Evaluation of Probabilistic Snow Forecasts for Winter Weather Operations at Intermountain West Airports

Dana M. Uden,[a,b] Matthew S. Wandishin,[b] Paul Schlatter,[c] and Michael Kraus[b]

[a] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*
[b] *NOAA/Global Systems Laboratory, Boulder, Colorado*
[c] *NOAA/NWS Boulder Weather Forecast Office, Boulder, Colorado*

(Manuscript received 3 October 2022, in final form 30 March 2023, accepted 9 May 2023)

ABSTRACT: This work set out to assess the performance of four forecast systems [the Short-Range Ensemble Forecast (SREF), High-Resolution Rapid Refresh Ensemble (HRRRE), the National Blend of Models (NBM), and the Probabilistic Snow Accumulation product (PSA) from the National Weather Service (NWS) Boulder, Colorado, Weather Forecast Office] when predicting snowfall events around the Intermountain West to advise winter weather decision-making processes at Denver International Airport. The goal was to provide airport personnel and the Boulder NWS Forecast Office with operationally relevant verification results on the timing and severity of these events so they are able to make better-informed decisions to minimize negative impacts of storms. Forecasts of snow events using various probability thresholds and a climatological snow-to-liquid ratio of 15:1 were evaluated against Meteorological Aerodrome Reports (METARs) for 24-h periods following four decision-making times spaced equally throughout the day. For the ensembles, a frequentist approach was used: the forecast probability equaled the percentage of ensemble members that predicted a snow event. The results show that the NBM had the best timing of snow events out of the products, while all the products tended to overforecast snow amount. Additionally, NBM had fewer snow events and rarely had high probabilities of snow, unlike the other forecast products.

KEYWORDS: Snowfall; Ensembles; Forecast verification/skill; Model evaluation/performance; Decision making

1. Introduction

Many industries depend heavily on weather forecasts to either increase revenue or prevent loss (e.g., tourism, agriculture, transportation, energy). The aviation sector, both from the airline and airport operations perspectives, requires accurate and timely forecasts to ensure customer and employee safety, airspace efficiency, and reduced costs (Morss et al. 2022). As forecast systems evolve, aviation decision-makers are presented with more and more options upon which to base operational decisions. Probabilistic weather forecasts are one possibly advantageous way to provide aviation decision-makers with the information they need. The Forecasting a Continuum of Environmental Threats (FACETs) (Rothfusz et al. 2018) is a prime example of probabilistic information being employed for a variety of hazards.

Aviation decision-makers must not only be able to interpret probabilistic forecasts, but understand how well they perform. However, these decision-makers do not necessarily use the forecast in the same way as another user and thus the verification methodology needs to reflect how the forecast product is actually used from day to day and what decisions are made based on the information. Murphy and Epstein (1967) refer to this type of evaluation as an "operational evaluation" where the focus is on the stakeholder and the verification approach varies with each decision-maker. Our study was crafted through this lens to determine the performance of several probabilistic forecast products for snowfall events at Denver International Airport (KDEN), with the airport operations and maintenance teams being the primary stakeholders. Extensive background research into the use case of the airport was conducted by Morss et al. (2022) and is summarized in section 2.

Evaluations of probabilistic snow forecast products have been performed (e.g., Stauffer et al. 2018; Scheuerer and Hamill 2019) as have verification studies of forecasts in the context of aviation operations (e.g., Mahringer 2008; Rudack and Ghirardelli 2010; Kim et al. 2011; Lee et al. 2020), but, to the authors' knowledge, not in a true impact-based nature for stakeholders (or decision-makers) concerned with runway snowfall and accumulation at airports. Additionally, this study not only assesses the performance of the products in terms of snow amount, but also how well the forecast systems capture the start and end times of the hazard. The time snow begins to accumulate on the runways is a critical part of the forecast for airport operations as it has implications for scheduling of staff and resources (Morss et al. 2022). National Weather Service forecasters have also reported that timing information is critical for providing impact-based decision support services (IDSS) to their core partners (with airports being a core partner for many offices), particularly in a map-based form (Demuth et al. 2020).

In the context of pre-storm preparations for snow events at the airport, this study aims to improve understanding of how

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

Michael Kraus: Retired.

Corresponding author: Dana M. Uden, dana.uden@noaa.gov

DOI: 10.1175/WAF-D-22-0170.1

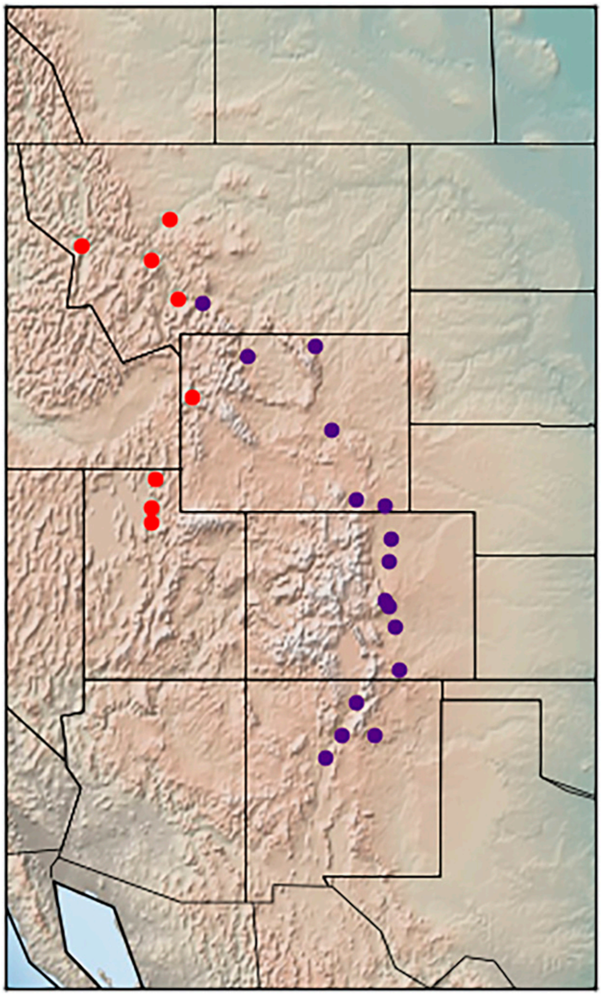


FIG. 1. Locations of the 24 airports used in this study. The red markers indicate airports not used in season 1 due to HRRRE domain limitations. The airports are listed in Table 1.

ensemble systems might be able to provide enhanced forecast information. The goal was to evaluate the performance of the Short-Range Ensemble Forecast (SREF), High-Resolution Rapid Refresh Ensemble (HRRRE), National Blend of Models (NBM), and the Probabilistic Snow Accumulation product (PSA) from the National Weather Service Boulder, Colorado, Weather Forecast Office (WFO) with an operational focus (the model systems will be explained in section 3). The forecast systems were evaluated against Meteorological Aerodrome Reports (METARs) at 24 airports around the Intermountain West (Fig. 1 and Table 1). The airports were selected for their similarity in elevation and climatology to KDEN and included in the evaluation to increase sample size. The time periods of analysis for this study are November 2018–April 2019 (excluding most of January 2019),[1] which will be referred to as season 1 and December 2019–April 2020 (season 2). Only the SREF and HRRRE data were available for season 1 and thus are the only two products used in the interannual comparison. The HRRRE domain was also limited in season 1 (stations depicted by the red points in Fig. 1 fell outside the domain). Background on the user research conducted by Morss et al. (2022) is in section 2, details on the forecast products are described in section 3, the event methodology is outlined in section 4, and results of the evaluation are presented in section 5.

TABLE 1. The 24 airports used in this study and their elevation. The airports in italics were excluded from the season 1 analysis due to domain limitations.

Airport	State	Code	Elevation (m)
Air Force	CO	KAFF	2003
CO Springs	CO	KCOS	1856
Denver	CO	KDEN	1640
Greeley	CO	KGXY	1420
Pueblo	CO	KPUB	1420
Trinidad	CO	KTAD	1756
Bozeman	*MT*	*KBZN*	*1361*
Great Falls	*MT*	*KGTF*	*1119*
Helena	*MT*	*KHLN*	*1182*
Livingston	MT	KLVM	1418
Missoula	*MT*	*KMSO*	*975*
Albuquerque	NM	KABQ	1618
Las Vegas	NM	KLVS	2091
Santa Fe	NM	KSAF	1930
Taos	NM	KSKX	2161
Logan	*UT*	*KLGU*	*1355*
Ogden/Hill AFB	*UT*	*KHIF*	*1459*
Salt Lake City	*UT*	*KSLC*	*1286*
Casper	WY	KCPR	1621
Cheyenne	WY	KCYS	1868
Cody	WY	KCOD	1553
Jackson	*WY*	*KJAC*	*1961*
Laramie	WY	KLAR	2216
Sheridan	WY	KSHR	1202

2. Background user research

During the winters of 2017/18 and 2018/19, Morss et al. (2022) from the National Center for Atmospheric Research (NCAR) conducted interviews, shadowed staff during snow events, and observed how the airport staff received and used weather information. Analysis of these data revealed two stages in which winter weather information was utilized at KDEN: pre-event planning and in situ tactical decision-making. This article focuses on forecast information for the pre-event stage, when the decision lead times best match with the capabilities of the modeling systems studied here. Before snow begins, the airport decision-makers gather forecast information about the timing and amount of snow, and other variables from multiple sources, including a private weather forecasting contractor and the Boulder WFO. They use this information to help make decisions about how to prepare for snow removal from runways and other paved surfaces, which is most important to airport operations, and staffing (Morss et al. 2022).

[1] Data for January 2019 were unavailable due to the federal government shutdown, which resulted in disruption to data feeds.

Table 2. Characteristics of the forecast systems used in the evaluation: the Short-Range Ensemble Forecast (SREF), the High-Resolution Rapid Refresh Ensemble (HRRRE), the National Blend of Models (NBM), and the Probabilistic Snow Accumulation product (PSA). Season 1 was the winter season from 2018 to 2019, and season 2 was the winter season from 2019 to 2020.

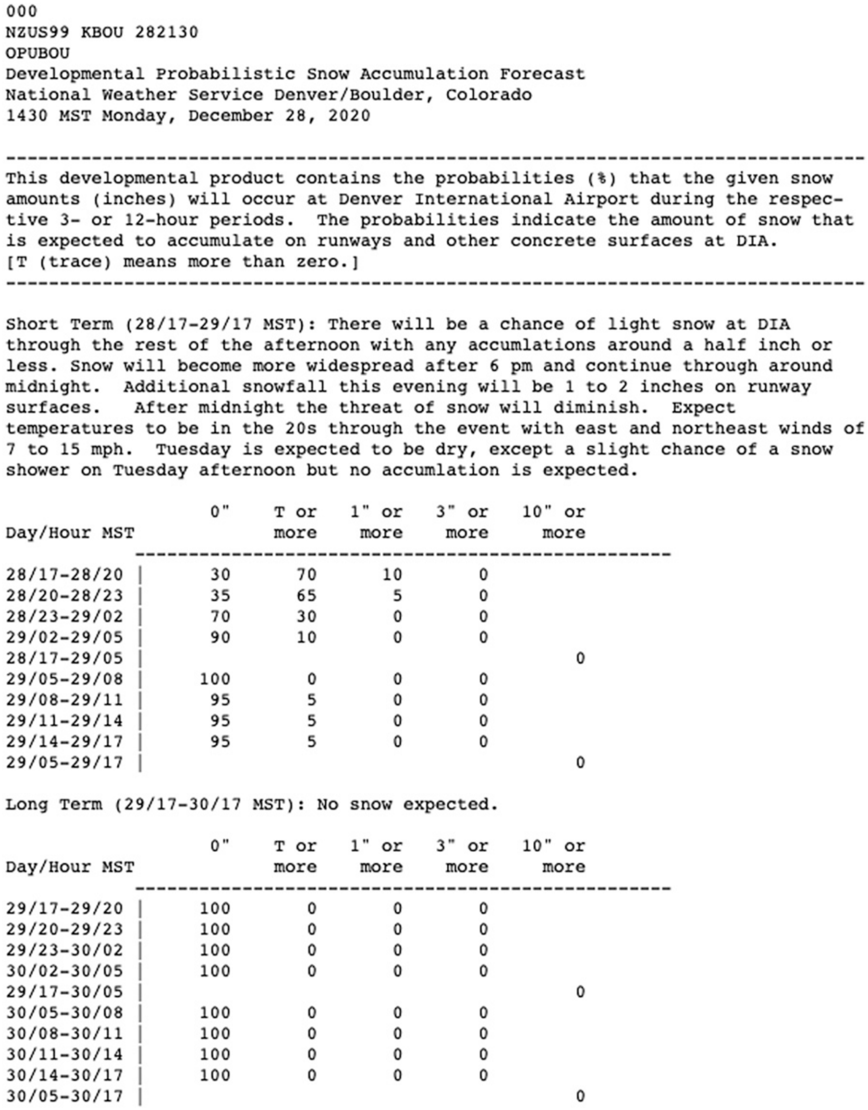
	SREF	HRRRE	NBM	PSA
Version	Operational	Experimental	3.2, text product	N/A, human-generated
Horizontal resolution	16 km	3 km	Point forecast	Point forecast
Time period	Seasons 1 and 2	Seasons 1 and 2	Season 2	Season 2
Latency	4 h	4 h	1 h	—
Initial conditions	Multianalysis, blended perturbation, multilateral boundary conditions from Global Ensemble Forecast System	The first nine members of the 36-member HRRR Data Assimilation System	Specific to NBM input models	—
Physics	NMMB and ARW cores, quasi-stochastic physics, multiphysics	See Benjamin et al. (2016)	Specific to NBM input models	—
Issuances	0300, 0900, 1500, 2100 UTC	0000, 1200 UTC	Hourly	Typically, four times per day
Forecast hours	87 h, every 3 h	36 h, every hour	25 h, every hour	48 h, every 3 h
Members	26	9	31, but not all available at each hour	—
Reference	Du et al. (2015)	Dowell et al. (2018)	Craven et al. (2020)	NWS WFO Boulder

The primary decision points prior to an event involve deciding which snow alert level to declare and at what time and if/when chemicals need to be applied to the runways and other paved areas. The snow alert level, based mostly on predicted snow accumulation amount, is especially important because it determines how many people will be on shift and how much snow removal equipment will be available to handle the first part of the storm. There are four alert levels in the KDEN Snow and Ice Control Plan: cautionary, A, B, and emergency. The following snow amounts correspond, respectively, to the alert levels and are the basis for the thresholds selected in this study as well as what is presented in the PSA: from a trace to 1 in., 1–3 in., 3–10 in., and greater than 10 in. The final decision on the alert level is typically determined at a 1000 local time meeting, usually on the day of or day before an event (Morss et al. 2022).

Morss et al. (2022) also looked at how the airport staff could use probabilistic information. The airport operations personnel have many years of experience with making decisions based on uncertain weather forecast information, but they do not have a sophisticated statistical understanding of probability theory. This study will include the colloquial terminology used by these decision-makers. Morss et al. (2022) reported personnel gaining a sense of forecast confidence from looking at many different sources of weather information and from talking with forecasters, both from the Boulder WFO and KDEN's paid forecast services contractor. They looked at forecast ranges (and sometimes probabilities) and alluded to the idea that "forecasts containing uncertainty information are more likely to be valuable if they are reliable, unbiased, and sufficiently sharp to provide information that is useful for their decisions" (Morss et al. 2022). The amount of uncertainty in a forecast (i.e., low probabilities spread across a wide range of storm total snowfall amounts) can sometimes cause the airport to elevate the snow alert level as a precaution, particularly if the snow overlaps with a time of day or year when airport traffic is high. Morss et al. (2022) suggested that an understanding of forecast skill and improved information about forecast uncertainty could be beneficial to the decision-makers in terms of both timing (onset and cessation) and severity (snow amount) of events.

The research conducted by Morss et al. (2022) provided many informative details for designing an impact-based assessment of probabilistic snow forecasts for airport decision-makers. The thresholds for snow amount were selected based off of those used in the snow alert declaration decisions, and freezing temperatures (2-m air temperature less than or equal to 32°F) were used as a filter to get closer to assessing snow accumulating on concrete/asphalt. The 1000 local time meeting time to determine the alert level was incorporated into the forecast issuance and lead times selected to be evaluated (see section 4).

Additionally, many of the results and conclusions in this study are presented from the conservative standpoint that an airport stakeholder would take, that is, choosing a probability threshold that produces high snowfall biases and longer events to ensure the airport is not underprepared. In their user interviews, Morss et al. (2022) heard statements like "we are always going to err on the side of caution" and "I feel better being a little more conservative than not." While safety is a primary goal for the airport, there are also other considerations like costs, predictability, and efficiency. Because of these complex factors and situational decisions that sometimes do not depend on the weather (i.e., air traffic congestion and airline decisions), Morss et al. (2022) ascertained that a cost-loss model would not be appropriate for this scenario; the model is



```
000
NZUS99 KBOU 282130
OPUBOU
Developmental Probabilistic Snow Accumulation Forecast
National Weather Service Denver/Boulder, Colorado
1430 MST Monday, December 28, 2020

--------------------------------------------------------------------------------
This developmental product contains the probabilities (%) that the given snow
amounts (inches) will occur at Denver International Airport during the respec-
tive 3- or 12-hour periods.  The probabilities indicate the amount of snow that
is expected to accumulate on runways and other concrete surfaces at DIA.
[T (trace) means more than zero.]
--------------------------------------------------------------------------------

Short Term (28/17-29/17 MST): There will be a chance of light snow at DIA
through the rest of the afternoon with any accumlations around a half inch or
less. Snow will become more widespread after 6 pm and continue through around
midnight.  Additional snowfall this evening will be 1 to 2 inches on runway
surfaces.   After midnight the threat of snow will diminish.  Expect
temperatures to be in the 20s through the event with east and northeast winds of
7 to 15 mph.  Tuesday is expected to be dry, except a slight chance of a snow
shower on Tuesday afternoon but no accumlation is expected.

                   0"    T or   1" or   3" or   10" or
Day/Hour MST             more    more    more     more
            --------------------------------------------------
28/17-28/20 |      30      70      10       0
28/20-28/23 |      35      65       5       0
28/23-29/02 |      70      30       0       0
29/02-29/05 |      90      10       0       0
28/17-29/05 |                                        0
29/05-29/08 |     100       0       0       0
29/08-29/11 |      95       5       0       0
29/11-29/14 |      95       5       0       0
29/14-29/17 |      95       5       0       0
29/05-29/17 |                                        0

Long Term (29/17-30/17 MST): No snow expected.

                   0"    T or   1" or   3" or   10" or
Day/Hour MST             more    more    more     more
            --------------------------------------------------
29/17-29/20 |     100       0       0       0
29/20-29/23 |     100       0       0       0
29/23-30/02 |     100       0       0       0
30/02-30/05 |     100       0       0       0
29/17-30/05 |                                        0
30/05-30/08 |     100       0       0       0
30/08-30/11 |     100       0       0       0
30/11-30/14 |     100       0       0       0
30/14-30/17 |     100       0       0       0
30/05-30/17 |                                        0
```

FIG. 2. An example of the PSA product from 28 Dec 2020.

too idealized to encompass all aspects of the decision-making process at the airport. Because the cost-loss model was not appropriate in this case, most traditional verification metrics for probabilistic forecasts (i.e., Brier scores, receiver operating characteristics, etc.) were not employed in this study and other metrics more in line with the airport's use case were used instead (see section 4).

3. Forecast systems

Table 2 depicts the four forecast systems employed in this study and their key characteristics. Latency is defined as the amount of processing time to produce output data; it is the difference between the issuance time and the time the data are available for operational use or distribution. Further details on each forecast system are provided in the following subsections.

a. SREF

The SREF is an operational ensemble forecast system run at the National Centers for Environmental Prediction (NCEP) at approximately 16-km horizontal resolution. It consists of 26 members; half of the members are made up of Weather Research and Forecasting Model (WRF) Advanced Research WRF core (ARW) members while the others come from the Nonhydrostatic Multiscale Model on the B-grid (NMMB) core. The model output is available in 3-h increments out to 87 h and is updated four times per day at 0300, 0900, 1500, and 2100 UTC (Du et al. 2015). The last major update to the SREF occurred in October 2015 (see https://www.nco.ncep.noaa.gov/pmb/changes/) and therefore the model did not change during the period of this assessment. Snowfall data from the individual ensemble members was used instead of the precalculated ensemble summary output. While the

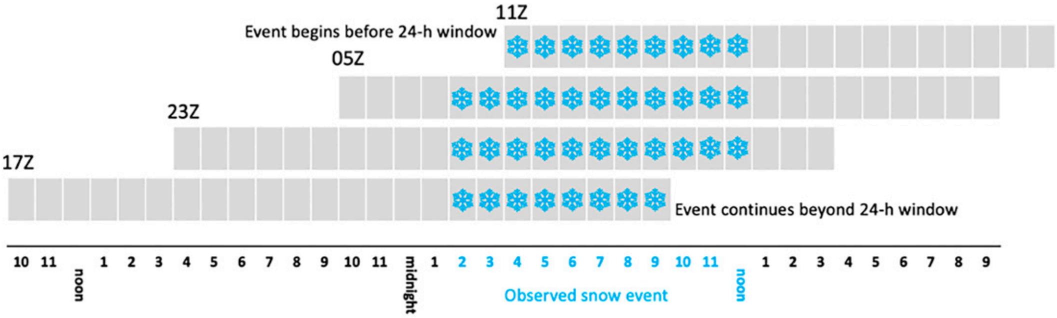

11Z
Event begins before 24-h window
05Z
23Z
17Z
Event continues beyond 24-h window
10 11 noon 1 2 3 4 5 6 7 8 9 10 11 midnight 1 2 3 4 5 6 7 8 9 10 11 noon 1 2 3 4 5 6 7 8 9
Observed snow event

FIG. 3. A schematic depicting the snow event methodology employed in this study. The four rows of gray cells represent the 24-h periods following the four decision-making times. An example observed snow event is illustrated by the snowflake icons; the event lasted from 0200 to 1200 local time (LT). For the 1700 UTC meeting time, the 24-h period ended prior to when snow actually stopped falling, while snow had already begun prior to the 1100 UTC meeting time.

SREF is slated to be phased out soon (see, e.g., NWS 2018), it is still being used in operations and therefore was included in this evaluation as a baseline.

b. HRRRE

Unlike the SREF, the HRRRE is not yet an operational product and is currently in development at the National Oceanic and Atmospheric Administration's Global Systems Laboratory. It employs only the WRF-ARW core (versions 3.8 and 3.9) in the ensemble system. The HRRRE uses the same ~3-km horizontal resolution as the deterministic HRRR and is available every hour out to 36 h. While the data-assimilation ensemble contains 36 members from the Global Data Assimilation System (GDAS), the HRRRE output contains only nine members (Dowell et al. 2018). The 0000 and 1200 UTC runs were used in this study as they were the only forecast issuances available at the time. The effect of the 3-h issuance offset with the SREF was deemed insignificant as performance was consistent across leads.

A number of changes were made during the 2018/19 winter season and between the 2018/19 and 2019/20 winter seasons that could have resulted in an impact on the snowfall forecast performance. First, the initialization changed to the Rapid Refresh (RAP; a 13-km North American model), which is itself initialized from the Finite Volume Cubed (FV3)-based Global Forecast System (GFS) version 15. The WRF was also updated to version 3.9 between seasons, and the domain was increased to include the full contiguous United States. Therefore, more airports were included in the season 2 analysis due to the westward expansion of the domain. Last, the ensemble spread was increased due to a couple of factors: the move to a rolling ensemble and the addition of stochastic physics (Global Systems Laboratory 2021). The HRRRE is used here as a proxy for the future Environmental Modeling Center (EMC) high-resolution regional ensemble; the SREF and the High-Resolution Ensemble Forecast, a multimodel, poor-man's ensemble system, are slated to be replaced by a new, single model core ensemble similar to the HRRRE (NWS 2018).

c. NBM

The NBM was created to provide a coherent and reliable national forecast product as the NWS forecasters' responsibilities evolve into more IDSS duties (Craven et al. 2020). It is a blend of many forecast models and systems, both deterministic and probabilistic, with different resolutions, temporal scales, and purposes. The approach involves bias correcting the individual model members [using the Unrestricted Real Time Mesoscale Analysis (URMA), as a truth set; see De Pondeca et al. 2011] prior to blending with an optimum weighting scheme. A final quality control check is then performed to ensure consistency. The weight each model has in the final product depends on the

TABLE 3. Data used for each product and 24-h period following each reference time. Note that the ideal leads for the SREF are 8–31, but are unavailable due to the 3-hourly nature of the product. Therefore, a linear adjustment was applied to the snow accumulation to account for the discrepancy.

	Reference time							
	1700 UTC		2300 UTC		0500 UTC		1100 UTC	
Product	Issue	Leads	Issue	Leads	Issue	Leads	Issue	Leads
SREF	0900 UTC	9–33 h	1500 UTC	9–33 h	2100 UTC	9–33 h	0300 UTC	9–33 h
HRRRE	1200 UTC	5–28 h	1200 UTC	11–34 h	0000 UTC	5–28 h	0000 UTC	11–34 h
NBM	1600 UTC	1–24 h	2200 UTC	1–24 h	0400 UTC	1–24 h	1000 UTC	1–24 h
METAR	0–23 h		0–23 h		0–23 h		0–23 h	

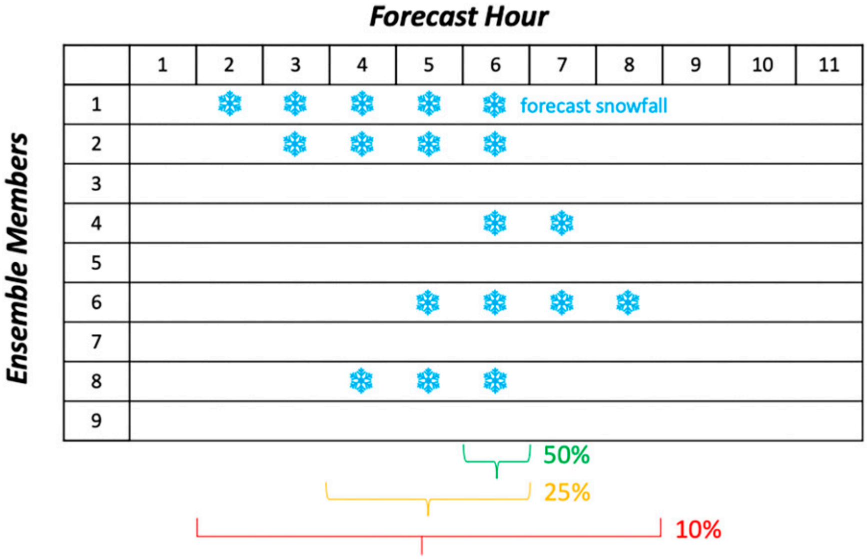


FIG. 4. Schematic highlighting the event threshold definition. The forecast hour/lead is on the horizontal axis, while the different forecast ensemble members are on the vertical axis. The blue snowflakes represent the forecast snowfall for each member. In this example, a 10% event occurred from forecast hour 2 to 8, a 25% event occurred from hour 4 to 6, and a 50% occurred at hour 6.

forecast parameter in question; weights for continuous parameters are based on a decaying average bias correction algorithm (Cui et al. 2012; Rudack 2020), while the discontinuous parameters have expert weights calculated from previous verification studies (Rudack 2020).

Version 3.2 of the NBM was utilized in this assessment (for season 2 only). It was experimental during the beginning of season 2, but became operational on 19 February 2020 (Meteorological Development Laboratory 2021). Version 3.2 of the NBM provided a new text product in addition to the gridded output, added more probabilistic fields, and improved probabilistic quantitative precipitation forecasts in the western mountains of the United States (Rudack 2020). As there were no probabilistic snow amount variables for accumulation periods less than 24 h (Craven et al. 2020) in the gridded output (which was deemed too coarse a temporal resolution for airport applications), the text products for each airport were used in this study. The NBM text products are available every hour on the hour and provide forecasts out to hour 25. However, the 31 model inputs to the ensemble are mostly older due to latency of the individual members; many hours contain a relatively small number of models (Rudack 2020). It is important to note that the NBM is not independent from the previous forecast products mentioned, both the SREF and HRRR (which is the foundation of the HRRRE) are inputs (Craven et al. 2020). Details on the parameters used from the NBM text files are provided in section 4.

d. PSA

The PSA forecast (e.g., Fig. 2) is a text product produced by the Boulder WFO specifically for public- and private-based airport operations at KDEN, with a goal of meeting the needs of several core partners that make high impact decisions regarding airport operations during snow events. It is typically updated four times per day (overnight, morning, afternoon, and evening—the exact times are not predetermined) if the probability of snow accumulation from the time of issuance out to 36 h at KDEN is at least 10%. Forecasters at Boulder WFO use all available numerical weather prediction (NWP)-based forecast data (including the NBM, SREF, convective-allowing models, global models, ensemble-based systems, and probabilistic data based on NWP) to generate a subjective probabilistic snow accumulation forecast for the airport, broken into operationally relevant 3-h increments. The WFO did not have access to the experimental HRRRE. The 3-h temporal resolution of the product is a direct result of feedback from KDEN decision-makers, as are the threshold columns, which represent accumulations needed for the different alert levels.[2]

Unlike the other three systems that predict snow accumulation on natural surfaces, the PSA predicts snow accumulation on concrete, which is most relevant for KDEN runway operations. This may explain some differences in the verification statistics in section 5. The PSA also includes both short-term and long-term narratives, which have been noted to be helpful to airport staff when making decisions (D. Cunningham 2019, personal communication), but are not evaluated in this quantitative study. Only the 3-hourly probabilities of snow in the short-term and long-term tables were assessed. The 12-hourly

[2] A newer version, not available during the experimental period, also contains an uncalibrated, automated "first guess."

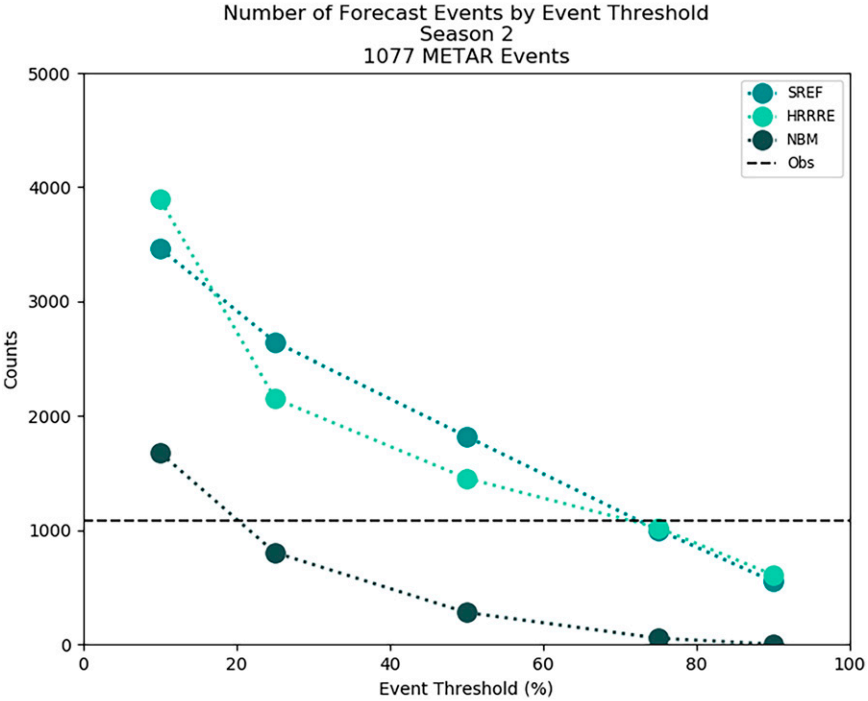


FIG. 5. The number of snow events (with no missing hours) by event threshold for season 2 (all airports). For the forecasts, the markers represent the number of forecast events over the season using the event threshold to define events; dotted lines are used to connect markers for easier visualization, but note that data are only available at the five thresholds. For SREF and HRRRE, a more restrictive event definition (higher event threshold) results in a closer match to the actual number of snow events that were observed. The number of observed events is independent of the forecast event threshold; there were a fixed number of events that occurred regardless of the forecast threshold.

10-in. or more forecasts were not evaluated due to sample size limitations.

4. Methodology

The methodology outlined in the following subsections was employed to examine the performance of the forecast products in the context of the airport snow removal operations. Of particular interest were the errors in event timing and magnitude, the reliability of the probabilistic forecasts, and the characteristics of misses and false alarms (i.e., losses and unnecessary costs).

a. Event definition

An event in this study is defined as accumulating snowfall with freezing temperatures (2-m air temperature less than or equal to 32°F). Snow events were constrained to a 24-h window following four operational decision-making/reference times: 1700, 2300, 0500, and 1100 UTC (corresponding to 1000, 1600, 2200, and 0400 local standard time). Additional times beyond the actual 1000 local decision time at KDEN were included for the following reasons: 1) to increase the sample size, 2) to represent a reasonable extension of the daily decision time (KDEN receives updates four times per day during snow storms from their private contractor) (Morss et al. 2022), and 3) to incorporate the availability of new data from the SREF and several NBM inputs. A 24-h period corresponds to two 12-h operational shifts at the airport (Morss et al. 2022). Finally, snow events were merged if the end of one and the start of another were within 6 h (within the constraints of the 24-h period).[3] The merging was employed because the events would be considered a single event from an airport planning perspective.

The schematic in Fig. 3 illustrates this forecast timeline for an example snow event. A 24-h window was selected to ensure the availability of all forecast products; the NBM has a maximum lead of 25 h with a 1-h latency. While the NBM has a short latency, 4 h was assumed for the SREF and HRRRE based on experiences with data acquisition. Table 3 identifies which leads compose the 24-h valid periods for the various products. Ideally, leads 8–31 would have been used for the SREF, but due to the lower temporal resolution (3-hourly), which only allowed for leads 9–33, a linear adjustment was

[3] The airport operations team did not define a specific time window for merging events. The 6-h window proposed here is an attempt to capture the nature of the decisions faced by the operations team.

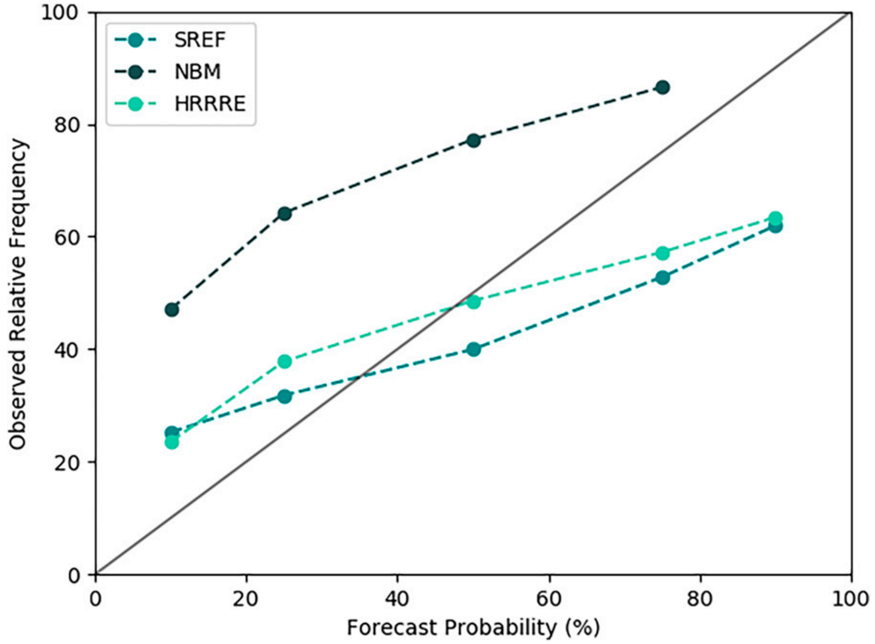


FIG. 6. Reliability diagram for the SREF, NBM, and HRRRE systems. The NBM did not have a 90% forecast. This figure shows data from season 2 and includes all airports.

applied to the snow accumulation amounts. All products and observations were required to be available at each of the four reference times. Forecast performance for snow accumulation was calculated only where all forecast data (i.e., every lead during the 24-h window) were present. For calculations of timing error, on the other hand, periods with missing data were included as long as the missing data did not cover consecutive hours.

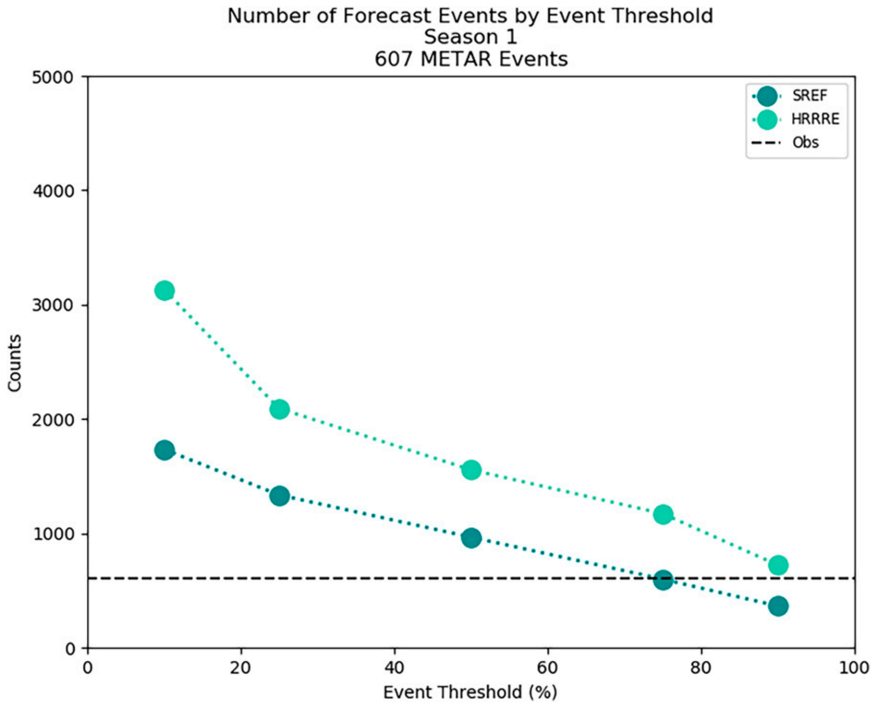


FIG. 7. As in Fig. 5, but for season 1 data (see Table 1 for list of airports used in season 1). The NBM was not available in season 1.

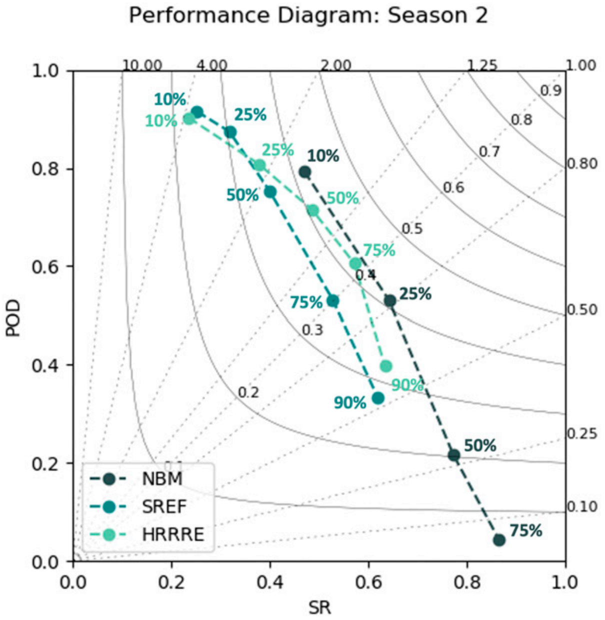


FIG. 8. A performance diagram comparing the probability of detection (POD) and success ratio (SR) of the systems. The SR can also be thought of as 1 − false alarm rate. The slanted dotted lines represent bias values with the diagonal implying an unbiased forecast. The curved gray lines represent the critical success index (CSI). An ideal forecast would be in the top-right corner of the plot. The markers indicate the forecast probability thresholds. These data are from season 2 and include all airports.

b. *Observed events*

Hourly precipitation and temperature measurements from METARs, as well as a mention of snow in the remarks, were used to determine if a snow event did or did not occur for any of the 24-h periods. Instances of mixed precipitation (i.e., remarks of rain and snow) were not included. To create snow amount from liquid precipitation records, a 15:1 snow-to-liquid ratio (SLR), the climatological value for Denver, representative of much of the study domain as well (Baxter et al. 2005), was applied. METAR liquid equivalent observations were used instead of human snow measurements because the latter were not available for all of the airports included in the analysis and the observations of snow depth at KDEN were deemed not sufficiently reliable. Furthermore, since the three automated forecasts give liquid equivalent, the effect of SLR errors is removed from the comparison of those products (SLR errors can affect the scale of the errors, but it will affect the scale equally for the three automated products), at the cost of greater uncertainty for the errors in PSA snow amount forecasts.

c. *Forecast events*

The probabilistic snow accumulation forecasts were evaluated at 10%, 25%, 50%, 75%, and 90% probability thresholds. The probability threshold was based on the number of members from the ensemble that met snow event criteria. For example, at each forecast hour, at least three out of the nine HRRRE members were required to have snow for a 25% event (Fig. 4). As illustrated in Fig. 4, each threshold, if reached, yields a subset of the lower threshold events. While there were very few 90% events, the airport personnel are greatly interested in forecast confidence and thus this threshold was included (Morss et al. 2022).

1) SREF AND HRRRE

The methodologies to determine if the SREF or HRRRE predicted an event were quite similar. Like the observed METAR events, these forecast events were defined by liquid-equivalent snow (converted to snowfall with a 15:1 SLR). The liquid-equivalent snowfall used from the postprocessed SREF output was a snow accumulation variable (accumulation over only the 3-h time step) while HRRRE events used a total snow depth variable.[4] Minimum, median, and maximum snowfall from the ensemble members that met event criteria were recorded for each event snow total. The forecast probability was determined by a simple frequentist approach, calculating the percentage of members meeting event criteria. To be included in the analysis, events were required to have three-quarters of the ensemble members available (20 out of 26 SREF members and 7 out of 9 HRRRE members).

2) NBM

While the NBM is an ensemble product, only the final blended output is available and the output variables are not an exact match for the SREF/HRRRE output. As a result, snowfall amounts and probabilities are computed differently, as follows. The NBM events were defined with an unconditional probability resulting from the product of two raw fields: the probability of 0.01 in. or greater precipitation within 1 h and the probability that if precipitation falls, it falls as snow. That is, $p(\text{snow}) = p(\text{quantitative precipitation forecast} \geq 0.01\text{ in.}) \times p(\text{snow}|\text{precipitation} = \text{yes})$; this probability of measurable snow is then binned using the 10%, 25%, 50%, 75%, and 90% thresholds. The event snow amounts were determined in two ways. First, the liquid precipitation was multiplied by a 15:1 SLR (fixed SLR method) to enable a consistent field across all forecast products. The NBM also includes a snow accumulation parameter based on a variable SLR. This second method [variable SLR method, see UCAR (2019) for a detailed explanation] was added to the study for comparison.

3) PSA

To create forecast events from the PSA, the snowfall columns in Fig. 2 were converted to discrete bins: 0 in., from a trace to 1 in., 1–3 in., and 3 in. or more. An event was created if the probability of any amount of snow was greater than the (forecast) probability threshold. For example, an event predicted to occur with a 25% probability was recorded if the probability of 0 in. of snow

[4] The HRRRE hourly snow accumulation variable was corrupted in postprocessing and, hence, only the positive snow depth differences (implying snow accumulated between time steps) between forecast hours were used to define hourly snowfall amount.

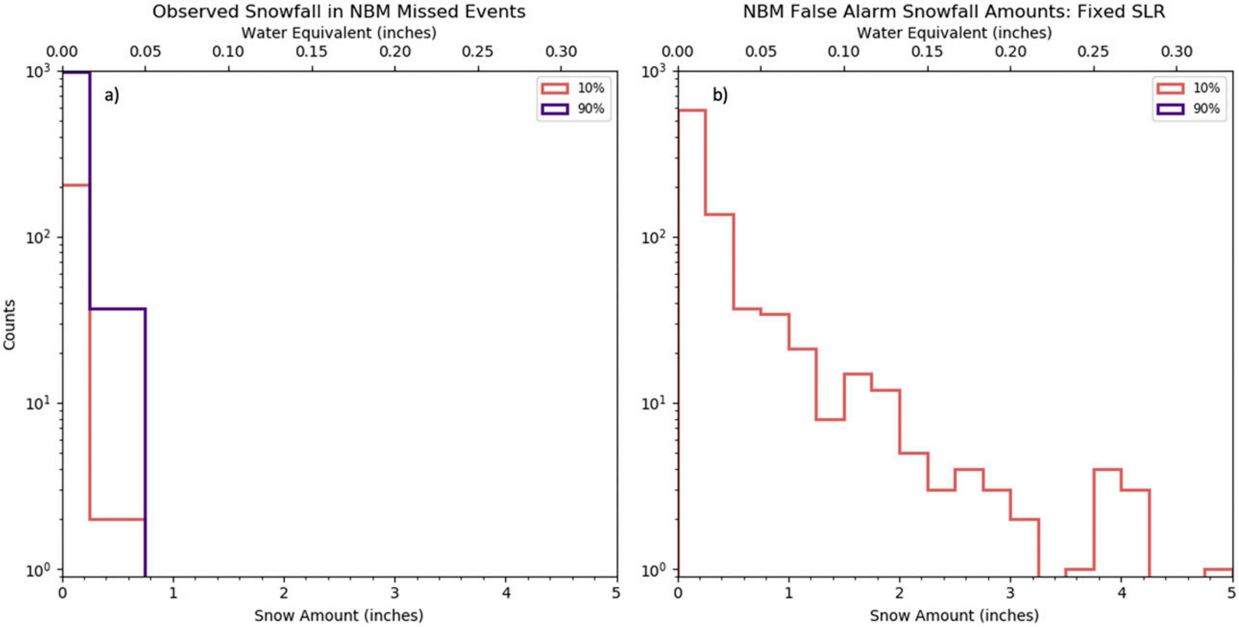


FIG. 9. (a) The number of observed NBM missed events and (b) the number of NBM false alarm events by event threshold and snow amount for all airports in season 2. There were no 90% false alarm events.

was 75% or less. The lower bound of each bin was used to determine snowfall amount as the highest bin was unbounded. The final snow amount for the event was then calculated based on a weighted average [Eq. (1)] across the bins:

$$\text{snow amount} = \frac{(\text{probability of trace–1 in.}) \times 0.01 + (\text{probability of 1–3 in.}) + (\text{probability of 3+ in.}) \times 3}{(\text{probability of trace–1 in.}) + (\text{probability of 1–3 in.}) + (\text{probability of 3+ in.})}. \quad (1)$$

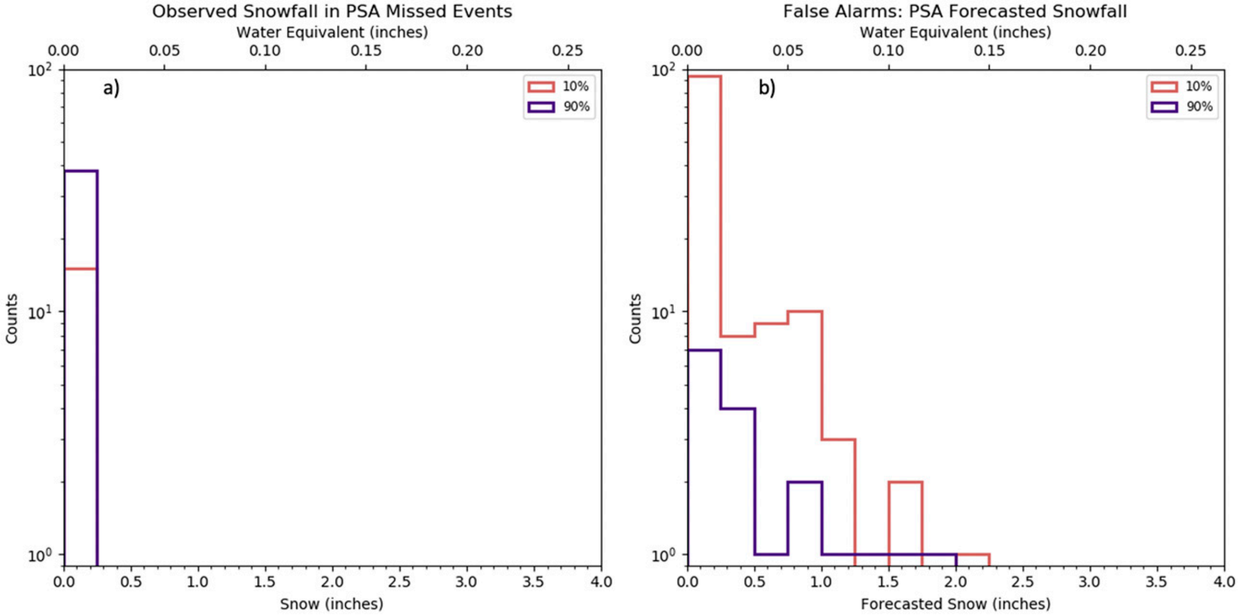


FIG. 10. (a) Observed snowfall in PSA missed events and (b) forecast snow amount in PSA false alarms by event threshold (colors) for season 2. Note, the PSA is only issued for KDEN.

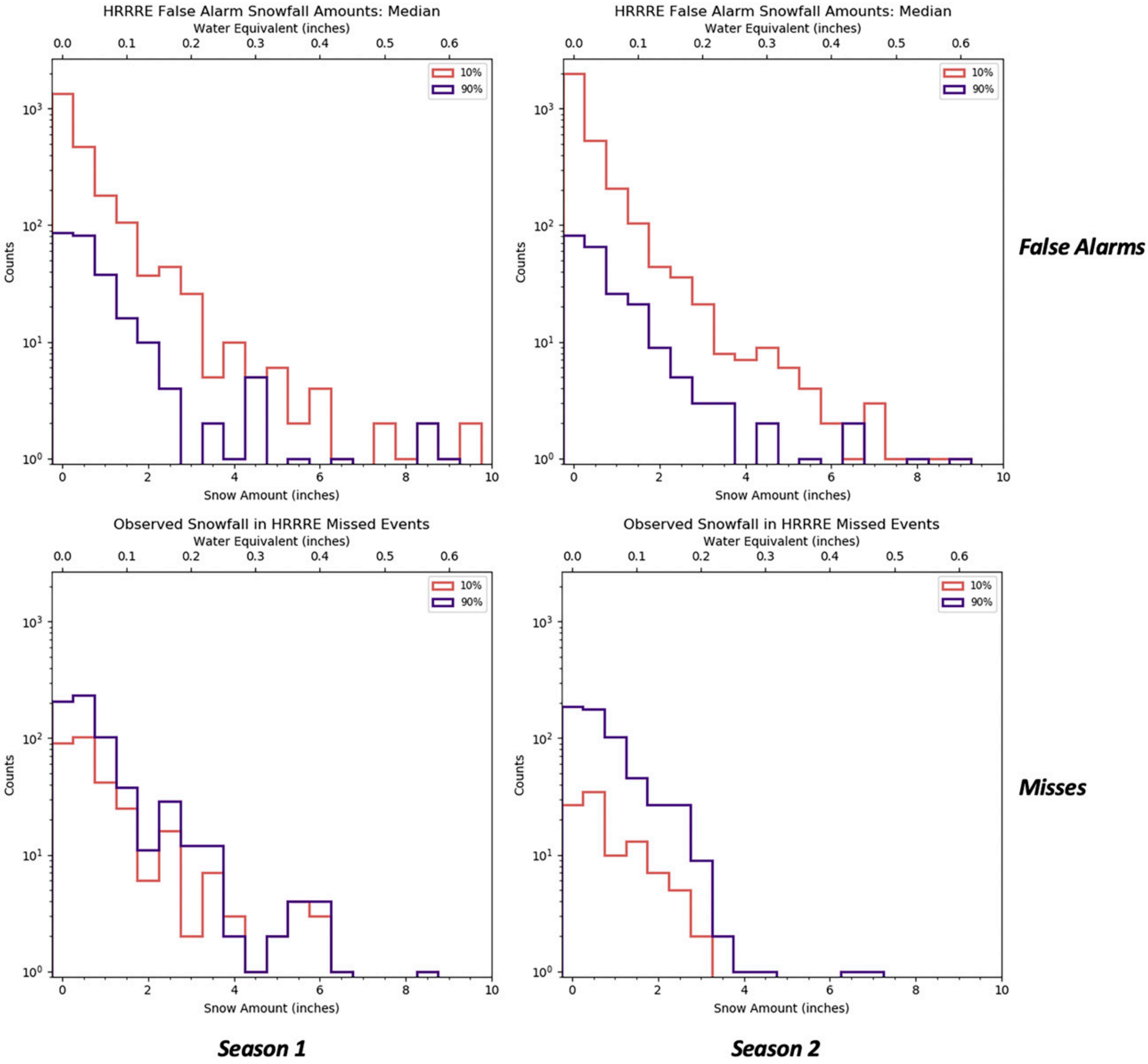


FIG. 11. HRRRE snow amounts in (top) false alarms and (bottom) missed events for (left) season 1 and (right) season 2. Note, season 2 contained data from eight additional airports. The HRRRE snowfall amounts shown in the top row represent the median amount.

Due to the irregular issuance of the PSA, the latest available product issued prior to the reference time was used; if the closest forecast was over 12 h old, it was not used and that reference time was skipped.

4) EVENT EQUALIZATION

In the interest of fairness, the forecast events were restricted to only instances when the SREF, HRRRE, and NBM were all available (i.e., the issuance was not missing); this check is referred to as event equalization in this study. The PSA was treated separately, however. Because it is only valid at one airport, removing issuances when one of the other products was missing resulted in too small of a sample size. For this reason, PSA events were created from all available PSA data.

d. *Pairing forecast and observed events*

Within each 24-h period corresponding to a given reference time, the forecast events were paired with the closest observed event based on start time. Therefore, matched events could not have start times more than 24 h apart. All forecast products were evaluated against hourly METAR events, even the 3-hourly SREF. The onset timing error was calculated as the difference between the forecast start time and the observed event start time. This was done for forecast events at each probability threshold. Similarly, the cessation error was calculated as the difference between the forecast end time and the last METAR hour that met event criteria. For example, if snow actually fell from hour 4 to 8 in Fig. 4, then there would be a negative 2-h onset error for

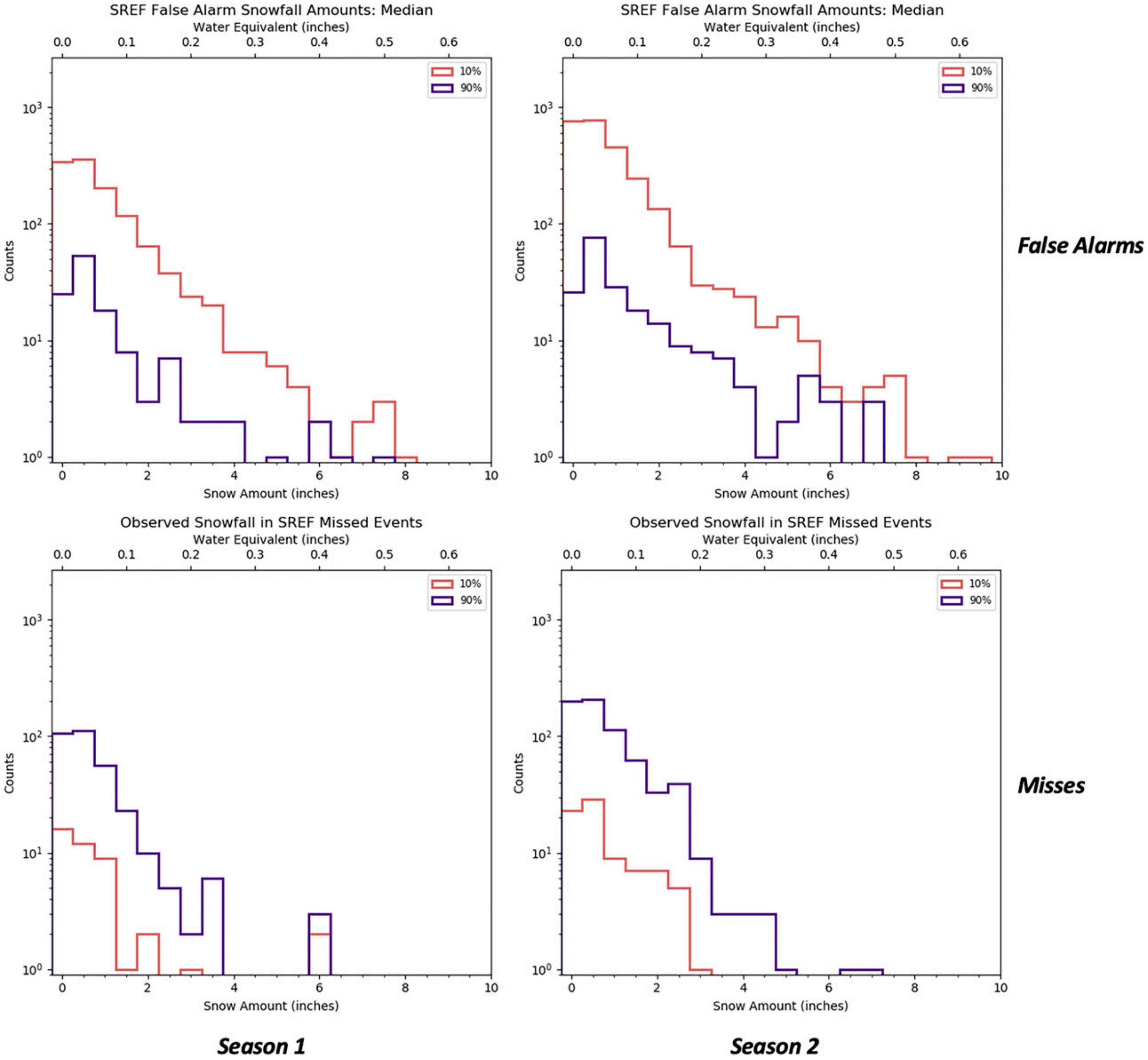


FIG. 12. As in Fig. 11, but for the SREF.

the 10% event and zero cessation error. Alternatively, the 25% event would have no onset error and a negative 2-h cessation error.

In addition to matched events (hits), false alarm events (forecast predicted snow when none occurred) and missed events (snow occurred but the forecast did not predict it) were recorded. It should be acknowledged here that the terms "false alarm" and "missed event" are generally out of place in the context of probabilistic forecasts (e.g., a calibrated 30% forecast is expected to "false alarm" 70% of time). The terms are retained in this paper, however, to represent the perspective of the decision-maker. For each of a set of probability thresholds, a forecast is classified as a false alarm when the probability exceeds the threshold (i.e., the decision-maker would take action) and no snow was observed. Similarly, when snow occurred but the forecast did not exceed the given threshold (i.e., the decision-maker would not take action), the event would be classified as a missed event. Thus, the terms false alarm and missed event refer only to the binary event of snow occurred or did not occur (for a given a forecast probability), irrespective of the forecast or observed amount.

5. Results

The figures in this section contain data from all event lead times as there was little change in forecast performance as a function of lead. Most figures depict season 2 data to allow for comparison with the NBM unless specified otherwise (i.e., seasonal comparisons of the SREF and HRRRE).

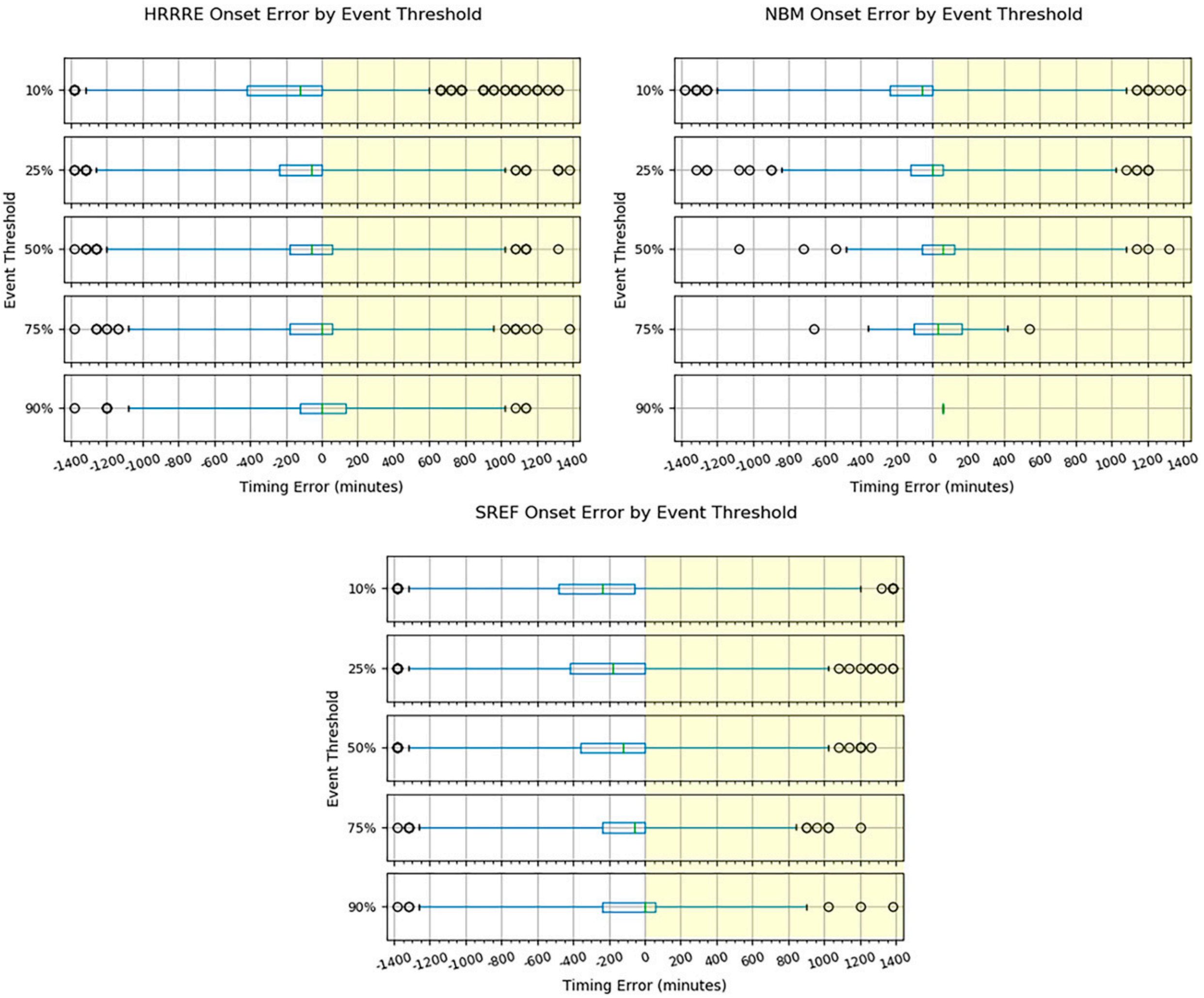


FIG. 13. Onset timing error by event threshold for the (top left) HRRRE, (top right) NBM, and (bottom) SREF for all airports in season 2. Negative values (white shaded area) imply the forecast started the event too early. Positive values (yellow shaded area) imply the forecast started the event too late. The whiskers represent the first and 99th percentiles, the boxes represent the 25th and 75th percentiles, the middle line represents the median, and the circles represent outliers.

a. Number of events

Perhaps the largest difference between the NBM and the other forecast products was the number of snow events for season 2 (Fig. 5). NBM predicted fewer events than the SREF and HRRRE. At the 10% probability threshold, NBM had more events than the 1077 observed, but the forecast probability never reached 90% throughout the season for all airports (only one 90% NBM event was recorded when allowing for nonconsecutive missing hours; see section 4a). The reliability diagram (Fig. 6) highlights over forecasting by the NBM, particularly at 50% and below, while the SREF and HRRRE display overconfidence. All three products suffer from reduced resolution (i.e., the slope of the reliability curves is less than one).

Since HRRRE and SREF data were available from both season 1 and season 2, an interannual comparison was conducted to determine if forecast performance was affected by the model system updates in the HRRRE. Changes in forecast difficulty between the two seasons was accounted for by comparison with the SREF which did not change. Figure 7 highlights the number of SREF, HRRRE, and observed events in season 1 for all airports. In season 1, the HRRRE had more events than SREF while the number of events between the two systems at each event threshold was more similar in season 2. At thresholds above 25%, the HRRRE actually had slightly more events in season 1 compared to season 2 despite fewer observed events taking place in season 1 (607 events observed in season 1 compared to 1077 observed in season 2). Note that the number of events in season 1 was dampened by the exclusion of the red airports in Fig. 1. The SREF had about the same number of 75% events as observed events in both seasons, but the HRRRE had too many events at all thresholds in season 1.

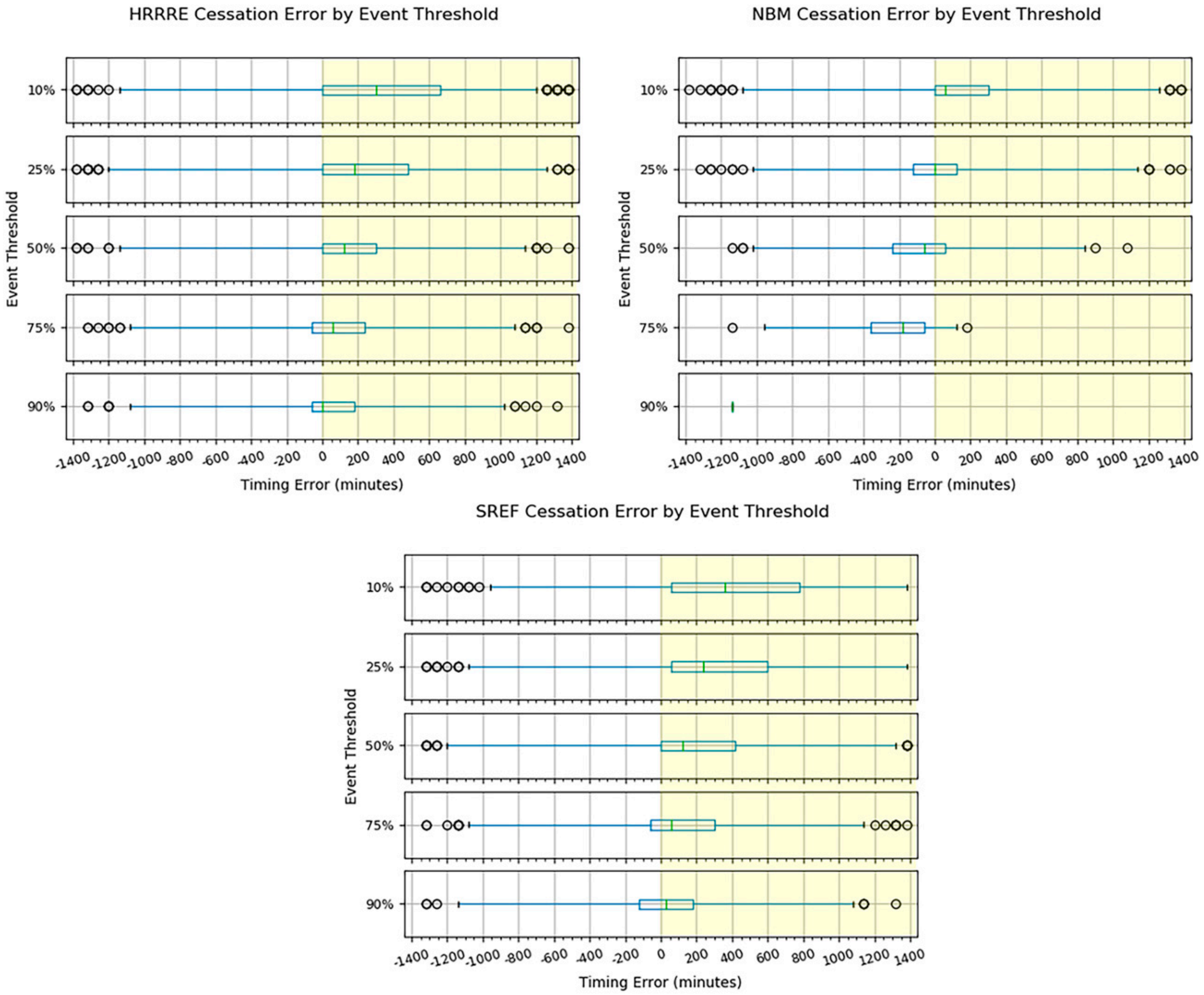


FIG. 14. Cessation timing error by event threshold for the (top left) HRRRE, (top right) NBM, and (bottom) SREF for all airports in season 2. Negative values (white shaded area) imply the forecast ended the event too early. Positive values (yellow shaded area) imply the forecast ended the event too late. The whiskers represent the first and 99th percentiles, while the circles represent outliers.

b. *False alarms and misses*

Figure 8 not only confirms the biases in the number of events from Fig. 5, but also depicts the probability of detection (POD) and success ratio (SR) of the forecast systems. It is evident that the NBM detected less events compared to the SREF and HRRRE, but had fewer false alarms (high success ratio). The 75% threshold for NBM produced a POD of only 4.5% while the SREF and HRRRE had 75% threshold POD values of 53.1% and 60.5%, respectively. The threshold with the highest critical success index (CSI) and best bias (closest to one) was 75% for HRRRE and SREF while only 25% for the NBM. The performance diagram highlights the significant difference in characteristics between the NBM and the other two systems.

The NBM false alarm and missed events had mostly negligible snowfall amounts. The operations teams at the airports considered in this study would have rarely been caught off guard if they used the NBM information in their decision-making, particularly concerning the missed events; nearly all missed events had a quarter of an inch of snowfall or less (Fig. 9a). While there were more cases of predicted higher snowfall amounts in the false alarm events, most of these events also had a quarter of an inch of snow or less predicted (Fig. 9b). There were no 90% false alarm events.

The PSA false alarm and missed events were comparable to NBM; there were no significant surprises during season 2. All missed events had less than or equal to a quarter inch of snow (Fig. 10a). Missed events included situations where there was no forecast snowfall within the matching window of an observed event and when there were two or more observed events that could potentially be paired with a forecast event: one was matched to the forecast while the other was counted as a miss. The false alarms were mostly 10% events with minimal snowfall accumulation (Fig. 10b).

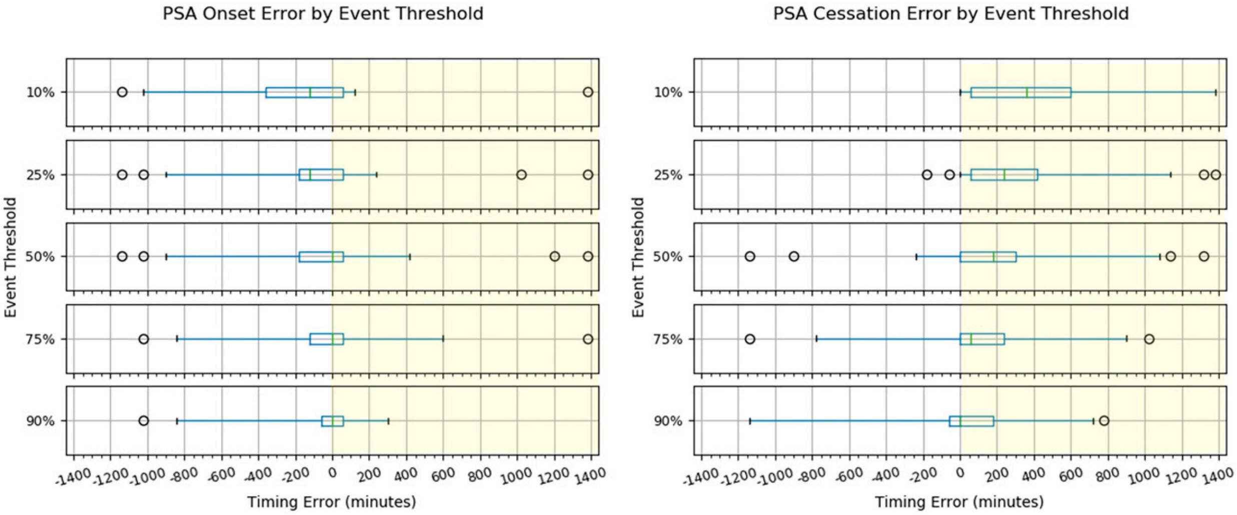


FIG. 15. PSA timing errors by event threshold: (left) onset and (right) cessation for KDEN in season 2.

Information on the false alarms and misses for the SREF and HRRRE was available for both winter seasons. The number of false alarms increased slightly for both HRRRE (Fig. 11) and SREF (Fig. 12) in season 2 (as expected due to the addition of eight airports), but the distribution of snow amount within those events was similar between seasons. The number of HRRRE missed events decreased though in season 2, particularly at the 10% probability threshold (Fig. 11, bottom row). It is worth noting that the HRRRE missed a number of events with more significant snowfall (in the 2–4-in. range) whereas the NBM and PSA only missed events with trivial amounts (Figs. 9 and 10). The SREF missed more events in season 2 compared to season 1 (Fig. 12).

c. *Timing errors*

Timing errors are important for the airport decision-makers' awareness, specifically the accuracy of the start time. It is critical that maintenance staff, equipment, and materials are ready for when snow begins accumulating. The timing of events is particularly important during busy hours when closing runways to clear snow has consequences for not only the airport, but also the Federal Aviation Administration and the airlines. The forecast end time of the storm is less critical to the airport personnel. They often will not remove a snow alert until they have visual confirmation that the snow has stopped. Therefore, weather forecasts of the end time are not utilized as frequently to determine when clean up can begin (Morss et al. 2022).

The optimal event threshold for producing the smallest timing errors was quite different for the NBM compared to the other products. Figure 13 depicts the onset errors for the forecast systems. While the median errors for the HRRRE and SREF were lowest for the higher event thresholds (75% and 90%), the NBM had the lowest median error at the 25% threshold. However, the KDEN airport operations team often like to err on the side of caution and choose a conservative threshold rather than being surprised with the early arrival of snow. In that case, the NBM 10% threshold for timing, which often started the events slightly too early (median error is 1 h early), would be an appropriate threshold to use so snow removal personnel and equipment would be more likely to be ready in time for the actual snow event.

Another difference between the two products was the variability in the errors. The NBM had a smaller interquartile range compared to HRRRE and SREF, especially for the 10% events; a narrower range means greater consistency which in turn means greater confidence in decision-making. In total, 50% of the NBM 10% events (i.e., the interquartile range represented by length of the box in the top right plot in Fig. 13) had start times that were within 230 min or less of the actual start time while 50% of the HRRRE events had start times that were within over ~400 min of the actual start time; both forecast systems predicted the onset too early. The cessation errors were similar in that the most accurate event threshold for NBM was the 25% threshold (Fig. 14). As with the onset of snow, the 10% threshold would be the conservative choice for cessation, often providing events that ended a little too late. Higher thresholds for the NBM produced events that ended too early whereas the HRRRE and SREF only ended on time or much later than observed.

It is worth noting that the 3-h lead steps for the SREF disadvantage that product relative to the other, 1-h lead products. Biases in the mean timing error are potentially exacerbated and the variability of the timing error distribution will be inflated. A subjective adjustment of the SREF distributions would suggest that SREF performance characteristics are more similar to those seen for the HRRRE.

Like the HRRRE and SREF (e.g., Fig. 13), the PSA timing errors were smallest for the higher event thresholds (Fig. 15). For onset of events, the PSA was on time (in terms of the median error) for thresholds of 50% or higher, otherwise it tended to be early. The PSA events were mostly late to end except for the 90% threshold. The cessation errors were generally larger than the onset errors and there was greater

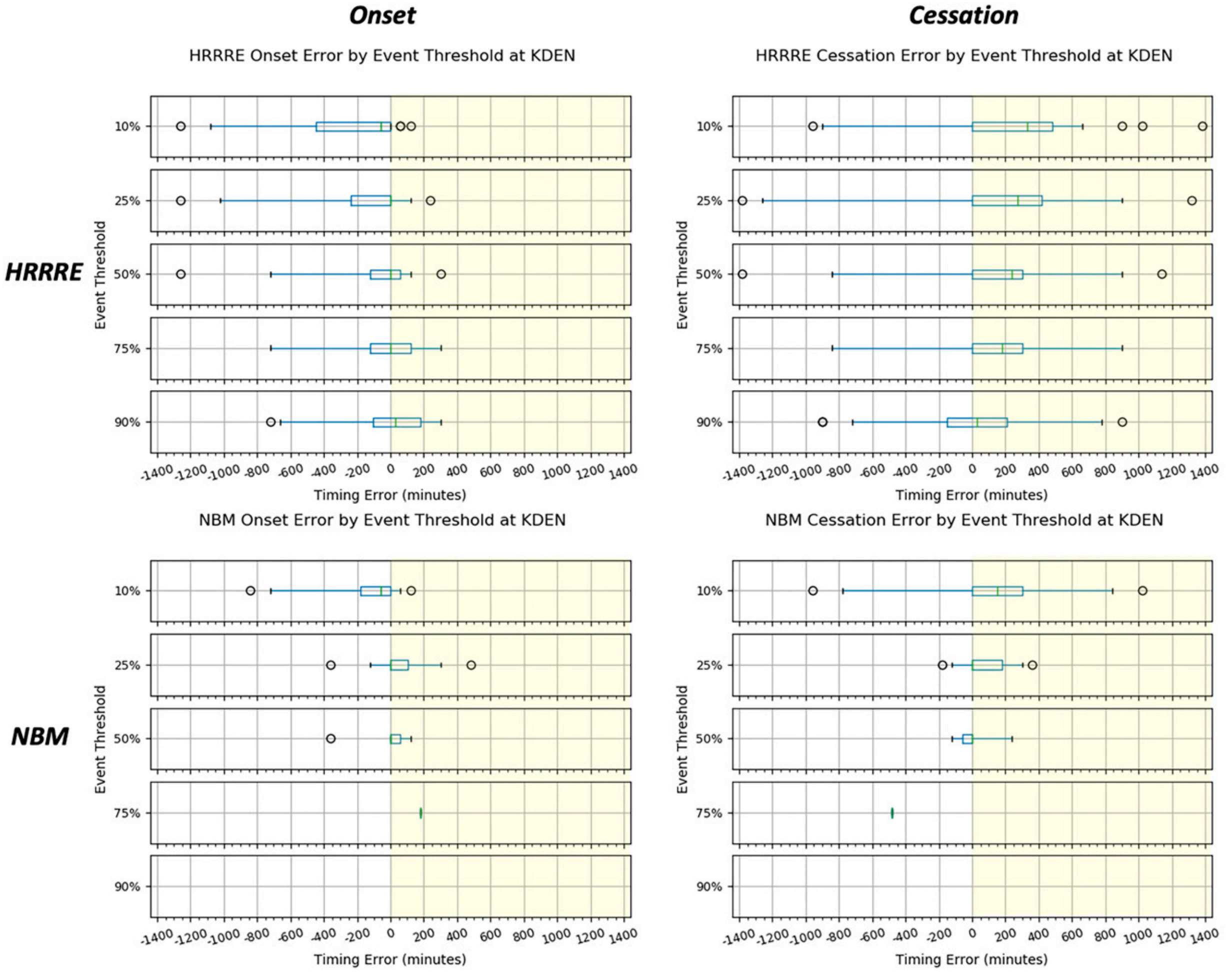


FIG. 16. (left) Onset and (right) cessation timing errors by event threshold at KDEN for the (top) HRRRE and (bottom) NBM in season 2.

variability in the cessation errors. The greater variability in cessation errors could be attributed to it being a longer-range forecast compared to the beginning of the event or that the forecasters generally are more conservative with cessation timing (they keep a small amount of snow in the forecast to be cautious). Compared to the NBM timing, the PSA provided several options if a conservative approach is desired (i.e., a longer event duration that likely captures the actual extent of the event). For the start of the event, either a 10% or 25% threshold would likely provide an earlier onset time, while any threshold below 90% would likely provide a later cessation time. A 75% threshold for cessation would still comply with a conservative approach without the much larger errors associated with the lower thresholds.

The HRRRE and NBM timing errors at KDEN are illustrated in Fig. 16. The NBM and HRRRE had similar onset errors, but compared to Fig. 15, the PSA had slightly higher errors, particularly at 10% and 25% thresholds where events started too early, comparatively. Alternatively, for cessation, the HRRRE had larger errors than NBM. Recall that the higher thresholds for NBM were very rare and therefore there was not enough of a sample for those analyses. As was seen with the timing plots using all airports, the NBM at KDEN had much less variability than the other two products as evidenced by the much smaller interquartile range.

d. Snow amount errors

As expected, the snow amount errors for NBM varied by event threshold (Fig. 17). While the presentation of the results in Fig. 17 may be untraditional, it can be used by the airport decision-maker to answer the following question: "If I base my decision on whether threshold *X* is exceeded, what sort of errors can I expect?" The largest biases occurred for the 75% events (90% was excluded due to lack of data). The NBM snow amount was closest to the observed [in terms of both mean error or bias and root-mean-square error (RMSE)] for the 10% events, but the product still had a slight high bias (Fig. 17a). The NBM had slightly higher biases at KDEN where the 10% events overforecast snowfall by 0.6 in. (Fig. 17b). There were not many events at KDEN that met the higher thresholds (50% and above). Therefore, the error curves beyond

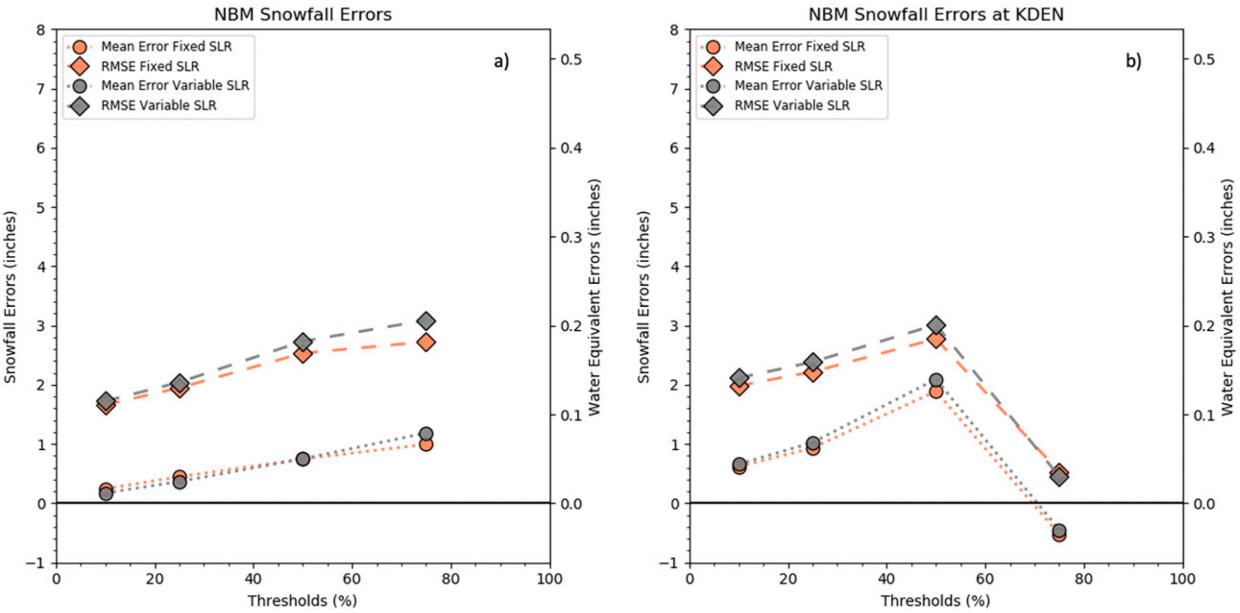


FIG. 17. NBM snow amount errors for (a) all airports and (b) KDEN only. The orange lines represent the fixed 15:1 SLR while the gray lines represent the snow amount directly from the NBM (variable SLR). Note, the sample size for the 50% events in (b) is only 10 events and only 2 for 75% events, hence, the noisiness of the curves at the upper thresholds. There were 37 10% events and 25 events at the 25% threshold in (b).

the 50% threshold are poorly defined. The fixed and variable SLR methods had similar biases suggesting that the climatological average (fixed) SLR was a decent approximation for the forecast (variable) SLR (which was not available from the SREF and HRRRE). The biases for the fixed and variable SLR were also stratified by observed snow amount and the two methods produced very similar results regardless of event intensity (not shown). It is unknown if the observed SLR was close to 15:1, but all forecast-observation comparisons for the SREF, HRRRE and NBM were done using the liquid-equivalent value and thus the SLR is irrelevant to the skill scores.

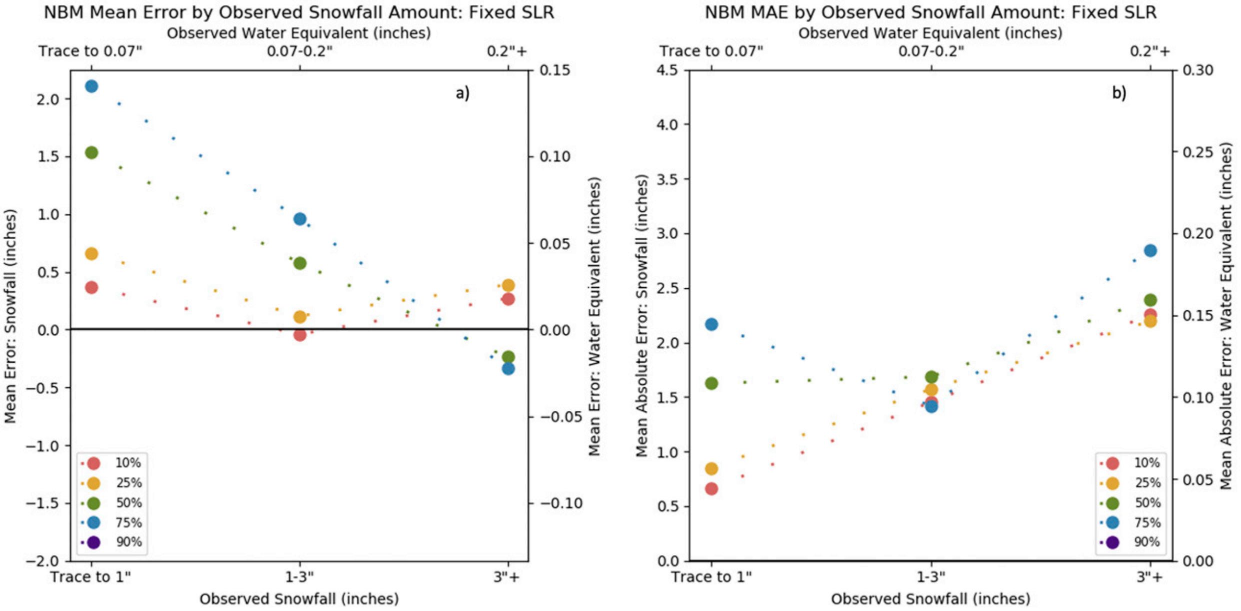


FIG. 18. NBM snow errors by observed snowfall amount for all airports in season 2: (a) bias and (b) mean absolute error. The observed snowfall was binned to the same bins as were used when evaluating the PSA.

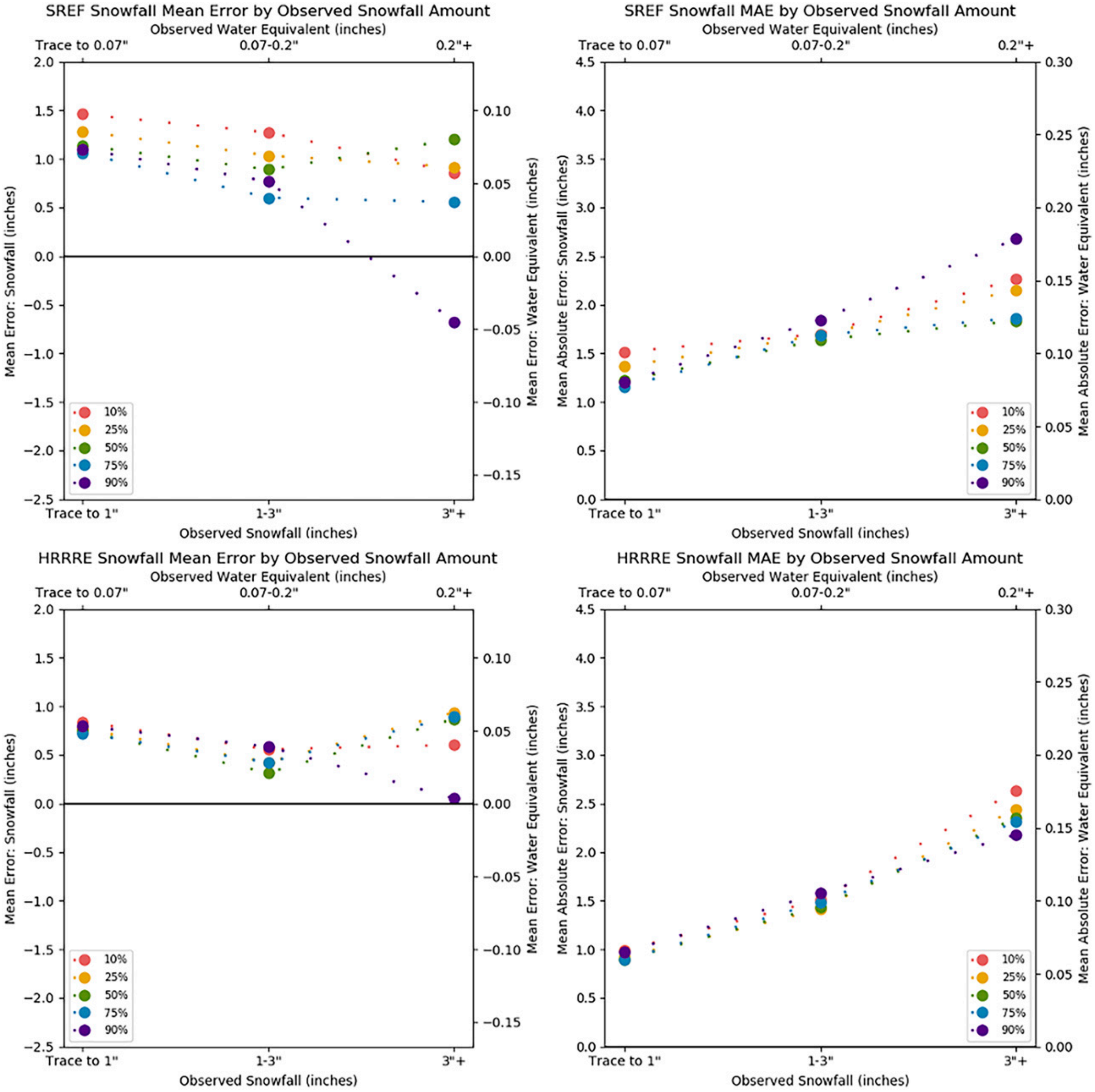


FIG. 19. As in Fig. 18, but for the (top) SREF and (bottom) HRRRE. The forecast snowfall presented is the median of the ensemble members. Data from all airports are included and are from season 2.

The observed severity of snow events was discovered to have an impact on the snowfall errors (Fig. 18). The NBM bias (mean error) was low for the 10% and 25% event thresholds regardless of observed snowfall amount. However, for higher thresholds, the bias improved with more observed snowfall. While the NBM biases were nearly zero for the 10% and 25% event thresholds, for the observed events that had 1–3 in. of snow, the higher mean absolute error (MAE) value (Fig. 18b) indicates a cancellation effect. The typical error was around 1.5 in., but there were instances where the NBM overforecast and underforecast snowfall.

Unlike the NBM, there was more consistency between forecast thresholds with the SREF and HRRRE, particularly the HRRRE (Fig. 19). While the NBM mean error varied by about 2 in. across forecast thresholds at the trace–1-in. category, the HRRRE mean error was around 0.75 in. for all thresholds. Using the median snow amount from the ensembles resulted in a positive bias, even for the lowest thresholds (which had a lower bias in the NBM). There was also less change in error across the observed snowfall bins for most SREF and HRRRE thresholds compared to the 50% and 75% NBM curves.

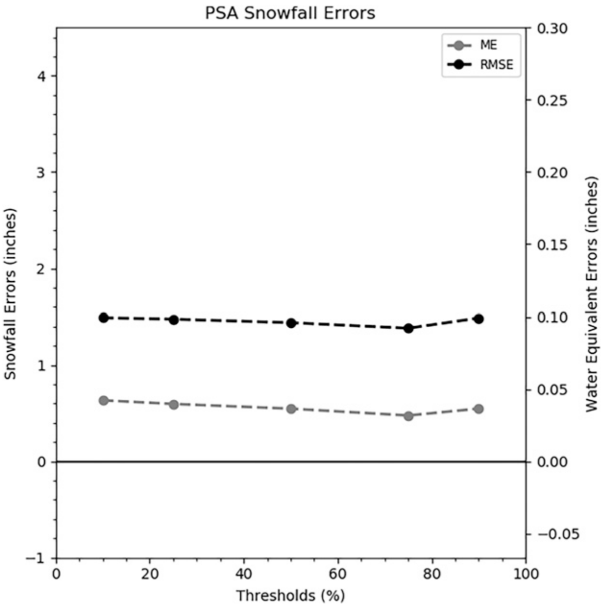


FIG. 20. PSA snowfall errors from season 2 at KDEN by event threshold.

Recall that the PSA product used in this study is specific to KDEN. The PSA (Fig. 20) had similar snowfall errors to the NBM (Fig. 17). Both had a ~0.6-in. bias for 10% events at KDEN. Unlike the NBM though, the PSA bias was nearly independent of the event threshold. Despite using the lower bound of the PSA bins to define snow amount, the product still slightly overforecast snowfall. Consider a scenario in which a forecaster was very confident that 2 in. of snow would accumulate. With the methodology used, the resulting forecast in the PSA would be for 1 in. (i.e., a 100% chance of exceeding 1 in. and 0% chance of exceeding 3 in.). If 3 in. of snow were observed, the resulting error would be −2 in. rather than the "true" (but unknowable, from the PSA perspective) error of −1 in. This method is similar to using the minimum amount from the ensemble products. If a product user assumed the high side of the forecast range, one would expect the snowfall errors in Fig. 20 to be greater.

The consistency in PSA performance across event thresholds is also evident when stratified by observed snow amount in Fig. 21. All thresholds had the lowest bias (least overforecasting) for 1–3-in. observed events. The larger MAE value for that bin indicates cancellation of errors. For example, 75% forecasts were typically off by about 1 in. (MAE), but it was almost as likely to forecast too little as too much snow resulting in an average of nearly zero (bias). The largest errors, both biases and MAE values, occurred for more significant observed events. There were three observed snow events that had 3 in. or more of snow resulting in 12 forecast–observation pairs due to multiple reference times and leads. Therefore, the larger errors could be a result of a small sample size.

As discussed in section 3, there were a number of HRRRE improvements made between seasons one and two that could have impacted the snowfall forecast performance. However, the following results suggest the changes had a negligible impact to the performance of the parameters included in this study. Changes in HRRRE snow amount errors were comparable to the SREF which can be treated as a baseline (Fig. 22). The RMSE values for both products were reduced in season 2 and the biases using the minimum snowfall from the ensemble slightly improved. One minor change to the HRRRE performance in

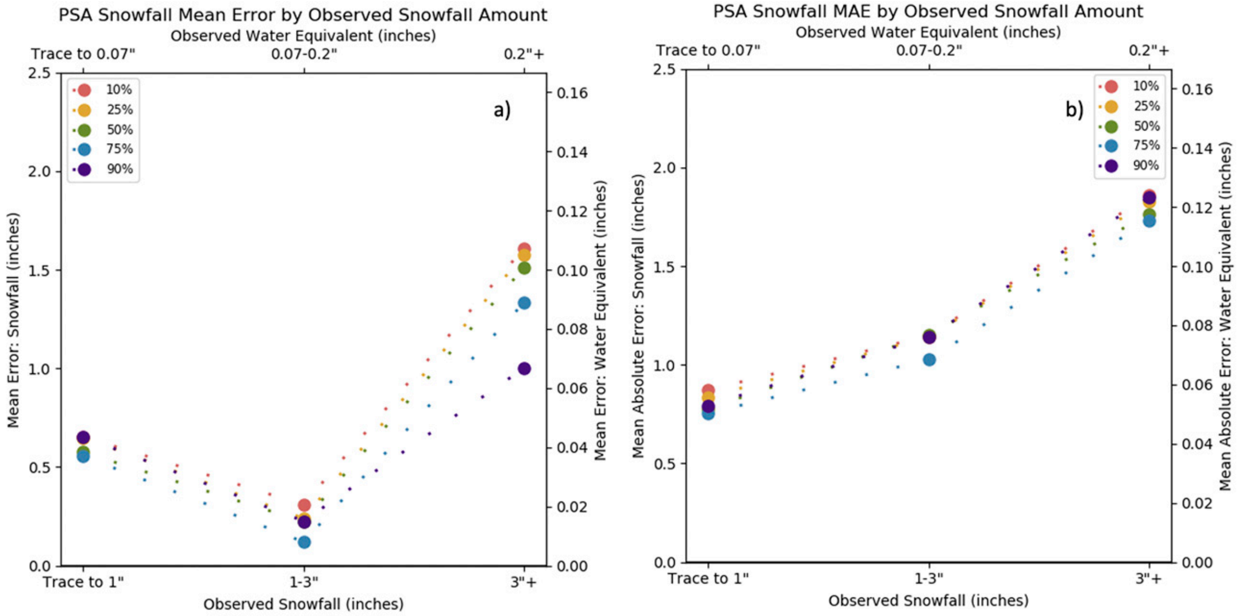


FIG. 21. PSA snowfall errors in season 2 by observed snow amount and event threshold at KDEN: (a) bias and (b) mean absolute error. Event thresholds are indicated by the different colors.

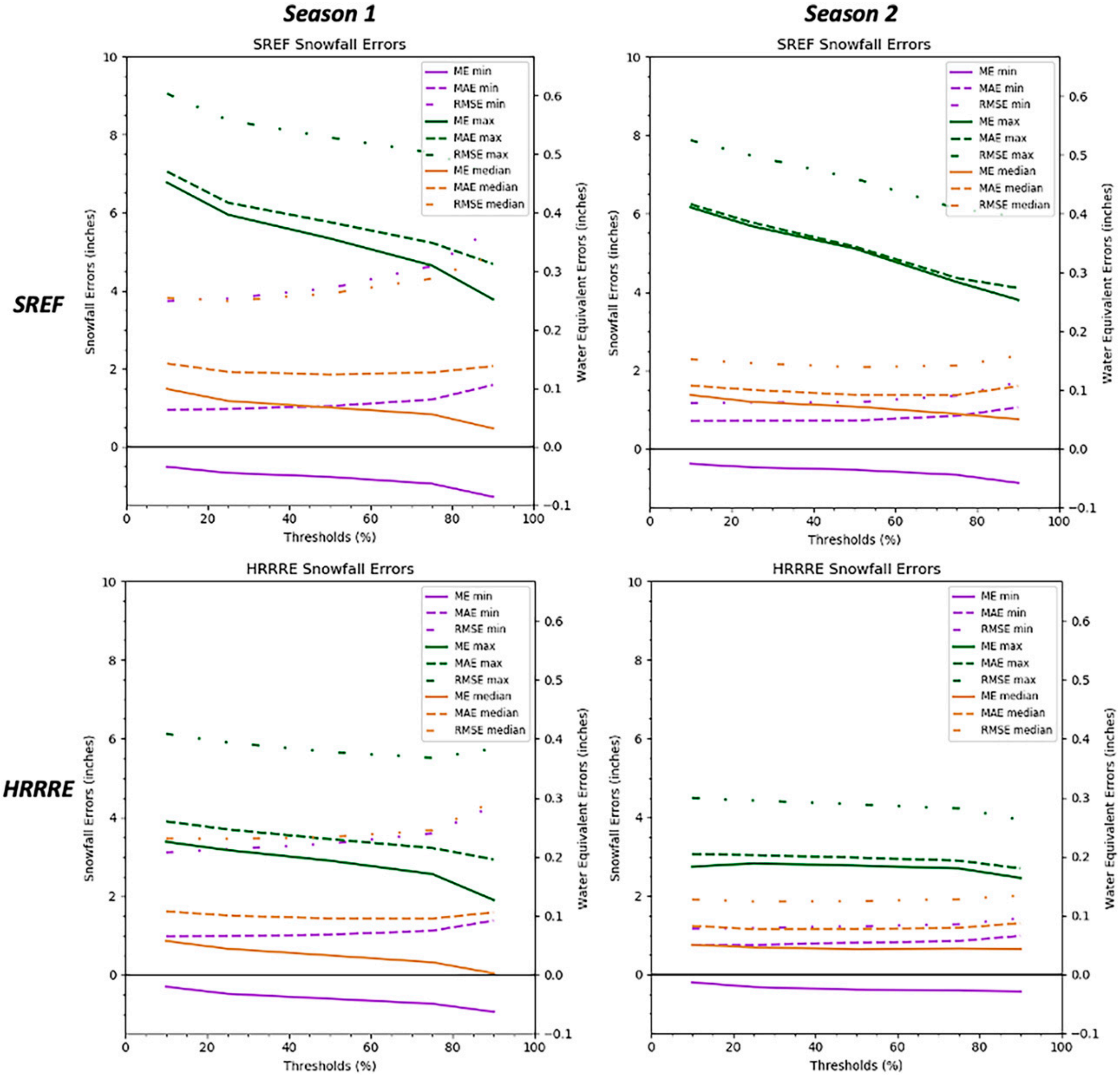


FIG. 22. Snowfall errors by event threshold for (left) season 1 and (right) season 2. Note, there were fewer airports included in season 1. (top) SREF results and (bottom) HRRRE results. The different colored lines represent the minimum, median, and maximum snow amounts from the members of the ensemble that had an event. The different line styles represent the different statistics: bias, MAE, and RMSE.

season 2 was that the errors were less influenced by the event threshold. In season 1, the bias curves had more of a negative slope highlighting the impact of changing the probability threshold (lower left plot in Fig. 22).

At KDEN, the HRRRE (Fig. 23) was on par with the PSA and the NBM 10% snow amount bias (Figs. 20 and 17). While the minimum snowfall (from the ensemble member with the least nonzero snow accumulation) from the HRRRE produced the smallest bias, it also underforecast snow amount. Choosing the median HRRRE snow amount resulted in a larger positive bias, but would allow airport decision-makers to be overprepared instead of underprepared, which was found to be their preferred mode (Morss et al. 2022). As was discussed previously with event timing, the airport also prefers to be conservative with snow amount and overestimate snowfall slightly to make sure they have enough personnel and supplies ready.

6. Summary and conclusions

In this study, the SREF, HRRRE, NBM, and PSA were evaluated with respect to snowfall forecast performance in the context of airport winter weather operations. The products were assessed using an impact-based event methodology and compared to METAR observations to determine snow

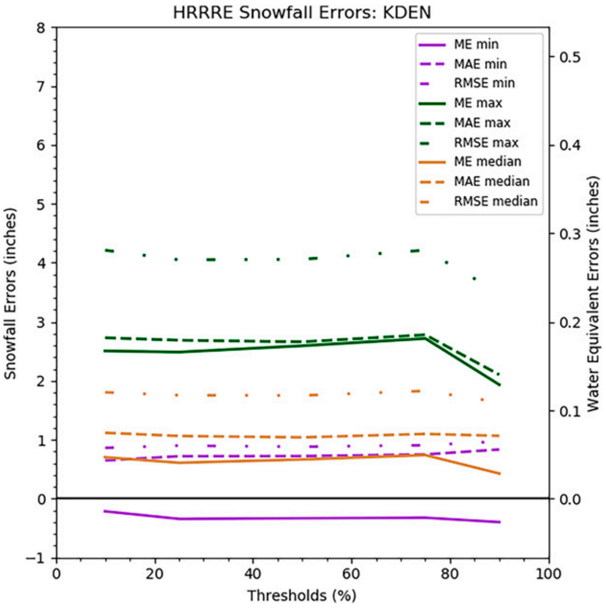


FIG. 23. As in Fig. 22, but for HRRRE snow amount errors at KDEN for season 2.

amount and timing errors. Overall, all of the probabilistic forecast products effectively provide information that would allow DIA operations to select forecast thresholds according to their needs: most accurate, most conservative, etc. The first specific conclusion from this assessment was that the NBM had quite different characteristics than the SREF, HRRRE, and PSA. The NBM had fewer events and rarely used higher probabilities, unlike the other products which more often had 75% and 90% events. The NBM had the best timing of snow events. While other forecast products also had a median timing error of zero, the range of NBM errors was narrower. With regard to snow amount, the NBM had mostly positive mean errors which increased with forecast probability threshold, particularly for low accumulation observed events (from a trace to 1 in.). The SREF and HRRRE median snow errors were more consistent across event threshold and observed snowfall amount. Additionally, the NBM had higher errors at KDEN compared to all airports in aggregate.

The PSA performed similarly to the NBM and HRRRE. Its snow amount bias, which was independent of event threshold, matched that of the optimal NBM and HRRRE thresholds. Like the NBM, the PSA did not have many missed events and those that did occur had minimal snowfall. The false alarms produced by the three products were also mostly minor events (i.e., had low forecast snow amounts). The similarities in performance among these forecast systems can partially be attributed to their dependence; the human-generated forecast is not independent of the others (except for the HRRRE as it was not operational) and the deterministic HRRR is an input to the NBM.

Additionally, an interannual comparison of the HRRRE and SREF highlighted that there were not meaningful changes in snowfall forecast performance between the 2018/19 and 2019/20 winter seasons. This conclusion was reached considering that SREF, whose configuration remained the same, had similar variations between the two seasons as the other systems did. It is worth mentioning that improvements may be seen if the successor to the HRRRE, the Rapid Refresh Forecast System (RRFS), is run more frequently than every 12 h as it was during this evaluation.

One shortcoming of this work should be noted, that despite expanding the domain to two dozen Intermountain West airports (with roughly similar weather to KDEN) over two winter seasons, there were few heavy snow events (nine events of 6+ in. in season 1 and 14 events in season 2). Therefore, the results are dominated by events accumulating a few inches or less, which are not the types of events that cause major problems for the KDEN operations, and so any conclusions about forecast performance for heavier snow can only be provisional. It is also possible that the apparent triviality of the false alarms and missed events (i.e., that they present little impact to airport decision-making) is at least in part due to this dominance of minor snow events. There is greater reason for confidence in the timing error results, as one might expect the smaller events to be preceded by a weaker signal and thus have greater uncertainty.

This study not only resulted in the determination of snowfall forecast errors of various probabilistic systems, but also prompted further product development and strengthened the relationship between KDEN and the NWS Boulder WFO. The PSA was refined after season 1 to better match the airport's needs, specifically with higher temporal resolution, and with accumulation thresholds consistent with those triggering KDEN's updated operational alert levels. Additionally, the airport operations team reported an increase in PSA usage.

After the SREF is retired and a HRRRE-like ensemble becomes operational within the RRFS framework, it will be important to reevaluate the capabilities of the available probabilistic forecast systems to provide useful information in the context of aviation and airport decision-making. A future evaluation could also incorporate a dynamic or temperature-adjusted SLR instead of a climatological value to account for the variations in conditions geographically, seasonally, etc.

Acknowledgments. The authors would like to acknowledge Rebecca E. Morss, Heather Lazrus, and Julie Demuth from the National Center for Atmospheric Research for their work investigating airport winter weather decision-making processes and views on probabilistic information, which helped inform the methodology selected for this study. Similarly, we thank the operations and emergency personnel from KDEN for their collaboration and feedback: Dave Cunningham, Jason Milam, and Mike Carlson. Last, the authors thank Zoltan Toth and three anonymous reviewers whose comments substantially improved the manuscript. Dana M. Uden is supported by funding from NOAA Award NA19OAR4320073.

Data availability statement. The data employed in this study were collected through a data feed and for the dates of the study, are no longer archived on a publicly accessible platform with the exception of METAR data and the PSA. METAR data can be obtained through the Iowa State University website: https://mesonet.agron.iastate.edu. The PSA text product is available at https://forecast.weather.gov/product.php?site=BOU&product=OPU&issuedby=BOU. Iowa State University has a long-term archive of the PSA (product OPUBOU) at https://mesonet.agron.iastate.edu/wx/afos/list.phtml. The HRRRE is not yet operational and thus data were obtained directly from the developers (contact: David Dowell at David.dowell@noaa.gov), while the SREF was retrieved from the NCEP Products Inventory, which maintains a rolling stream of several days of data.

REFERENCES

Baxter, M. A., C. E. Graves, and J. T. Moore, 2005: A climatology of snow-to-liquid ratio for the contiguous United States. *Wea. Forecasting*, **20**, 729–744, https://doi.org/10.1175/WAF856.1.

Benjamin, S., 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.

Craven, J. P., D. E. Rudack, and P. E. Shafer, 2020: National blend of models: A statistically post-processed multi-model ensemble. *J. Oper. Meteor.*, **8**, 1–14, https://doi.org/10.15191/nwajom.2020.0801.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410, https://doi.org/10.1175/WAF-D-11-00011.1.

Demuth, J. L., and Coauthors, 2020: Recommendations for developing useful and usable convection-allowing model ensemble information for NWS forecasters. *Wea. Forecasting*, **35**, 1381–1406, https://doi.org/10.1175/WAF-D-19-0108.1.

De Pondeca, M. S. F., and Coauthors, 2011: The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612, https://doi.org/10.1175/WAF-D-10-05037.1.

Dowell, D., C. Alexander, T. Alcott, and T. Ladwig, 2018: HRRR Ensemble (HRRRE) guidance 2018 HWT Spring Experiment. Global Systems Laboratory, 6 pp., https://rapidrefresh.noaa.gov/internal/pdfs/2018_Spring_Experiment_HRRRE_Documentation.pdf.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Regional ensemble forecast systems at NCEP. *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/Manuscript/Paper273421/NWP2015_NCEP_Regional-Ensembles_paper.pdf.

Global Systems Laboratory, 2021: CHANGE LOG: Real-time, experiment HRRR Data-Assimilation System (HRRRDAS) and ensemble forecast (HRRRE) on jet. Global Systems Laboratory, accessed 5 January 2021, https://rapidrefresh.noaa.gov/internal/log/change/HRRRE_changes.txt.

Kim, J., H. Chun, R. D. Sharman, and T. L. Keller, 2011: Evaluations of upper-level turbulence diagnostics performance using the Graphical Turbulence Guidance (GTG) System and Pilot Reports (PIREPs) over East Asia. *J. Appl. Meteor. Climatol.*, **50**, 1936–1951, https://doi.org/10.1175/JAMC-D-10-05017.1.

Lee, D., H. Chun, and J. Kim, 2020: Evaluation of multimodel-based ensemble forecasts for clear-air turbulence. *Wea. Forecasting*, **35**, 507–521, https://doi.org/10.1175/WAF-D-19-0155.1.

Mahringer, G., 2008: Terminal aerodrome forecast verification in Austro Control using time windows and ranges of forecast conditions. *Meteor. Appl.*, **15**, 113–123, https://doi.org/10.1002/met.62.

Meteorological Development Laboratory, 2021: About the National Blend of Models (NBM). National Weather Service, accessed 6 January 2021, https://vlab.noaa.gov/web/mdl/nbm-versions.

Morss, R. E., H. Lazrus, J. L. Demuth, and J. Henderson, 2022: Improving probabilistic weather forecasts for decision making: A multi-method study of the use of forecast information in snow and ice management at a major U.S. airport. NCAR Tech. Note NCAR/TN-573+STR, 43 pp., https://doi.org/10.5065/wyv1-wq11.

Murphy, A. H., and E. S. Epstein, 1967: Verification of probabilistic predictions: A brief review. *J. Appl. Meteor.*, **6**, 748–755, https://doi.org/10.1175/1520-0450(1967)006<0748:VOPPAB>2.0.CO;2.

NWS, 2018: NCEP Environmental Modeling Center (EMC) 3-year implementation plan: FY2018-2020. NWS, 82 pp., https://www.weather.gov/media/sti/nggps/EMC%20Implementation%20Plan%20FY18-20%20v1.docx.pdf.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: FACETs: A proposed next-generation paradigm for high-impact weather forecasting. *Bull. Amer. Meteor. Soc.*, **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

Rudack, D. E., 2020: An historical overview of NOAA's National Blend of Models (NBM). *26th Conf. on Probability and Statistics*, Boston, MA, Amer. Meteor. Soc., 7.3, https://ams.confex.com/ams/2020Annual/webprogram/Manuscript/Paper364390/AMS Extended Abstract _2020_NBM-History-11-25-19._Figures-atEnd_DB.pdf.

——, and J. E. Ghirardelli, 2010: A comparative verification of Localized Aviation Model Output Statistics Program (LAMP) and numerical weather prediction (NWP) model forecasts of ceiling height and visibility. *Wea. Forecasting*, **25**, 1161–1178, https://doi.org/10.1175/2010WAF2222383.1.

Scheuerer, M., and T. M. Hamill, 2019: Probabilistic forecasting of snowfall amounts using a hybrid between a parametric and an analog approach. *Mon. Wea. Rev.*, **147**, 1047–1064, https://doi.org/10.1175/MWR-D-18-0273.1.

Stauffer, R., G. Mayr, J. Messner, and A. Zeileis, 2018: Hourly probabilistic snow forecasts over complex terrain: A hybrid ensemble postprocessing approach. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **4**, 65–86, https://doi.org/10.5194/ascmo-4-65-2018.

UCAR, 2019: NBM v3.2 winter weather guidance. The COMET Program, accessed 4 February 2023, https://www.meted.ucar.edu/mobile/moddescription.php?id=1460.