NNPFs in context, we generated CG1 surrogate severe probability forecasts (SSPFs) with six different HRRR diagnostics (Table 3). The surrogate diagnostics were chosen based on their assumed relationship to lightning occurrence. For example, LTG is computed using updraft speed and graupel mixing ratio, while Dye et al. (1989) noted that radar reflectivity ≥ 40 dBZ at the −10°C level was necessary for electrified convection. The procedure to create CG1-SSPFs was similar to that of Sobash et al. (2020), summarized in their Fig. 2. While Sobash et al. (2020) chose a single optimal threshold for UH, we examined a range of thresholds based on the forecast bias. Also, we computed the CG1-SSPFs on a 40-km grid, although only verify at the 80-km grid points to be consistent with the CG1-NNPFs. Using all 40-km grid points for verification does not appreciably change the results.

TABLE 2. Hyperparameter choices for optimal neural network configuration determined through five-fold cross-validation experiment.

Hyperparameter	Value
No. of hidden layers	1
No. of neurons in hidden layer	1024
Dropout rate	0
Learning rate	0.001
No. of training epochs	30
Hidden layer activation function	Rectified linear unit
Output layer activation function	Sigmoid
Optimizer	Adam
Loss function	Binary cross-entropy
Batch size	1024
Regularization	0
Batch normalization	On

To create the CG1-SSPFs, the six diagnostics were upscaled onto the 40-km grid by computing the maximum value within each 40-km grid box. Then, a range of thresholds were selected for each diagnostic based on the forecast bias. For example, for a forecast bias of 1.0, an equivalent number of forecast "yes" 40-km grid boxes are generated as the number of "yes" 40-km CG1 grid boxes when summed across all individual hourly HRRR forecasts in 2021. Thresholds for each forecast diagnostic were determined for biases between 0.5 and 1.5, in 0.05 increments; thresholds associated with a subset of these biases are provided in Table 3. The thresholds associated with these fields are substantially smaller than those chosen when compared to observed severe weather reports [e.g., UH thresholds range from 2 to 9 $m^2 s^{-2}$ when compared to CG1, whereas the optimal thresholds in Sobash et al. (2020) are typically an order of magnitude larger].

Finally, 1- and 4-h CG1-SSPFs were generated by applying a Gaussian smoother, σ, to the binary grid of forecast "yes" points for each of the 21 thresholds. A 4-h maximum was applied to the binary grid of hourly forecast points to produce the 4-h CG1-SSPFs. CG1-SSPFs were generated for σ ranging from 20 to 300 km, in 20-km increments. The CG1-SSPFs were then used as a non-ML forecast for comparison to the CG1-NNPFs, as well as to get a sense of the underlying skill of the HRRR at predicting lightning using the six surrogate lightning diagnostic fields. The maximum Brier skill score (BSS) across all 21 thresholds and 15 σ values was used to compare the differences in BSS for each of the six

TABLE 3. Surrogate diagnostic thresholds for forecast bias values of 0.5, 0.75, 1.0, 1.25, and 1.5 using 40-km grid boxes where lightning ≥ 1 ENTLN CG flash per hour (1 095 321 total observed grid boxes) from 1 Jan 2021 to 31 Dec 2021. HM indicates hourly maximum field.

Abbreviation	Diagnostic	Bias/threshold				
		1.50	1.25	1.00	0.75	0.50
UH	HM 2–5 km AGL cyclonic updraft helicity (m^2 s^{-2})	2.79	3.27	4.18	5.96	8.67
UP	HM updraft speed below 400 hPa (m s^{-1})	6.01	6.70	7.75	9.28	11.65
LTG	HM lightning threat No. 3 (flashes per square kilometer per 5 min)	0.081	0.097	0.118	0.149	0.199
GRPL	HM vertically integrated graupel (mm)	1.93	2.68	3.84	5.71	8.90
REF10	HM reflectivity at −10°C (dBZ)	36.85	37.87	39.57	41.84	44.19
CREF	Composite reflectivity (dBZ)	44.19	45.58	47.15	49.11	52.04

diagnostics and to compare to the CG1-NNPFs, as discussed in section 3a.

e. Storm Prediction Center Thunderstorm Outlooks

To further place CG1-NNPF performance in context, we compared the performance of the CG1-NNPFs to the SPC Thunderstorm Outlooks. These outlooks consist of CG1 probabilities within approximately 20 km of a point across the CONUS (Harrison et al. 2022), matching the event definition used in the creation of the CG1-NNPFs and CG1-SSPFs (Fig. 3). SPC Thunderstorm Outlooks are issued five times a day for different valid periods and consist of probability contours ≥ 10%, 40%, and 70%. For instance, at 0600 UTC, SPC Thunderstorm Outlooks are issued for 4-h valid periods of 1200–1600, 1600–2000, and 2000–0000 UTC. A summary of the different issuance times and valid periods is shown in Fig. 4a, along with the number of available outlooks during 2021 for each issuance and valid time combination.

We compared the SPC Thunderstorm Outlooks to the 4-h 0000 UTC-initialized CG1-NNPFs. Specifically, the 1200–1600, 1600–2000, 2000–0000, and 0000–0400 UTC SPC forecasts were paired with the 4-h CG1-NNPFs centered on HRRR forecast hours 14, 18, 22, and 26. For the 8-h 0400–1200 UTC time period, we combined the CG1-NNPF for 0400–0800 UTC and 0800–1200 UTC, from HRRR forecast hours 30 and 34, to get the equivalent 8-h CG1-NNPF for 0400–1200 UTC, assuming the two probabilities were independent:

$$1 - [(1 - \mathrm{NNPF}_{0400-0800}) \times (1 - \mathrm{NNPF}_{0800-1200})] = \mathrm{NNPF}_{0400-1200}. \quad (1)$$

In this way, the SPC forecasts issued at these five different issuance times were validated against the most recent 0000 UTC NNPF (assuming that the 0000 UTC CG1-NNPF would be unavailable until after 0130 UTC). Figures 4b and 4c provide a comparison of lead times as a function of issuance and valid period. We restricted the verification to time periods with a valid SPC Thunderstorm Outlook, valid GLM flash counts, and valid ENTLN lightning observations. Due to sporadic missing Thunderstorm Outlooks, GLM files, and ENTLN time periods, the number of cases ranges from 330 to 378 (Fig. 4).

While CG1-NNPFs range continuously from 0% to 100%, the SPC Thunderstorm Outlook is constrained to four probability levels. For a fair comparison of the performance of the two probabilistic forecasts, we discretized the CG1-NNPFs to the same four probability levels of the SPC Thunderstorm Outlook (Table 4), demonstrated in Fig. 5.

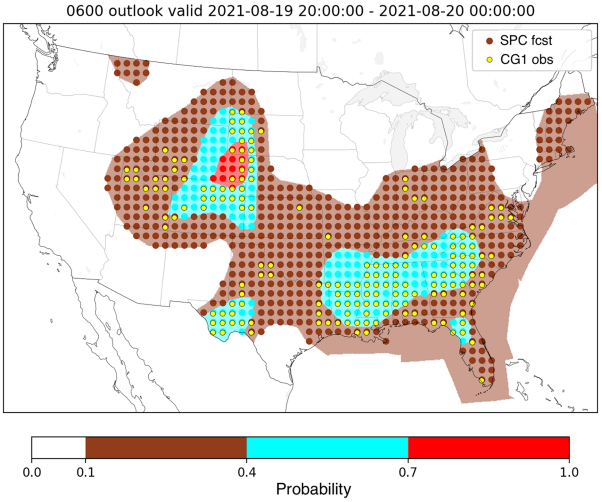

FIG. 3. SPC Thunderstorm Outlook (filled polygons) issued at 0600 UTC 19 Aug 2021 valid from 2000 UTC 19 Aug 2021 through 0000 UTC 20 Aug 2021. The 80-km gridbox centers within the polygon areas (brown, blue, and red circles) and locations of CG1 (yellow circles) are also shown. SPC outlook extends into coastal waters, while the grid is restricted to land points.

f. Forecast verification

Objective verification was performed on the 1- and 4-h CG1-SSPFs and CG1-NNPFs using the BSS, area under the relative operating characteristic curve (ROCA), and attributes diagrams (Wilks 2006). These complementary metrics evaluate different aspects of probabilistic forecast performance, including both forecast calibration and discrimination. The reference forecast in the BSS was the sample climatology within the entire testing dataset or grouped by forecast hour, in the case of hourly BSSs. For some forecasts, performance diagrams were also generated to examine binary performance metrics such as probability of detection, false alarm ratio, and critical success index (CSI) across multiple probabilistic forecast thresholds (Roebber 2009). CG1-NNPFs were only generated at 80-km grid boxes within the CONUS; while observations of

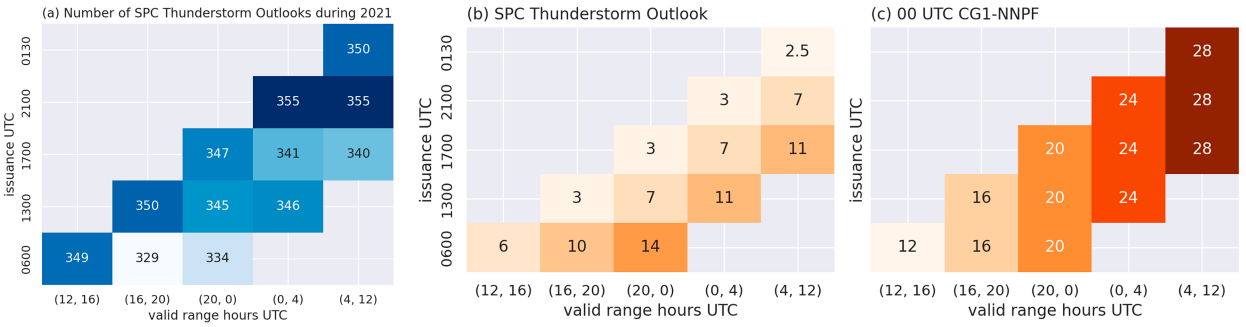

FIG. 4. (a) Number of SPC Thunderstorm Outlooks by issuance and valid range during 2021, (b) lead time (hours) of SPC Thunderstorm Outlook, and (c) lead time (hours) of 0000 UTC CG1-NNPFs.

lightning exist outside of the CONUS boundaries, these were not evaluated in this work.

Where appropriate, statistical significance was assessed by computing 90% bootstrapped confidence intervals (CIs) for skill differences between forecast pairs. The CIs were constructed by computing statistics (e.g., BSS) for 10 000 different forecast datasets, with each dataset consisting of a different set of 365 HRRR forecasts chosen from the original set of 365 HRRR forecasts with replacement (Hamill 1999).

3. Results

a. *Skill of HRRR surrogate lightning forecasts in 2021*

As described in the methodology, 4- and 1-h CG1-SSPFs were generated using six surrogate diagnostics in the HRRR (Table 3). For each diagnostic, we compared hourly BSSs using the CG1-SSPFs which produced the largest BSSs across all 21 thresholds and 15σ values. The UP CG1-SSPFs were statistically significantly better than the CG1-SSPFs generated with the other five diagnostics for nearly all forecast hours (Fig. 6). For most forecast hours, the CREF and UH CG1-SSPFs had the lowest BSS. These results were similar for both the 4- and 1-h time scales (Fig. 6). CG1-SSPFs generated with GRPL, LTG, and REF10, had similar BSSs, with skill indistinguishable across most forecast hours. One period when the LTG and UH CG1-SSPFs were among the most skillful SSPFs was between forecast hours 27–31, yet during these hours, the LTG and UH CG1-SSPF BSSs were not statistically significantly different than the UP CG1-SSPF BSSs.

The forecast bias, and thus surrogate threshold, that produced the highest BSS varied among forecast hour for all diagnostics (not shown). For UP, this "optimal" bias varied between 1.2 and 1.4 for forecast hours 1–6, and between 0.9 and 1.15 for forecast hours 7–48. The larger "optimal" bias at the start of the forecast is likely due to the impact of the HRRR initialization and spinup on the diagnostics (Dowell et al. 2022). Relatedly, the CREF BSS and "optimal" bias varies differently than the other five diagnostics during the first few hours of the forecast (Fig. 6), potentially due to radar data assimilation that influences the microphysical fields during initialization. Finally, CG1-SSPFs generated with all six diagnostics had similar diurnal cycles of skill, with the BSS peaking in the late afternoon and decreasing overnight, and the first diurnal skill peak being larger than the second diurnal peak (Fig. 6).

TABLE 4. Generation of discretized CG1-NNPFs from the continuous CG1-NNPFs.

Continuous CG1-NNPF	Discretized CG1-NNPF
$0\% \leq \text{NNPF} < 10\%$	0%
$10\% \leq \text{NNPF} < 40\%$	10%
$40\% \leq \text{NNPF} < 70\%$	40%
$70\% \leq \text{NNPF} < 100\%$	70%

While LTG was designed to identify potential lightning activity in CAM forecasts, it did not consistently outperform UP CG1-SSPFs. Further, the UH CG1-SSPFs were occasionally outliers in terms of skill, especially during the afternoon and early evening (e.g., forecast hours 15–24). It is hypothesized that the low magnitudes of UH (Table 3) were identifying nonconvective updrafts that did not go on to generate lightning compared to the other diagnostics, reducing the effectiveness of UH for lightning prediction. UH could also fail to identify simulated convection without any appreciable rotation that did produce lightning, e.g., monsoon convection. This result is noteworthy since UH is often the most skillful at anticipating the presence of other severe hazards, e.g., hail or severe convective winds, given the propensity of supercells to generate severe weather events. Here, it is the least skillful diagnostic.

While the relative BSS rankings among the six diagnostics did not change substantially with σ, optimizing the BSS requires varying σ as a function of forecast hour. For instance, for forecast hours 1–6, UP CG1-SSPFs generated by smoothing with σ values between 40 and 60 km had the largest BSS (Fig. 7). Using larger values of σ at short lead times substantially reduced the BSS, indicating that the CG1-SSPFs at these lead times had skill on space scales of 40–60 km. As forecast hour increased, the σ that maximized the BSS increased, reflecting the larger space–time uncertainty present in the HRRR forecasts. For instance, by forecast hours 24–36, the σ = 40 km CG1-SSPFs had BSS magnitudes 0.05–0.1 smaller than CG1-SSPFs using σ between 80 and 120 km. Overall, the values of σ that optimized the BSS were less than those of the maximized skill when generating forecasts for other severe weather hazards (e.g., the maximum BSS in Fig. 6b in

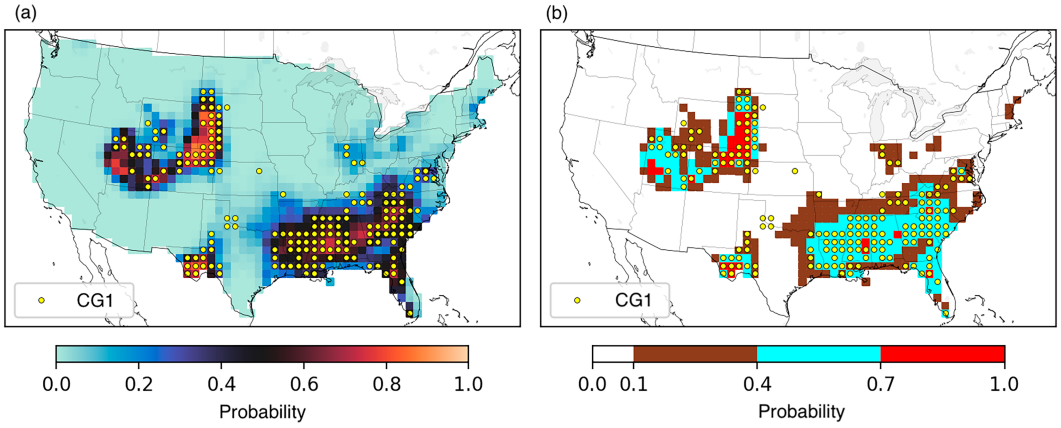

FIG. 5. (a) Continuous CG1-NNPF and (b) discretized CG1-NNPF for same valid time as the SPC Thunderstorm Outlook in Fig. 3. Yellow circles indicate grid boxes of CG1 within the 4-h valid time.

SA24 occurred with $\sigma \geq 160$ km). It is clear that forecasts of lightning and thunderstorms are more predictable than convective hazards such as hail and tornadoes, in part due to the greater frequency of events. CG1 forecast guidance derived from CAM output should reflect this by using smaller neighborhood sizes and smoothing length scales.

b. *Skill of HRRR ML lightning forecasts in 2021*

First, we compare the UP CG1-SSPF BSSs, generated with σ and UP thresholds that maximized the BSS at each forecast hour (Fig. 6), to the CG1-NNPFs. Using this baseline, the 4- and 1-h CG1-NNPF BSSs were statistically significantly larger than each forecast's surrogate counterpart across nearly all forecast hours (Fig. 8). Both the 4- and 1-h CG1-NNPFs improved upon the UP CG1-SSPFs more so during the day than overnight. Differences were largest at forecast hours 1–6, likely due to the reliance of the CG1-SSPFs on UP which was still spinning up during the start of the forecast, while the CG1-NNPFs incorporated both surrogate and environmental fields. Beyond the initial spinup period, BSS differences were

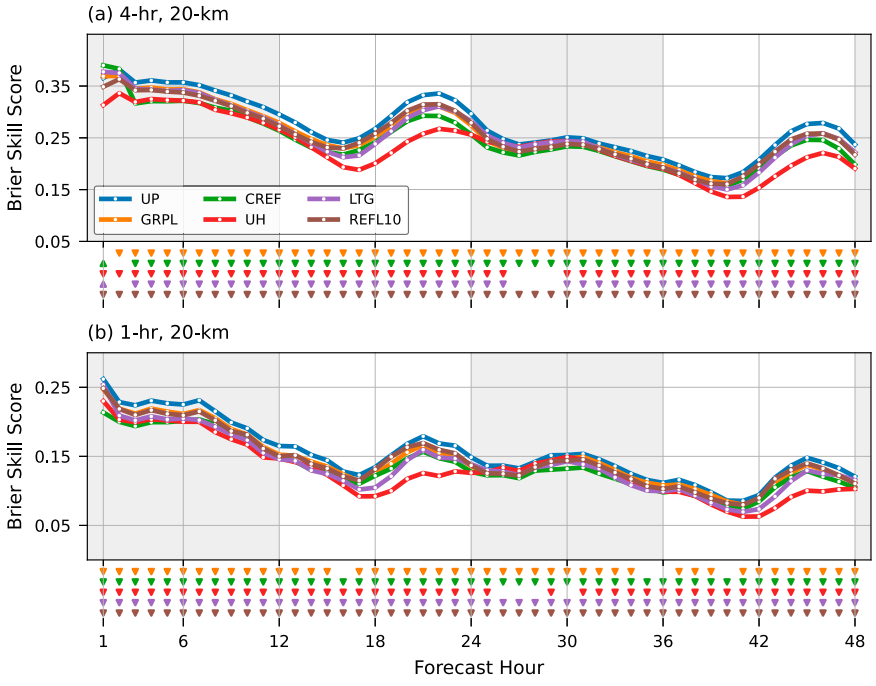

FIG. 6. Hourly BSS for (a) 4-h, 20-km and (b) 1-h, CG1-SSPFs generated using six HRRR surrogate diagnostics (abbreviations found in Table 3). Skill was aggregated over all 2021 CG1-SSPFs. The largest BSS among all σ and diagnostic threshold pairs is shown at each forecast hour. Triangles denote if the GRPL, CREF, UH, LTG, and REFL10 BSSs are statistically significantly larger (upward triangle) or smaller (downward triangle) than the UP forecast BSSs at the 90% confidence level.

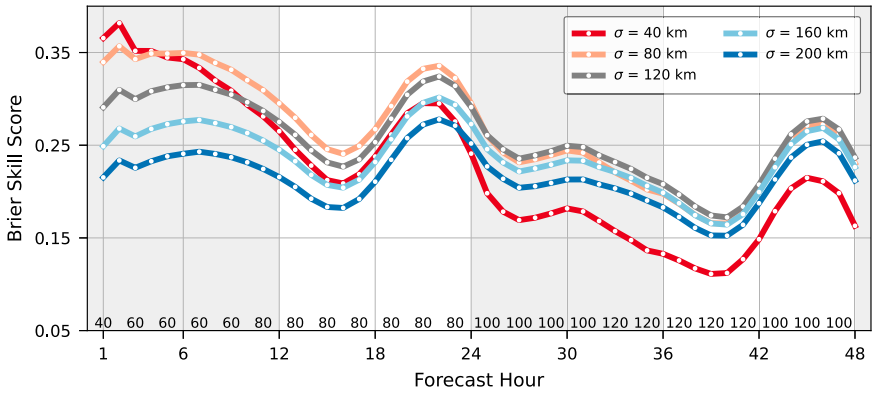

FIG. 7. As in Fig. 6, but for the 4-h, 20-km UP CG1-SSPFs using σ = 40, 80, 120, 160, and 200 km. Text along the x axis indicates the σ that generates forecasts with the largest BSS for every other forecast hour.

largest between 1800 and 1900 UTC, indicating a lag in CG1-SSPF skill compared to the CG1-NNPFs. Overnight, BSS differences approached zero, indicating that the CG1-NNPFs were similar in skill to the CG1-SSPFs. This was especially true overnight during the second diurnal cycle, i.e., forecast hours 33–39, where the BSS differences were not statistically significantly different from zero (Fig. 8). Many of these skill characteristics are similar to those seen in Sobash et al. (2020) when comparing their ML and surrogate forecasts for severe weather reports.

When aggregated across forecast hours, the 4- and 1-h CG1-NNPF and UP CG1-SSPFs both exhibited excellent reliability (Figs. 9a,b). The CG1-NNPFs had slightly better reliability compared to the UP CG1-SSPFs, primarily for probabilities >50%, although these differences were small. While the distribution of forecast probabilities was similar between the 4-h CG1-NNPF and CG1-SSPFs, the 1-h CG1-NNPFs generated larger probability values compared to the 1-h UP CG1-SSPFs. This is likely due to the impact of the spatial smoothing on the generation of the UP CG1-SSPFs, which reduced forecast sharpness. Variations in reliability were more evident when the CG1-NNPFs were evaluated by forecast hour (Figs. 9c,d). Both the 4- and 1-h CG1-NNPFs were most reliable at forecast hour 1, with reliability decreasing overnight. For example, at forecast hour 12, the 4- and 1-h CG1-NNPFs possessed an overforecasting bias at high probabilities. This was also somewhat evident during the day 2 overnight period at forecast hour 36. Combined with the BSS results (Fig. 8), this indicates that the decreased BSS overnight is due in part to overconfident forecast probabilities.

Finally, in addition to the CG1-NNPFs, the NNs output two additional forecasts: the probability of ≥1 ENTLN IC (IC1) flash and the probability of ≥1 GLM (GLM1) flash for both the 4- and 1-h time windows. The BSSs for the IC1 and GLM1 NNPFs were nearly identical for all forecast hours (Fig. 10), which was not surprising given their similar base rates (Fig. 2) and the fact that IC flashes dominate the GLM flash detections. The IC1 and GLM1 NNPFs had slightly

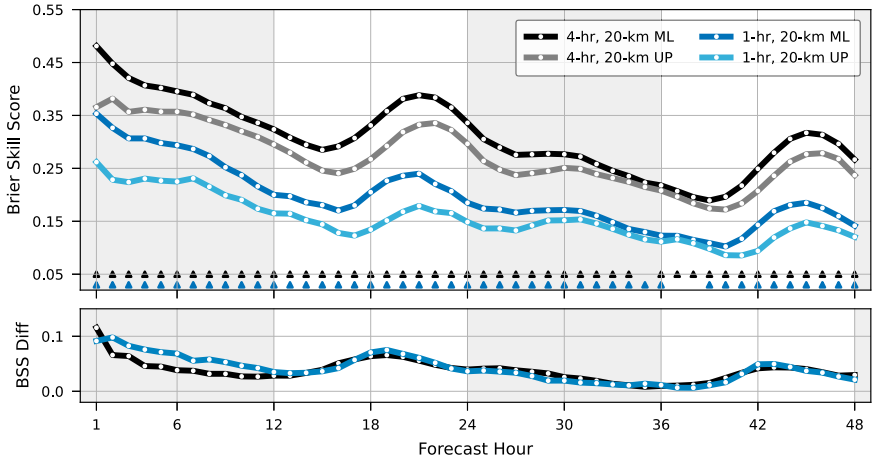

FIG. 8. As in Fig. 6, but for the 4- and 1-h CG1-NNPFs and UP CG1-SSPFs. Triangles denote if the CG1-NNPF BSS is statistically significantly larger (upward triangle) or smaller (downward triangle) than the UP CG1-SSPFs at the 90% confidence level.

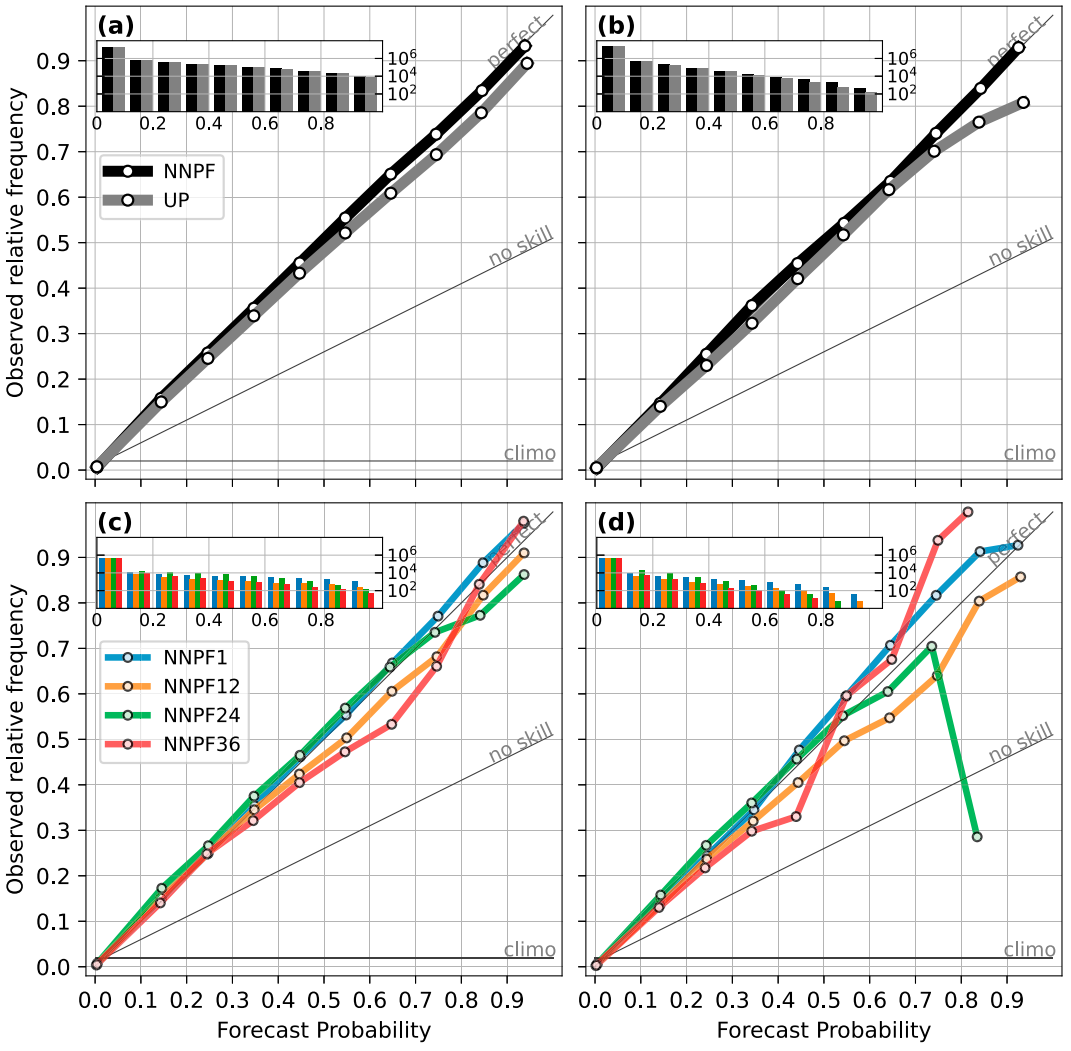

FIG. 9. Attributes diagrams for (a) 4-h CG1-NNPFs and UP CG1-SSPFs and (b) 1-h CG1-NNPFs and UP CG1-SSPFs aggregated over all forecast hours, as well as (c) 4- and (d) 1-h CG1-NNPFs for forecast hours 1, 12, 24, and 36.

larger BSSs than the CG1-NNPFs for both the 4- and 1-h forecast time windows, which is likely due to the smaller base rate and greater challenge of predicting CG1 compared to IC1 or GLM1. All forecasts exhibited a similar diurnal cycle of skill.

c. *NNPF skill relative to SPC thunderstorm outlooks in 2021*

In this section, we compare the performance of the 4-h discretized CG1-NNPFs to the SPC probabilistic CG1 Thunderstorm Outlooks (CG1-SPC). First, we examine the aggregate verification statistics across all CG1-SPC outlooks, including all four issuance times and five valid time ranges, and compare to the corresponding 4-h CG1-NNPFs. While the discretized CG1-NNPFs and CG1-SPC appear to be underconfident (Fig. 11a), this is largely an artifact of the discretization, given that the continuous CG1-NNPFs were quite reliable (e.g., Fig. 9a). For example, the observed frequency of CG1 events within the 40%–70% forecast areas was ≈49% and 53% for the CG1-SPC and discretized CG1-NNPFs, respectively (Fig. 11a). Since both forecasts' observed frequencies fell within the range of expected observed frequencies, we assume both forecasts have similar reliability; this was also true for forecasts in the 10%–40% and 70%–100% bins.

One notable difference is that the CG1-SPC outlooks contained a larger number of 10% and 40% probabilities and a smaller number of 70% probabilities (e.g., Figs. 11b,c). Because of this, the forecast bias was smaller for the CG1-NNPFs compared to the CG1-SPC forecasts at the 10% and 40% thresholds (i.e., the aggregated CG1-SPC outlook areas at these thresholds were larger than the CG1-NNPF areas). On the other hand, the bias was slightly smaller at the 70% threshold (e.g., Fig. 11b). Even though the reliability was slightly better for the CG1-SPC outlooks, the CG1-SPC BSS was smaller (0.27 vs 0.23) with lower CSIs at the three nonzero forecast thresholds (Fig. 11b). The larger area of coverage for

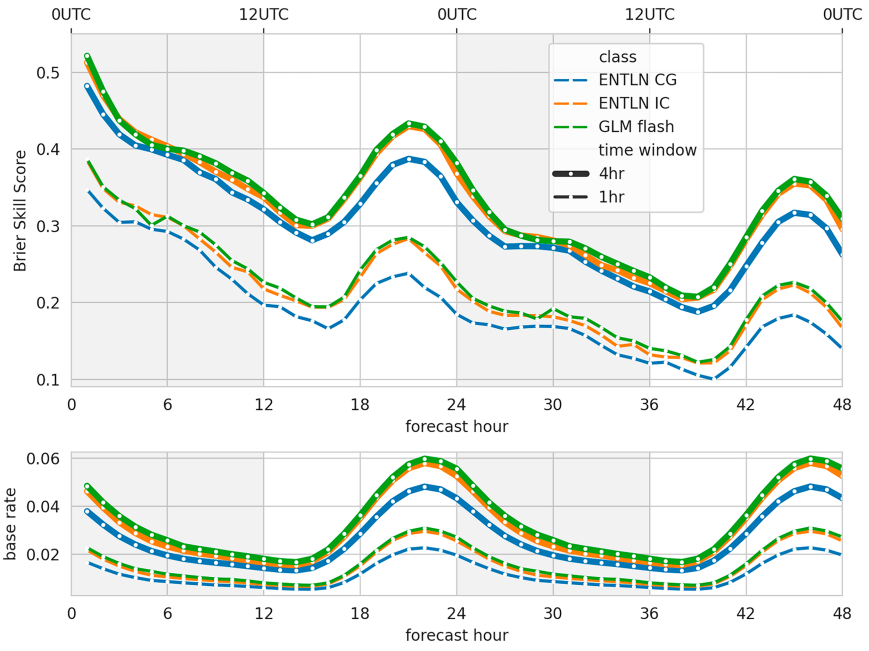

FIG. 10. (top) BSS and (bottom) observation base rate as a function of forecast hour for ≥1 ENTLN CG (blue), ENTLN IC (orange), and GLM flash (green) NNPFs for both 4-h (solid) and 1-h (dashed) forecasts.

the 10% probability may be due to a prioritization of the probability of detection in the CG1-SPC forecasts. Doing so reduces the aggregate BSS and CSI, although the forecasts remain reliable at these lower probability thresholds.

Next, the skill of the CG1-SPC outlooks is presented as a function of issuance and valid time to deconstruct the influences of forecast lead time and time of day, given that valid time influences forecast skill (e.g., Fig. 8). The CG1-SPC outlooks were most skillful at the 2000–0000 UTC valid time range, for all three issuance times (i.e., 0600, 1300, and 1700 UTC; Fig. 12a), similar to when the peak occurred in CG1-NNPF skill (Fig. 8). For the forecasts valid at 2000–0000 UTC, the CG1-NNPFs were of similar skill to the CG1-SPC outlooks with only small differences in BSS. This was also true for the 0000–0400 UTC valid range, although during this window, the CG1-SPC outlooks were more skillful at the 1700 and 2100 UTC issuance times. Larger BSS differences in favor of the CG1-NNPFs existed overnight (0400–1200 UTC) and for the morning and early afternoon valid times (1200–1600 and 1600–2000 UTC). For all of these overnight and early morning forecasts, the CG1-NNPFs were statistically significantly more skillful. For instance, during the 0400–1200 UTC valid time, the CG1-NNPF BSS was 0.04 larger than the CG1-SPC outlook BSS issued at 1700 UTC; BSS differences ≥ 0.05 occurred at the 1200–1600 and 1600–2000 UTC valid times as well.

In general, the difference between CG1-NNPF and CG1-SPC skill decreased as the SPC issuance time became closer to the valid time (i.e., lead time is reduced, moving upward in the panels within Fig. 12). For instance, the skill benefit of the CG1-NNPFs during the 0400–1200 UTC forecast window when using the 1700 UTC issued SPC forecasts was nearly removed when compared to the 0130 UTC issued CG1-SPC outlook. The CG1-SPC forecast has a lead time of only 2.5 h, compared to the CG1-NNPF which was issued based on the 0000 UTC HRRR initialized 25.5 h prior. This trend of improved CG1-SPC outlook skill as lead time decreased occurred at all valid times.

d. CG1-NNPF, CG1-SPC, and HREF-CT skill in 2022

We further examine skill differences between the CG1-NNPFs and CG1-SPC outlooks by extending the analysis to 2022. In addition, we compared CG1-NNPF skill to the HREF-CT CG1 forecasts. While HREF-CT forecasts were available in 2021, we chose to compare the HREF-CT forecasts to the CG1-NNPFs only in 2022 since the composition of the operational HREF was not fixed throughout 2021. Similar to the analysis in the previous section, we used all available 2022 CG1-SPC 4-h forecasts and paired those with the 4-h 0000 UTC CG1-NNPFs, while for the HREF-CT comparison, we used the 0000 UTC HREF-CT and 0000 UTC CG1-NNPFs, both for the 4- and 1-h time windows.

Some improvement of the CG1-SPC forecasts over the CG1-NNPFs was noted during 2022 (Fig. 13) relative to 2021, although many trends remain the same. For example, most CG1-SPC forecasts outperformed the CG1-NNPFs for the 2000–0000 UTC and 0000–0400 UTC valid times with these differences being statistically significant, whereas in 2021, these BSS differences, while of similar magnitude, favored the CG1-NNPFs. Yet, the overnight CG1-NNPFs were still more skillful than the CG1-SPCs (Fig. 13). Other trends seen in the 2021 forecasts were also similar, e.g., the BSS differences between

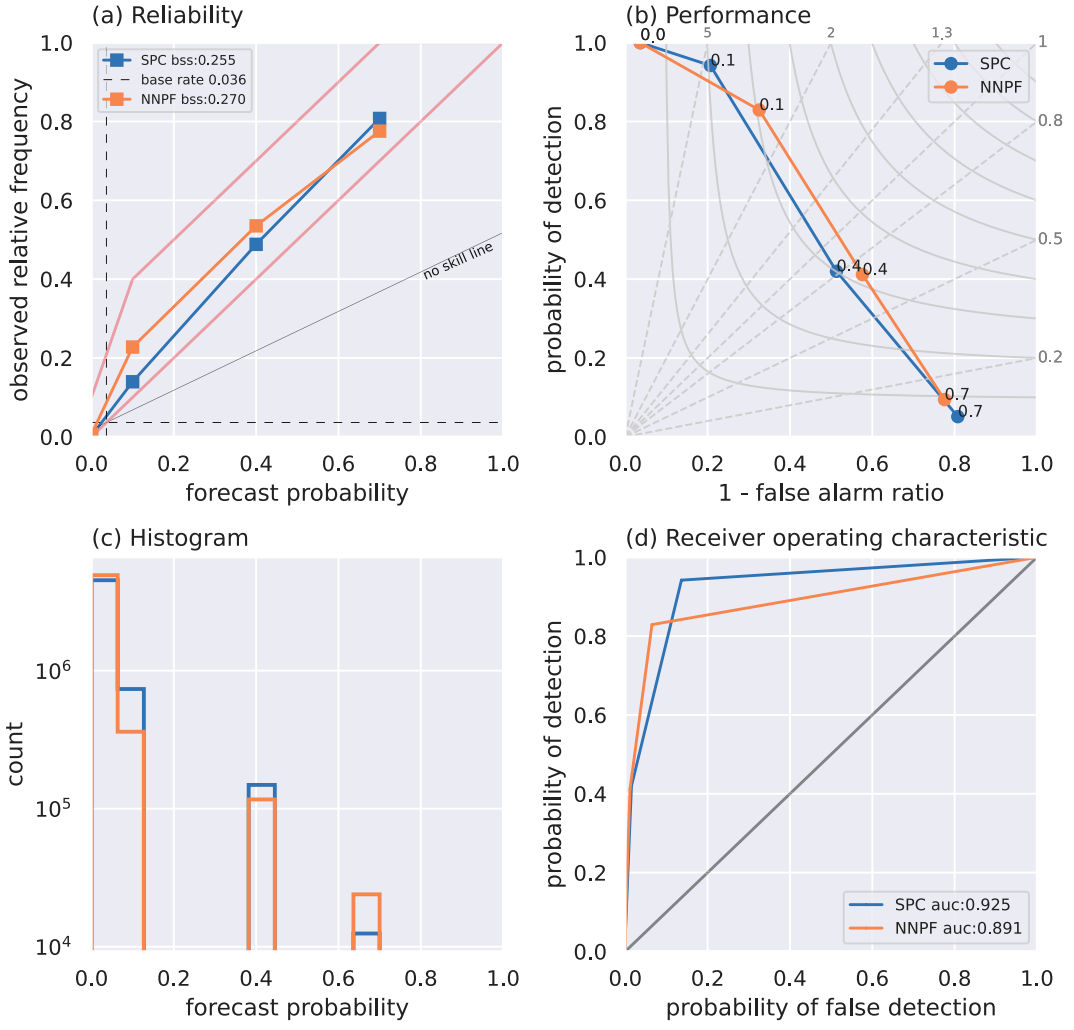

FIG. 11. (a) Reliability diagram, (b) performance diagram, (c) forecast count histogram, and (d) receiver operating characteristic curves for the CG1-SPC (blue) and discretized CG1-NNPFs (orange) aggregated across all available 2021 forecasts. Red lines in (a) indicate range of observed frequencies that are considered reliable for the discretized forecasts.

the CG1-SPC and CG1-NNPFs decreased with shorter lead times as the CG1-SPC forecasts became more skillful.

To assess the skill of the HREF-CT guidance relative to the CG1-NNPFs, we compared these two sets of forecasts in 2022. As described in Harrison et al. (2022), the HREF-CT CG1 probabilistic forecasts are generated using a weighted combination of three HREF ensemble probabilities: the probability of 4-km AGL reflectivity exceeding 40 dBZ, accumulated QPF exceeding 0.08 in., and most-unstable lifted index ≤ -1. These probabilities are then calibrated to be statistically reliable using National Lightning Detection Network (NLDN) CG1 lightning strikes. Both the 4- and 1-h HREF-CT CG1 forecasts were competitive with the CG1-NNPFs, although for most forecast hours the CG1-NNPFs had larger BSSs than the HREF-CT forecasts; these differences were nearly always statistically significant (Fig. 14). The largest BSS differences occurred during the first ≈18 h of the forecast when the CG1-NNPF BSSs were larger by as much as 0.1–0.2. Beyond 18 h, the BSS differences were smaller, with nearly identical BSSs for the 4- and 1-h forecasts for forecast hours 21–36, even though these small differences were often statistically significant. BSS differences increased during the start of the second diurnal cycle, i.e., from 36 to 45 h, favoring the CG1-NNPFs. The increase in forecast skill by using ML, even when only using a single deterministic CAM compared to using simpler calibration with an ensemble, is impressive and suggests that ML-based ensemble thunderstorm guidance could improve forecast skill even further.

4. Conclusions and discussion

In this work, we generated 4- and 1-h probabilistic forecasts of lightning across the CONUS by postprocessing the HRRR with neural networks (NNs). These NNPFs were produced

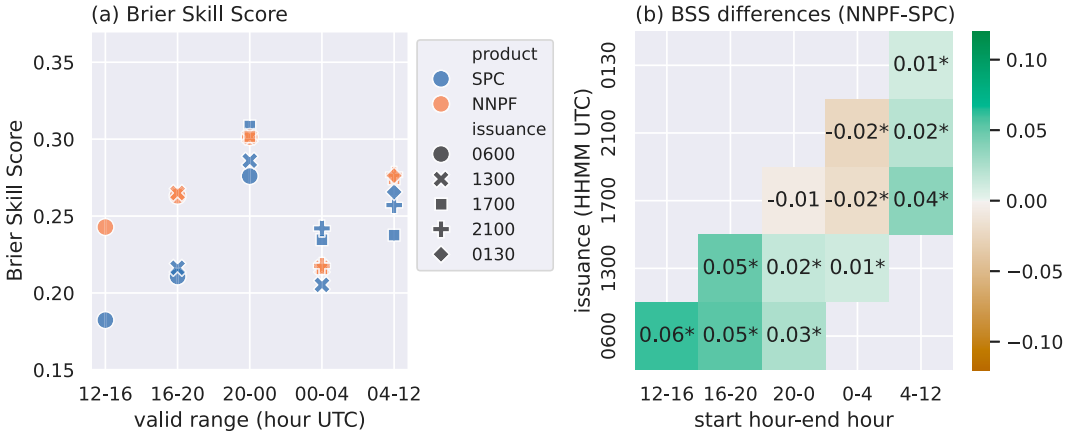

FIG. 12. (a) BSS of CG1-SPC and discretized CG1-NNPFs aggregated by valid (x axis) and issuance time (marker style). The CG1-NNPF BSS may vary slightly for different issuance times due to differences in the number of missing CG1-SPC forecasts. (b) BSS differences between the discretized CG1-NNPFs and CG1-SPC forecasts. BSS differences are starred if they are statistically significant at the 90% confidence level using an $n = 1000$ bootstrap resampling across forecast dates. For both panels, forecast skill was aggregated across all forecasts issued between 1 Jan 2021 and 31 Dec 2021.

for each 0000 UTC HRRR forecast in 2021 using NNs trained with 0000 UTC HRRR forecasts from 2019 to 2020. For training, we used *GOES-16* GLM flashes as well as ground-based lightning sensors from the ENTLN, which can discriminate between IC and CG flashes. NN output included the probability of ≥1 CG (CG1), IC, or GLM lightning flash within 20 km of a point, within a 4- or 1-h window centered on HRRR forecast hours between 1 and 48. We evaluated the CG1-NNPFs using standard verification metrics, such as BSS and reliability. In addition, the skill of the CG1-NNPFs was compared to several non-ML baselines, including surrogate severe probabilistic forecasts (SSPFs) for CG1 using six lightning surrogates, operational SPC Thunderstorm Outlooks, and calibrated HREF thunderstorm guidance.

Among the six lightning surrogates, hourly maximum updraft speed (UP) resulted in the CG1-SSPFs with the largest BSS across nearly all forecast hours, while midlevel updraft helicity (UH) was the least skillful. CG1-SSPFs using the HRRR lightning diagnostic, designed to be a surrogate for lightning flashes in CAMs, was less skillful than the UP CG1-SSPFs. The skill of the CG1-SSPFs was found to be sensitive to the degree of spatial smoothing. The optimal smoothing length increased steadily with lead time from 40 km (i.e., little-to-no smoothing) at lead times of 1–6 h, increasing to 120 km by forecast hour 30. The CG1-NNPFs were compared to the UP CG1-SSPFs and were found to be statistically significantly more skillful than the CG1-SSPFs, especially during the first 6–12 h of the forecast and during the early afternoon part of the diurnal cycle, when convection often initiates. These differences were reduced overnight, especially during the second diurnal cycle (forecast hours 30–36). The CG1-NNPFs exhibited excellent reliability, although overforecasting was noted overnight. CG1-NNPF skill had a strong diurnal cycle, similar to the cycle noted in the forecasts of other severe hazards in

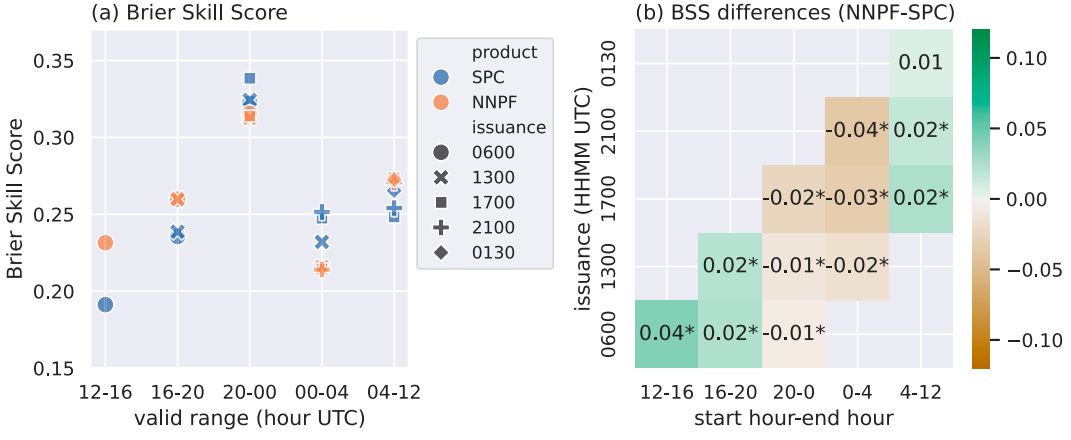

FIG. 13. As in Fig. 12, but for CG1-NNPF and CG1-SPC forecasts aggregated across all forecasts issued between 1 Jan 2022 and 31 Dec 2022.

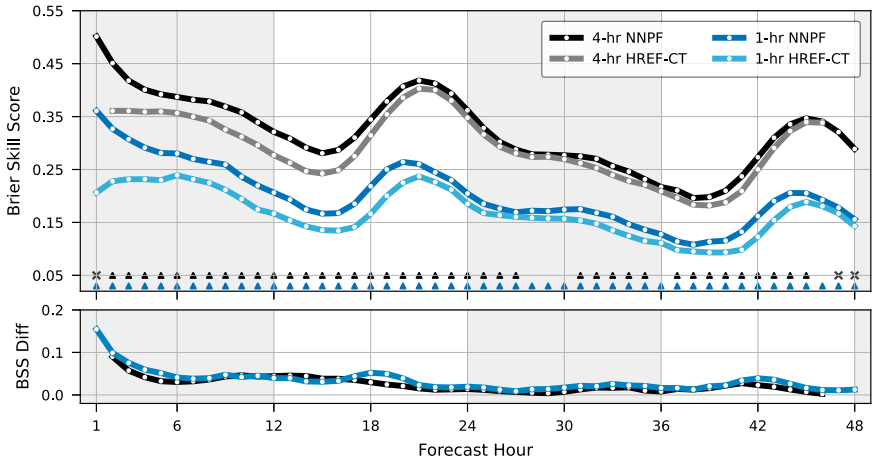

FIG. 14. As in Fig. 8, but for CG1-NNPFs and HREF-CT forecasts using all 0000 UTC initialized HRRR and HREF forecasts between 1 Jan 2022 and 31 Dec 2022.

SA24. All of these results were similar for both the 4- and 1-h CG1-NNPFs, with the 1-h CG1-NNPFs having lower skill than the 4-h CG1-NNPFs.

The CG1-NNPFs, discretized to match the SPC outlook probability intervals, were of similar skill to the SPC Thunderstorm Outlooks during late afternoon and evening valid times (i.e., 2000–0000 and 0000–0400 UTC) but were superior to the SPC outlooks overnight (i.e., 0400–1200 and 1200–1600 UTC) and into the early afternoon (1600–2000 UTC). This was true for both 2021 and 2022. For all valid periods, the differences between the CG1-NNPFs and CG1-SPC outlooks decreased as the issuance time approached the valid time, reflecting the increase in skill of the CG1-SPC outlooks as lead time decreased. The CG1-NNPFs were also more skillful than CG1 guidance generated from the HREF (HREF-CT). Differences between these two guidance products were greatest during the first 12–18 h of the forecast, after which skill differences approached zero. This was true at both the 1- and 4-h forecast time scales.

Further exploration and discussion of the reasons for the differences between the CG1-NNPFs and the three forecast baselines is warranted. First, in both 2021 and 2022, the CG1-NNPFs were statistically significantly more skillful than the CG1-SPC outlooks, especially overnight. Additionally, it is surprising that the SPC forecasts did not drastically improve upon the CG1-NNPFs at short lead times (e.g., the 2000–0000 and 0000–0400 UTC valid times issued at 1300 and 1700 UTC), although there is some signal of this in the 2022 forecasts at short afternoon forecast lead times. Thus, using the CG1-NNPFs as a first guess tool across all forecast issuance times may be of value to forecasters. One explanation for the overnight and early morning skill differences is that SPC forecasters may have relied on SREF-based CT guidance to generate the CG1-SPC outlooks in 2021 and 2022 (I. Jirak 2023, personal communication), which perform poorly overnight, as noted by Harrison et al. (2022). The decrease in the skill gap in 2022 compared to 2021 may be due to increased utilization of the HREF-CT guidance. Future work should examine more recent CG1-SPC forecasts to see if these skill differences remain.

It is noteworthy that the CG1-NNPFs, based on a single deterministic CAM forecast, had larger BSSs than the HREF-CT forecasts, which use a full 10-member ensemble to derive forecast uncertainty. One possibility is that the NNs were better able to learn uncertainty information rather than using simple calibration tables based on the observed frequencies of CG1 in certain predicted environmental conditions. A second possibility is that the HREF-CT used NLDN CG1 for calibration while the CG1-NNPFs used ENTLN CG1 flashes for training. This may favor the CG1-NNPFs since these use the same data source for training and verification. Melick et al. (2015) noted that NLDN CG1 flash counts are slightly larger than ENTLN CG1 counts, although their analysis was performed at a finer spatial scale (4 km), and recent network upgrades (e.g., Murphy et al. 2021; Zhu et al. 2022) may have changed these biases. Since our forecasts predict CG1 on relatively coarse grids over 1- or 4-h time periods, we expect the differences in skill metrics due to lightning dataset to be small, although this should be confirmed in the future.

In any case, training the NNs with an ensemble of CAMs, such as the HREF or RRFS, would likely result in NNPFs with even greater skill. The biggest differences between the HREF-CT and CG1-NNPFs were during model spinup and during the early morning and afternoon hours, similar to the differences noted when comparing the CG1-NNPFs to the CG1-SSPFs (e.g., Fig. 8). This may suggest that similar biases, such as delayed convection initiation, exist among all CAM forecasts comprising the HREF, and that this bias is better ameliorated with ML. Other avenues of future work include examining CG1 forecasts outside of CONUS, e.g., offshore regions of the CONUS, as well as adjacent countries within the HRRR domain, such as northern Mexico and southern Canada, where remotely sensed observations can be exploited to improve predictions of thunderstorms and their hazards, as well as the utility of the NNPFs in specific forecasting domains, e.g., for planning and routing in aviation.

Acknowledgments. We thank Dr. Terra Ladwig of NOAA/Global Systems Laboratory, who provided the HRRRX dataset, as well as Dr. David Harrison of NOAA/SPC who provided us with archived HREF-CT forecasts. Discussions with Dr. Israel Jirak of NOAA/SPC also improved aspects of the manuscript. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977. This research was also supported by NOAA OAR Grants NA19OAR4590128 and NA21OAR4590163 and the NCAR Short-term Explicit Prediction Program. Supercomputing support was provided by NCAR Cheyenne and Casper (Computational and Information Systems Laboratory 2020). MRMS column-max reflectivity was converted from Mdv with Py-ART (Helmus and Collis 2016).

Data availability statement. Operational HRRR data were retrieved from Google Cloud. GLM flash centroids came from Amazon S3 bucket noaa-goes16 product GLM-L2-LCFA. Cloud-to-ground and in-cloud flash centroids were originally produced by Weather Bug (formerly Earth Networks) and generally may not be redistributed, except for research or noncommercial purposes.

REFERENCES

Blaylock, B. K., and J. D. Horel, 2020: Comparison of lightning forecasts from the high-resolution rapid refresh model to Geostationary Lightning Mapper observations. *Wea. Forecasting*, **35**, 401–416, https://doi.org/10.1175/WAF-D-19-0141.1.

Bright, D. R., M. S. Wandishin, R. E. Jewell, and S. J. Weiss, 2005: A physically based parameter for lightning prediction and its calibration in ensemble forecasts. *Conf. on Meteorological Applications of Lightning Data*, San Diego, CA, Amer. Meteor. Soc., 4.3, https://ams.confex.com/ams/pdfpapers/84173.pdf.

Bruning, E., 2019: Deeplycloudy/glmtools: Glmtools release to accompany publication. Zenodo, accessed 15 December 2021, https://doi.org/10.5281/zenodo.2648658.

Bruning, E. C., and Coauthors, 2019: Meteorological imagery for the Geostationary Lightning Mapper. *J. Geophys. Res. Atmos.*, **124**, 14 285–14 309, https://doi.org/10.1029/2019JD030874.

Charba, J. P., F. G. Samplatsky, A. J. Kochenash, P. E. Shafer, J. E. Ghirardelli, and C. Huang, 2019: Lamp upgraded convection and total lightning probability and "potential" guidance for the conterminous United States. *Wea. Forecasting*, **34**, 1519–1545, https://doi.org/10.1175/WAF-D-19-0015.1.

Cintineo, J. L., and Coauthors, 2018: The NOAA/CIMSS ProbSevere model: Incorporation of total lightning and validation. *Wea. Forecasting*, **33**, 331–345, https://doi.org/10.1175/WAF-D-17-0099.1.

——, M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: Noaa Probsevere v2.0—Probhail, Probwind, and Probtor. *Wea. Forecasting*, **35**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

——, ——, and ——, 2022: Probsevere lightningcast: A deep-learning model for satellite-based lightning nowcasting. *Wea. Forecasting*, **37**, 1239–1257, https://doi.org/10.1175/WAF-D-22-0019.1.

Clark, A. J., and Coauthors, 2022: The third real-time, virtual spring forecasting experiment to advance severe weather prediction capabilities. *Bull. Amer. Meteor. Soc.*, **104**, S456–S458, https://doi.org/10.1175/BAMS-D-22-0213.1.

Computational and Information Systems Laboratory, 2020: HPE/SGI ICE XA—Cheyenne. National Center for Atmospheric Research, accessed 13 December 2024, https://doi.org/10.5065/D6RX99HX.

Dowell, D. C., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Du, J., and Coauthors, 2014: NCEP regional ensemble update: Current systems and planned storm-scale ensembles. *26th Conf. on Weather Analysis and Forecasting/22nd Conf. on Numerical Weather Prediction*, Atlanta, GA, Amer. Meteor. Soc., J1.4, https://ams.confex.com/ams/94Annual/webprogram/Paper239030.html.

Dye, J. E., W. P. Winn, J. J. Jones, and D. W. Breed, 1989: The electrification of New Mexico thunderstorms: 1. Relationship between precipitation development and the onset of electrification. *J. Geophys. Res.*, **94**, 8643–8656, https://doi.org/10.1029/JD094iD06p08643.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazards in the Warn-on-Forecast system. *Mon. Wea. Rev.*, **149**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

Goodman, S. J., and Coauthors, 2013: The GOES-R Geostationary Lightning Mapper (GLM). *Atmos. Res.*, **125–126**, 34–49, https://doi.org/10.1016/j.atmosres.2013.01.006.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Harrison, D. R., M. S. Elliott, I. L. Jirak, and P. T. Marsh, 2022: Utilizing the high-resolution ensemble forecast system to produce calibrated probabilistic thunderstorm guidance. *Wea. Forecasting*, **37**, 1103–1115, https://doi.org/10.1175/WAF-D-22-0001.1.

——, ——, ——, and ——, 2023: Predicting probabilistic lightning flash density from the HREF calibrated thunder guidance. *22nd Conf. on Artificial Intelligence for Environmental Science*, Denver, CO, Amer. Meteor. Soc., 6A.1, https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/416071.

Helmus, J. J., and S. M. Collis, 2016: The Python Arm Radar Toolkit (Py-ART), a library for working with weather radar data in the python programming language. *J. Open Res. Software*, **4**, e25, https://doi.org/10.5334/jors.119.

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

——, R. S. Schumacher, and I. L. Jirak, 2023: A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. *Wea. Forecasting*, **38**, 251–272, https://doi.org/10.1175/WAF-D-22-0143.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

——, S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, https://doi.org/10.1175/2010WAF2222430.1.

Lapierre, J. M. Hoekzema, M. Stock, C. Merrill, and S. C. Thangaraj, 2019: Earth networks lightning network and dangerous thunderstorm alerts. *2019 11th Asia-Pacific Int. Conf. on Lightning (APL)*, Hong Kong, China, Institute of Electrical and Electronics Engineers, 1–5, https://doi.org/10.1109/APL.2019.8816032.

Leinonen, J., U. Hamann, U. Germann, and J. R. Mecikalski, 2022: Nowcasting thunderstorm hazards using machine learning: The impact of data sources on performance. *Nat. Hazards Earth Syst. Sci.*, **22**, 577–597, https://doi.org/10.5194/nhess-22-577-2022.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Wea. Forecasting*, **35**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0258.1.

——, ——, and A. McGovern, 2022: Comparing and interpreting differently designed random forests for next-day severe weather hazard prediction. *Wea. Forecasting*, **37**, 871–899, https://doi.org/10.1175/WAF-D-21-0138.1.

McCaul, E. W., Jr., S. J. Goodman, K. M. LaCasse, and D. J. Cecil, 2009: Forecasting lightning threat using cloud-resolving model simulations. *Wea. Forecasting*, **24**, 709–729, https://doi.org/10.1175/2008WAF2222152.1.

Melick, C. J., P. Marsh, A. Dean, I. L. Jirak, and S. J. Weiss, 2015: Lightning characteristics and relationship to preliminary local storm reports. *40th National Weather Association Annual Meeting*, Oklahoma City, OK, National Weather Association, 1–10, https://www.spc.noaa.gov/publications/melick/ltg-lsr.pdf.

Murphy, M. J., J. A. Cramer, and R. K. Said, 2021: Recent history of upgrades to the U.S. National Lightning Detection Network. *J. Atmos. Oceanic Technol.*, **38**, 573–585, https://doi.org/10.1175/JTECH-D-19-0215.1.

Ortland, S. M., M. J. Pavolonis, and J. L. Cintineo, 2023: The development and initial capabilities of thundercast, a deep learning model for thunderstorm nowcasting in the United States. *Artif. Intell. Earth Syst.*, **2**, e230044, https://doi.org/10.1175/AIES-D-23-0044.1.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Price, C., and D. Rind, 1992: A simple lightning parameterization for calculating global lightning distributions. *J. Geophys. Res.*, **97**, 9919–9933, https://doi.org/10.1029/92JD00719.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258 https://doi.org/10.1175/BAMS-D-18-0041.1.

——, B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Wea. Forecasting*, **35**, 2293–2316, https://doi.org/10.1175/WAF-D-20-0069.1.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Sobash, R. A., and D. A. Ahijevych, 2024: Evaluating machine learning–based probabilistic convective hazard forecasts using the HRRR: Quantifying hazard predictability and sensitivity to training choices. *Wea. Forecasting*, **39**, 1399–1415, https://doi.org/10.1175/WAF-D-23-0221.1.

——, J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

——, G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Wea. Forecasting*, **35**, 1981–2000, https://doi.org/10.1175/WAF-D-20-0036.1.

Song, G., S. Li, and J. Xing, 2023: Lightning nowcasting with aerosol-informed machine learning and satellite-enriched dataset. *npj Climate Atmos. Sci.*, **6**, 126, https://doi.org/10.1038/s41612-023-00451-x.

Thompson, K. B., M. G. Bateman, and L. D. Carey, 2014: A comparison of two ground-based lightning detection networks against the satellite-based Lightning Imaging Sensor (LIS). *J. Atmos. Oceanic Technol.*, **31**, 2191–2205, https://doi.org/10.1175/JTECH-D-13-00186.1.

Wang, X., K. Hu, Y. Wu, and W. Zhou, 2023: A survey of deep learning-based lightning prediction. *Atmosphere*, **14**, 1698, https://doi.org/10.3390/atmos14111698.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd Edition, Elsevier, 627 pp.

Yair, Y., B. Lynn, C. Price, V. Kotroni, K. Lagouvardos, E. Morin, A. Mugnai, and M. del Carmen Llasat, 2010: Predicting the potential for lightning activity in mediterranean storms based on the Weather Research and Forecasting (WRF) model dynamic and microphysical fields. *J. Geophys. Res.*, **115**, D04205, https://doi.org/10.1029/2008JD010868.

Zhou, K., Y. Zheng, W. Dong, and T. Wang, 2020: A deep learning network for cloud-to-ground lightning nowcasting with multisource data. *J. Atmos. Oceanic Technol.*, **37**, 927–942, https://doi.org/10.1175/JTECH-D-19-0146.1.

Zhu, Y., M. Stock, J. Lapierre, and E. DiGangi, 2022: Upgrades of the Earth Networks Total Lightning Network in 2021. *Remote Sens.*, **14**, 2209, https://doi.org/10.3390/rs14092209.