On the Evaluation of Probabilistic Thunderstorm Forecasts and the Automated Generation of Thunderstorm Threat Areas during Environment Canada Pan Am Science Showcase

DOMINIQUE BRUNET, DAVID SILLS,^a AND NORBERT DRIEDGER

Environment and Climate Change Canada, Toronto, Ontario, Canada

(Manuscript received 21 January 2019, in final form 16 July 2019)

ABSTRACT

An object-based forecasting, nowcasting, and alerting system prototype was demonstrated during the summer 2015 Environment Canada Pan Am Science Showcase (ECPASS) in Toronto. Part of this demonstration involved the generation of experimental thunderstorm threat areas by both automated NWP postprocessing algorithms and by a pair of human forecasters. This paper first develops a rigorous verification methodology for the intercomparison of continuous as well as categorical probabilistic thunderstorm forecasts. The methodology is then applied to the intercomparison of thunderstorm forecasts made during ECPASS. Statistical postprocessing of forecasts by smoothing with optimal bandwidth followed by recalibration is found to improve the skill scores of all thunderstorm forecasts studied at all lead times between 6 and 48 h. In addition, the calibrated ensemble mean forecasts are found to be better than the calibrated deterministic thunderstorm forecasts for all lead times considered, though postprocessing of the convective rain-rate forecast gives results that are statistically comparable with the ensemble mean forecast. Thunderstorm threat areas that were automatically generated by thresholding the output of NWP-based postprocessed algorithms have better scores than those generated by human forecasters for most lead times beyond 9 h, indicating that they could be integrated as an automated tool for providing high-quality “first-guess” thunderstorm threat areas in an object-based forecasting, nowcasting, and alerting system. A unique contribution of this paper is a novel verification methodology for the fair comparison between continuous and categorical probabilistic forecasts, a methodology that could be used for other experiments involving human- and automatically generated object-based forecasts derived from probabilistic forecasts.

1. Introduction

In southern Canada, one of the main meteorological threats to life and property in summer is the severe thunderstorm. In addition to lightning, such storms have the potential to produce damaging winds (gusts to 90 km h^{-1} or more), large hail (2 cm or more in diameter), flash floods, and tornadoes. Environment and Climate Change Canada (ECCC) has a mandate to warn Canadians about such threatening weather phenomena.

In addition to severe weather watches and warnings, ECCC also issues public forecasts that refer to the probability of thunderstorms.

The Environment Canada Pan Am Science Showcase (ECPASS) took place during the 2015 Pan Am/ParaPan Am Games in Toronto (Joe et al. 2018). As part of ECPASS, prototype “next generation” forecasting and nowcasting tools employing an object-based approach (Sills et al. 2009) were demonstrated and evaluated. At two Research Support Desks (RSDs; Sills and Taylor 2008), meteorologists created analyses, nowcasts, and forecasts focused on thunderstorm and severe weather threats (see Joe et al. 2018 for details). Part of the demonstration involved the development and evaluation of automated tools for processing object-based forecasts such as thunderstorm threat areas. For example, methods for the automated time interpolation of threat areas have been reported in Brunet and Sills (2015).

Denotes content that is immediately available upon publication as open access.

^a Current affiliation: Western Engineering, University of Western Ontario, London, Ontario, Canada.

Corresponding author: Dominique Brunet, dominique.brunet@canada.ca

In this paper, we focus on the automated generation of thunderstorm threat areas. We also report on the statistical postprocessing and verification of gridded numerical weather prediction (NWP) guidance, as they are the sources for the automated thunderstorm threat areas. Specifically, we perform 1) an intercomparison of raw and postprocessed NWP guidance as well as 2) an intercomparison of automatically generated and human-generated threat areas. In both cases, lightning observations are used for the “truth” dataset.

The idea of postprocessing forecasts by smoothing has been proposed by several authors. In converting a deterministic forecast to a probabilistic one using a neighborhood approach, Theis et al. (2005) use the terminology “pseudo ensemble” to describe smoothed NWP output, a good interpretation of such derived products. They note some advantages in terms of computation cost compared to running actual mesoscale ensemble models, and compared to ongoing subjective interpretation by forecasters. Indeed, neighborhood methods (see Ebert 2008) have been used to enhance sample size even given multiple ensemble members to work with. Schwartz and Sobash (2017) provide a review of such methods, and of relevance to our study, provide useful commentary on the interpretation of relative operating characteristic (ROC) curves. Furthermore, Adams-Selin et al. (2019) used the same upscaling technique as Ben Bouallègue and Theis (2014) during their WRF-HAILCAST study using an 80-km grid chosen to match severe weather outlook production at the U.S. Storm Prediction Center (SPC). Probabilities were averaged across ensemble members, and then further smoothed with a Gaussian smoother with standard deviation of 120 km, thus highlighting another project with considerable smoothing. The amount of smoothing proposed in the different methods referred above is in line with a skillful spatial scale of more than 50 km for various WRF precipitation models as found by Mittermaier and Roberts (2010) using the fractions skill score of Roberts and Lean (2008).

Two previous works on the verification of thunderstorm forecasts in Canada are Casati and Wilson (2007) and Sills et al. (2012). The generation of automated thunderstorm and severe weather threat areas has also been investigated by a few authors (see e.g., Mills 2004; Dance et al. 2010; Karstens et al. 2015). The unique contribution of this paper is to combine smoothing, dilation, recalibration and thresholding methods for both the generation of automatic thunderstorm threat areas and for the comparison with human forecaster generated threat areas. Of particular interest for the demonstration was useful feedback on the associated forecaster workload and the automatically generated “first-guess”

thunderstorm threat areas. We do not discuss here results related to forecaster workload and workflow, but suggest that high-quality automated guidance should improve both (e.g., Karstens et al. 2018). The problem of verifying and interpreting warnings is also investigated in Sharpe (2016), which provides a complementary discussion to what is presented in this paper.

The paper is divided into nine sections. In the next section, we describe the different forecasts and observations we use, as well as preprocessing methods to map data to a common verification domain. Methods for simple and efficient statistical postprocessing of forecasts are described in section 3. Training results are presented in section 4 and the verification results on the effect of postprocessing are shown in section 5. The NWP guidance intercomparison as well as the human- and automatically generated thunderstorm threat areas intercomparison are presented in section 6 and section 7, respectively. We follow with a discussion of the methods and results in section 8 and conclusions of the paper in section 9.

2. Data sources

An intensive observation and experimental forecast campaign was carried out during the 2015 Pan Am (10–26 July) and ParaPan Am (7–15 August) Games in Toronto. Enhanced monitoring and experimental NWP were in place for all of July, August, and September across the Greater Toronto Area and in southern Ontario, Canada. This is also a time of year when thunderstorms occur frequently in this region of Canada (see Burrows et al. 2002).

a. NWP-based thunderstorm forecasts

Methods by which thunderstorm forecasts and nowcasts are generally produced include the following: 1) thunderstorm advection from radar or lightning observations (Dixon and Wiener 1993; Mueller et al. 2003; Germann and Zawadzki 2004; Bowler et al. 2006; Meyer et al. 2013, etc.), 2) ingredient-based methods derived from thunderstorm environment variables (e.g., Bright et al. 2005), 3) ingredient-based methods derived from cloud-resolving model parameters (McCaul et al. 2009; Yair et al. 2010; Barthe et al. 2010; Lynn et al. 2015) or 4) explicit modeling of electrification (Helsdon et al. 2001; Mansell et al. 2005; Barthe and Pinty 2007). Furthermore, ingredient-based methods can either be based on heuristic rules derived from expert knowledge or empirically learned from the data (e.g., Ukkonen et al. 2017; Simon et al. 2018).

For the verification experiment, we focus on experimental convective products derived from two regional

TABLE 1. Comparison of the characteristics of the different thunderstorm forecasts used. The time step represents the time interval between each available forecast output. For probabilistic/index forecasts, it also represents the temporal window over which probabilities/indices are computed, whereas the deterministic forecast field is considered instantaneous. The sliding calibration window (in grid cells) varies from $45 \text{ km} \times 45 \text{ km}$ for REPS-TI to $75 \text{ km} \times 75 \text{ km}$ for RDPS-BL.

Name	Model	Grid spacing	Time step	Forecast type	Field (units)	Calibration window
REPS-TI	Regional ensemble	15 km	3 h	Probabilistic	Lightning occurrence (%)	3×3
RDPS-IN	Regional deterministic	10 km	1 h	Probabilistic	Lightning occurrence (%)	5×5
RDPS-KF	Regional deterministic	10 km	1 h	Deterministic	Convective rain rate (mm h^{-1})	None
RDPS-BL	Regional deterministic	15 km (upscaled)	3 h	Index	Lightning occurrence (unitless)	5×5

ECCC operational NWP models: the Regional Deterministic Prediction System (RDPS; see Caron and Anselmo 2014) and the Regional Ensemble Prediction System (REPS; see Lavaysse et al. 2013). One advantage of focusing only on ECCC operational NWP models is that they all share a common dynamical and physical core, thus allowing an easy and fair way to compare postprocessing methods.

RDPS-BL is a statistical postprocessing of the (upscaled) 15-km RDPS and was developed by Burrows et al. (2005). It is based on the conditional climatology of lightning over a $75 \text{ km} \times 75 \text{ km}$ window using the “random forest” machine learning technique by Breiman (2001). It is thus a thunderstorm environment ingredient-based method learned from the data.

RDPS-IN is a calibrated postprocessing of the 10-km RDPS developed by Taylor et al. (2014). It uses rules to combine four ingredients important for thunderstorm initiation. These ingredients are as follows: 1) most unstable convective available potential energy (MUCAPE), 2) most unstable convective inhibition (MUCIN), 3) CAPE between most unstable lifted parcel level and 3 km (MULPL–3 km CAPE), and 4) integrated vertical velocity below the most unstable equilibrium level (IVVMUEL). The Cloud Physics Thunder Parameter (CPTP) developed by Bright et al. (2005) is used as a mask ($\text{CPTP} \geq 25$) to exclude regions with low likelihood of thunderstorm electrification. It is calibrated over a $50 \text{ km} \times 50 \text{ km}$ window. It is thus a thunderstorm environment ingredient-based method derived from expert knowledge.

RDPS-KF is the output from the well-known Kain–Fritsch deep moist convection parameterization (Kain 2004) used operationally with the 10-km RDPS. It is not a thunderstorm forecast per se, but it can be used as a proxy since there is usually a strong correlation between convective precipitation and lightning. However, it should be noted that this correlation breaks down in different climates or conditions (e.g., dry thunderstorms, tropical cyclones). We could say that RDPS-KF is an ingredient-based method with a single ingredient. We

transform this forecast into a probabilistic one by converting rain rates into thunderstorm probabilities (details discussed in section 3c).

Finally, REPS-TI (R. Frénette 2019, personal communication) is a calibrated thunderstorm forecast derived from the 15-km REPS ensemble forecast, combining the CPTP (≥ 25) and Kain–Fritsch rain rate [$\geq 2.5 \text{ mm (3 h)}^{-1}$]. It is calibrated on a $45 \text{ km} \times 45 \text{ km}$ neighborhood using the ensemble mean. It is again a thunderstorm environment ingredient-based method.

All the compared forecasts were run at 1200 UTC¹ each day for at least 48 h. Table 1 summarizes some information on the specifications of the different models that will be compared. In Fig. 1, we show the bounding box of the forecasts we want to compare, as well as a customized verification domain. All the forecast domains use a polar stereographic projection, but they differ in the location, orientation and size of their bounding box.

For a uniform weighting between grid cells, we define a verification grid with $2 \text{ km} \times 2 \text{ km}$ grid cells using an equal-area projection. The specific choice we make is to use the Lambert Azimuthal Equal Area projection, but any equal-area projection will give equivalent results. The 1-h forecasts were converted to 3-h forecasts by taking the maximum probability/value over the 3-h time interval.

b. Human-generated forecasts

The two RSDs, located in the Ontario Storm Prediction Centre’s operations area, were operated by four research meteorologists with widely varying experience levels in the area of thunderstorm prediction. During the Games periods, they used the interactive Convective Analysis and Storm Tracking (iCAST) prototype (Sills et al. 2009) and the Aurora workstation (Greaves et al. 2001) to view weather data and create prognoses via drawing and editing of MetObjects. MetObjects as defined at ECCC are object-based graphical representations of

¹ Local time (EDT) is UTC – 4 h.

FIG. 1. Bounding boxes of the domains for different original NWP outputs (REPS-TI in blue, RDPS-BL in green, RDPS-IN and RDPS-KF both in red) and verification domain to which the NWP outputs are reprojected (in pink). Note that the original domain for RDPS-KF and RDPS-IN is exactly the same. The mesh grid represents a group of 40×40 grid cells (each with a nominal resolution of $2 \text{ km} \times 2 \text{ km}$). The equal-area map projection clearly illustrates that, contrary to the domains of the different forecasts, the cells of the verification grid all have the same area, thus avoiding any geographical bias.

meteorological features (e.g., jets, fronts, high and low pressure centers) and threat areas that can include attributes (e.g., intensity, likelihood, type). Available weather data included past weather, current observations (radar data, satellite imagery, and lightning and surface observations enhanced for ECPASS), and various NWP-derived guidance.

To begin defining thunderstorm threat areas via MetObjects, forecasters chose between numbers of different model-generated first-guess fields, used the existing thunderstorm threat area from the previous shift, or started new areas from scratch. Once an initial area was defined, various statistical and NWP guidance products were examined—and conceptual models were employed using an ingredient-based approach (e.g., Doswell et al. 1996)—to help the forecaster make more detailed adjustments. Behind-the-scenes work on refining the first-guess fields continued through the project based on forecaster feedback. While first-guess fields were not often used at the beginning of the demonstration, they were used more as their quality improved.

Thunderstorm forecasts for day 1 were issued in the morning at 3-h intervals (1800, 2100, and 0000 UTC), and for days 2 and 3 were issued in the afternoon at 6-h intervals (0600, 1200, 1800, and 0000 UTC). While the

forecast features (fronts, jets, etc.) were valid at the nominal time, all thunderstorm area forecasts were valid for the following three hours. A four-category probabilistic approach was used for the thunderstorm threat areas, with the qualitative categories “None,” “Chance,” “Likely,” and “Certain.” We will describe how these qualitative categories can be related into equivalent NWP quantitative forecasts in section 3d. Examples of other types of MetObjects used as part of the thunderstorm nowcasts and forecasts are discussed in Joe et al. (2018).

c. Lightning observations

In-cloud (IC) and cloud-to-ground (CG) lightning observations were obtained from the operational North American Lightning Detection Network (NALDN; see Orville et al. 2011). The NALDN detection range covers the United States and most of Canada and the data are available 365 days a year, 24 h a day. The reported accuracy is of the order of 200 m spatially and less than a microsecond temporally. The detection efficiencies for ground and cloud flashes for Ontario and neighboring areas in the United States in 2015 were estimated to be 95% and 40%–50%, respectively (see Nag et al. 2014).

One difficulty for verification is that observations are latitude–longitude points indicating the location of the lightning flash, but forecasts are generally represented on a grid. It is necessary to relax the observations spatially in order to make them comparable to forecasts and to be less stringent on what constitutes a good forecast. Therefore, we take 25 km as the radius of a circular buffer, which we call the radius of relaxation. To match observations with forecasts, lightning flashes were thus binned every three hours on the 2-km verification grid and then expanded using the 25-km buffer.

The circular buffer means that any grid point within 25 km from the lightning observation is verified as an occurrence of thunderstorm. Note that even without any buffering or upscaling each forecast grid intrinsically introduces some kind of relaxation by binning lightning observations. The circular buffer around lightning observations can be interpreted as a component of a neighborhood method as described by Ebert (2008), except that we do not necessarily apply the same buffer to forecast outputs. The choice of 25 km for the radius of relaxation can be considered as an arbitrary human safety buffer zone, meaning that 25 km is considered as a safe distance to be from lightning flashes. The radius of relaxation is a fixed parameter defined before any calibration and validation of the probabilistic thunderstorm forecasts.

3. Postprocessing of forecasts

We separate our dataset into two subsets: a training set (July 2015) and a validation set (August–September 2015), since some parameters of the postprocessing methods will be learned from the training dataset. Moreover, we will also use the training dataset to apply a calibration to the forecasts in order to remove any forecasting bias. By performing these operations on the training set, we avoid the risk of any overfitting and thus ensure that the obtained verification scores are not influenced by the tuning and calibration of the forecasts.

Two standard probabilistic verification measures are used for both the tuning and the intercomparison of the forecasts: the area under the receiver operating characteristic (ROC) curve and the Brier skill score (BSS). The ROC curve compares the trade-off between the probability of detection (POD) and the probability of false detection (POFD) for different forecast probabilities, and is independent of the actual values of the probabilities and thus insensitive to miscalibration. This is why we use the area under the ROC curve (ROC-AUC) as an objective function that can be used before performing any forecast calibration.

FIG. 2. Illustration of the dilation operation for a given radius of r . The original probabilistic forecast can be separated between the region for which the probability of lightning (POL) is greater than 0 (inside the blue area) and the region for which the probability of lightning is zero (outside the blue area). For any radius r greater than 0, the forecast is dilated by assigning a value of $-r$ to all locations at a distance r from the $\text{POL} > 0$ area. This new “pseudo”-forecast will have a more negative value as the geographic distance from the $\text{POL} > 0$ forecast area increases.

The BSS provides a simple measure of probabilistic forecast accuracy. The decomposition of the BSS into calibration and refinement terms is equivalent to the bias-variance decomposition for the mean square error. After calibration of the forecast, the calibration term is expected to be close to zero, so the BSS will essentially measure the difference in refinement (or sharpness) between forecasts.

a. Dilation

Forecasting with zero probability is often incorrect as there could be some small residual probability of thunderstorms. When computing the ROC-AUC, some forecasts can be severely penalized because they have many misses when they predict a zero probability of thunderstorms (or no convective rain in the case of RDPS-KF). A simple way to dramatically improve these forecasts is to create new predictions forecasts by dilation of the nonzero lightning probability forecast area. A dilation (also known as a buffer in geographic information science) inflates a forecast area by a margin of thickness (or radius) r (see Fig. 2). By varying the value of the radius of dilation, several new forecasts can be created and the ROC curve can be completed.

The new forecast at a location, which previously had zero probability, is temporally assigned to a pseudovalue

of $-r$, where r is the minimal dilation necessary to include a given location in the dilated forecast. These pseudovalues can be computed efficiently using the negative of the distance transform (first introduced for image processing by Rosenfeld and Pfaltz 1966) of the nonzero lightning probability forecast area. Note that the negative pseudovalues will be transformed back to probabilities at the forecast calibration step. A more mathematical discussion on the use of the distance transform and its generalization to postprocessed thunderstorm forecast areas can be found in Brunet and Sills (2017).

b. Optimal smoothing

We use a smoothing operation to transform deterministic or probabilistic forecasts into smoother forecasts that take into account the spatial uncertainty of the forecasts. Given a probabilistic forecast at a given location, the smoothing will have the effect of spreading out the possible location of the forecast, thus reflecting the uncertainty on the location of the forecast. Since calibration of the forecasts is expected to bring the calibration term of the Brier score close to zero and since the ROC curve is independent of the calibration of the forecast, we can find the optimal smoothing by either minimizing the refinement term of the Brier score or by maximizing the area under the ROC curve. The latter option is selected in this paper.

We choose an isotropic Gaussian filter as the smoothing filter and perform an exhaustive search for the optimal smoothing bandwidth for each of the forecasts, aggregating all valid times and lead times of the training set (cases from July 2015). The optimal smoothing parameter is found for both raw and dilated forecasts, but before performing the calibration step. The forecasts are then recalibrated on the verification grid so that a fair comparison of the potential of each forecast can be made.

c. Calibration of probabilistic forecasts

Calibration of a probabilistic forecast is done by applying a transfer function to the original forecast values (e.g., dilated pseudoforecasts or raw probabilities) such that the new probabilities match the observed frequencies of lightning. This calibration procedure reduces the calibration term to zero without changing the refinement term, thus giving a better overall Brier score. To ease computation, we quantize the forecast into $N = 100$ bins, where the bins are chosen uniformly (according to the arc length) along the ROC curve. This method of binning ensures a better approximation of the ROC curve than, for example, dividing the bins uniformly by percentage values. The choice of $N = 100$

is empirically observed to be a good trade-off between computational speed and accuracy. We count the relative frequency of lightning observations as defined in section 2c for each of $N = 100$ quantized bins for all valid times and lead times. We then fit a curve that will define the transfer function between forecast values and the (re)-calibrated probability of lightning. To restrict the choice of the fitted curve to a monotonically increasing one, we perform an isotonic least squares regression. We estimate the uncertainty of the transfer function by block-bootstrapping over the July 2015 cases (one block per day), resampling 1000 times. The 2.5 percentile and the 97.5 percentile are used, respectively, as the lower and upper bounds of a 95% confidence interval.

d. Thresholding continuous forecasts

Since the human-generated forecasts have only four qualitative categories (None, Chance, Likely, and Certain), we need to reduce the NWP-based forecasts to similar categories for a fair comparison. Such an evaluation will help determine the difference in quality between the human-generated forecasts and those that are automatically generated using thresholding of postprocessed NWP.

The choice of threshold is guided by the principle that we want to generate comparable forecasts. This principle means that the balance between probability of detection (POD) and probability of false detection (POFD) should be the same for all the thresholded forecasts. This balance is accomplished by drawing a straight line between the perfect forecast ($\text{POD} = 1$, $\text{POFD} = 0$) and the no-skill forecast ($\text{POD} = \text{POFD}$) that passes by the point representing POD and POFD for the human-generated forecast (see Fig. 3). All dichotomous forecasts with (POD, POFD) scores on this line have a constant ratio of $(1 - \text{POD})/\text{POFD}$. Since

$$\frac{1 - \text{POD}}{\text{POFD}} = \frac{sM}{(1 - s)F},$$

where s is the base rate, M is the number of misses, and F is the number of false alarms, the ratio of $(1 - \text{POD})/\text{POFD}$ only depends on the ratio of misses and false alarms given two forecasts verified against the same observations. Keeping $(1 - \text{POD})/\text{POFD}$ constant is thus equivalent to comparing forecasts with the same “bias ratio” F/M . Note that this is a different measure of bias than the frequency bias which is the ratio of forecasted over observed events. However, both types of bias will be one when $M = F$.

We select the threshold value for which the ROC curve intersects the line for each of the categories

FIG. 3. Diagram of the model ROC curve and of the line used to choose a comparable probability threshold with human-generated forecasts. Starting from the P_{perfect} forecast ($\text{POD} = 1, \text{POFD} = 0$), a line is drawn passing by the P_{human} forecast. The slope of this line is $(\text{POD}_{\text{human}} - 1)/\text{POFD}_{\text{human}}$. The intersection point of this line with the ROC curve for a probabilistic forecast is called P_{model} whereas the intersection of this line with the no skill line ($\text{POD} = \text{POFD}$) is called P_{noskill} . The probability threshold is the probability value at the location of P_{model} on the ROC curve. The relative skill of the thunderstorm forecast is defined as the ratio of the line segment $\overline{P_{\text{model}}P_{\text{noskill}}}$ over the line segment $\overline{P_{\text{perfect}}P_{\text{noskill}}}$. After a simple computation, the relative skill simplifies to $\text{POD}_{\text{model}} - \text{POFD}_{\text{model}} = \text{PSS}_{\text{model}}$.

(Chance/Likely/Certain). Since the forecast is better as the ROC curve approaches the upper left corner of the diagram, we compare the relative skill of the forecasts by computing the ratio of the length between the intersection point and the no-skill point over the length between the perfect point and the no-skill point. A geometric argument can show that this ratio of lengths corresponds exactly to the Peirce skill score (PSS).

4. Results: Training

a. Optimal smoothing

The ROC-AUC as a function of the bandwidth of the smoothing parameter for forecasts at all lead times up to 48 h is shown in Fig. 4 both with and without extending the forecast by dilation. For all forecasts, the smoothing operation has a positive effect, with the most marked effect for RDPS-KF, RDPS-IN, and RDPS-BL. Extending the forecast by dilation yields large gains in the ROC-AUC score, except for REPS-TI where it has no effect. However, directly smoothing the forecast for RDPS-KF and RDPS-IN eventually leads to slightly

better ROC-AUC than smoothing the dilated forecast. For all four forecasts, the ROC-AUC score for dilated forecasts without smoothing is similar to the smoothing of forecasts without dilation using a bandwidth of 20 pixels (40 km). Only for RDPS-BL does smoothing the dilated forecast lead to a better ROC-AUC score for the optimal bandwidth.

The optimal smoothing bandwidth ranges from 50 km for REPS-TI to 100 km for RDPS-KF and RDPS-BL (with dilation). Notice how these bandwidths are much larger than the model forecast horizontal grid spacings, which are between 10 and 15 km, and the relaxation parameter for lightning observations, which is 25 km. For each of the eight cases (RDPS-IN, RDPS-BL, RDPS-KF, and REPS-TI, with or without dilation), the ROC-AUC as a function of the smoothing bandwidth is relatively flat around the optimal bandwidth. A change of bandwidth of ± 20 km could thus be allowed without reducing the ROC-AUC by much.

In Fig. 5, the ROC curves aggregated for all lead times up to 48 h are shown for four different types of post-processing: (i) none, (ii) dilation only, (iii) smoothing only, and (iv) dilation then smoothing. Each case of smoothing uses the optimal bandwidth. It can be seen that the dilation of the forecast “completes” the ROC curve for areas of high POD and POFD. Indeed, the raw forecast is interpolated as a straight line between the point on the ROC curve corresponding to the smallest positive value of the forecast and the point $(\text{POD}, \text{POFD}) = (1, 1)$ while the ROC curve for the dilated forecast follows the shape of the remaining lowest part of the ROC curve for the raw (i.e., without dilation) forecast. Smoothing of the forecast also contributes to forecasts with high POD and POFD and thus has a similar effect on the ROC curve. It also has a positive effect in other areas of the ROC curve, pushing the curve to the left (lower POFD for constant POD), thus increasing the ROC-AUC. Finally, combining dilation and smoothing has a similar effect to smoothing only, with a noticeable positive impact for the ROC curve for RDPS-BL. Otherwise, the difference of ROC-AUC for smoothing with or without dilation is less than 0.05 for RDPS-IN and RDPS-KF (and null for REPS-TI since the dilation does not change the forecast). In light of these results, optimally smoothed forecasts will be used in the validation step for REPS-TI, RDPS-IN and RDPS-KF while optimally smoothed forecasts *after dilation* will be used for RDPS-BL.

b. Calibration

In Figs. 6 and 7, we present the different calibrating functions that were found, as well as the frequency of

FIG. 4. Area under the ROC curve for all lead times combined as a function of the bandwidth (km) of the smoothing parameter for (a) REPS-TI, (b) RDPS-KF, (c) RDPS-IN, and (d) RDPS-BL. Orange: with dilation of the forecast. Blue: without dilation of the forecast. Open squares indicate the value of the ROC-AUC when no smoothing is applied. Filled squares indicate the value of the ROC-AUC for the optimal smoothing value.

observations obtained for each forecast value using the July 2015 training dataset for all lead times up to 48 h. As shown by the green line in Fig. 6, only REPS-TI is already calibrated for the relaxed lightning observations. This is explained by the fact that REPS-TI was calibrated for a $45 \text{ km} \times 45 \text{ km}$ lightning forecast binning, which is roughly equivalent in size to a 25-km disk radius. For lower probabilities/values, the sample size is large enough that a good curve fitting can be found for all of the forecasts. As the forecast values increase and the number of events diminishes, it is harder to find a good fit and the width of the confidence intervals becomes larger. Finally, for the two highest values of RDPS-KF in Fig. 6c, the frequency of observations is actually decreasing. Similarly, for

RDPS-BL (Fig. 6d) and in particular for its postprocessed case (Fig. 7d), the frequency of observations peaks right after zero to about 40% before decreasing to 30%.

The calibration functions were applied to both the raw NWP forecasts and the postprocessed NWP forecasts. An example of the resulting calibrated forecasts with or without postprocessing is shown in Fig. 8 for the (typical) case of 1200 UTC 12 July 2015 with lead time of 12 h.

c. Thresholding

Table 2 provides the threshold values for the four raw forecasts and the four postprocessed forecasts along with the corresponding calibrated probability values

FIG. 5. ROC curves aggregated for all lead times for four postprocessing cases—None, dilation only (D), smoothing only (S), dilation, then smoothing (D + S)—and for four numerical models: (a) REPS-TI, (b) RDPS-KF, (c) RDPS-IN, and (d) RDPS-BL. The ROC-AUC is indicated in the legend for each case. Notice how dilation “completes” the ROC curve of the raw forecasts for regions of high POD/POFD, dramatically improving the ROC-AUC.

computed over the whole dataset. For RDPS-KF and RDPS-BL, there is no forecast matching intersecting the “Chance” category line, so the smallest positive forecast is taken instead. For RDPS-BL-PP, a negative forecast value indicates that the dilated

forecast is used. Note that the forecast probabilities are the predicted probability of lightning at a location (within a 25-km buffer) given a forecast value, not the probability of lightning inside the area over the threshold.

FIG. 6. Calibrating functions (orange lines) obtained by isotonic least squares regression of the proportion of lightning observations within a 25 km radius for raw NWP forecasts values (at all lead times up to 48 h) grouped into 100 bins (red points). The 95% confidence intervals are obtained by block-bootstrapping (dashed lines). (a) REPS-TI, (b) RDPS-KF, (c) RDPS-IN, and (d) RDPS-BL.

5. Results: Postprocessing methods

Forecasts are compared against observations for the months of August and September 2015. Results are stratified by lead time (relative to the model run). The 95% confidence intervals and significance tests are computed via block-bootstrapping using available forecasts for each day as a block (up to 61 blocks), resampling 1000 times.

a. Discriminative power

The comparison of the ROC-AUC by lead time for each type of forecast with or without further postprocessing is presented in Fig. 9. For all forecasts and all lead times, postprocessing leads to statistically significant improvement in the discriminative power of the forecasts. The effect of the postprocessing is the most

FIG. 7. Calibrating functions (orange lines) obtained by isotonic least squares regression of the proportion of lightning observations within a 25-km radius for postprocessed NWP forecasts values (at all lead times up to 48 h) grouped into 100 bins (red points). The 95% confidence intervals are obtained by block-bootstrapping (dashed lines). (a) REPS-TI-PP, (b) RDPS-KF-PP, (c) RDPS-IN-PP, and (d) RDPS-BL-PP.

dramatic for RDPS-KF and RDPS-IN, bringing average forecasts with a ROC-AUC score between 0.7 and 0.8 to excellent forecasts with a ROC-AUC score over 0.9. Poor forecasts such as RDPS-BL at 24-h lead time are also significantly improved by a difference in ROC-AUC of more than 0.2. The improvement in forecast quality for RDPS-IN by smoothing is slightly reduced, but is still between 0.05 and 0.1, making a good forecast

excellent. The contribution of smoothing for REPS-TI is more subtle as the forecasts were already excellent, but the smoothing led to a statistically significant improvement of 0.01 for the ROC-AUC at all lead times.

b. Skill

The comparison of the BSS by lead time for each type of forecast with or without further postprocessing is

FIG. 8. Examples of calibrated forecasts (probability of lightning) with (-PP) or without postprocessing for 1200 UTC + 12 h 12 Jul 2015. (a) REPS-TI, (b) RDPS-KF, (c) RDPS-IN, (d) RDPS-BL, (e) REPS-TI-PP, (f) RDPS-KF-PP, (g) RDPS-IN-PP, and (h) RDPS-BL-PP. Recalibration after smoothing increases the maximum attainable calibrated probability of lightning in three (RDPS-KF, RDPS-IN, and RDPS-BL) of the four cases presented.

presented in Fig. 10. For all forecasts, postprocessing leads to statistically significant improvement in the skill of the forecast for most of the lead times. The exceptions are for RDPS-IN with a lead time of 27 h and for RDPS-BL with a lead time between 21 and 27 h, as well as for a lead time of 45 h where there is no statistically significant difference between the BSS with or without postprocessing.

A diurnal cycle is observed for the BSS for all forecast types, with generally lower skill for lead times corresponding to local times between the middle of the night (0200 LT) and late morning (1100 LT) and higher skill for early afternoon (1400 LT) to late evening (2300 LT). The lower skill is believed to be related to the

inherent difficulties in NWP handling of overnight convection and its remains. However, the skill during this period is noticeably worse for the RDPS-IN and RDPS-BL forecasts. These forecasts were designed based on storm environment considerations as opposed to being based on the generation of precipitation by the model. As such, they tend to overforecast in areas where sufficient vertical velocity is unavailable to initiate convection. It appears that this overforecasting problem is worst in the early to late morning. Notice also that the cases for which postprocessing does not improve the forecast correspond to cases for which the original forecast did not have much skill to start with.

TABLE 2. Threshold values for Chance, Likely and Certain categories for raw and postprocessed (-PP) automatically generated forecasts and corresponding calibrated forecast probabilities.

	Chance	Likely	Certain
REPS-TI (%)	8.4 (9%)	28.3 (28%)	51.7 (52%)
RDPS-KF (mm h^{-1})	>0 (2%)	3.2 (25%)	16.0 (39%)
RDPS-IN (%)	17.2 (9%)	53.0 (23%)	82.3 (36%)
RDPS-BL (index)	>0 (2%)	0.2 (28%)	71.5 (29%)
REPS-TI-PP (smoothed %)	9.7 (10%)	25.8 (29%)	44.3 (52%)
RDPS-KF-PP (smoothed mm h^{-1})	0.5 (8%)	2.5 (28%)	8.3 (42%)
RDPS-IN-PP (smoothed %)	16.3 (10%)	36.4 (26%)	57.4 (35%)
RDPS-BL-PP (smoothed and dilated index)	-12.4 (10%)	0.2 (35%)	16.8 (35%)

FIG. 9. Comparison of ROC-AUC vs lead time for forecast with postprocessing (-PP) and without postprocessing. The 95% confidence intervals of the difference of ROC-AUC between each forecast and postprocessed forecast are represented as upward and downward pointing triangles that are centered around the ROC-AUC of the postprocessed forecasts. (a) REPS-TI. (b) RDPS-KF. (c) RDPS-IN. (d) RDPS-BL.

6. Results: Forecast intercomparison

a. Discriminative power

For each forecast (four choices of thunderstorm forecasts and with or without postprocessing), the ROC curves are computed for each lead time (from 6 to 45 h). The ROC-AUC scores are then computed as well as the

difference between the ROC-AUC score for REPS-TI and the ROC-AUC score for each of the other thunderstorm forecasts. A 95% confidence level for the difference in the ROC-AUC scores is added to each forecast to test for significance. Figure 11 summarizes the results.

For the raw forecasts, REPS-TI is clearly the better forecast (significant difference at 95%), followed by

FIG. 10. Comparison of BSS vs lead time for forecast with postprocessing (-PP) and without postprocessing. The 95% confidence intervals of the difference of BSS between each forecast and postprocessed forecast are represented as upward and downward pointing triangles that are centered around the BSS of the postprocessed forecasts. (a) REPS-TI. (b) RDPS-KF. (c) RDPS-IN. (d) RDPS-BL.

RDPS-IN as a distant second. RDPS-KF and RDPS-BL are tied in third place. For all the cases, we can observe a slight decrease in discriminative power as a function of lead time. A diurnal cycle can also be observed with a marked dip for RDPS-BL having a 24-h lead time, corresponding to early morning.

For postprocessed forecasts, the difference is not as marked. REPS-TI-PP is the better forecast for most of

the lead times, but it is statistically equivalent to RDPS-KF-PP between 6 and 27-h lead time and between 39- and 45-h lead time. RDPS-BL-PP does well between 9- and 15-h lead time and between 39- and 42-h lead time with no statistically significant difference with REPS-TI-PP, but it performs significantly worse than any other forecast at 24-h lead time. While RDPS-IN is the second best raw forecast, RDPS-IN-PP is the worst of

FIG. 11. Area under the ROC curve as a function of the lead time for four different NWP-based thunderstorm forecasts. (a) Raw forecasts. (b) Postprocessed forecasts. The 95% confidence intervals of the difference of ROC-AUC between each forecast and REPS-TI in (a) and REPS-TI-PP in (b) are represented as upward and downward pointing triangles that are centered around the ROC-AUC of the other forecasts.

the postprocessed forecasts for most of the lead times and REPS-TI-PP is significantly better than RDPS-IN-PP for all lead times, except for 42-h lead time.

b. Skill of the calibrated forecasts

For each forecast (four choices of NWP-based forecasts with or without postprocessing), the BSS is computed for each lead time (from 6 to 45 h). The skill of the three other forecasts without (with) postprocessing are compared to REPS-TI (REPS-TI-PP) by computing a 95% confidence interval of the difference in BSS. Figure 12 summarizes the results.

REPS-TI is significantly more skillful statistically than the three other deterministic forecasts for all lead times, except for RDPS-IN at 42-h lead time. The three other forecasts have a similar performance, except for between 24- and 27-h lead time when RDPS-KF does not dip in performance as much as RDPS-IN and RDPS-BL. For the postprocessed forecasts, REPS-TI-PP and RDPS-KF-PP are the two leading forecasts with a statistical tie for most of the lead times, except at 18-h lead time when RDPS-KF-PP is significantly better statistically. The BSS of RDPS-IN-PP is not as good as REPS-TI-PP: only from 18- to 21-h lead time as well as between 42- and 45-h lead time is RDPS-IN-PP not significantly worse statistically compared to REPS-TI-PP. Similarly, RDPS-BL-PP is not as skillful as REPS-TI-PP, except between 6- and 12-h lead time, at 18-h lead time and

between 33- and 36-h lead time, when it is a statistical tie. Overall, the relative ranking of the forecasts in term of skill and discriminative power are very similar. This is not a surprising result as all forecasts were recalibrated using the same dataset (July 2015 data).

c. Comparison between postprocessed deterministic forecasts and ensemble forecasts

An interesting observation can be made if we compare the postprocessed deterministic forecast with the REPS-TI ensemble mean forecast. In Fig. 13, we show both the ROC-AUC versus lead time and the BSS versus lead time for the three postprocessed deterministic forecasts (RDPS-IN-PP, RDPS-BL-PP and RDPS-KF-PP), as well as REPS-TI ensemble mean forecast, with confidence intervals.

For several lead times, not only is REPS-TI not significantly better statistically than the other postprocessed forecasts, but it is even significantly worse statistically than RDPS-KF-PP for half of the first 24-h lead time in terms of both skill and discriminative power. Indeed, for the ROC-AUC comparison, RDPS-KF-PP is better than REPS-TI for lead times between 9 and 15 h, whereas it is a statistical tie for the remainder of the lead times. During afternoons and evenings, RDPS-BL-PP is also statistically tied with REPS-TI and is better than REPS-TI with 9-h lead time, but it is doing worse during nights and mornings. RDPS-IN-PP

FIG. 12. BSS as a function of the lead time for four different NWP-based thunderstorm forecasts. (a) Raw forecasts. (b) Postprocessed forecasts. The 95% confidence intervals of the difference of BSS between each forecast and REPS-TI in (a) and REPS-TI-PP in (b) are represented as upward and downward pointing triangles that are centered around the BSS of the other forecasts.

manages a statistical tie with REPS-TI between 9- and 21-h lead time and between 39- and 42-h lead time, but otherwise does worse. This can be compared with results shown in Fig. 11a, where all deterministic forecasts are significantly worse statistically than REPS-TI by a wide margin.

For the BSS, the story is similar. RDPS-KF-PP is significantly better statistically than REPS-TI between 12- and 27-h lead time (except for 21-h lead time which is a statistical tie by a small margin) and for 42-h lead time while it is a statistical tie for the other lead times. The results for RDPS-BL-PP are equivalent to those for

FIG. 13. (a) Area under the ROC curve as a function of the lead time and (b) BSS as a function of the lead time for the three postprocessed probabilistic thunderstorm forecasts (RDPS-IN-PP, RDPS-BL-PP, and RDPS-KF-PP) as well as REPS-TI ensemble mean forecast. The 95% confidence interval of the difference of ROC-AUC between each postprocessed forecast and REPS-TI are represented as upward and downward pointing triangles that are centered around the ROC-AUC of the other forecasts.

FIG. 14. Comparison of automatically generated forecasts with human-generated categorical forecasts. (a) PSS for Chance category as a function of lead time with confidence intervals of the differences between scores from automated forecasts and human-generated forecasts obtained by bootstrapping. (b) ROC curve and PSS scores for each category for all lead times. The PSS scores for Chance/Likely/Certain categories for each forecast are indicated in the legend.

ROC-AUC with a statistical tie on afternoons and evenings and significantly worse statistically on nights and mornings. RDPS-IN-PP also has a few cases with a statistical tie with REPS-TI (9–21-h lead time as well as 36–45-h lead time). Once again, this should be compared with Fig. 12b for which only RDPS-IN between 42 and 45 h does not do significantly worse than REPS-TI.

7. Evaluation of categorical probabilistic forecasts

a. Peirce skill score for comparable categories

The postprocessed forecasts are thresholded into monotonic categorical forecasts using the thresholding method described in section 3d. We call these postprocessed and thresholded forecasts for the Chance/Likely/Certain categories “automatically generated forecasts” in contrast to the human-generated forecasts they emulate. The PSS is then computed for these automatically generated and human-generated forecasts for each lead time as well as 95% confidence intervals for the difference of PSS between automatically and human-generated forecasts. A score difference greater than 0 for more than 97.5% or less than 2.5% of the forecasts is deemed to be statistically significant.

Figure 14a shows the PSS as a function of lead time for automatically generated forecasts obtained by

smoothing and thresholding as well as for human-generated forecasts, whereas Fig. 14b aggregates the ROC curve when the human forecast is available for all lead times combined.

For the Chance category, there is no lead time for which the human-generated forecasts perform significantly better statistically than any other automatically generated forecasts obtained by thresholding postprocessed forecasts. Actually, for all lead time beyond 6-h lead time, the automatically generated forecasts REPS-TI-PP and RDPS-KF-PP are significantly better statistically than the human-generated forecasts. RDPS-BL-PP is also significantly better statistically for lead times of 9, 12, 36, and 42 h, whereas RDPS-IN-PP is significantly better statistically for lead times of 18 and 42 h. For 42-h lead time, all automatically generated forecasts obtained from thresholding postprocessed forecasts perform significantly better statistically than the human-generated forecasts.

The aggregated PSS scores tell a similar story. The automatically generated forecasts obtained from thresholding postprocessed forecasts are better than the human-generated forecasts for the three categories (Chance, Likely and Certain). In case of the Certain category, the PSS score is about twice as good for RDPS-IN-PP, RDPS-KF-PP, and REPS-TI-PP. However, significance intervals cannot be computed for aggregated

TABLE 3. Difference of PSS between human-generated forecasts and automatically generated forecasts for best (maximum), typical (median) and worst (minimum) cases for each tercile (T1: easy, T2: average, T3: difficult) of PSS for human-generated “Chance” forecasts. Positive numbers indicate that the human-generated forecasts are relatively better than automatically generated forecasts.

	T1 (easy)			T2 (average)			T3 (difficult)		
	Worst	Typical	Best	Worst	Typical	Best	Worst	Typical	Best
REPS-TI	−0.15	0.03	0.99	−0.31	−0.05	0.74	−0.99	−0.09	0.42
RDPS-IN	−0.30	0.04	0.99	−0.43	0.06	0.51	−1.05	−0.07	0.50
RDPS-KF	−0.21	0.05	0.99	−0.43	−0.04	0.74	−0.97	−0.07	0.44
RDPS-BL	−0.32	0.04	1.12	−0.34	0.00	0.65	−0.40	−0.03	0.48

scores due to the lack of independence between forecasts of successive lead times.

b. Subjective evaluation of forecasts

Because of the inherent limitation of objective scores to fully reproduce what is the perceived (subjective) skill of each forecast, a visual comparison of different forecasts is also necessary. However, a full subjective comparison of each forecast for all days and all lead times is time consuming and often impractical. Usually, a few selected (possibly cherry-picked) examples are shown to illustrate the typical behavior of each forecast.

We propose automating this selection, thus avoiding any selection bias by using the objective scores to provide guidance on which forecasts are typical, best, and worst relative to each other. Specifically, we divide all cases into three groups depending on the PSS for the human-generated forecasts. These terciles can be considered to represent easy (T1), average (T2), and difficult (T3) forecasts for human forecasters. For each tercile, we find the case for which the human-generated forecasts did best, worst, or had a typical skill relative to the other forecasts with respect to the Chance category. The case selection algorithm has two steps. First, for each tercile and each automatically generated forecast, we compute the minimum (worst case), median (typical case), and maximum (best case) difference of PSS between the human-generated forecasts and automatically generated forecasts. The results are summarized in Table 3. Second, we select a specific case of best, worst, and typical forecasts for each tercile by searching for the forecast time and lead time which yields the smallest mean squared error between the difference of PSS between human-generated forecasts and automatically generated forecasts at a specific time and lead time and the difference of PSS for the best, typical, and worst case for each tercile obtained in the first step (i.e., each column of Table 3). Figures 15–17 present the corresponding cases for the subjective comparison with automatically generated threat areas from thresholding postprocessed forecasts.

For the cases in Fig. 15 (best case), the human-generated forecasts perform best relative to postprocessed forecasts when there is little thunderstorm activity. For 10 August, three of the four automated forecasts failed to predict any thunderstorm activity on the boundary between Ontario and Quebec provinces, whereas the human-generated forecast had a perfect probability of detection for this area. Likewise, three of the four automated forecasts failed to predict thunderstorm activity west of Lake Superior on 22 July. Only RDPS-BL forecasted some thunderstorms, but it was penalized for missing the exact location. The 8 August case illustrates a good example of blending, with the human forecaster integrating well the information provided by the different forecasts. RDPS-KF performed better, however, since the number of false alarms is smaller. The thunderstorm activity over Lake Superior is missed by all forecasts, including the human-generated forecast.

Figure 16 (worst case) shows examples of automatically generated probabilistic threat areas that are very similar in smoothness than the human-generated areas, but are sharper and more skillful. While the human-generated forecasts have a tendency to either underforecast (as for the 20 July case) or to overforecast (as for the 16 July case), the automatically generated forecasts are more consistent. Note that this might reflect different forecasting tendencies between multiple human forecasters.

Figure 17 (typical case) provides a summary of the relative performance of the human-generated forecasts against automatically generated forecasts. The 25 July case represents a more difficult case with widespread thunderstorm activity in northern Ontario, western Quebec and east of Lake Erie. The human-generated forecast provides a good approximation of the location of thunderstorms, but misses several occurrences. The automated forecasts better cover the majority of thunderstorm activity but do not give a precise location for their occurrence. For the 13 August case, the human-generated forecast has less misses than the other forecasts, but it is accompanied with a higher

Best case for T3 (difficult) forecasts: 2015-08-08 12:00 UTC +24H

Best case for T2 (average) forecasts: 2015-08-10 12:00 UTC +24H

Best case for T1 (easy) forecasts: 2015-07-22 12:00 UTC +18H

FIG. 15. Cases for which the human-generated forecasts (HF) do best for Chance category relative to automatically generated forecasts obtained by thresholding postprocessed forecasts. Yellow: Chance category, orange: Likely category, magenta: Certain category, and gray: relaxed lightning observations. Each row represents a tercile of the performance of the human forecaster. The PSS for Chance category is indicated under each forecast.

number of false alarms. Finally, all forecasts performed very well for the 9 July case; RDPS-IN and REPS-TI have a better POD at the cost of a higher POFD.

8. Discussion

a. Discussion on continuous probabilistic forecasts

The extension of the forecast by dilation has a dramatic effect on the ROC-AUC as it is able to fill the “missing” part of the ROC curve. When enough smoothing is applied, the effect is similar since the forecast eventually spreads out. When the authors first noticed that smoothing improves the ROC-AUC by “filling in” the ROC curve, it was hypothesized that

combining dilation with smoothing could further improve the forecast by allowing the dilation to cover the high POD/POFD areas of the forecast while not unnecessarily over smoothing other parts. However, results from this study indicate that smoothing alone is generally better. For RDPS-BL, smoothing combined with dilation improves the forecast, but the optimal smoothing bandwidth is not reduced compared to smoothing only. One possible explanation could be that the smoothing bandwidth necessary to cover more high POD/POFD cases is the same as the one that optimally improves other parts of the forecast.

The optimal smoothing bandwidth itself is surprisingly large. While the horizontal grid spacing of the

Worst case for T3 (difficult) forecasts: 2015-07-20 12:00 UTC +24H

Worst case for T2 (average) forecasts: 2015-07-16 12:00 UTC +9H

Worst case for T1 (easy) forecasts: 2015-07-13 12:00 UTC +36H

FIG. 16. Cases for which the human-generated forecasts (HF) do worst for Chance category relative to automatically generated forecasts obtained by thresholding postprocessed forecasts. Yellow: Chance category, orange: Likely category, magenta: Certain category, and gray: relaxed lightning observations. Each row represents a tercile of the performance of the human forecaster. The PSS for Chance category is indicated under each forecast.

forecasts studied is between 10 and 15 km, the optimal smoothing bandwidth ranges from 50 (for REPS-TI) to 100 km (for RDPS-BL). One apparent drawback of smoothing could be the loss of sharpness of the forecasts. While smoothed forecasts are less precise, they are more skillful and their recalibration can in fact lead to sharper probabilistic forecasts as seen by comparing Fig. 6 with Fig. 7. The fact that a smoothed and calibrated deterministic forecast (RDPS-KF) does on average better than a calibrated ensemble forecast (REPS-TI) and is statistically equivalent to a smoothed and calibrated ensemble forecast has the obvious consequence that one could obtain the same quality of forecast for a fraction of the computational cost. Beyond

this practical consideration, this result could indicate that a better way to combine ensemble members is needed instead of simply taking the calibrated ensemble mean forecast. See for example Ben Bouallègue and Theis (2014) for a possible approach. The current work could be extended to ensemble forecasts by finding the optimal smoothing and calibrating each ensemble member before computing an aggregated (e.g., average) probability.

b. Discussion on categorical probabilistic forecasts

The results on the performance of the human forecaster relative to automated forecasts based on numerical predictions were somewhat surprising for the

Typical case for T3 (difficult) forecasts: 2015-08-13 12:00 UTC +30H

Typical case for T2 (average) forecasts: 2015-07-12 12:00 UTC +12H

Typical case for T1 (easy) forecasts: 2015-08-08 12:00 UTC +9H

FIG. 17. Typical cases for human-generated forecasts with typical (median) performance for Chance category relative to automatically generated forecasts obtained by thresholding postprocessed forecasts. Yellow: Chance category, orange: Likely category, magenta: Certain category, and gray: relaxed lightning observations. Each row represents a tercile of the performance of the human forecaster. The PSS for Chance category is indicated under each forecast.

authors based on a preliminary verification experiment by Sills et al. (2012) where the human-generated forecasts performed better than the other forecasts evaluated for a lead time of 6 h. However, differences with the current experiment are several: 1) introduction of REPS-TI forecasts (although in 2011 the forecasters already had access to calibrated U.S. Short-Range Ensemble Forecast guidance developed by Du et al. 2015), 2) different forecasters with widely varying levels of experience, 3) introduction of new postprocessing methods, 4) evaluation at lead times up to 48 h, and 5) revision of every step of the verification methodology. This result echoes Novak et al. (2014) for the verification of postprocessed and calibrated rain and

temperature forecasts. However, caution is advised in interpreting the relative performance of the human forecasters from the ECPASS demonstration since there are several limitations in the experimental design, including the fact that we did not reproduce the setting of an operational forecast office.

The 6-h lead time forecast is calculated from the run time, but for a fair comparison the lead time should be computed from the time the forecast was generated. For the NWP models presented in this paper, a delay of up to ten hours has to be accounted for (up to 4 h depending on the computational schedule, plus wait time between runs every 6 h). It could thus be argued that forecasts initiated 6 and 12 h earlier should also be

compared. In contrast, the human forecasters could generate forecasts almost continuously and have the advantage of access to the latest observations. This advantage is reduced with the introduction of rapid update cycle models (see e.g., Benjamin et al. 2016) and could be reduced to nothing with the introduction of a near-real-time statistical postprocessing scheme.

Categorical probabilistic forecasts can either have quantitative probabilities derived from the quantization of continuous probabilities or qualitative probabilities described by a human forecaster. The passage from qualitative to quantitative categories can either be done by defining the meaning of each category a priori (e.g., “Chance” is a probability between 20% and 50% of lightning within a 25-km radius and in a 3-h window) or by computing the probabilities a posteriori by counting the frequency of occurrence for each category. Note also that probability of occurrence is always with respect to an area and a time interval, and these parameters should be defined explicitly. Originally, the probabilities were defined a priori, but without specifying the area to which these probabilities were associated. Once the 25-km radius of relaxation is defined, it could have been possible to recalibrate the human-generated forecasts so that they match quantitative limits such as 20%–50%–80% for the four categories, but it would involve interpolating between the geometric areas. The method used in the paper to choose a threshold for probabilistic forecasts based on the ROC curve is equivalent to the second method of quantifying the probabilities of the categorical forecasts even though we never explicitly compute these probabilities.

c. Discussion on the verification method

Modern spatial verification techniques have recently been introduced for high-resolution deterministic forecasts (see Gilleland et al. 2009), but we argue that point-to-point comparisons are totally appropriate for continuous probabilistic forecasts. Indeed, in contrast to deterministic forecasts, a good probabilistic forecast should not have the double penalty problem, otherwise it would mean that the probabilistic forecast was too sharp. A certain level of smoothness is a desirable feature to reflect an appropriate degree of spatial uncertainty.

The comparison between continuous and categorical forecasts provides a particular challenge. Two different interpretations of categorical (discrete) forecasts lead to different approaches in interpolating the ROC curve. A hypothetical forecast system could be interpreted as having a smooth ROC curve, but because of operational limitations (e.g., number of ensemble members or, as in this study, number of categorical forecasts), only

discrete points are measured. The goal would then be to estimate the underlying smooth ROC curve, using for example a normal deviate space assumption (Mason 1982; Wilson 2000). Another interpretation is that as a forecast user only has the discrete points available, the ROC curve should be interpolated with straight lines between these points to reflect the limited amount of information available. We find neither option palatable. While the former interpretation would allow estimation of the potential performance of a human forecaster if they would perform a continuous forecast, this is not a realistic setting. On the other hand, the latter interpretation penalizes the human-generated forecasts even if each of the individual categorical forecasts would perform better than a comparable categorical forecast derived from NWP. By reducing the continuous forecast to a categorical one, we are not assessing if the human-generated forecast is better or worse than the overall continuous forecast, but rather measuring how the automatically generated categorical forecasts fare for a similar balance of POD and POFD.

The choice of the intersection of the line $(1 - \text{POD})/\text{POFD}$ as the thresholding criterion as proposed in this study is not the only possibility. For example, Swets (1996), chapter 5, proposes to use a curve of constant ROC slope, whereas Drummond and Holte (2006) transform ROC curves into cost curves, which allows for interpolation of the ROC curve at points with the same probability cost. We prefer our method because of its simplicity and its relationship to the PSS. Manzato (2007) also found relationships between the PSS and the ROC curves. He proposed the point at the maximum distance from the no-skill bisector line $\text{POD} = \text{POFD}$ as an optimal threshold maximizing the PSS. The line perpendicular to the no-skill bisector could then be seen as another thresholding criterion. The comparison between these different methods of selecting a comparable categorical forecast is out of scope for this paper, but the effect of choosing one method over the other should have a negligible influence on the verification results.

9. Conclusions

We evaluated four different probabilistic thunderstorm forecasts in Ontario, Canada, and neighboring regions. Simple postprocessing methods such as smoothing, dilation and calibration by least squares isotonic regression were introduced. Categorical probabilistic forecasts were generated automatically as first-guess thunderstorm threat areas with $1 - \text{POD}$ over POFD ratio similar to threat areas generated by human forecasters.

Of the four probabilistic thunderstorm forecasts, the ensemble forecast (REPS-TI) performs best for all lead times between 6 and 48 h. However, forecasts that are postprocessed via smoothing are even better, raising the possibility that the ensemble mean may not be the best use of ensemble forecasts. Skillful thunderstorm threat area MetObjects can be generated by thresholding postprocessed forecasts, but the raw NWP outputs should not be directly thresholded as they do not lead to high quality threat areas.

Since the ECPASS demonstration in 2015, there have been some new developments on Canadian NWP forecasts. The 2.5-km HRDPS has since become operational, the Town Energy Balance urban canopy model has been added, and the Milbrandt and Yau double-moment microphysics parameterization has been replaced by the P3 scheme (J. Milbrandt 2019, personal communication). Starting from summer 2019 the REPS-TI code and the calibration will no longer be used operationally. A simpler approach using only the KF convective rain rate is being implemented. The 10-km REPS has also been updated with a new boundary layer formulation (R. Frénette 2019, personal communication). While these improvements are expected to yield to better verification scores for raw NWP models, the results presented in this paper remain valid and relevant. However, the specific optimal smoothing bandwidth and the calibration function would need to be updated for these new forecasts.

The performance of the human-generated forecasts over the experimental period was not as good as postprocessed and calibrated deterministic or ensembles forecasts in term of BSS and ROC-AUC, but there are several caveats and those prevent us from arriving at a more broadly applicable conclusion. Nevertheless, subjective evaluation of the forecasts reveals that human forecasters generally performed better than any of the automated forecasts in marginal cases (low thunderstorm activity). Methods developed here for the comparison of continuous forecasts with categorical human-generated forecasts will allow future and better experiments to more accurately gauge the relative contributions of human forecasters and automation. Ultimately, these verification methods could guide the design of an evolving optimal man-machine mix.

Acknowledgments. The authors thank Anna-Belle Filion, Helen Yang, and Neil Taylor for their efforts as RSD forecasters for the ECPASS demonstration. We would also like to acknowledge the valuable technical support of Emma Hung, for helping the Research Support Desk to keep running as well as for helping to program an early version of the verification routines.

We are also grateful to Ronald Frénette, William Burrows, and Neil Taylor for providing numerical forecast data. Finally, we thank Harold Brooks for an interesting discussion on the interpretation of the ROC curve, which inspired the discussion on the verification method in this paper. Between 2013 and 2016, D. Brunet was funded by a NSERC Visiting Fellowship in a Canadian government laboratory.

REFERENCES

- Adams-Selin, R. D., A. J. Clark, C. J. Melick, S. R. Dembek, I. L. Jirak, and C. L. Ziegler, 2019: Evolution of WRF-HAILCAST during the 2014–16 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34**, 61–79, <https://doi.org/10.1175/WAF-D-18-0024.1>.
- Barthe, C., and J.-P. Pinty, 2007: Simulation of a supercellular storm using a three-dimensional mesoscale model with an explicit lightning flash scheme. *J. Geophys. Res.*, **112**, D06210, <https://doi.org/10.1029/2006JD007484>.
- , W. Deierling, and M. C. Barth, 2010: Estimation of total lightning from various storm parameters: A cloud-resolving model study. *J. Geophys. Res.*, **115**, D24202, <https://doi.org/10.1029/2010JD014405>.
- Ben Bouallègue, Z., and S. E. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: From verification results to probabilistic products. *Meteor. Appl.*, **21**, 922–929, <https://doi.org/10.1002/met.1435>.
- Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, <https://doi.org/10.1175/MWR-D-15-0242.1>.
- Bowler, N. E., C. E. Pierce, and A. W. Seed, 2006: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quart. J. Roy. Meteor. Soc.*, **132**, 2127–2155, <https://doi.org/10.1256/qj.04.100>.
- Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, <https://doi.org/10.1023/A:1010933404324>.
- Bright, D., R. Jewell, M. Wandishin, and S. Weiss, 2005: A physically-based parameter for lightning prediction and its calibration in ensemble forecasts. Preprints, *Conf. on Meteorological Applications of Lightning Data*, San Diego, CA, Amer. Meteor. Soc., 4.3, https://ams.confex.com/ams/Annual2005/techprogram/paper_84173.htm.
- Brunet, D., and D. Sills, 2015: An implicit contour morphing framework applied to computer-aided severe weather forecasting. *IEEE Signal Process. Lett.*, **22**, 1936–1939, <https://doi.org/10.1109/LSP.2015.2447279>.
- , and —, 2017: A generalized distance transform: Theory and applications to weather analysis and forecasting. *IEEE Trans. Geosci. Remote Sens.*, **55**, 1752–1764, <https://doi.org/10.1109/TGRS.2016.2632042>.
- Burrows, W. R., P. King, P. J. Lewis, B. Kochtubajda, B. Snyder, and V. Turcotte, 2002: Lightning occurrence patterns over Canada and adjacent United states from lightning detection network observations. *Atmos.–Ocean*, **40**, 59–80, <https://doi.org/10.3137/ao.400104>.
- , C. Price, and L. Wilson, 2005: Warm season lightning probability prediction for Canada and the northern United States. *Wea. Forecasting*, **20**, 971–988, <https://doi.org/10.1175/WAF895.1>.

- Caron, J.-F., and D. Anselmo, 2014: Regional deterministic prediction system (RDPS). Tech. Rep., Environment and Climate Change Canada, 40 pp., http://collaboration.cmc.ec.gc.ca/cmc/CMOI/product_guide/docs/lib/technote_rdp400_20141118_e.pdf.
- Casati, B., and L. J. Wilson, 2007: A new spatial-scale decomposition of the Brier score: Application to the verification of lightning probability forecasts. *Mon. Wea. Rev.*, **135**, 3052–3069, <https://doi.org/10.1175/MWR3442.1>.
- Dance, S., E. Ebert, and D. Scurrah, 2010: Thunderstorm strike probability nowcasting. *J. Atmos. Oceanic Technol.*, **27**, 79–93, <https://doi.org/10.1175/2009JTECHA1279.1>.
- Dixon, M., and G. Wiener, 1993: TITAN: Thunderstorm identification, tracking, analysis, and nowcasting: A radar-based methodology. *J. Atmos. Oceanic Technol.*, **10**, 785–797, [https://doi.org/10.1175/1520-0426\(1993\)010<0785:TTITAA>2.0.CO;2](https://doi.org/10.1175/1520-0426(1993)010<0785:TTITAA>2.0.CO;2).
- Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting*, **11**, 560–581, [https://doi.org/10.1175/1520-0434\(1996\)011<0560:FFFAIB>2.0.CO;2](https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2).
- Drummond, C., and R. C. Holte, 2006: Cost curves: An improved method for visualizing classifier performance. *Mach. Learn.*, **65**, 95–130, <https://doi.org/10.1007/s10994-006-8199-5>.
- Du, J., G. DiMego, D. Jovic, B. Ferrier, B. Yang, and B. Zhou, 2015: Short range ensemble forecast (SREF) system at NCEP: Recent development and future transition. *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 2A.5, <https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273421.html>.
- Ebert, E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, <https://doi.org/10.1002/met.25>.
- Germann, U., and I. Zawadzki, 2004: Scale dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. *J. Appl. Meteor.*, **43**, 74–89, [https://doi.org/10.1175/1520-0450\(2004\)043<0074:SDOTPO>2.0.CO;2](https://doi.org/10.1175/1520-0450(2004)043<0074:SDOTPO>2.0.CO;2).
- Gilleland, E., D. Ahijevych, B. Brown, B. Casati, and E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, <https://doi.org/10.1175/2009WAF2222269.1>.
- Greaves, B., R. Trafford, N. Driedger, R. Paterson, D. Sills, D. Hudak, and N. Donaldson, 2001: The AURORA nowcasting platform—Extending the concept of a modifiable database for short range forecasting. Preprints, *17th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Albuquerque, NM, Amer. Meteor. Soc., 236–239.
- Helsdon, J. H., Jr., W. A. Wojcik, and R. D. Farley, 2001: An examination of thunderstorm-charging mechanisms using a two-dimensional storm electrification model. *J. Geophys. Res.*, **106**, 1165–1192, <https://doi.org/10.1029/2000JD900532>.
- Joe, P., and Coauthors, 2018: The Environment Canada Pan and Parapan American Science showcase project. *Bull. Amer. Meteor. Soc.*, **99**, 921–953, <https://doi.org/10.1175/BAMS-D-16-0162.1>.
- Kain, J. S., 2004: The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.*, **43**, 170–181, [https://doi.org/10.1175/1520-0450\(2004\)043<0170:TKCPAU>2.0.CO;2](https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2).
- Karstens, C. D., and Coauthors, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30**, 1551–1570, <https://doi.org/10.1175/WAF-D-14-00163.1>.
- , and Coauthors, 2018: Development of a human-machine mix for forecasting severe convective events. *Wea. Forecasting*, **33**, 715–737, <https://doi.org/10.1175/WAF-D-17-0188.1>.
- Lavaysse, C., M. Carrera, S. Bélair, N. Gagnon, R. Frénette, M. Charron, and M. K. Yau, 2013: Impact of surface parameter uncertainties within the Canadian Regional Ensemble Prediction System. *Mon. Wea. Rev.*, **141**, 1506–1526, <https://doi.org/10.1175/MWR-D-11-00354.1>.
- Lynn, B. H., G. Kelman, and G. Ellrod, 2015: An evaluation of the efficacy of using observed lightning to improve convective lightning forecasts. *Wea. Forecasting*, **30**, 405–423, <https://doi.org/10.1175/WAF-D-13-00028.1>.
- Mansell, E. R., D. R. MacGorman, C. L. Ziegler, and J. M. Straka, 2005: Charge structure and lightning sensitivity in a simulated multicell thunderstorm. *J. Geophys. Res.*, **110**, D12101, <https://doi.org/10.1029/2004JD005287>.
- Manzato, A., 2007: A note on the maximum Peirce skill score. *Wea. Forecasting*, **22**, 1148–1154, <https://doi.org/10.1175/WAF1041.1>.
- Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30** (4), 291–303.
- 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>.
- Meyer, V., H. Höller, and H. Betz, 2013: Automated thunderstorm tracking: Utilization of three-dimensional lightning and radar data. *Atmos. Chem. Phys.*, **13**, 5137–5150, <https://doi.org/10.5194/acp-13-5137-2013>.
- Mills, G. A., 2004: Verification of operational cool-season tornado threat-area forecasts from mesoscale NWP and a probabilistic forecast product. *Aust. Meteor. Mag.*, **53** (4), 269–277.
- Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, <https://doi.org/10.1175/2009WAF2222260.1>.
- Mueller, C., T. Saxen, R. Roberts, J. Wilson, T. Betancourt, S. Dettling, N. Oien, and J. Yee, 2003: NCAR auto-nowcast system. *Wea. Forecasting*, **18**, 545–561, [https://doi.org/10.1175/1520-0434\(2003\)018<0545:NAS>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<0545:NAS>2.0.CO;2).
- Nag, A., M. Murphy, K. Cummins, A. Pifer, and J. Cramer, 2014: Recent evolution of the U.S. National Lightning Detection Network. *Proc. 23rd Int. Lightning Detection Conf./Fifth Int. Lightning Meteorology Conf.*, Tucson, AZ, Vaisala, Paper 1-6.
- 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**, 489–504, <https://doi.org/10.1175/WAF-D-13-00066.1>.
- Orville, R., G. Huffines, W. Burrows, and K. Cummins, 2011: The North American Lightning Detection Network (NALDN) analysis of flash data: 2001–09. *Mon. Wea. Rev.*, **139**, 1305–1322, <https://doi.org/10.1175/2010MWR3452.1>.
- Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, <https://doi.org/10.1175/2007MWR2123.1>.
- Rosenfeld, A., and J. L. Pfaltz, 1966: Sequential operations in digital picture processing. *J. Assoc. Comput. Mach.*, **13**, 471–494, <https://doi.org/10.1145/321356.321357>.
- Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations.

- Mon. Wea. Rev.*, **145**, 3397–3418, <https://doi.org/10.1175/MWR-D-16-0400.1>.
- Sharpe, M. A., 2016: A flexible approach to the objective verification of warnings. *Meteor. Appl.*, **23**, 65–75, <https://doi.org/10.1002/met.1530>.
- Sills, D., and N. Taylor, 2008: The Research Support Desk (RSD) initiative at Environment Canada: Linking severe weather researchers and forecasters in a real-time operational setting. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 9A.1, https://ams.confex.com/ams/24SLS/techprogram/paper_142033.htm.
- , N. Driedger, B. Greaves, E. Hung, and R. Paterson, 2009: iCAST: A severe storm nowcasting prototype focused on optimization of the human-machine mix. *25th Conf. on Severe Local Storms*, Denver, CO, Amer. Meteor. Soc., 2.9, https://ams.confex.com/ams/25SLS/techprogram/paper_175963.htm.
- , —, and W. Burrows, 2012: Verification of forecaster-generated iCAST thunderstorm nowcasts and comparison to automated thunderstorm forecasts: Preliminary results. *Proc. Third World Weather Research Programme Symposium on Nowcasting and Very Short Range Forecasting*, Rio de Janeiro, Brazil, WMO, 11.4, yorku.ca/pat/research/dsills/papers/WSN12/WSN12_iCAST_Verif_FINAL.pdf.
- Simon, T., P. Fabsic, G. J. Mayr, N. Umlauf, and A. Zeileis, 2018: Probabilistic forecasting of thunderstorms in the Eastern Alps. *Mon. Wea. Rev.*, **146**, 2999–3009, <https://doi.org/10.1175/MWR-D-17-0366.1>.
- Swets, J. A., 1996: *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Psychology Press, 324 pp.
- Taylor, N., W. Burrows, and D. Sills, 2014: Post-processing of Canadian regional-scale NWP to develop first-guess forecasts of thunderstorm and severe weather threat areas. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 7, <https://ams.confex.com/ams/27SLS/webprogram/Paper254376.html>.
- Theis, S., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, <https://doi.org/10.1017/S1350482705001763>.
- Ukkonen, P., A. Manzato, and A. Mäkelä, 2017: Evaluation of thunderstorm predictors for Finland using reanalyses and neural networks. *J. Appl. Meteor. Climatol.*, **56**, 2335–2352, <https://doi.org/10.1175/JAMC-D-16-0361.1>.
- Wilson, L. J., 2000: Comments on “Probabilistic predictions of precipitation using the ECMWF ensemble prediction system.” *Wea. Forecasting*, **15**, 361–364, [https://doi.org/10.1175/1520-0434\(2000\)015<0361:COPPOP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0361:COPPOP>2.0.CO;2).
- Yair, Y., B. Lynn, C. Price, V. Kotroni, K. Lagouvardos, E. Morin, A. Mugnai, and M. C. 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. Atmos.*, **115**, D04205, <https://doi.org/10.1029/2008JD010868>.