Evaluation of a Probabilistic Subfreezing Road Temperature Nowcast System Based on Machine Learning

MICHAEL E. BALDWIN,[a,b] HEATHER D. REEVES,[a,b] AND ANDREW A. ROSENOW[a,b]

[a] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/National Severe Storms Laboratory, Norman, Oklahoma*

(Manuscript received 9 August 2023, in final form 30 October 2023, accepted 31 October 2023)

ABSTRACT: Road surface temperatures are a critical factor in determining driving conditions, especially during winter storms. Road temperature observations across the United States are sparse and located mainly along major highways. A machine learning–based system for nowcasting the probability of subfreezing road surface temperatures was developed at NSSL to allow for widespread monitoring of road conditions in real time. In this article, these products were evaluated over two winter seasons. Strengths and weaknesses in the nowcast system were identified by stratifying the evaluation metrics into various subsets. These results show that the current system performed well in general, but significantly underpredicted the probability of subfreezing roads during frozen precipitation events. Machine learning experiments were performed to attempt to address these issues. Evaluations of these experiments indicate reduction in errors when precipitation phase was included as a predictor and precipitating cases were more substantially represented in the training data for the machine learning system.

SIGNIFICANCE STATEMENT: The purpose of this study is to better understand the strengths and weaknesses of a system that predicts the probability of subfreezing road surface temperatures. We found that the system performed well in general, but underpredicted the probabilities when frozen precipitation was predicted to reach the surface. These biases were substantially improved by modifying the system to increase its focus on situations with falling precipitation. The updated system should allow for improved monitoring and forecasting of potentially hazardous conditions during winter storms.

KEYWORDS: Statistics; Forecast verification/skill; Transportation meteorology; Machine learning

1. Introduction

Winter weather–related vehicle crashes have resulted in approximately 1000 fatalities per year in the United States (Tobin et al. 2022a). Numerous studies have noted an elevated risk of vehicle crashes during winter precipitation and slippery road conditions (e.g., Tobin et al. 2019; Malin et al. 2019; Theofilatos and Yannis 2014; Strong et al. 2010; Andrey et al. 2003). Significant efforts are made by road management agencies to reduce these risks, such as de-icing chemical application, plowing, and road sign messaging, with snow and ice removal expenditures from state and local agencies recently totaling over \$4 billion yr^{-1} (Federal Highway Administration 2022). Road weather information systems (RWIS) can provide valuable guidance to transportation managers for decisions regarding when and how to maintain roadways and minimize icy conditions. RWIS stations collect and communicate various weather and road surface parameters using a set of environmental sensors at a given location (Manfredi et al. 2008). Studies of the cost effectiveness of RWIS have found substantial benefits of RWIS implementation, affecting issues such as improved efficiency of road maintenance operations and estimated safety enhancements (e.g., Sharma 2022; Veneziano et al. 2014).

Road surface temperature T_R is a key factor in determining road conditions, especially when precipitation is reaching the surface. RWIS observations of T_R are collected in many states from fixed stations and mobile sensors, although station coverage is sparse and sensors tend to be located on major highways. To allow for more universal monitoring and analysis of potentially hazardous conditions across the United States, including lower capacity roadways or other locations without RWIS stations, a machine learning–based system for nowcasting the probability of subfreezing T_R known as ProbSR (Handler et al. 2020) was developed. This system has been running in experimental mode at the National Severe Storms Laboratory (NSSL) in real time for several winter seasons, has been available to operational forecasters, and has been used in experimental products related to weather impacts on transportation that are under development (Tobin et al. 2022b).

In this article, ProbSR was evaluated over two recent winter seasons and results were analyzed to better understand the performance characteristics of the system. In general, comprehensive evaluation of weather analysis and forecast products is necessary to monitor the performance of the prediction system. Stratifying the verification information (Murphy 1995) using variables relevant to the forecasting process can provide additional insights into the quality of the forecasts, allowing for further investigation and potential improvements in the forecast system. Relevant subsets of the evaluation results were compiled in this work, which allowed for identification of deficiencies in the system,

Corresponding author: Michael Baldwin, michael.baldwin@noaa.gov

DOI: 10.1175/WAF-D-23-0137.1

particularly when frozen precipitation was predicted to reach the surface. Machine learning (ML) experiments were performed to attempt to address issues identified by this evaluation. The various datasets that were analyzed are described in section 2. Evaluation results from the operational system as well as the machine learning experiments are discussed in section 3. Section 4 presents an example case demonstrating the performance of the updated machine learning system. Section 5 provides a summary of this work and a discussion of future directions for this research.

2. Data

Gridded probabilities of subfreezing road surface temperatures have been generated in near real time within the experimental Multi-Radar Multi-Sensor (MRMS) system at NSSL (Zhang et al. 2016) over the past several winter seasons. ProbSR values were computed by a random-forest algorithm that used HRRR 2-h forecast parameters as predictors (Handler et al. 2020). A modified version of this system was implemented prior to the 2021/22 cold season (October–March) and was trained on data from the 2018/19 and 2019/20 cold seasons using T_R observations from RWIS stations. This version included several modifications from the system described by Handler et al. (2020), which are described briefly here. Hyperparameters of the random-forest model were tuned using cross validation over each month of the cold season using the average precision score as the performance metric [equivalent to the area under the performance diagram curve (AUPDC; Flora et al. 2021)]. The random forest was trained using 16 predictors (Table 1) from version 3 of the HRRR model (Fovell and Gallagher 2020). Note that these predictors were a simplified subset of the 30 predictors used in the original ProbSR system (Handler et al. 2020). The cross-validated probabilities were calibrated using isotonic regression over the full training dataset. This tuning/training/calibration was performed using the Scikit-learn Python package (Pedregosa et al. 2011). The output from the random-forest system was interpolated from the HRRR grid to the MRMS grid (0.01° grid spacing) for display and archive purposes. Figure 1a presents the average ProbSR over the domain across two winter seasons (October–March 2021/22 and 2022/23). This shows that larger ProbSR values were found more often toward the north and in higher terrain.

RWIS road surface temperature observations were collected from MesoWest (Horel et al. 2002) across the United States over two cold seasons (October–March) during 2021/22 and 2022/23. Observations taken within ±15 min of top of each hour were averaged together. Values at RWIS stations with multiple sensors (i.e., multiple lanes, bridge decks, and ramps) were averaged across all available road temperature sensors at a station. These analysis procedures are identical to those processed by Handler et al. (2020) to generate hourly T_R observations for training and testing of the machine learning system. RWIS stations were quality controlled to remove stations with large numbers of persistent (>24 h) or missing values (>250 h) within each month or large differences between the RWIS T_R and values at nearby stations (absolute value of monthly mean difference between nearby stations >10°C). 1404 stations were included during the 2021/22 season and 1128 stations were included in the 2022/23 season. The average frequency of subfreezing road temperatures at each station is displayed in Fig. 1b, indicating higher frequency at more northern locations as well as stations at higher elevation. The station distribution varies considerably across the United States and several northern states appear to have no road temperature observations. These states may in fact have a network of RWIS stations deployed, but did not routinely communicate their observations or make them publicly available to MesoWest, so these regions are missing in our analysis.

TABLE 1. Input predictors from 2-h HRRR forecast fields to the ProbSR random-forest algorithm.

Input predictors	Input predictors
Surface temperature (T_{sfc})	2-m temperature (T_2)
Friction velocity	10-m wind speed (gust)
Latent heat flux	Sensible heat flux
Consecutive hours below freezing T_{sfc}	Consecutive hours above freezing T_{sfc}
Consecutive hours below freezing T_{2m}	Consecutive hours above freezing T_{2m}
Downward shortwave radiation flux	Downward longwave radiation flux
2-m dewpoint	Mid–cloud cover percentage
No. of days from 10 Jan	Urban land use/land cover flag

To evaluate ProbSR nowcasts, hourly gridded fields were acquired from the NSSL archive throughout the 2021/22 and 2022/23 October–March periods and values from the nearest MRMS grid point were paired with each RWIS station. Operational HRRR (version 4; Dowell et al. 2022) 2-h forecasts were obtained from the NOAA AWS archive throughout this period (Blaylock 2022), nearest grid point values of several forecast parameters (surface temperature, 2-m temperature/dewpoint, 10-m wind gusts, sensible/latent/ground heat fluxes, low/mid/high/total cloud cover, precipitation rate, and surface radiation fluxes) were also paired with each RWIS T_R observation.

For this work, road temperature observations from the previous 5-yr period (2016–21) were used to develop a station-based hourly "climatology" of the frequency of subfreezing road surface conditions. RWIS T_R observations were collected across this 5-yr period. For each station and date/hour, the average frequency of subfreezing road surface temperature was calculated using rolling windows of ±10 days and ±2 h. An example is provided in Fig. 2 for RWIS station KSKL (Skyline WY 230). For a given date (12 December in this example) there were 21 days across 5 years of observations within the ±10-day window, 105 observations in total. The ±2-h window results in 525 possible temperature observations for analysis at that particular date and hour. For KSKL at 1800 UTC 12 December, 308 of the 525 RWIS T_R values were below 0°C, resulting in a value of 0.59 for the climatology for this particular station, hour, and day. This hourly, station-based climatology will be used in this work as a baseline for comparison with ProbSR nowcasts and machine learning experiments in the sections that follow.

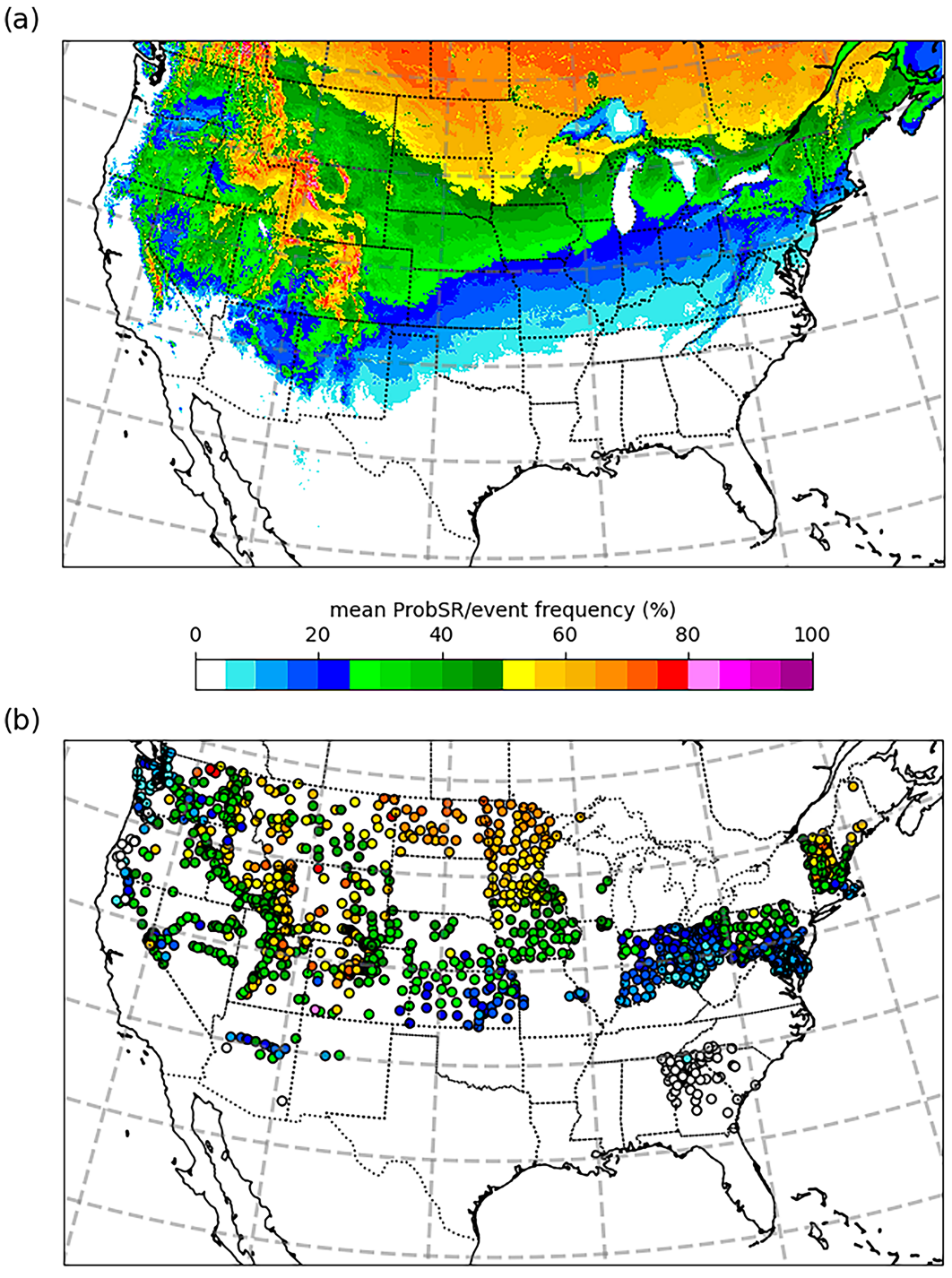

FIG. 1. (a) Mean ProbSR (%) over two cold periods (October–March of 2021/22 and 2022/23). (b) RWIS station distribution and subfreezing road temperature event frequency for the quality-controlled set of stations active during the same periods.

3. ProbSR evaluation results

a. Summary verification metrics

Bias and mean-square error (MSE) are widely used metrics in forecast evaluation. For forecasts expressed in terms of probabilities, MSE is commonly referred to as the Brier score (Brier 1950). Probability forecasts can have any value from 0 to 1. Following the notation of Murphy (1988) the forecast probability value for the ith case will be denoted as f_i and the observed probability value for the ith case will be denoted as

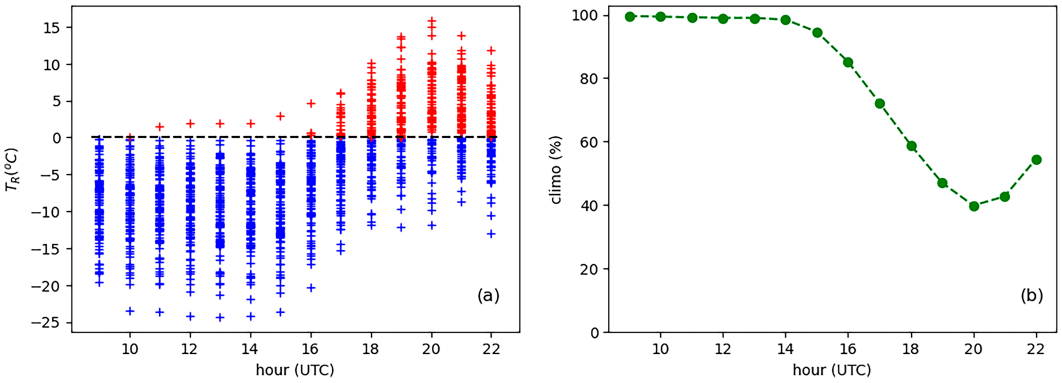

FIG. 2. (a) Observed T_R values (°C) at RWIS station KSKL ("WY 230–Skyline") for a rolling window of ±10 days centered at 12 Dec over a 5-yr period (2016–21) (red symbols: $T_R \geq 0$°C; blue symbols: $T_R < 0$°C). (b) Station-based hourly climatology of subfreezing T_R (%) at KSKL corresponding to these observations.

x_i (in our case, $x_i = 1$ when the observed $T_R < 0$°C and $x_i = 0$ for $T_R \geq 0$°C). Bias and MSE from the set of all cases in the quality-controlled dataset are defined as

$$\text{bias} = \frac{1}{N}\sum_{i=1}^{N}(f_i - x_i) = \overline{f} - \overline{x} \quad \text{and}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(f_i - x_i)^2,$$

where N is the number of cases in the dataset. For the full two-season evaluation period, 9.3 million matched pairs of forecasts and observations were available for evaluation. ProbSR had a bias of −0.013, indicating a slight underprediction of the probability of subfreezing road surfaces, or a slight warm bias. ProbSR also had an MSE of 0.055, which indicates a high level of overall accuracy since this is near the expected value of MSE for a perfect forecast (=0.0). To provide some context for a particular value of MSE, a skill score can be calculated by comparing the MSE of the forecast system with one obtained from a "reference" forecast system, such as the above-defined climatology or a competing forecast. Traditionally for probability forecasts, the mean observation (base rate) is used as the reference forecast, making the reference forecast unbiased and single valued. This version of the skill score is known as the Brier skill score (BSS) and is defined as

$$\text{BSS} = \frac{\text{MSE} - \text{MSE}_{\text{ref}}}{\text{MSE}_{\text{perfect}} - \text{MSE}_{\text{ref}}} = 1 - \frac{\text{MSE}}{\text{MSE}_{\text{ref}}}, \quad \text{with}$$

$$\text{MSE}_{\text{ref}} = \frac{1}{N}\sum_{i=1}^{N}(\overline{x} - x_i)^2$$

A perfect forecast system will have BSS = 1, negative values of skill are found for forecast systems with larger MSE values than the reference forecast. For ProbSR across the two seasons the traditional form of BSS was 0.76, indicating at first glance a high level of skill.

It is important to note that a very simple reference forecast is used in the traditional BSS calculation to establish a baseline for determining skill. Mason (2004) showed how using a constant reference forecast to determine skill can provide misleading results. Hamill and Juras (2006) also highlighted several issues with using single-valued reference forecasts in determining skill, particularly when the verification dataset pools temporal and spatial subsets with widely varying climatologies. More realistic and informative estimates of skill can be obtained using reference forecasts with higher fidelity. Murphy (1988) discussed alternatives to using a single-valued reference forecast in MSE-related skill scores, such as multivalued reference forecasts that are generated using data external to the verification dataset. Following Murphy (1988), we use the station-based hourly climatology described in the previous section as a multivalued reference forecast. This is consistent with the prior information that would be available to a typical decision-maker. MSE of this climatological reference forecast was calculated as follows (μ_i represents the hourly station-based climatological value corresponding to the observed value x_i):

$$\text{MSE}_{\text{climo}} = \frac{1}{N}\sum_{i=1}^{N}(\mu_i - x_i)^2.$$

In this case, $\text{MSE}_{\text{climo}}$ for the 2-yr evaluation period was 0.117, using this value for the reference forecast provides a modified Brier skill score of 0.53, indicating a moderately high level of skill overall.

The diagrams in Fig. 3 provide additional information regarding the performance of ProbSR across the 2021/22 and 2022/23 cold seasons. The attributes diagram (Hsu and Murphy 1986; Fig. 3a) displays the observed frequency of the predicted event conditioned on the forecasted probability against the forecast probability. This indicates that ProbSR was highly reliable, subfreezing road surface temperatures were observed at nearly the same frequency as the predicted probabilities. Evidence of excellent discrimination capability can be found in the receiver operating characteristic (ROC) curve (Mason 1982; Harvey et al. 1992; Fig. 3b), which displays probability of detection (POD; ratio of correct "yes" forecasts to observed "yes" events) against probability of false detection (POFD; ratio of incorrect "yes" forecasts to observed "no" events) using a set of increasing forecast thresholds. This

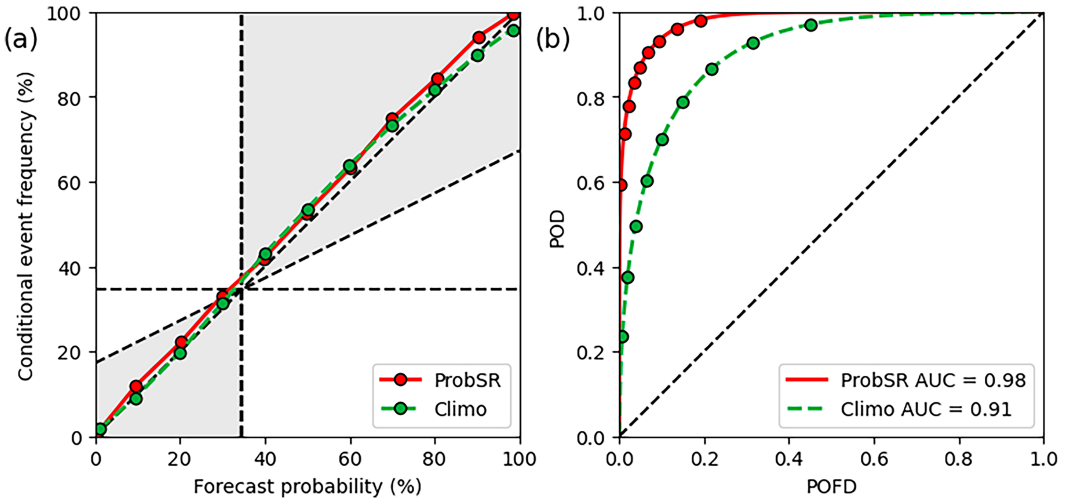

FIG. 3. Traditional verification metrics for the full domain and the 2021–22 and 2022–23 seasons (quality-controlled stations displayed in Fig. 1): (a) attributes diagram and (b) ROC curve (ProbSR: red; station-based hourly climatology: green). Here, AUC indicates area under the curve.

chart indicates that ProbSR was able to separate subfreezing and above-freezing observed events very successfully. Overall, the various summary measures of performance show that ProbSR provided nowcast information of very high quality during these two winter seasons.

b. *Evaluation results stratified by HRRR surface temperature and time of day*

The summary performance measures indicate that ProbSR provided higher quality information than a station-based hourly climatology for nowcasting subfreezing road surface conditions. Weaknesses in a forecasting system can be challenging to discover, particularly when evaluation metrics consist of summaries of large sample sizes. More informative evaluation results can be obtained by conditioning the verification statistics on additional variables (or covariates), a process known as stratification (Murphy 1995). Selecting covariates that are related to the forecasting process and assessing the conditional distributions of the verification metrics can provide useful insight into the performance of the forecast system, such as identifying conditional biases and situations that are more (or less) challenging for the forecast system.

We will begin this analysis by computing subsets of ProbSR evaluation statistics across a range of categories of HRRR surface temperature T_{sfc}. The verification dataset was divided into subsets using 1°-wide T_{sfc} categories, and verification metrics were computed for each of these subsets (Fig. 4). MSE and bias values are displayed at the center of each bin. For example, the value shown at 2.5°C represents all of the forecast–observation pairs found within the $2° \leq$ HRRR $T_{sfc} < 3$°C category. This allows for discovery of changes in behavior of the forecast system as a function of a covariate (T_{sfc} in this case). ProbSR displayed slight changes in bias

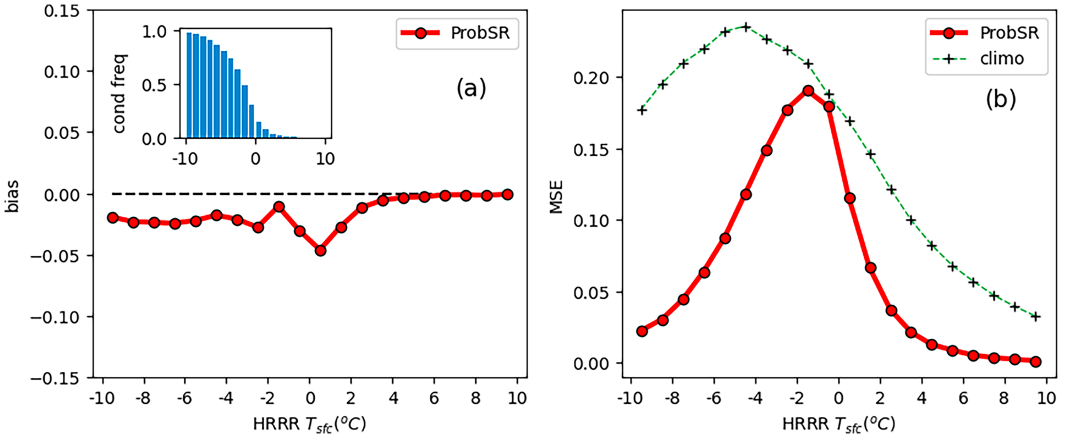

FIG. 4. (a) ProbSR bias and conditional frequency of $T_R < 0$°C (histogram inset) stratified by HRRR T_{sfc} (°C). (b) MSE stratified by HRRR T_{sfc} (°C).

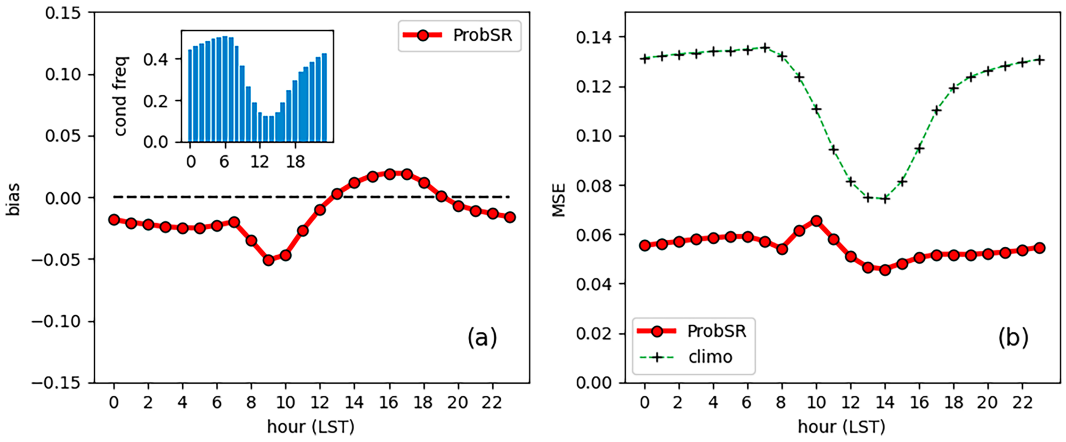

FIG. 5. (a) ProbSR bias and conditional frequency of $T_R < 0$°C (histogram inset) stratified by hour of the day [approximate local solar time (LST)]. (b) MSE stratified by hour of the day.

across the range of HRRR T_{sfc} categories, with low (warm) biases at colder temperatures and nearly zero biases for warmer temperatures. As HRRR T_{sfc} approaches values near and just below 0°C, the conditional frequency of $T_R < 0$°C approaches 0.5, indicating that these situations represent the most uncertain portion of the forecasting domain. Corresponding MSE values peaked in these situations and decreased considerably as T_{sfc} values moved farther away from 0°C. Some context can be provided for these MSE values by comparing ProbSR MSE values with those obtained from using the station-based hourly climatology as a reference forecast. MSE values approaching the climatological reference were found for ProbSR at HRRR T_{sfc} values near and just below 0°C. This indicates that ProbSR was providing very little additional information beyond the hourly climatology and had nearly zero skill in these situations. However, for warmer and colder T_{sfc} values, ProbSR MSE values were considerably lower than the climatology reference, indicating that nowcasts in these situations had significant positive skill relative to the local climatology.

The diagrams in Fig. 5 present evaluation metrics stratified by time of day. At each station, local solar time (LST) was approximated by subtracting the station longitude divided by 15° from UTC time and rounding the result to the nearest hour. This analysis allows for discovery of changes in behavior of the nowcast system as a function of time of day. These results indicate a slight diurnal cycle in bias for ProbSR (Fig. 5a). During the overnight hours ProbSR bias was consistent at a value slightly below zero. The negative (warm) bias became most pronounced in the morning hours, before quickly switching to a positive (cold) bias during the afternoon hours. The afternoon positive bias diminshed rapidly at around local sunset (1800 LST). ProbSR MSE values (Fig. 5b) were fairly consistent throughout the evening and overnight hours, with a slight peak in MSE occurring around 1000 LST, coinciding with the most negative (warm) bias values. MSE values obtained from using the station-based hourly climatology as a reference indicate the lowest MSE values during the early afternoon hours, coinciding with the time of day when the conditional frequency of $T_R < 0$°C reaches a minimum. These results indicate that ProbSR was most *skillful* (greatest reduction in MSE relative to the local climatological reference) during the overnight hours and most *accurate* (lowest MSE) during the early afternoon.

c. *Evaluation results stratified by precipitation type*

To analyze the performance of ProbSR in situations critical to winter road maintenance decision-making, such as when frozen precipitation reaches the surface or when liquid precipitation falls onto a relatively cold surface, results are further stratified by precipitation phase. Specifically, HRRR 2-h forecasts of the "percent of frozen precipitation" (CPOFP; Benjamin et al. 2020) were used to categorize the phase of precipitation predicted by the model to reach the surface. A CPOFP threshold of 5% was used to separate mainly liquid precipitation from situations containing some frozen precipitation. Dry conditions (defined as hourly precipitation rate less than 0.0001 mm h^{-1}) were also evaluated as a separate subset. During the 2-yr evaluation period a large majority of locations and times were in the "dry" subset (95.4%), with the precipitating locations and times roughly evenly split between the "liquid" (2.4%) and "frozen" (2.2%) phases. Figure 6 shows the bias and MSE statistics stratified by HRRR precipitation type, HRRR T_{sfc}, and time of day. Since the sample sizes for the precipitating subsets were diminshed, especially for unusual situations (i.e., liquid precipitation with very cold T_{sfc} or frozen precipitation with very warm T_{sfc}) 95% confidence intervals were estimated for each statistic using the Student's t distribution and curves were truncated for sample sizes less than 0.01% of the full dataset. These results clearly indicate that ProbSR had substantially different performance characteristics when frozen precipitation reached the surface than it displayed during either liquid precipitation or dry conditions. For dry conditions, bias and MSE as a function of both HRRR T_{sfc} and time of day behaved nearly the same as what was found previously with the full dataset, with a slight diurnal cycle in bias and peak MSE values for HRRR T_{sfc} conditions near and slightly below 0°C. For liquid

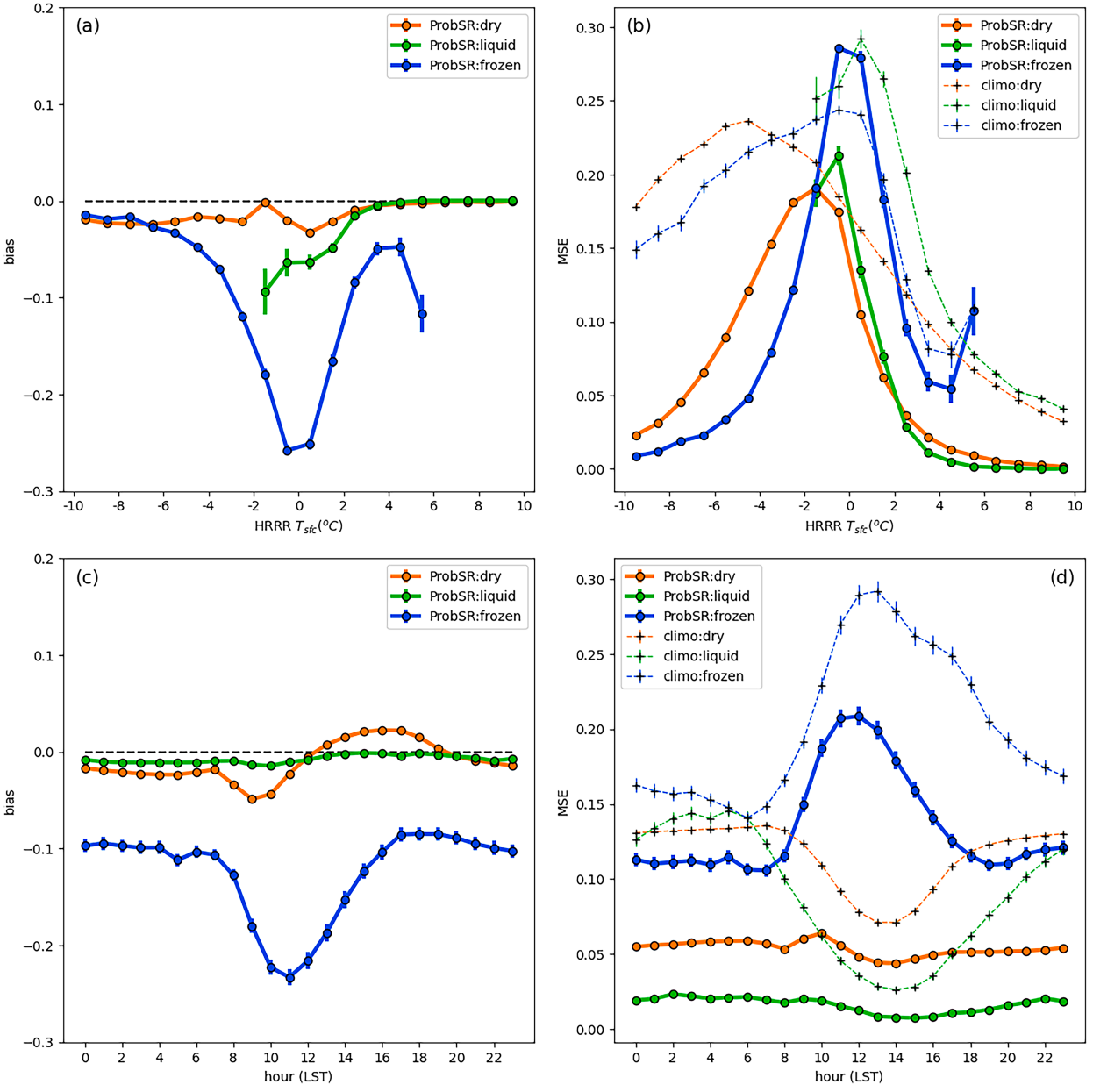

FIG. 6. The 2021–22 and 2022–23 ProbSR performance statistics stratified by precipitation type (orange: no precipitation "dry," green: CPOFP $<$ 5% "liquid," and blue: CPOFP $\geq$ 5% "frozen"): (left) bias and (right) MSE statistics by (a),(b) HRRR T_{sfc} (°C) values and (c),(d) hour of day (LST). Error bars represent the 95% confidence interval for each statistic estimated using the Student's t distribution. Statistics are not plotted for subsets of sample size less than 0.01% of the full dataset (N = 9 298 791).

precipitation, ProbSR biases (Fig. 6a) tended to become more negative with colder HRRR T_{sfc} values, indicating an underestimation of the probability of subfreezing road surface temperatures when liquid precipitation was falling onto relatively cold HRRR T_{sfc} conditions. MSE values for liquid precipitation were similar to those for dry conditions, reaching a peak for the HRRR T_{sfc} category between −1° and 0°C. ProbSR performance for liquid precipitation did not appear to have a diurnal cycle (Figs. 6c,d). For frozen precipitation situations, ProbSR showed a large negative (warm) bias for HRRR T_{sfc} values in a several degree neighborhood of 0°C. MSE values also peaked in this range of HRRR T_{sfc} conditions, exceeding those obtained using the hourly station-based climatology as a reference forecast, indicating that the forecasts were not skillful in those situations. ProbSR behavior during frozen precipitation also contained a significant diurnal signal, with negative (warm) biases as well as larger MSE values occurring around solar noon. These deficiencies that were discovered within precipitating cases are especially important when considering critical scenarios in winter road

maintenance decision-making, such as freezing rain and snowfall accumulation on road surfaces.

Without a more thorough evaluation of the HRRR model, it is difficult to connect the performance issues of the random-forest system to specific error characteristics in the HRRR. The results that we have obtained show that the connections between HRRR input features and RWIS T_R behave differently during dry conditions than in precipitating situations. It is not clear from these results if this is due to issues with physical parameterizations in the HRRR or differences in the behavior of road surface temperatures in these situations. The relatively small sample of precipitating cases in the original training data most likely limited the system's ability to learn about these differences.

d. Machine learning experiments

Given the characteristics of these errors and the importance of determining whether or not road surfaces are subfreezing when precipitation is reaching the surface, we have great motivation to improve the performance of ProbSR. Potential sources of error for this nowcasting system are proposed next. The version of ProbSR evaluated in this work was trained using data during 2018/19 and 2019/20 winter seasons, prior to the HRRR version-4 upgrade (implemented 4 December 2020) using the 16 predictors listed in Table 1. Changes to the HRRR have likely affected the input predictors enough to require updated training of the system. None of the predictors were directly related to precipitation phase. In addition, the training dataset consisted mainly of dry conditions, while situations with precipitation reaching the surface represented only a small sample of cases in the training data, which may have limited the system's ability to effectively capture differences in the behavior of the input features between dry and precipitating situations. For instance, data from the 2021/22 and 2022/23 seasons show that HRRR T_{sfc} values were 2.7°C lower on average than corresponding RWIS T_R values during dry conditions, while only 1.0°C lower when precipitation was predicted to reach the surface. These factors suggest multiple modifications to ML training for experimentation: increasing the relative proportion of precipitating cases in the training data and adding predictors related to precipitation phase reaching the surface.

For the following experiments, random-forest algorithms were tuned/trained using the same procedures that were used to train the most recent version of the system (outlined in section 2: hyperparameter tuning using cross validation over each month, average precision score for the performance metric, isotonic calibration over the training dataset) with training data from the 2021/22 season (HRRR version 4). Evaluation of these experiments was performed using the full quality-controlled dataset from the 2022/23 season. The training data for the "control" experiment were taken as a random sample of size 400 000 from the full quality-controlled RWIS data from the 2021/22 season (October–March). Dry conditions represented the vast majority of cases (95%) in the training data in the control experiment with liquid and frozen precipitation evenly split in the remaining cases. Although it was trained on a new dataset, the control experiment was intended to mimic the performance of the operational version of ProbSR, using the same 16 predictors and similar distributions of dry/precipitating training cases as were used by the most recent version of the system. Results from evaluation (using 2022/23 data) of the control experiment are shown in Fig. 7, indicating that the control experiment captured the general behavior of the errors in the operational ProbSR system. While biases and MSE values of the control experiment do not match ProbSR exactly, such differences are not surprising since the control experiment was trained using data from a different year and version of the HRRR model than the operational ProbSR system.

Three additional machine learning experiments ("training," "predictor," and "final") were run to test the proposed hypothesis that a low proportion of precipitating cases in the training data and lack of predictors related to precipitation phase were factors in the reduced performance of ProbSR. For the "training" experiment, an alternate training dataset was collected to test the impact of the relative distribution of precipitating cases within the training data on the ML. The alternate training dataset consisted of a sample size of 400 000 randomly drawn from the quality-controlled 2021/22 RWIS data, with 25% of those cases drawn from liquid precipitating cases, 25% from frozen precipitation, and the remaining 50% from dry conditions. The mean frequency of $T_R < 0$°C in this alternate training dataset was similar to what was obtained in the "control" experiment (32.9% vs 31.3% in control). The "training" experiment used this dataset along with the same 16 predictors as the control experiment (and operational ProbSR). The impact of adding predictors to the ML training was tested in the "predictor" experiment by including CPOFP to the predictor set, increasing the number of predictors to 17, while using the same training dataset as the control experiment. The combination of both of these factors (use of the alternate training dataset and the additional CPOFP predictor) was tested in the "final" experiment.

Results from evaluation of these ML experiments using 2022/23 data are shown in Fig. 8, stratifying the results by precipitation phase and HRRR T_{sfc}. These results show a substantial amount of sensitivity to the training data. For example, for dry conditions and HRRR T_{sfc} values near 0°C (Fig. 8a), biases switch from slightly negative in the control to slightly positive in the "training" experiment. There is a similar positive shift in bias for liquid (Fig. 8c) and frozen (Fig. 8e) precipitation subsets. The additional predictor appears to provide a benefical effect on the ML in these experiments, producing only slight changes in bias for dry and liquid precipitation cases and a significant positive change in bias for frozen precipitation. The combination of both factors ("final") appears to make the best adjustment to biases overall, especially in the frozen precipitation situations where the large negative bias in the control experiment for HRRR T_{sfc} values near 0°C was substantially reduced (Fig. 8e). The impact of these factors on MSE is fairly minor (and beneficial) for dry and liquid precipitation cases (Figs. 8b,d). For frozen precipitation situations (Fig. 8f), MSE values show

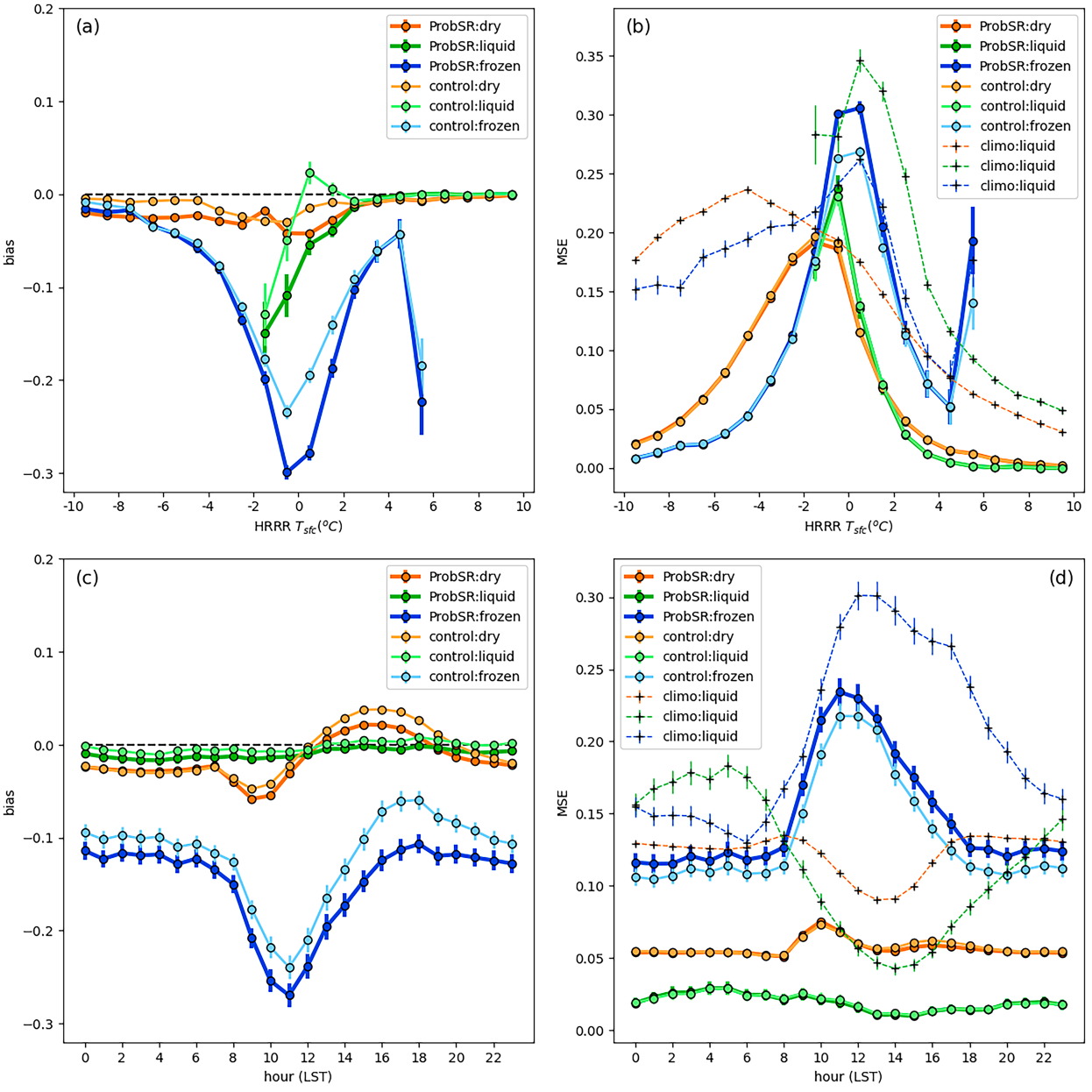

FIG. 7. The 2022–23 ProbSR and ML control experiment performance statistics stratified by precipitation type (orange: no precipitation "dry," green: CPOFP < 5% "liquid," and blue: CPOFP ≥ 5% "frozen"): (left) bias and (right) MSE statistics by (a),(b) HRRR T_{sfc} (°C) values and (c),(d) hour of day (LST). Error bars represent the 95% confidence interval for each statistic estimated using the Student's *t* distribution. Statistics are not plotted for subsets of sample size less than 0.01% of the full dataset ($N = 3\,800\,679$).

significant improvements for each factor in these experiments, especially for HRRR T_{sfc} values near 0°C. The combination of both factors ("final") results in the lowest MSE values overall.

Results from evaluation of these ML experiments stratifying the results by precipitation phase and time of day are shown in Fig. 9. The effect of the experimental training data appears to be a positive shift in biases, especially for dry and frozen precipitation cases (Figs. 9a,e). The additional predictor produced a similar positive shift of biases for frozen precipitation situations (Fig. 9e). The combination of both factors resulted in minor changes in bias for dry and liquid precipitation cases along with the best overall reduction in bias for frozen precipitation. Again, the impact of these factors on MSE is slightly beneficial for dry and liquid precipitation cases (Figs. 9b,d) and more significant for frozen precipitation situations (Fig. 9f). The combination of both factors results in the lowest MSE values overall, although the diurnal signals in MSE remained fairly consistent across the ML experiments.

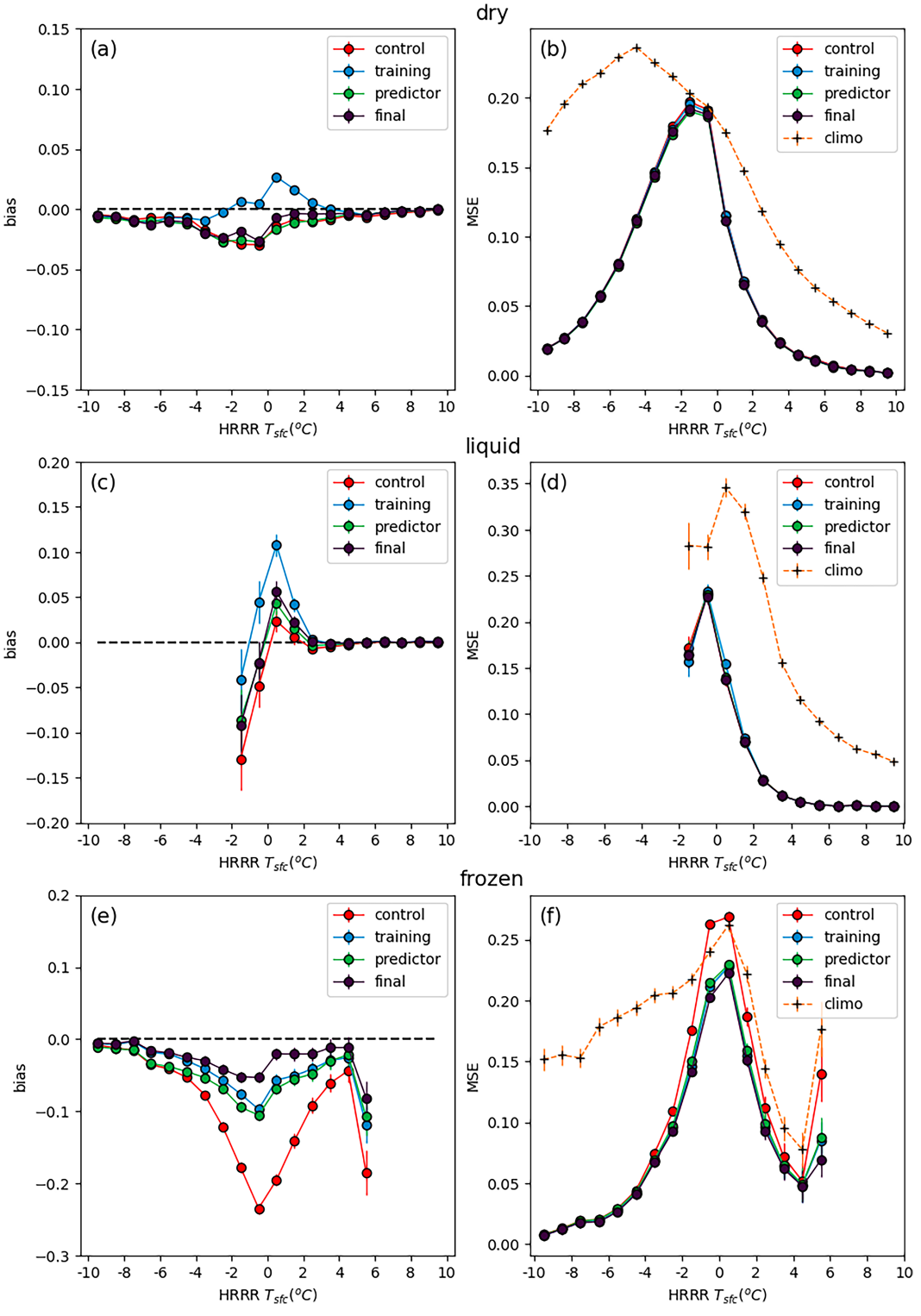

FIG. 8. Verification metrics [(left) bias and (right) MSE] for ML experiments as a function of HRRR T_{sfc} (°C), with results stratified by HRRR percent of frozen precipitation (CPOFP): (a),(b) no precipitation (dry); (c),(d) CPFOP $<$ 5% (liquid); and (e),(f) CPOFP $\geq$ 5% (frozen). Error bars represent the 95% confidence interval for each statistic estimated using the Student's t distribution. Statistics are not plotted for subsets of sample size less than 0.01% of the full dataset ($N = 3\,800\,679$).

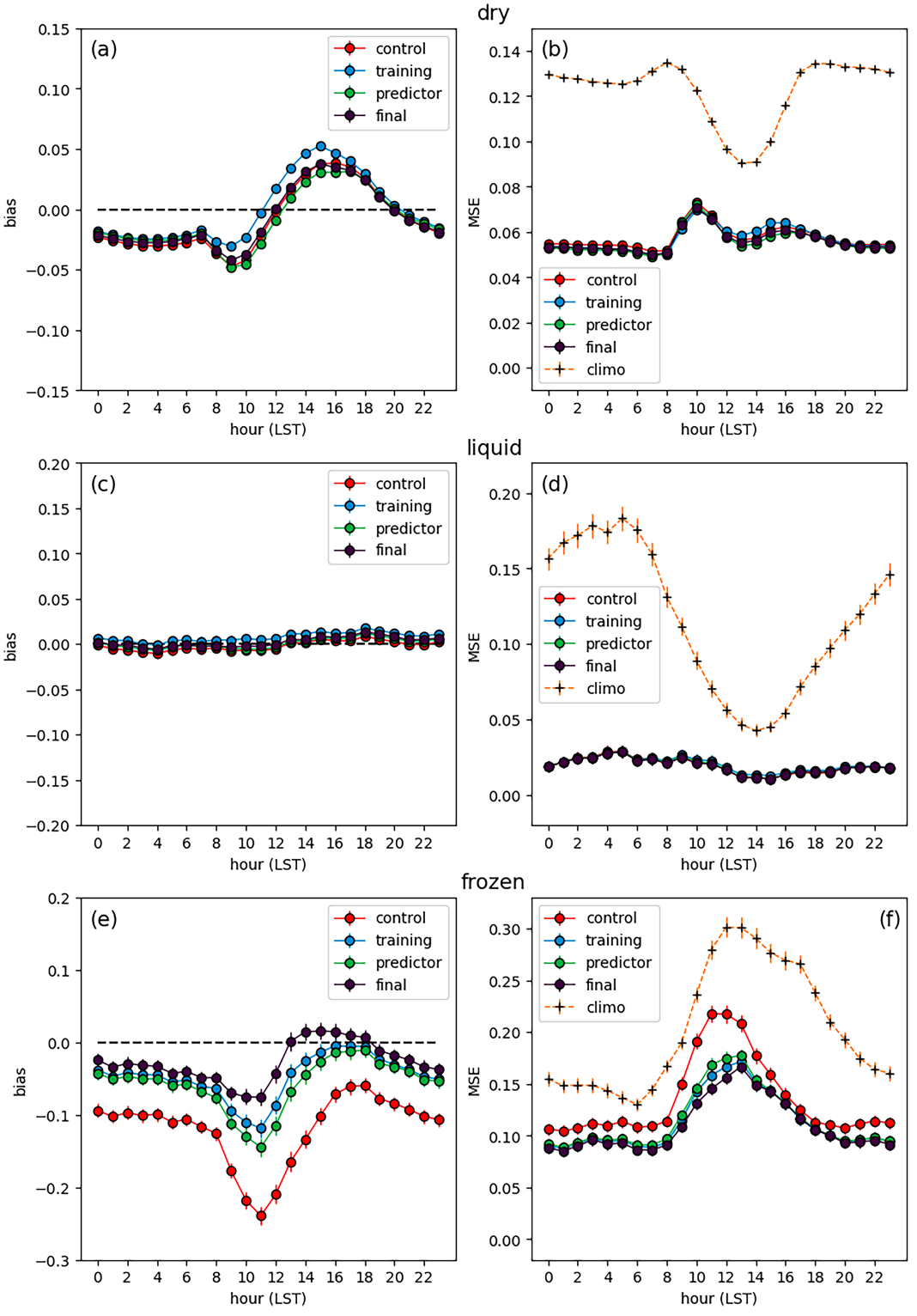

FIG. 9. Verification metrics [(left) bias and (right) MSE] for ML experiments as a function of hour of the day (LST), with results stratified by HRRR percent of frozen precipitation (CPOFP): (a),(b) no precipitation (dry); (c),(d) CPFOP $< 5\%$ (liquid); and (e),(f) CPOFP $\geq 5\%$ (frozen). Error bars represent the 95% confidence interval for each statistic estimated using the Student's t distribution.

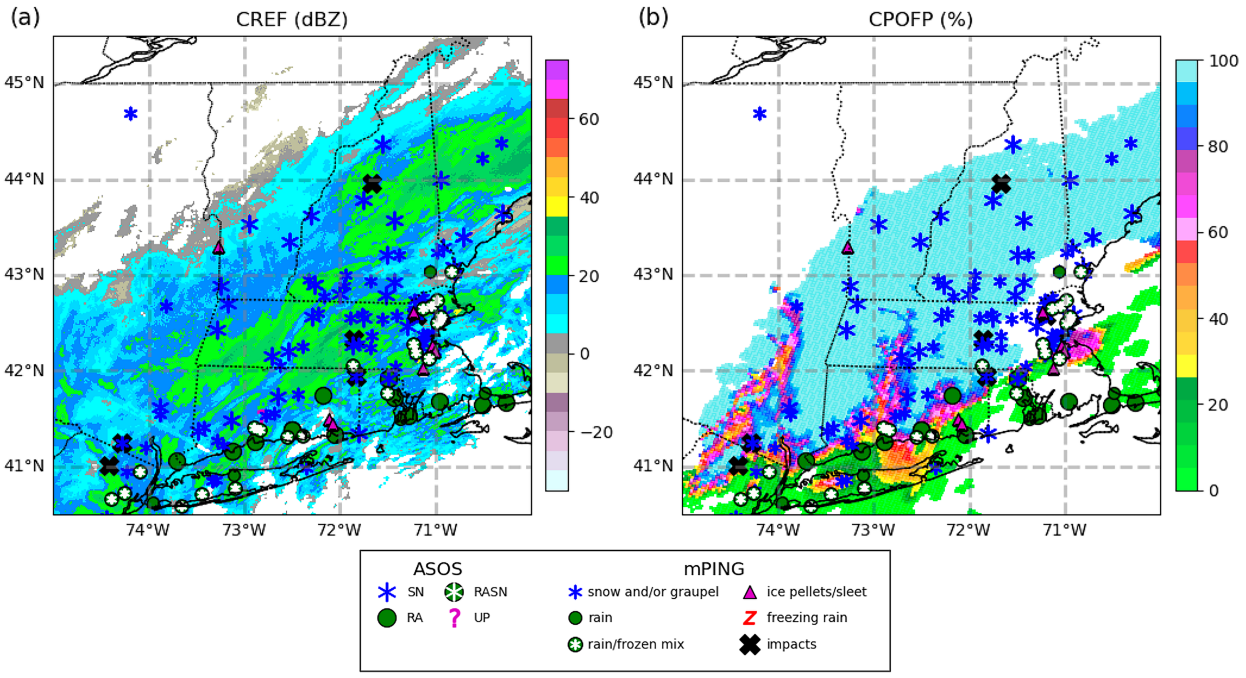

FIG. 10. (a) Composite reflectivity (CREF; dBZ) from NSSL MRMS, and (b) HRRR 2-h forecast of percentage of frozen precipitation (CPOFP; %) valid at 1800 UTC 23 Jan 2023. ASOS and mPING reports of precipitation/present weather within ±30 min of 1800 UTC 23 Jan 2023 are plotted as symbols [liquid (rain or drizzle): green circles, frozen (snow or ice pellets/sleet): blue asterisks, freezing (rain or drizzle): red letter Zs, mixed liquid/frozen: white asterisks inside a circle, and ASOS UP: magenta question marks]. ASOS stations are larger symbols, and mPING reports are smaller size symbols. The mPING reports of "Snow accumulating on roads and sidewalks" within ±30 min of 1800 UTC are plotted as large black X symbols.

4. Case study

A sample case will be presented to demonstrate the behavior of the modified system when frozen precipitation was falling onto relatively warm surfaces. While conclusive evidence of improved performance of a probabilistic forecast system cannot be determined from analysis of a single case, useful information regarding the behavior of the updated nowcast system in comparison with the previous system can often be obtained from such an analysis. On 23 January 2023, a developing extratropical cyclone moved along the Atlantic coast across the northeastern United States. A region of wintry precipitation was observed to the north of the cyclone track (Fig. 10a), at 1800 UTC observations from ASOS and mPING (Elmore et al. 2014) indicated frozen precipitation reaching the surface from eastern New York into central Maine. A transition zone from snow to a mix of rain and snow to liquid precipitation was observed across southern Connecticut into eastern Massachusetts. Multiple reports of snow accumulating on roads/sidewalks from mPING ("impacts") were scattered across this region, such as central New Hampshire, eastern Massachusetts, and southern New York. The HRRR 2-h forecast CPOFP field (Fig. 10b) represented the observed precipitation phase at the surface fairly well.

Hourly RWIS (average of multiple sensors per station and all reports within ±15-min window) T_R observations at 1800 UTC 23 January 2023 indicated subfreezing values over much of the region where frozen precipitation was reaching the surface (Fig. 11), such as central/southern New Hampshire and Vermont and western Massachusetts. RWIS T_R values transitioned to above-freezing values to the south of this area, mainly across Connecticut and eastern Massachusetts. ProbSR values from the "control" experiment were generally low, less than 25% across a large portion of the region where frozen precipitation was observed at the surface and T_R values were subfreezing. This behavior is consistent with the low (warm) bias of ProbSR during frozen precipitation situations that was determined in the previous section. The "final" ML experiment produced significantly higher values (generally greater than 60%) across the region where frozen precipitation reaching the surface and subfreezing T_R values were observed. Moderately high ProbSR values (30%–40%) were also generated in the "final" experiment across western Connecticut where HRRR CPOFP indicated frozen precipitation was reaching the surface, although observed T_R values were at or above 0°C. Similar "final" ProbSR values in the 30%–40% range were generated in southern New York/northern New Jersey in the area where mPING reports of "snow accumulating on roads and sidewalks" were found. While we do not have RWIS T_R observations in those locations, the mPING reports of accumulating snow appear to provide some justification for nonzero ProbSR values.

5. Summary

In this work, probabilistic nowcasts of subfreezing road surface temperatures generated from a machine learning–based

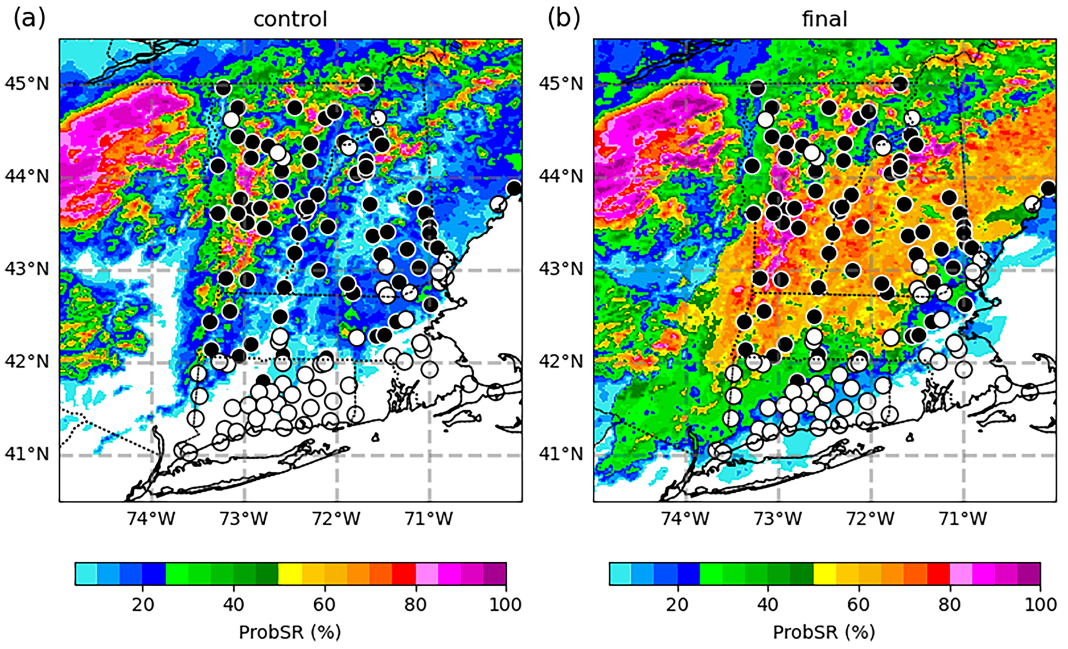

FIG. 11. Probability of subfreezing road surface temperature (%; color shades) valid at 1800 UTC 23 Jan 2023 [(a) "control" experiment and (b) "final" experiment]. Circle symbols indicate RWIS observations from quality-controlled stations, and sites with multiple sensors and/or multiple reports within ±15 min of 1800 UTC are averaged together (black: RWIS temperature < 0°C; white: RWIS temperatures ≥ 0°C).

system known as ProbSR were evaluated over two recent winter seasons. RWIS T_R observations were collected and analyzed over the previous five winter seasons to generate an hourly station-based climatology that was used as a reference forecast and a baseline for measuring the skill of ProbSR nowcasts. Standard summary forecast verification metrics found ProbSR to be highly accurate in general, with excellent reliability and ability to discriminate between observed subfreezing and above-freezing cases. Error metrics were stratified by HRRR T_{sfc}, time of day, and precipitation phase to help to identify deficiencies in the nowcast system. These results indicated that ProbSR had nearly zero skill in situations where the input HRRR T_{sfc} values were near and slightly below 0°C, but had significant positive skill for warmer and colder T_{sfc} values. ProbSR was generally the most accurate during the early afternoon hours, coinciding with the typical timing of maximum T_R values in the diurnal cycle. ProbSR had substantially reduced performance when frozen precipitation reached the surface relative to its performance during either liquid precipitation or dry conditions. A significant underprediction (warm) bias was found for frozen precipitation events, especially for input T_{sfc} values near 0°C and during afternoon hours. It was suggested that different versions of the HRRR model and a sparsity of precipitating events in the training data, as well as a lack of predictors related to precipitation phase were factors leading to the reduced performance during frozen precipitation situations. After updating the system to use HRRR, version 4, along with a higher proportion of precipitating cases in the training dataset as well as adding the percent of frozen precipitation as a predictor in the random forest, the deficiencies identified in our evaluation were substantially ameliorated.

The modifications to the ProbSR system found in the "final" machine learning experiment will be implemented at NSSL in the experimental real-time MRMS system in time for the 2023/24 winter season. Development of this system and other products related to the impacts of weather on the transportation system is expected to continue. Future work may include deep learning techniques, applications for extended forecast periods, enhancements to surface temperature predictions for bridge decks and other elevated surfaces, products related to the accumulation and melting rates of frozen precipitation on roadways, and using camera/image data to validate and build new models. We anticipate that this research will result in improved situational awareness and safety outcomes during high-impact winter storm events.

Acknowledgments. The authors thank three anonymous reviewers for their thoughtful suggestions that helped to substantially improve this paper. We also thank Patrick Adrian Campbell (CIWRO/NSSL), Alan Gerard (NSSL), and David Harrison (CIWRO/SPC) for providing many helpful comments and suggestions in their informal reviews of this paper. This study was made possible in part by the data made available by the governmental agencies, commercial firms, and educational institutions participating in MesoWest, which were obtained using the Synoptic Data PBC Mesonet API (https://github.com/mesowx/MesoPy). Archived HRRR forecasts in this work were obtained using code generously provided by Brian Blaylock's Herbie Python package (https://doi.org/10.5281/zenodo.4567540). Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement

NA21OAR4590162 and NA22OAR4590169, U.S. Department of Commerce.

Data availability statement. RWIS observations are archived and accessible online (https://mesowest.utah.edu). NOAA High-Resolution Rapid Refresh (HRRR) Model output (https://registry.opendata.aws/noaa-hrrr-pds) and output from ProbSR and the machine learning experiments analyzed herein (https://data.nssl.noaa.gov/thredds/catalog/WRDD/TAT/Data/ProbSR_eval/catalog.html) are available online.

REFERENCES

Andrey, J., B. Mills, M. Leahy, and J. Suggett, 2003: Weather as a chronic hazard for road transportation in Canadian cities. *Nat. Hazards*, **28**, 319–343, https://doi.org/10.1023/A:1022934225431.

Benjamin, S. G., E. P. James, J. M. Brown, E. J. Szoke, J. S. Kenyon, R. Ahmadov, and D. D. Turner, 2020: Diagnostic fields developed for hourly updated NOAA weather models. NOAA Tech Memo. OAR GSL-66, 55 pp., https://doi.org/10.25923/f7b4-rx42.

Blaylock, B. K., 2022: Herbie: Retrieve numerical weather prediction model data (version 2022.9.0). Zenodo, accessed 13 July 2022, https://doi.org/10.5281/zenodo.4567540.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

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

Elmore, K. L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, H. D. Reeves, V. Farmer, and L. P. Rothfusz, 2014: MPING: Crowd-sourcing weather reports for research. *Bull. Amer. Meteor. Soc.*, **95**, 1335–1342, https://doi.org/10.1175/BAMS-D-13-00014.1.

Federal Highway Administration, 2022: Highway Statistics Publications, Highway Finance Tables SF-4C and LGF-2, 2011 to 2021. Accessed 25 May 2023, https://www.fhwa.dot.gov/policyinformation/statistics.cfm.

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

Fovell, R. G., and A. Gallagher, 2020: Boundary layer and surface verification of the High-Resolution Rapid Refresh, version 3. *Wea. Forecasting*, **35**, 2255–2278, https://doi.org/10.1175/WAF-D-20-0101.1.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

Handler, S. L., H. D. Reeves, and A. McGovern, 2020: Development of a probabilistic subfreezing road temperature nowcast and forecast using machine learning. *Wea. Forecasting*, **35**, 1845–1863, https://doi.org/10.1175/WAF-D-19-0159.1.

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863–883, https://doi.org/10.1175/1520-0493(1992)120<0863:TAOSDT>2.0.CO;2.

Horel, J., and Coauthors, 2002: MesoWest: Cooperative Mesonets in the western United States. *Bull. Amer. Meteor. Soc.*, **83**, 211–226, https://doi.org/10.1175/1520-0477(2002)083<0211:MCMITW>2.3.CO;2.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram a geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, https://doi.org/10.1016/0169-2070(86)90048-8.

Malin, F., I. Norros, and S. Innamaa, 2019: Accident risk of road and weather conditions on different road types. *Accid. Anal. Prev.*, **122**, 181–188, https://doi.org/10.1016/j.aap.2018.10.014.

Manfredi, J., and Coauthors, 2008: Road weather information system environmental sensor station siting guidelines, version 2.0. FHWA Tech. Rep. FHWA-HOP-05-026, FHWA-JPO-09-012, 85 pp., https://rosap.ntl.bts.gov/view/dot/3290.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Mason, S. J., 2004: On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **132**, 1891–1895, https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424, https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

——, 1995: A coherent method of stratification within a general framework for forecast verification. *Mon. Wea. Rev.*, **123**, 1582–1588, https://doi.org/10.1175/1520-0493(1995)123<1582:ACMOSW>2.0.CO;2.

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

Sharma, D., 2022: Safety Effects of Road Weather Information Systems (RWIS)—A large-scale empirical investigation. M.S. thesis, Dept. of Civil and Environmental Engineering, University of Alberta, 81 pp., https://era.library.ualberta.ca/items/31c5a489-a8cd-4d39-9320-58c2ee98110b.

Strong, C. K., Z. Ye, and X. Shi, 2010: Safety effects of winter weather: The state of knowledge and remaining challenges. *Transp. Rev.*, **30**, 677–699, https://doi.org/10.1080/01441640903414470.

Theofilatos, A., and G. Yannis, 2014: A review of the effects of traffic and weather characteristics on road safety. *Accid. Anal. Prev.*, **72**, 244–256, https://doi.org/10.1016/j.aap.2014.06.017.

Tobin, D. M., M. R. Kumjian, and A. W. Black, 2019: Characteristics of recent vehicle-related fatalities during active precipitation in the United States. *Wea. Climate Soc.*, **11**, 935–952, https://doi.org/10.1175/WCAS-D-18-0110.1.

——, H. D. Reeves, M. N. Gibson, and A. A. Rosenow, 2022a: Weather conditions and messaging associated with fatal winter-weather-related motor-vehicle crashes. *Wea. Climate Soc.*, **14**, 835–848, https://doi.org/10.1175/WCAS-D-21-0112.1.

——, J. Kastman, and J. A. Nelson, 2022b: Developing an impact-based NWS product for surface-transportation hazards. *2022 Fall Meeting*, Chicago, IL, Amer. Geophys. Union, Abstract A52F-03, https://agu.confex.com/agu/fm22/meetingapp.cgi/Paper/1116144.

Veneziano, D., X. Shi, L. Ballard, Z. Ye, and L. Fay, 2014: A benefit-cost analysis toolkit for road weather management technologies. *Climate Effects on Pavement and Geotechnical Infrastructure*, American Society of Civil Engineers, 217–230, https://doi.org/10.1061/9780784413326.022.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.