Development of a Probabilistic Subfreezing Road Temperature Nowcast and Forecast Using Machine Learning

Shawn L. Handler* and Heather D. Reeves

Cooperative Institute for Mesoscale Meteorological Studies, The University of Oklahoma, and NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

Amy McGovern

University of Oklahoma, School of Computer Science, School of Meteorology, Norman, Oklahoma

*Corresponding author address: Cooperative Institute for Mesoscale Meteorological Studies, 120 David Boren Blvd Suite 2100, Norman, Oklahoma.

E-mail: shawn.handler@noaa.gov

Early Online Release: This preliminary version has been accepted for publication in *Weather and Forecasting*, may be fully cited, and has been assigned DOI 10.1175/WAF-D-19-0159.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

In this study, a machine learning algorithm for generating a gridded CONUS-wide probabilistic road-temperature forecast is presented. A random forest is used to tie a combination of HRRR model surface variables and information about the geographic location and time of day/year to observed road temperatures. This approach differs from its predecessors in that road temperature is not deterministic (i.e., provides a forecast of a specific road temperature), but rather it is probabilistic, providing a 0-100% probability that the road temperature is subfreezing. This approach can account for the varying controls on road temperature that are not easily known or able to be accounted for in physical models, such as amount of traffic, road composition, and differential shading by surrounding buildings and terrain. The algorithm is trained using road temperature observations from one winter season (Oct-Mar 2016-17) and calibrated/evaluated using observations from the following winter season (Oct-Mar 2017-18). Case-study analyses show the algorithm performs well for various scenarios and captures the temporal and spatial evolution of the probability of subfreezing roads reliably. Statistical evaluation for the predicted probabilities shows good skill as the mean area under the receiver operating characteristics curve is 0.96 and the Brier Skill Score is 0.66 for a 2-hr forecast and only degrades slightly as lead time is increased. Additionally, the algorithm produces well-calibrated probabilities, and consistent discrimination between clearly above-freezing and subfreezing environments.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

1. Introduction

On average, adverse road weather causes 5897 fatalities per year, making this the leading cause of weather-related fatalities in the United States (Pisano et al. 2018; Walker et al. 2018). Among the primary road hazards are slippery conditions associated with accumulating snow/ice, which only occur when the road temperature (T_R) is either subfreezing, T_R is less than the freezing point of de-icing chemicals that may have been applied, or if the rate of accumulating precipitation exceeds the rate of melting at the surface. This makes knowledge of T_R an important first step in anticipating whether icy/snowy roads may constitute a human health and safety threat.

The first instinct for estimating T_R may be to use numerical weather prediction (NWP) analyses and forecasts of soil temperature. However, soil temperature in a model will be dictated by the dominant land-use category in each grid box. Over most of the CONUS, this is some form of vegetation. But even in urban areas, where the land-surface model parameterizes processes typical of various impervious surfaces, soil-temperature forecasts may not reflect actual road temperatures. Recent work demonstrates this is true, showing that HRRR soil-temperatures typically underestimate the actual road temperature between 1-7°C, but differences as great as 14°C have been noted (Downing et al. 2020).

Several physical road models which predict T_R have been developed. The most well-known road-weather model in the United States is the Model of the Environment and Temperature of Roads (METRo; Crevier and Delage 2001). METRo is a physically-based approach for providing deterministic predictions of T_R developed by Environment Canada. It works in a similar manner to land-surface models in that the energy balance at the road/atmosphere interface is explicitly computed. It accounts for the various forcings that affect T_R, including the heat diffusivity and conductivity unique to the road composition, insolation, longwave radiation, and precipitation.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

METRo has been adapted for experimental use within the NWS and evaluated by Rutz and Gibson (2013). Using RWIS observations from western Montana and northern Idaho, they show that for T_R ranging from $\pm 5^{\circ}$C, METRo has mean errors ranging from -2.5 to 0.5°C suggesting it has a slight cold bias. Several other physical road models have been developed with similar performance statistics (e.g., Rayer 1987; Jacobs and Raatz 1996; Shao and Lister 1996; Sass 1997; Bouilloud et al. 2009; Yang et al. 2012; Fujimoto et al. 2012; Kangas et al. 2015). A range of statistical techniques for predicting T_R have also been attempted with varying degrees of success and applicability (e.g., Hertl and Schaffar 1998; Shao 1998; Juga et al. 2013; Toms et al. 2017).

The primary limitation of the the METRo model, or its counterparts, for CONUS-wide implementation is the need for in-pavement T_R observations to initialize the forecast. As will be shown, platforms providing in-pavement observations are not evenly distributed across the CONUS, with some states having no observations at all. Therefore, an algorithm independent of those observations which produces gridded CONUS-wide nowcasts and forecasts of the probability that T_R is subfreezing is presented.

This paper is organized as follows: Section 2 describes existing technology for measuring T_R and the justification for a probabilistic approach. Section 3 discusses the use of machine learning along with the data and methods used in this study. Section 4 highlights algorithm output along with comparisons to available observations through case-study analyses. Section 5 provides statistical results for nowcast and forecasting algorithms, and lastly Section 6 concludes with a summary and discussion.

2. Measurements and limitations of observed T_R

T_R is directly measured in some states via the Road Weather Information System (RWIS) network at a temporal resolution of 5 minutes. RWIS sites typically have sensors embedded flush

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

with the roadway surface with each sensor installed in the innermost wheel well of a lane on the roadway (Boselly 1993). These are the sensors for which T_R is defined in this study. Sensors can also be installed up to depths of 2-m underground (thus providing subsurface temperature data), an example being North Dakota RWIS sites, but these supplemental observations are not used in this study due to their limited deployment. The installation of RWIS sensors is at the discretion of individual state Departments of Transportation (DOTs) and, hence, they are not evenly distributed across the CONUS (Fig 1a). Note that there is very poor to no coverage in most southern-tier states. Even in those states that do have RWIS sites, the spatial distribution is nonuniform, with more urbanized areas/heavily-trafficked roads being better monitored than secondary or tertiary roadways. The most egregious example of a nonuniform distribution is Missouri, wherein sensors are only deployed along Interstates 70 and 44.

In those states with a sufficiently uniform distribution of RWIS sites, it is possible to perform an objective analysis of the T_R observations to create a gridded analysis, but the spatial representativeness of the observations is questionable. A typical distribution of T_R in the wake of a winter storm is provided over northern Ohio at 1100 UTC 14 November 2018 (Fig. 1b). T_R observations range from -5°C to 5°C across the region shown and from -4°C to 2°C in just the Cleveland metropolitan area. A high degree of spatial variability, such as in Fig. 1b, occurs frequently. This is demonstrated through comparison of T_R observations for RWIS sites that are within 5 km of each other during the winter months (Oct-Mar) of the $2016-17$ and $2017-18$ winter seasons. Only those sites indicated as red in Fig. 1a are used for this evaluation. This assessment includes 600,606 pairings. For those RWIS sites with multiple sensors (i.e., a sensor for each lane of traffic), the minimum T_R is chosen for the analysis. While the median absolute difference in T_R is 1.44°C, the absolute difference in T_R exceeds 3°C for 22.16% of the pairs. Another 11% of pairs have one site with an above-freezing T_R while the neighboring site is subfreezing. This variability indicates

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

there can be critically-important subgrid-scale variability that most high-resolution, deterministic, objective analyses and NWP analyses will not be able to capture.

Since RWIS sites can be equipped with separate thermometers for each lane of traffic, it is also possible to quantify the intersite variability, an example of which is provided at the OH112 RWIS site for 14 November 2018 (see Fig. 1b for location). One sensor has above-freezing T_R observations throughout the day, while the other has subfreezing T_R from 0500 to 1500 UTC (Fig. 1c). The difference in T_R between the two thermometers ranges from 1 to 6°C over the course of the day. Expanding this variability experiment to the same T_R observations used above shows that the median absolute difference in T_R between any two lanes is 1°C while the 75^{th} percentile difference is 2°C. Hence, temperature differences of 2°C or more between any two lanes at a single RWIS site is not uncommon and can happen 25% of the time. This opens up the possibility of one lane being subfreezing while an adjacent lane is above freezing for the same RWIS installation. An assessment of the frequency with which one lane has above-freezing T_R while a neighboring lane has subfreezing T_R is performed using the same dataset but only including those installations where all sensors indicated T_R within +/-5°C. This assessment includes 578,173 observation pairs. Individual sensors from a single installation disagree on whether T_R is subfreezing 14% of the time. These single-site and state-wide variations in T_R are not unexpected. Differing amounts of traffic, differential shading from nearby vegetation/construction, and even road preparation activities in advance of or during winter weather can all cause one lane of the roadway to have relatively higher or lower T_R.

In addition to the above variability and uncertainty, there is also the potential for the RWIS sensors to provide inaccurate measurements of T_R. Quality-control tests of the various in-pavement sensors used by the RWIS network in Ohio show a mean absolute error of 1°C compared to baseline tests (Scott et al. 2005). Field tests conducted by Scott et al. (2005) show that sensor accuracy

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

can be greatly impacted for different daily thermal cycles. Road temperature measurements from infrared (IR) thermometry and surface-mounted sensors can differ significantly, especially during snowy and icy road conditions (Jonsson and Riehm 2012). Since the installation and maintenance of RWIS sites is at the discretion of state DOTs, it is possible that varying degrees of accuracy can be exhibited from one state to another, especially if the sensors are provided by different manufacturers. The sensor readings are also affected by the work crew that installed them, which can differ even within the same DOT.

So while it is optimal to use only the best observations, it's important to realize that there is uncertainty in the RWIS temperature observations. Differences in the amount of shading by surrounding vegetation/buildings, road composition, and traffic counts can all cause T_R to change rapidly in short distances. None of these forcings can be reliably parameterized within an NWP model and, while they may be reasonably captured within physical road models that provide a single-site deterministic forecast, the site-to-site and intersite variability suggests that the application of deterministic road models to the entire CONUS may not be trustworthy. Hence, a probabilistic approach may be more appropriate.

3. Data and Methods

All of the above factors provide good justification for stepping away from deterministic approaches, and instead, viewing T_R in a probabilistic frame of reference. The scientific objective of this study is to produce a gridded, reliable analysis/forecast of the probability that T_R is subfreezing (T_{Rprob}). This is possible with machine learning (ML). A random forest (RF; Breiman 2001) ML model serves as the foundation for the algorithm. RFs have been shown to be successful in several different meteorological disciplines (e.g., Gagne et al. 2014; Elmore and Grams 2016;

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

Ahijevych et al. 2016; McGovern et al. 2017; Herman and Schumacher 2018), but to the best of our knowledge, have not been used for this purpose.

a. Random forest algorithm

RFs are most simply described as an ensemble of decision trees. Decision trees (Quinlan 1986, 1993), in turn, can be thought of as a mapping of possible outcomes to a series of yes-or-no questions. Individual decision trees on their own are prone to overfitting and high variance, and subsequently do not generalize well to new unseen data. RFs alleviate this issue by performing bootstrap aggregation, also called bagging. Only a certain number of features and training examples are used to train each individual tree. By doing this, the variance of a single tree collapses while only suffering a small increase in bias. One can use the mean from the forest for a deterministic solution, but by preserving the solutions from individual trees, a probabilistic value can be computed.

b. Algorithm inputs

The algorithm ingests 15 near-surface variables from the High Resolution Rapid Refresh (HRRR; Weygandt et al. 2009) and an additional 15 derived and static variables. Table 1 lists short descriptions for the features used in this study. These variables have been demonstrated to be important for dictating the road temperature in previous research (e.g., Crevier and Delage 2001). For the purposes of training, the 02-hr forecast is used because the latency in obtaining the data in real time($\sim$1 hr) allows for a 1-hr nowcast. For clarity, a 1200 UTC T_{Rprob} nowcast released at 1100 UTC is determined using the 1000 UTC 02-hr HRRR forecast. The T_{Rprob} nowcast is the primary focus of this study. However, it will be shown that these methods can be applied to longer lead times and thus the algorithm can be used prognostically (see section 5b). Since this study

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

is concerned with determining whether a road surface is subfreezing or not, only the cool season months, specifically 01 Oct - 31 Mar, are examined.

c. RWIS observations

The observed T_R from select RWIS sites (red dots in Fig. 1a) are used as the required target variable, which is needed for training the RF. These sites are used to give a representative sample of the various weather and climate regions within the CONUS. Some of the remaining RWIS sites (grey dots in Fig. 1) are used for case-study analyses. Only observations of T_R within 15 minutes prior to the top of the hour are considered in order to best match with the HRRR model outputs. A nearest-neighbor technique is used to pair each T_R observation to its corresponding HRRR grid point. Each RWIS T_R observation is binary encoded as 0 (1) to indicate T_R is greater than (less than or equal to) 0°C.

The complete dataset spans two cool seasons: 01 Oct 2016 to 31 Mar 2017 and 01 Oct 2017 to 31 Mar 2018 with a total of 8,616,744 1-hr observations collected. Because some of the RWIS observations are missing or are of insufficient quality, they are quality controlled as follows. First, all times across all sites with missing T_R observations are discarded from the dataset. Second, instances where the change in T_R exceeds 30°C hr^{-1} or where the change in T_R exceeds 50°C day^{-1} are also removed as they are most likely errant. Third, if the maximum and minimum T_R differed by less than 1°C day^{-1}, those days are also removed as even on overcast days, T_R should vary by more than this, according to manual analyses of typical T_R ranges on cloudy days. Lastly, given the uncertainties in T_R observations, RWIS sites reporting multiple T_R observations must all be either above or below freezing to be used in training.

After data preprocessing, the total number of T_R observations over the two-year period is 5,994,591 of which $\sim$27% of the samples are subfreezing. The distribution of the quality-

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

controlled T_R observations over the two seasons (Fig. 2) reveals that the highest number of subfreezing observations belongs to the months of December and January. While October has the fewest number of total subfreezing T_R observations at 21,173 ($\sim$2% of the entire month's distribution), all but two months suffer from class imbalance (i.e., one target response is more frequently observed than the other). Class imbalance can cause ML algorithms not to learn as effectively because the model will be biased to predict the most frequent class (Batista et al. 2004). The problem of class imbalance is addressed through the use of class weights such that the weights are inversely proportional to class frequencies (i.e., incorrect predictions of the minority class are penalized more heavily), along with restricting the training set only to include examples where T_R is between -5 and 5°C. This reduces the training set size in the 2016$-$17 winter season to 1,098,029 observations. Performing this additional step does help alleviate some of the class imbalance issue as now 43.7% of the total training samples have subfreezing T_R, but there is the caveat of not including potentially useful information from the neglected data. For the testing/validation set (2017$-$18 winter season), no subsampling is performed.

d. Dataset splitting and experiment design

Figure 3 provides an illustration of the experiment design. The overall dataset is split into tuning/training, testing, and probability-calibration datasets. The 2016$-$17 season is used for training and tuning the hyperparameters of the RF model. The most important hyperparameters to tune include the depth of each decision tree (Max_D), the minimum number of samples required to be at a leaf node (Min_{sl}), the minimum number of samples needed to make a split (Min_{ss}), the maximum number of features to be used for splitting (Max_{feat}), weighting of the class labels (Class_w), and the total number of trees in the forest (N_{est}). In total, 50 iterations of K=6$-$fold cross validation are performed over a predefined hyperparameter space (see Table 2 for range of values) to identify

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-19-0159.1.

the optimal parameters for the RF. Here, K=6 because cross validation is performed over each month of the winter season. This cross validation approach is more appropriate then randomly sampling points within the domain since time series data are typically autocorrelated and, therefore, one observation is not completely independent from previous timesteps. The area under the receiver operating characteristics (ROC; Metz 1978) curve (AUC) is chosen as the performance metric since it is insensitive to the class label distribution. The set of hyperparameters with the highest AUC averaged over all the 6 folds is used to construct the optimal base RF model (see bold values in Table 2). The 1,098,029 observations from the 2016−17 winter year are then used to train the base RF model using the best hyperparameter values.

Probability calibration and testing are performed using data from the 2017−18 winter season. Typically, the raw probabilities from ML models are not well-calibrated (Niculescu-Mizil and Caruana 2012). Isotonic regression (Niculescu-Mizil and Caruana 2012) is used for calibrating the probabilities from the base model. For every two weeks of calibration data, there is one week of testing data. A 72−hour (3−day) gap between the end of a two-week calibration block and the beginning of the one-week testing block is enforced to minimize data leakage from the calibration set into the test set. The 72−hour gap was chosen based on manual examination of T_R partial autocorrelation function plots from various RWIS sites, and the 72^{nd} lag corresponds to a correlation of zero (not shown). The exact splits of calibration and testing weeks are provided in Table 3. While the calibration and testing data are split from the same winter season, it is important to realize that those data are *completely* independent of the training set.

The final calibrated probabilities are produced in the following steps.

1. Produce uncalibrated probabilities for both calibration and testing sets using the base RF model.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

2. Train the calibration model (i.e, the isotonic regression model) using the 18 weeks of uncalibrated probabilities.

3. Calibrate the test set probabilities by passing the uncalibrated probabilities associated with the testing set through the calibration model (generated from step 2).

4. Output from the T_{Rprob} algorithm

In this section, three case studies are presented that highlight model performance for various regions within the CONUS and compare the output to available observations. Also examined are the features from the random forest considered most important in influencing T_R.

a. 04 February 2018: Multi-car pile-up in Missouri

A winter storm brought snow and ice to central Missouri on 4 February 2018 that resulted in over 130 roadway incidents on Interstate 44. One of these is a 12 car pile-up, which resulted in one fatality east of Lebanon, Missouri. This accident occurred near 2300 UTC that day.

Observations from the nearest Automated Surface Observing Station (ASOS) site in Lebanon, Missouri (KLBO; indicated in Fig. 4d) show generally-decreasing two-meter temperature (T_{2m}) leading up to the event (Fig. 4a). T_{2m} first becomes subfreezing around 1500 UTC. Also shown on Fig. 4a is the observed T_R from the MO009 RWIS site. T_R generally follows T_{2m} until sunrise, after which T_R increases to $\sim$5°C. Rapid cooling of T_R begins at 1800 UTC. At 1900 UTC, snow is reported at KLBO and continues until 0000 UTC the next day. The HRRR T_{2m} agrees well with the observed T_{2m}, showing a gradual decrease throughout the day (Fig. 4b). Like the observed T_R, the HRRR surface temperature (T_{sfc}), experiences a temporary increase after sunrise and then decreases after 1800 UTC. The HRRR solar radiation (S) is also included in Fig. 4b. It shows a

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

gradual increase between 1300 and 1800 UTC, followed by intermittent increasing and decreasing values associated with changes in cloud cover (not shown).

A time series of T_{Rprob} at the location of the accident shows reasonable correspondence to the HRRR input variables (Fig. 4c). Namely, probabilities increase as temperatures decrease or as cloud cover increases. Plan views of the T_{Rprob} valid at 2000 and 2300 UTC show generally increasing probabilities over Missouri resulting in T_{Rprob} ranging from 95-100% over most of the state by 2300 UTC (Figs. 4d,e). The observed T_R (overlaid in Figs. 4d,e) shows reasonable agreement with the algorithm output — areas with elevated probabilities have subfreezing T_R and areas with low-zero probabilities have mostly above-freezing T_R.

b. 02 April 2018: Snow in complex terrain

The second example occurs in early spring in Washington state. During the overnight to early morning hours of 1-2 April 2018, snowfall in the Cascade mountain range forced the closure of I-90, a major artery transecting the Cascade Mountains and into the Seattle metropolitan area.

The observed T_{2m} is below freezing for much of the period at the KSMP ASOS site (Fig. 5a; for location of ASOS and RWIS sites, see Fig. 5d). The observed T_R from the nearest RWIS site (TFRAN) was slightly above freezing in the early evening hours before becoming subfreezing overnight. The TFRAN site is located near Snoqualmie Pass for which Washington State Department of Transportation camera verification shows snow on the roadway during the overnight hours (not shown). Similar to the Missouri case, solar insolation causes T_R to increase dramatically after 1700 UTC (Fig. 5b). The HRRR T_{2m} has a slight warm bias relative to the observed temperature, but otherwise, shows a similar trend (Fig. 5b). HRRR T_{sfc} increases in response to strong increases in insolation starting near 1400 UTC. However, between 1400 and 1700 UTC, T_R remains

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

fairly constant despite nonzero solar radiation flux, thus it may be likely that during this time, the insolation was melting the accumulated snow.

The T_{Rprob} time series shows the same anticorrelation with T_R as the Missouri event (Fig. 5c). Plan views of T_{Rprob} are displayed for 0900 UTC and 1200 UTC 2 April 2018, both of which fall within the time window for when the portion of I-90 was closed (Fig. 5d,e). All RWIS sites that traversed through the Cascades along I-90 report subfreezing temperatures at these times. The T_{Rprob}-algorithm output compares well with observations, having probabilities that, at times, exceed well past 80%, indicating that subfreezing roads and, hence, accumulating snow, are likely if the roadway has not been previously treated.

c. 03-04 March 2019: Transition season case

The last example represents a transition season event in Maryland. On 3-4 March 2019, a wintery mix was anticipated for much of the state and nearby neighboring areas. Varying amounts of snowfall, with tight snowfall gradients, were forecast across the region.

The observed T_{2m} from the Gaithersburg, Maryland ASOS and observed T_R from RWIS site MD056 is provided in Fig. 6a. T_R and T_{2m} reach their maximum value between 1500 and 1800 UTC and begin falling soon after once the initial precipitation began falling at 1700 UTC. For much of the event, observed T_{2m} is between 0 and 5°C. Snow begins to fall near 1800 UTC lasting until 2100 UTC after which a mix of winter precipitation was reported. HRRR model fields show an increase in T_{sfc} and T_{2m} from 1200 UTC to 1600 UTC coincident with an increase in incoming solar radiation (Fig. 6b). As cloud cover increases, incoming radiation decreases as do the other HRRR model temperature fields and the observed T_R and T_{2m}. The T_{Rprob} time series accordingly shows very low values for much of the event, only peaking at 50% by the end of the forecast period, coincident with the HRRR temperature fields dipping below 0°C (Fig. 6c).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

Plan views of the T_{Rprob} output are provided at 2100 3 March 2019 UTC and 0600 UTC 4 March 2019 (Fig. 6d,e). At 2100 UTC, T_{Rprob} is low across much of the region with the exception of extreme northwest Maryland. A handful of RWIS sites report subfreezing roads with T_{Rprob} ranging from $40-80\%$ in that area. As the precipitation moves out of the region at 0600 UTC, T_{Rprob} remains low over the interior portions of Maryland (i.e., $T_{Rprob} < 25\%$) and higher in regions of northwest Maryland in which a larger consensus of RWIS sites report subfreezing T_R. Overall, T_{Rprob} performs well considering temperatures are close to freezing for much of the event.

d. Feature importance

From the preceding case-study analyses, it appears that some of the HRRR input features may be more important than others. For example, in Figs. 4b,c, local spikes in T_{Rprob} occur coincident with local spikes in insolation. The RF algorithm can provide information on which features are most important for the model to distinguish between class labels (i.e., subfreezing roadway versus not a subfreezing roadway). For this study, the rank of each predictor's importance is determined using two variations of permutation importance, labeled as single-pass and multi-pass (see McGovern et al. 2019). For each feature within the testing dataset, all instances are permuted, the model is scored with the new values, and then compared to the baseline score to quantify the reduction in skill (single-pass; Breiman 2001). The second method is similar, but aims to alleviate the potential issue of strongly correlated/redundant predictors (multi-pass; Lakshmanan et al. 2015). For the multi-pass method, once a feature is identified as important, its values stay permuted and the procedure outlined above is repeated until all features have been examined. Here, the skill score metrics used to assess importance are the Brier Skill Score (BSS) and AUC.

Figure 7 displays the top 10 most important predictors for both skill score metrics and both permutation approaches. Starting with the single-pass results, most of the temperature-based fields

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

are among the top ten most important features, as are some radiation features, such as the incoming solar radiation and the upward longwave radiation flux. However, the reduction in skill does not differ considerably from the original score for most of these features. It is possible that information in one feature may be strongly correlated with other features, and thus any correlated feature on its own could be deemed unimportant. When examining the multi-pass results, the top four most important features are the same between skill metrics, which are the surface temperature, upward longwave radiation flux, T_{2m}, and the number of hours T_{sfc} is subfreezing. The reduction in skill drops off dramatically if these predictors are removed such that the RF output is no better than a random forecast. These results are consistent with what one might intuitively expect as all of these top-most important features are obvious controls on T_R.

5. Statistical Analysis

Probabilistic forecasts are typically evaluated using the ROC curve, the attributes diagram (Hsu and Murphy 1986), and the performance diagram. Statistical metrics associated with these diagrams, such as the BSS and AUC, will be further discussed when appropriate. The focus here is on the nowcast prediction, but a forecasting perspective of the algorithm (using longer NWP model lead times), and sensitivity tests where the "freezing" threshold is modified will also be discussed.

a. Nowcast Performance

The ROC curve shows the probability of detection (POD) versus the probability of false detection (POFD). Ideally, the ROC curve should be as close to the upper left hand corner of the figure as possible (i.e, high POD, low POFD). The AUC is a single metric typically used to assess model performance. An AUC > 0.9 is considered "excellent" whereas AUC $<= 0.5$ is considered no

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

better than a random forecast (Luna-Herrera et al. 2003; Muller et al. 2005; Mehdi and Ahmadi 2011).

From Fig. 8a, the average AUC for the seven-week test set is ∼0.96, which falls within the "excellent" range. Each of the testing week curves hug the upper-left corner of the diagram, which is desirable. The individual ROC curves do show subtle variability, with week one (18−25 October) having the lowest individual AUC of 0.93 and the last testing period (29−31 March) having the highest individual AUC of 0.98. The higher AUC for the last testing period is attributed to the fact that there are many more true negatives (i.e., above freezing observations) relative to other weeks (and fewer observations), and thus the POFD is inherently smaller for this week. Even though the AUC for the first week is comparatively low, it still falls within the range of values considered to be "excellent".

The performance diagram shows POD versus success ratio (SR) with contours of frequency bias (FB) and critical success index (CSI) overlaid (Fig. 8b). SR and CSI can change significantly based on the distribution of class labels, which makes the performance diagram difficult to judge. Typically, maximum CSI should occur where the FB is close to 1 (Roebber 2009), and thus ideally, the curves should generally be in the upper right hand corner of the diagram. The performance diagram for the algorithm depicts maximum CSI of 0.69 at a FB of 1.04. The FB of 1.04 does indicate a subtle overforecasting bias, but this number is still close to zero and thus the algorithm is deemed to perform well. The week of 18−25 October is the only week that is far removed from the grouping in the upper right hand corner. There are fewer subfreezing T_R observations for the month of October compared to other months (see Fig. 2) and, thus, its performance may not be as strong owing to fewer samples (i.e., the class imbalance problem). Further inspection reveals that October remains underrepresented in the training set as only 3.2% of all training examples arise in October, of which only 4,470 examples are subfreezing. Also, over 85% of the observations

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

during the month of October were within regions of complex terrain, which may not be completely resolved by the HRRR model.

While the algorithm's performance over the climatological distribution of T_R is good, the reader may wish to know how well the algorithm performs for T_R within a range close to 0°C. Figures 8c,d show the ROC curves and performance diagram, respectively, of the algorithm for T_R observations in the range of -5 to +5°C. Generally speaking, the curves in each diagram do deviate from their "ideal" configurations, however the algorithm performs very well within this more confined T_R range. The mean AUC is 0.91, which is still considered excellent, and thus, the algorithm is able to discriminate between classes effectively.

The attributes diagram (Fig. 9a) shows the average forecast probabilities versus the conditional event frequency. The diagram features the reliability curve, a climatology line, the no-resolution line, the perfect-reliability line, and a no-skill line. If the reliability curve passes through the region bounded by the no-skill line and the climatological line (gray region in Fig. 9a), the classifier is deemed to have skill better than climatology. The BSS is typically used to compare the model Brier Score, which measures the accuracy of probabilistic forecasts (Wilks 2006), to a climatology forecast. For a model to be considered better than climatology, the BSS must be greater than zero.

The reliability curves for all weeks follow the ideal reliability curve and are considered skillful for all of the probability bin ranges. The calibrated output probabilities for most weeks appear not to have any significant under or over forecasting biases. The only exception being that of the week of 14−21 November, which does have a slight underforecasting bias for all probabilities less than 80%. The mean BSS for all weeks is 0.66 indicating the skill is better than climatology. Here, climatology is defined as the proportion of T_R observations below freezing for each respective month that a test week falls within. The week of 18−25 October has the lowest BSS of 0.44

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

compared to the highest week of 03−10 February with a BSS of 0.75 which agrees with the conclusions from the performance diagram.

Alongside the attributes diagram is a histogram of the forecast probabilities (Fig. 9b). The forecast probabilities from the the calibrated model are considered sharp with values maximized near 0 and 1 for low T_{Rprob} and high T_{Rprob}, respectively. T_{Rprob} in the mid-range (i.e., 30−70%) are less prominent.

Overlaid on Fig. 9b are a few of the most important features from the RF as determined from the feature importance section. When atmospheric temperatures are below (above) −8°C (5°C), the RF is very likely to assign high (low) T_{Rprob}. However, when atmospheric temperatures are between 0°C and −4°C, the mid-range probabilities tend to be most prominent. This may be attributed to error within the NWP model and thus the less certain mid-range probabilities are more common. A study by Reeves et al. (2014) showed that uncertainty in the temperature forecasts from a NWP model (on the order of a few degrees Celsius) when predicting certain precipitation types can greatly impact the validity of the forecasts. While forecasting precipitation type is not the goal or within the scope of this paper, their results support these and indicate that subtle NWP model errors on the order of just a few degrees can substantially impact any type of forecast, including the prediction of T_{Rprob}.

Also shown within Fig. 9b are the two derived features: the number of hours T_{sfc} and T_{2m} are subfreezing. As the number of hours of subfreezing temperatures increases, it is more likely the model will produce higher T_{Rprob}. When T_{sfc} has experienced subfreezing temperatures for ∼20 hours, T_{Rprob} is high, and almost certain at ∼50 hours. However, there are instances where the surface of the road can be forecast to be subfreezing only a handful of hours after T_{sfc} or T_{2m} become subfreezing. These situations may arise when air temperatures have been near freezing (but not subfreezing) for extended periods of time.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

Similar to Fig. 8, the attributes diagram and forecast probability distribution are shown for the T_R range of -5 to +5°C. The attributes diagram (Fig. 9c) is similar to that of Fig. 9a implying that the probabilities are still reliable and well-calibrated for this confined T_R range. The climatology has increased, which is evident by comparing the vertical dashed line in Fig. 9a and Fig. 9c. This increase acts to decrease the BSS, in part, from 0.66 to 0.51. However, this BSS value is still skillful. The forecast probability distribution (Fig. 9d) shows similar results to that of (Fig. 9b) as the forecast probabilities are considered sharp with values maximized near 0 and 1 for low T_{Rprob} and high T_{Rprob}, respectively.

Typically, a machine learning model is compared to a baseline model to see if the new model adds additional skill. One potential baseline model would be to use the top HRRR soil-temperature as a proxy for T_R, similar to Downing et al. (2020). Using soil temperature as a proxy for road temperature gives a POD of 0.8, FAR of 0.4, FB of 1.33, and a CSI of 0.52 over the test set. This kind of statistical evaluation is not as straightforward with T_{Rprob} output, because the output is not binary. Thus, it would be inappropriate do a direct comparison to show relative skill of one method over the other. However, the statistics presented in Downing et al. (2020), as well as this rather high FAR and FB computed over the test set, provide sufficient justification for the added skill of this algorithm over the soil-temperature approach.

b. Forecast Performance

Herein, we transition from a nowcasting to forecasting perspective. The 18-hr HRRR, 36-hr North American Mesoscale Forecast System (NAM) and Global Forecast System (GFS) data are used to produce 18-hr and 36-hr T_{Rprob} forecasts. The 18-hr HRRR forecast is the longest lead

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

time available at hourly intervals for the HRRR model.[1] The 36-hr NAM and GFS forecasts are chosen since state DOTs typically decide on a treatment plan for roadways 1-2 days in advance of a weather event. The 18-hr HRRR forecast data are evaluated using the previously discussed 1-hr nowcast algorithm. Because the NAM and GFS have different model variables, such as the number and depth of soil temperature layers, grid spacing, and biases compared to the HRRR model, individual RFs were trained for each NWP model, but using the same approach as was discussed in Section 3d. The variables included within each forecast model are chosen to maximize skill and are not meant to be identical to the nowcast algorithm.

Results are positive as each forecast maintains an adequate amount of skill when compared with the 1-hr nowcast. Loss of skill compared to the nowcast is represented by a decrease in BSS of $\sim$0.06 for the 18-hr HRRR forecast, and $\sim$0.09 for both the 36-hr GFS forecast and 36-hr NAM forecast. The probabilities are well calibrated and reliable as seen in the attributes diagram for the 18-hr HRRR forecast (Fig. 10a). However, the GFS and NAM 36-hr probabilities are not as well calibrated (10b-c). The performance diagram (Fig. 11a-c) also highlights the loss of skill with the forecasts compared to the nowcast. Fewer training examples for the NAM and GFS, due to fewer model runs (i.e., 6 hourly runs compared to hourly for the HRRR), does impact the performance of the model and less than optimal performance is more noticeable in the months of October and March compared to the nowcast. The 18-hr HRRR forecast is superior to both NAM and GFS 36-hr forecasts. The skillful forecasts are encouraging and could serve as a tool for forecasters when trying to assess short-range impacts on roads.

[1] As of October. 2019, the 36-hr HRRR forecast is now the longest lead time. At the time of this study, not enough 36-hr data were available for reliable ML.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

c. *Sensitivity to the freezing threshold for the 1-hr nowcast*

There is motivation to modify the threshold used to differentiate a subfreezing road from a non-subfreezing road since public-works personnel may treat roadways in advance of expected winter storms. These treatments act to lower the freezing point of water. To examine this sensitivity and gauge the algorithm's performance, two new RFs are trained using "freezing" thresholds of -3 and -6°C. All other aspects of these experiments are identical to the original nowcast (see Section 3d), and the following results are only applied to the 1-hr nowcast.

Perturbing the freezing threshold drastically changes the number of subfreezing road events, thus worsening the class imbalance problem. Overall occurrence of subfreezing roads drops from $\sim$29% for the 0°C threshold to less than 10% for the -6°C threshold. The attributes diagrams for the two perturbed freezing thresholds reveal a less skillful nowcast compared to the baseline 0°C threshold (Fig. 12), although it is still better than climatology with the exception of the first week, which actually has negative skill for the -6°C threshold. The mean BSS for the weeks are 0.60 and 0.40 for the -3°C and -6°C thresholds, respectively, compared to the mean BSS of 0.67 for the standard 0°C threshold. The -6°C threshold mean BSS is skewed by the poor performance of week 1. Neglecting this week results in a mean of BSS of 0.59. Again, these differences are believed to be attributed to class-imbalance problems, and also to the fact that such cold T_R observations are not likely during October. Addressing the class imbalance issue using other resampling techniques may improve the algorithm, but this is left to future research.

6. Summary and Discussion

The goal of this study was to produce an algorithm to provide CONUS-wide probabilities that road temperatures are subfreezing that is both accurate and efficient. This was accomplished through use of machine learning (i.e., a random forest). The algorithm is trained on the 2016$-$17

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-19-0159.1.

winter year and verified using seven weeks of the 2017−18 winter season. Results indicate the algorithm outperforms climatology with a mean Brier Skill Score of 0.66, while the mean AUC score of ∼0.96 is considered excellent. The algorithm performs well over then entire climatological distribution of T_R, and also for instances when T_R is within a range close to freezing, defined herein as -5 to +5°C. The algorithm does well during the winter months, but can struggle at times during the transition or "shoulder" months (i.e., months of October and March) of the cool season and possibly in regions of complex terrain. The histogram of forecast probabilities is desirable with a higher frequency of probabilities assigned in the low (0-10%) and high (90-100%) range compared to the intermediate probabilities. Probabilities assigned in the 40-60% range were shown to be associated with instances where the forecast T_{sfc} and T_{2m} are within close proximity to 0°C. This is to be expected given inherent NWP model uncertainty, consistent with previous studies.

Class imbalance may be one of the root causes for decreases in skill during the months of October and March, and for the "freezing" threshold sensitivities. The other reason may be that the NWP model on which the algorithm is trained cannot fully resolve the complex terrain, leading to less reliable skill. Experiments with more sophisticated resampling techniques, such as creating synthetic minority training examples (SMOTE; Chawla et al. 2002), may improve the skill for transition months. It is also possible that more observations are needed to better calibrate the probabilities.

From a forecasting perspective, using a random forest to model the probability that T_R is subfreezing provides adequate skill out to 36-hrs using various NWP model parameters as input. The forecasting algorithm could be used to provide emergency managers, state DOTs, and forecasters ample time to prepare for the most appropriate road treatment plan and/or messaging for a winter event. The output could also be useful for local law enforcement to help prevent or better respond to accidents. For different "freezing" thresholds, lower freezing thresholds coincide with reduced

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

skill. However, the algorithm's skill still outperforms climatology. Lastly, it is demonstrated that the algorithm produces reliable and accurate predictions for various winter events in various geographic regions as discussed in the three case-study analyses. These strong results suggest the algorithm can add value to the forecasting process.

There are some considerations the reader may wonder about that are worthy of discussion before closing. The first is that the algorithm does not account for precipitation. Less than 8% of all observations occur coincident with precipitation and early iterations of the algorithm that incorporated precipitation show that it had negligible influence on the results. Therefore, this feature was removed to improve the algorithm's computational efficiency. Nor does the algorithm account for the amount of traffic (or other anthropogenic effects), which can modulate T_R, or any previous road treatment from state public works vehicles. Since this algorithm is targeted for implementation within the National Weather Service (NWS), only those products that are available within NWS operations are included as inputs to this algorithm. But, this could be folded in as a feature in the future. Unlike METRo, or other physical road models, this algorithm currently does not provide information about hazards like accumulating snow or ice. Work is underway to connect this algorithm's output to forecasts and analyses of hydrometeor phase and quantitative precipitation amounts to create a road hazards product. Last, this algorithm has only been applied to deterministic model output so far. There are several approaches to applying machine learning to ensemble output (e.g., Gagne et al. 2014, 2017; Loken et al. 2019). The best approach for this application is a topic currently under investigation.

Acknowledgments. The authors would like to thank the three anonymous reviewers as their suggestions greatly improved the manuscript. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agree-

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

ment #NA11OAR4320072, U.S. Department of Commerce. Special thanks to Mesowest, the University of Utah's HRRR archive (Blaylock et al. 2017), researchers in the McGovern research group at the University of Oklahoma for providing feature importance code (https://github.com/gelijergensen/PermutationImportance), and the scikit-learn Python library (Pedregosa et al. 2011).

References

Ahijevych, D., J. O. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Wea. Forecasting*, **31**, 43–50, doi:10.1175/waf-d-15-0113.1.

Batista, G. E. A. P. A., R. C. Prati, and M. C. Monard, 2004: A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, **6 (1)**, 20–29, doi:10.1145/1007730.1007735.

Blaylock, B. K., J. D. Horel, and S. T. Liston, 2017: Cloud archiving and data mining of high-resolution rapid refresh forecast model output. *Computers & Geosciences*, **109**, 43–50, doi:10.1016/j.cageo.2017.08.005.

Boselly, E. S., 1993: Road weather information systems: What are they and what can they do for you? *Transportation Research Record*, **1387**, 191–195.

Bouilloud, L., and Coauthors, 2009: Road surface condition forecasting in france. *J. Appl. Meteor. Climatol.*, **48**, 2513–2527, doi:10.1175/2009jamc1900.1.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, doi:10.1023/A.1010933404324.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002: Smote: Synthetic minority over-sampling technique. *JAIR*, **16**, 321–357, doi:10.1613/jair.953, URL http://dx.doi.org/10.1613/jair.953.

Crevier, L.-P., and Y. Delage, 2001: Metro: A new model for road-condition forecasting in canada. *J. Appl. Meteor.*, **40 (11)**, 2026–2037, doi:10.1175/1520-0450(2001)040⟨2026:manmfr⟩2.0.co;2.

Downing, L., H. Li, J. Desai, M. Liu, D. M. Bullock, and M. E. Baldwin, 2020: Evaluation of the high-resolution rapid refresh model for forecasting roadway surface temperatures. *36th Conference on Environmental Information Processing Technologies*, Boston, MA, Amer. Meteor. Soc.

Elmore, K., and H. Grams, 2016: Using mping data to generate random forests for precipitation type forecasts. *14th Conf. on Artificial and Computational Intelligence and Its Applications to the Enviornmental Sciences*, New Orleans, LA, Amer. Meteor. Soc.

Fujimoto, A., A. Saida, and T. Fukuhara, 2012: A new approach to modeling vehicle-induced heat and its thermal effects on road surface temperature. *J. Appl. Meteor. Climatol.*, **51**, 1980–1993, doi:10.1175/jamc-d-11-0156.1.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, 1819–1840, doi:10.1175/waf-d-17-0010.1, URL http://dx.doi.org/10.1175/waf-d-17-0010.1.

Gagne, D. J. I., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043,

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

doi:10.1175/waf-d-13-00108.1.

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

Hertl, S., and G. Schaffar, 1998: An autonomous approach to road temperature prediction. *Meteor. Appl.*, **5**, 227–238, doi:10.1017/s1350482798000838.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2 (3)**, 285–293, doi: 10.1016/0169-2070(86)90048-8.

Jacobs, W., and W. E. Raatz, 1996: Forecasting road-surface temperatures for different site characteristics. *Meteor. Appl.*, **3**, 243–256, doi:10.1002/met.5060030306.

Jonsson, P., and M. Riehm, 2012: Infrared thermometry in winter road maintenance. *J. Atmos. Oceanic Technol.*, **29**, 846–856, doi:10.1175/JTECH-D-11-00071.1.

Juga, I., P. Nurmi, and M. Hippi, 2013: Statistical modelling of wintertime road surface friction. *Meteor. Appl.*, **20**, 318–329, doi:10.1002/met.1285.

Kangas, M., M. Heikinheimo, and M. Hippi, 2015: Roadsurf: a modelling system for predicting road weather and road surface conditions. *Meteor. Appl.*, **22**, 544–553, doi:10.1002/met.1486.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which polarimetric variables are important for weather/no-weather discrimination? *Journal of Atmospheric and Oceanic Technology*, **32**, 1209–1223, doi:10.1175/jtech-d-13-00205.1.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

595 Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Post-processing
596 next-day ensemble probabilistic precipitation forecasts using random forests. *Wea. Forecasting*,
597 **in press**, doi:10.1175/waf-d-19-0109.1.

598 Luna-Herrera, J., G. Martínez-Cabrera, R. Parra-Maldonado, J. A. Enciso-Moreno, J. Torres-
599 López, F. Quesada-Pascual, R. Delgadillo-Polanco, and S. G. Franzblau, 2003: Use of re-
600 ceiver operating characteristic curves to assess the performance of a microdilution assay for
601 determination of drug susceptibility of clinical isolates of mycobacterium tuberculosis. *Euro-*
602 *pean Journal of Clinical Microbiology and Infectious Diseases*, **22 (1)**, 21–27, doi:10.1007/
603 s10096-002-0855-5.

604 McGovern, A., K. L. Elmore, D. J. G. II, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith,
605 and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision making for
606 high-impact weather. *Bull. Amer. Meteor. Soc.*, doi:10.1175/bams-d-16-0123.1.

607 McGovern, A., R. Lagerquist, D. J. G. II, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and
608 T. Smith, 2019: Making the black box more transparent: Understanding the physical impli-
609 cations of machine learning. *Bulletin of the American Meteorological Society*, **in press**, doi:
610 10.1175/bams-d-18-0195.1.

611 Mehdi, T., N., and M. Ahmadi, 2011: Kernel smoothing for roc curve and estimation for
612 thyroid stimulating hormone. *Int. J. Public Health Res. Spec. Issue*, 239–242, doi:10.1175/
613 bams-d-14-00173.1.

614 Metz, C. E., 1978: Basic principles of roc analysis. *Seminars in Nuclear Medicine*, **8 (4)**, 283–298,
615 doi:10.1016/s0001-2998(78)80014-2.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

Muller, M. P., G. Tomlinson, T. J. Marrie, P. Tang, A. McGeer, D. E. Low, A. S. Detsky, and W. L. Gold, 2005: Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia? *Clinical Infectious Diseases*, **40 (8)**, 1079–1086, doi:10.1086/428577.

Niculescu-Mizil, A., and R. A. Caruana, 2012: Obtaining calibrated probabilities from boosting. *Proc. 21st Conf. on Uncertainty in Artificial Intelligence*, Edinburgh, Scotland, Association for Uncertainty in Artificial Intelligence, 413–420.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pisano, P., G. Guevara, R. Alfelor, R. Murphy, and B. C. Boyce, 2018: Communicating road weather impacts to the traveling public. *34th Conference on Environmental Information Processing Technologies*, Austin, TX, Amer. Meteor. Soc.

Quinlan, J. R., 1986: Induction of decision trees. *Machine Learning*, **1 (1)**, 81–106, doi:10.1007/bf00116251.

Quinlan, J. R., 1993: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 302 pp.

Rayer, J., 1987: The meteorological office forecast road surface temperature model. *Meteor. Mag.*, **116**, 180–191.

Reeves, H. D., K. L. Elmore, A. Ryzhkov, T. Schuur, and J. Krause, 2014: Sources of uncertainty in precipitation-type forecasting. *Wea. Forecasting*, **29 (4)**, 936–953, doi:10.1175/waf-d-14-00007.1.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24 (2)**, 601–608, doi:10.1175/2008waf2222159.1.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

Rutz, J. J., and C. V. Gibson, 2013: Integration of a road surface model into nws operations. *Bull. Amer. Meteor. Soc.*, 1495–1500, doi:10.1175/bams-d-12-00037.1.

Sass, B. H., 1997: A numerical forecasting system for the prediction of slippery roads. *J. Appl. Meteor.*, **36**, 801–817, doi:10.1175/1520-0450(1997)036⟨0801:anfsft⟩2.0.co;2.

Scott, B., E. Minge, and S. Peterson, 2005: *The Aurora Consortium: Laboratory and field studies of pavement temperature sensors*. Minnesota Department of Transportation Rep. MN/RC–2005-44, 110 pp.

Shao, J., 1998: Improving Nowcasts of Road Surface Temperature by a Backpropagation Neural Network. *Wea. Forecasting*, **13**, 164–171, doi:10.1175/1520-0434(1998)013⟨0164:inorst⟩2.0.co.

Shao, J., and P. J. Lister, 1996: An automated nowcasting model of road surface temperatures and state for winter road maintenance. *J. Appl. Meteor.*, **35**, 1352–1361.

Toms, B. A., J. B. Basara, and Y. Hong, 2017: Usage of Existing Meteorological Data Networks for Parameterized Road Ice Formation Modeling. *J. Appl. Meteor. Climatol.*, **56**, 1959–1976, doi:10.1175/jamc-d-16-0199.1.

Walker, C., D. Steinkruger, P. Gholizadeh, B. Dao, S. Hasanzadeh, M. R. Anderson, and B. Esmaeili, 2018: Developing a winter severity index to improve safety and mobility. *34th Conference on Environmental Information Processing Technologies*, Austin, TX, Amer. Meteor. Soc.

Weygandt, T. S., T. Smirnova, S. Benjamin, K. Brundage, S. Sahm, C. Alexander, and B. Schwartz, 2009: The high resolution rapid refresh (hrrr): An hourly updated convection resolving model using radar reflectivity assimilation from the ruc/rr. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, Nebraska, Amer. Meteor.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

Soc., 15A.6, [Available online at https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154317.html.].

Wilks, D. S., 2006: *Statistical methods in the atmospheric sciences*. 2nd ed., Academic Press, 627 pp.

Yang, C. H., D.-G. Yun, and J. G. Sung, 2012: Validation of a road surface temperature prediction model using real-time weather forecasts. *KSCE J. Civ. Eng.*, **16**, 1289–1294, doi: 10.1007/s12205-012-1649-7.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

LIST OF TABLES

Table 1. Input features to the RF algorithm. 32

Table 2. Hyperparameter ranges used for model tuning. Bold values identify the optimal hyperparameter values. 33

Table 3. List of dates from the 2017-2018 winter season used for probability calibration and testing 34

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

TABLE 1. Input features to the RF algorithm.

Feature
Hours T_{2m} <=0°C
Incoming short wave radiation flux (S)
Hours T_{sfc} <= 0°C
Surface temperature (T_{sfc})
Visible beam downward solar flux (V_{bd})
Hours T_{2m} >0°C
Upward long wave radiation flux ($\lambda_{\uparrow}$)
2-m air temperature (T_{2m})
Hours T_{sfc} > 0°C
T_{2m} and T_{sfc} difference ($HRRR_{dT}$)
Absolute difference between current date and 10 Jan
Friction velocity at the surface (V_{fric})
S and $\lambda_{\uparrow}$ difference
2-m dewpoint temperature (T_d)
Latent heat flux at the surface (L_{hf})
Simulated brightness temperature (T_{irbt})
S and $\lambda_{\downarrow}$ difference
Downward long wave radiation flux ($\lambda_{\downarrow}$)
G and S_{hf} difference
Ground flux (G)
Sensible heat flux at the surface (S_{hf})
Surface roughness (S_R)
10-m wind speed (U_{10m})
Visible diffuse downward solar flux (V_{dd})
Total cloud cover percentage (C_{total})
Low cloud cover percentage (C_{low})
Mid cloud cover percentage (C_{mid})
High cloud cover percentage (C_{high})
Urban (L_U) and rural (L_R) HRRR land classification

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

673 TABLE 2. Hyperparameter ranges used for model tuning. Bold values identify the optimal hyperparameter
674 values.

Parameter	Range of values
N_{est}	[100,150,**300**,400,500]
Max_D	[6,8,10,15,**20**]
Max_{feat}	[**5**,6,8,10]
Min_{ss}	[4,5,8,10,15,20,**25**,50]
Min_{sl}	[4,**5**,8,10,15,20,25,50]
$Class_w$	[**"balanced"**; 0:0.25,1:3; 0:0.5,1:2.5; 0:0.75,1:3.5;0:1,1:5;0:1,1:10;0:1,1:5]

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

TABLE 3. List of dates from the 2017-2018 winter season used for probability calibration and testing

Calibration	Testing
01-15 Oct 2017	18-25 Oct 2017
28-11 Oct-Nov 2017	14-21 Nov 2017
24-08 Nov-Dec 2017	11-18 Dec 2017
21 Dec 2017 - 04 Jan 2018	07-14 Jan 2018
17-31 Jan 2018	03-10 Feb 2018
13-27 Jan-Feb 2018	02-09 Mar 2018
12-26 Feb-Mar 2018	29-31 Mar 2018

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

LIST OF FIGURES

Fig. 1. (a) Spatial distribution of RWIS locations across the CONUS along with the US Interstate system. All grey or red circles indicate valid RWIS sites. Red circles indicate sites used in training/testing, but gray circles could be used for case study analyses. (b) RWIS sites in northern Ohio at 1100 UTC 14 November 2018. Sites are color coded according to their lowest T_R. Roadways are indicated as blue for interstates and grey for city and county roads. (c) Observed T_R at the two sensors at OH112 (See Fig. 2 red arrow for location) on 14 November 2018. 37

Fig. 2. Distribution of RWIS road temperature observations for each month of the cool season. Subfreezing (above-freezing) T_R observations are shown in orange (blue). 38

Fig. 3. Flow chart illustrating the methods of this study. The left side outlines methods for training data, whereas the right side outlines the methods for testing and calibration. 39

Fig. 4. Analysis from the 04 February 2018 Missouri multi-car pile up showing time series of (a) observed T_{2m} and precipitation type from the KLBO ASOS and T_R from the MO009 RWIS sites (locations of these are indicated in panel d), (b) the HRRR 02-hr forecast T_{2m}, T_{sfc}, and insolation ($HRRR_s$), (c) the T_{Rprob} output, and (d,e) plan views of T_{Rprob} at 2000 and 2300 UTC, respectively. RWIS observations are overlaid in (d,e). Sites indicated as black (white) have subfreezing (above-freezing) T_R. 40

Fig. 5. Analysis from the 01-02 April 2018 Washington snow in complex terrain case-study showing time series of (a) observed T_{2m} and precipitation type from the KSMP ASOS and T_R from the TFRAN RWIS sites (locations of these are indicated in panel d), (b) the HRRR 02-hr forecast T_{2m}, T_{sfc}, and insolation ($HRRR_s$), (c) the T_{Rprob} output, and (d,e) plan views of T_{Rprob} at 0900 and 1200 UTC, respectively. RWIS observations are overlaid in (d,e). Sites indicated as black (white) have subfreezing (above-freezing) T_R. 41

Fig. 6. Analysis from the 03-04 March 2019 Maryland transition season case-study showing time series of (a) observed T_{2m} and precipitation type from the KGAI ASOS and T_R from the MD056 RWIS sites (locations of these are indicated in panel d), (b) the HRRR 02-hr forecast T_{2m}, T_{sfc}, and insolation ($HRRR_s$), (c) the T_{Rprob} output, and (d,e) plan views of T_{Rprob} at 2100 and 0600 UTC, respectively. RWIS observations are overlaid in (d,e). Sites indicated as black (white) have subfreezing (above-freezing) T_R. 42

Fig. 7. Feature importance plots determined by the two methods of permutation. The single-pass (multi-pass) method is shown on the left (right) column, and the the BSS (AUC) skill metric is shown on the top (bottom) row. 43

Fig. 8. a) Receiver operator characteristic (ROC) curve for each of the 7 testing weeks. Each color represents a testing week as denoted in the inset of panel (a). The solid black line represents the mean of the entire testing set. Dashed line represents the no-skill line. b) Performance diagram with each individual testing week shown along with the overall mean. Contoured dashed lines represent the frequency bias, whereas the colored contours represent the critical success index (CSI) values. c) Similar to (a) except for -5 to +5°C T_R range only. d) Similar to (b) except for -5 to +5°C T_R range only. 44

Fig. 9. (a) Attributes diagram for the testing set. The (1:1) diagonal dashed line represents the perfect reliability curve, the vertical dashed line represents the climatology line, the horizontal dashed line is the no-resolution line, and the second quasi-horizontal line that intersects with the perfect reliability line is the no-skill line. The shaded region corresponds to a skillful

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.

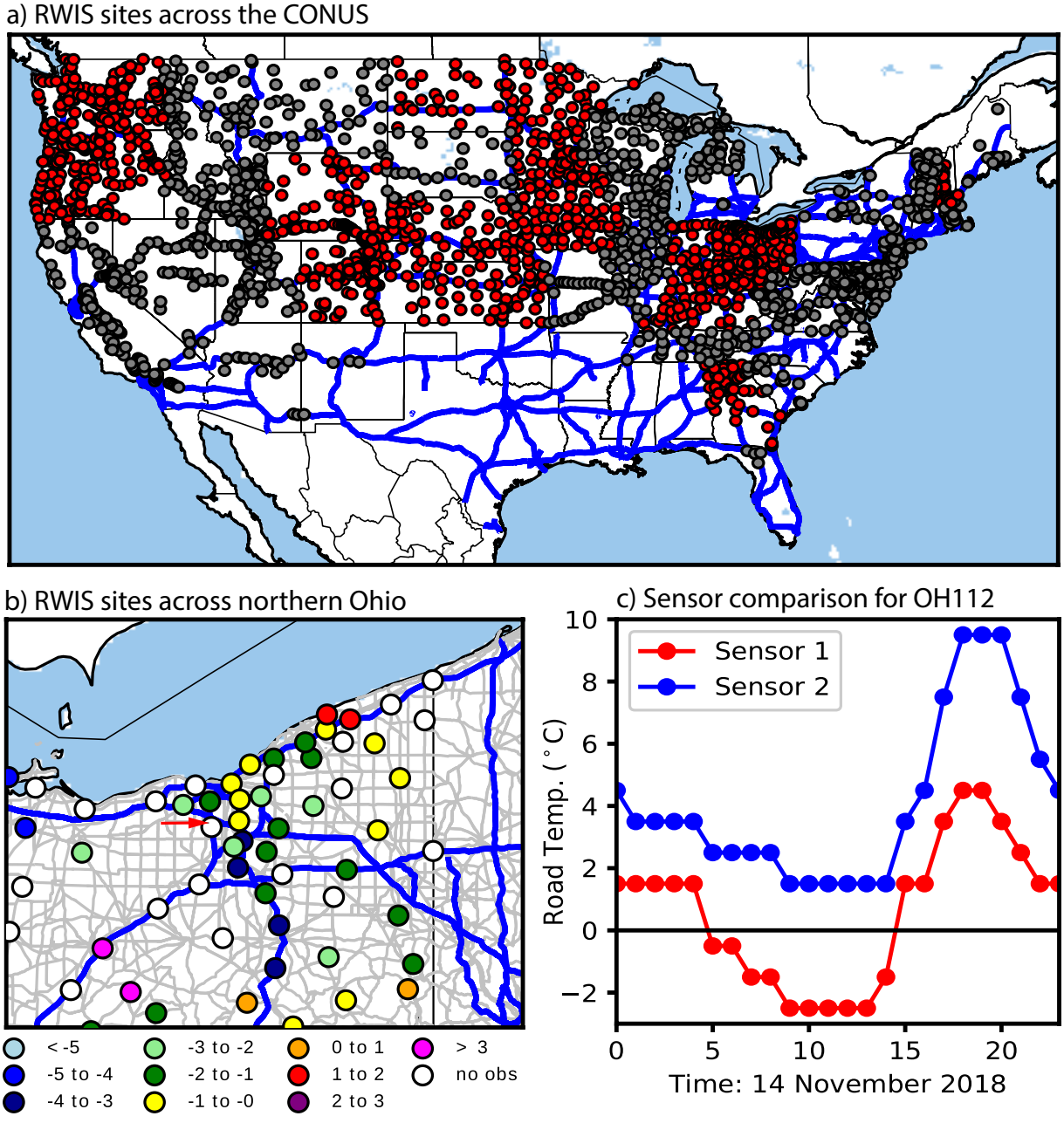
forecast (i.e., greater than climatology). (b) Histogram of forecast probabilities. The height of each probability bin corresponds to the number of counts in each bin. Solid blue line represents the mean 2-m temperature (top left), surface temperature (top right), number of hours the 2-m temperature is subfreezing (bottom left), and number of hours the SFC temperature is subfreezing (bottom right). Dashed lines correspond to 10th and 90th percentiles of each variable. c) Similar to (a) except for -5 to +5°C T_R range only. d) Similar to (b) except for -5 to +5°C T_R range only. 45

Fig. 10. As in figure 9a, except for each forecast model. (a) HRRR 18-hr forecast (b) NAM 36-hr forecast, and (c) GFS 36-hr forecast. 46

Fig. 11. As in figure 8b, except for each forecast model. (a) HRRR 18-hr forecast (b) NAM 36-hr forecast, and (c) GFS 36-hr forecast. 47

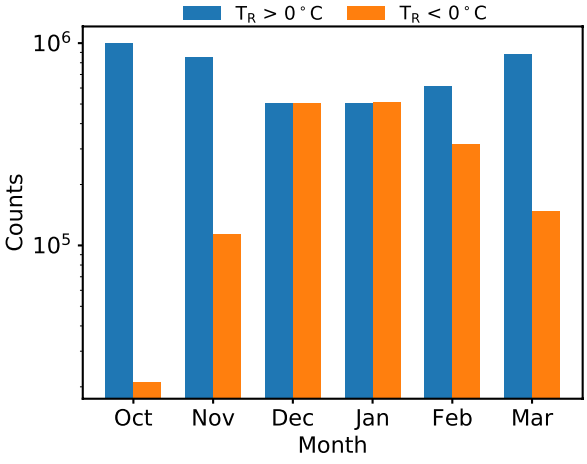
Fig. 12. As in figure 9a, except for (a) perturbed freezing threshold of -3°C and (b) perturbed freezing threshold of -6°C. 48

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.



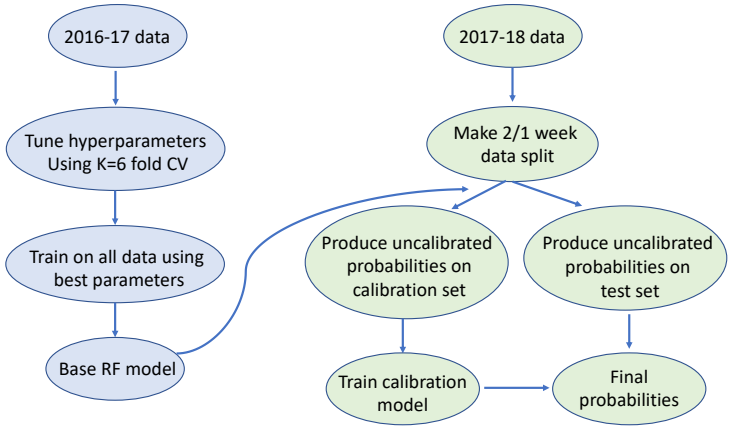
732 FIG. 1. (a) Spatial distribution of RWIS locations across the CONUS along with the US Interstate system. All
733 grey or red circles indicate valid RWIS sites. Red circles indicate sites used in training/testing, but gray circles
734 could be used for case study analyses. (b) RWIS sites in northern Ohio at 1100 UTC 14 November 2018. Sites
735 are color coded according to their lowest T_R. Roadways are indicated as blue for interstates and grey for city and
736 county roads. (c) Observed T_R at the two sensors at OH112 (See Fig. 2 red arrow for location) on 14 November
737 2018.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.



738 FIG. 2. Distribution of RWIS road temperature observations for each month of the cool season. Subfreezing
739 (above-freezing) T_R observations are shown in orange (blue).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.



740 FIG. 3. Flow chart illustrating the methods of this study. The left side outlines methods for training data,
741 whereas the right side outlines the methods for testing and calibration.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-19-0159.1.

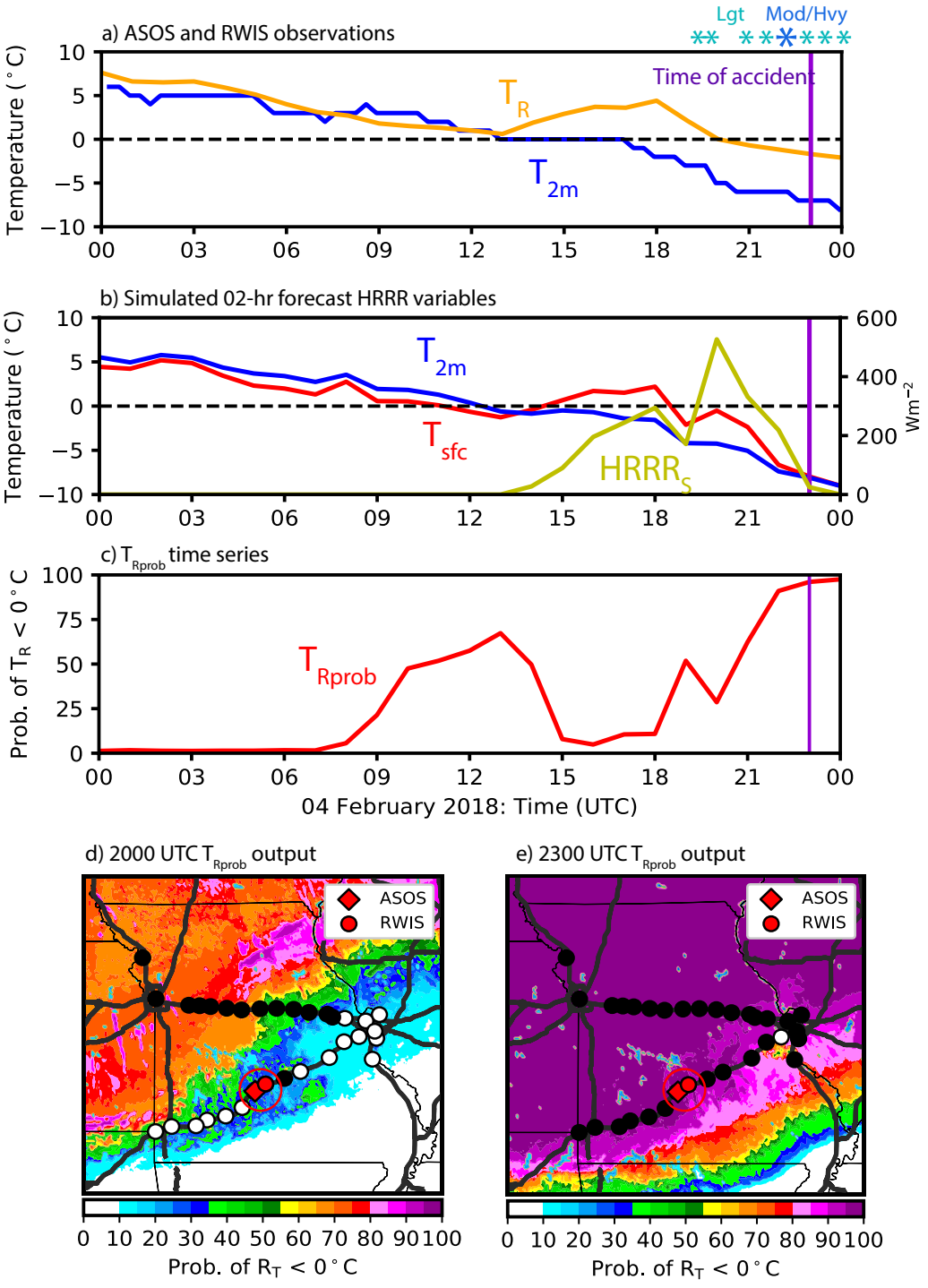
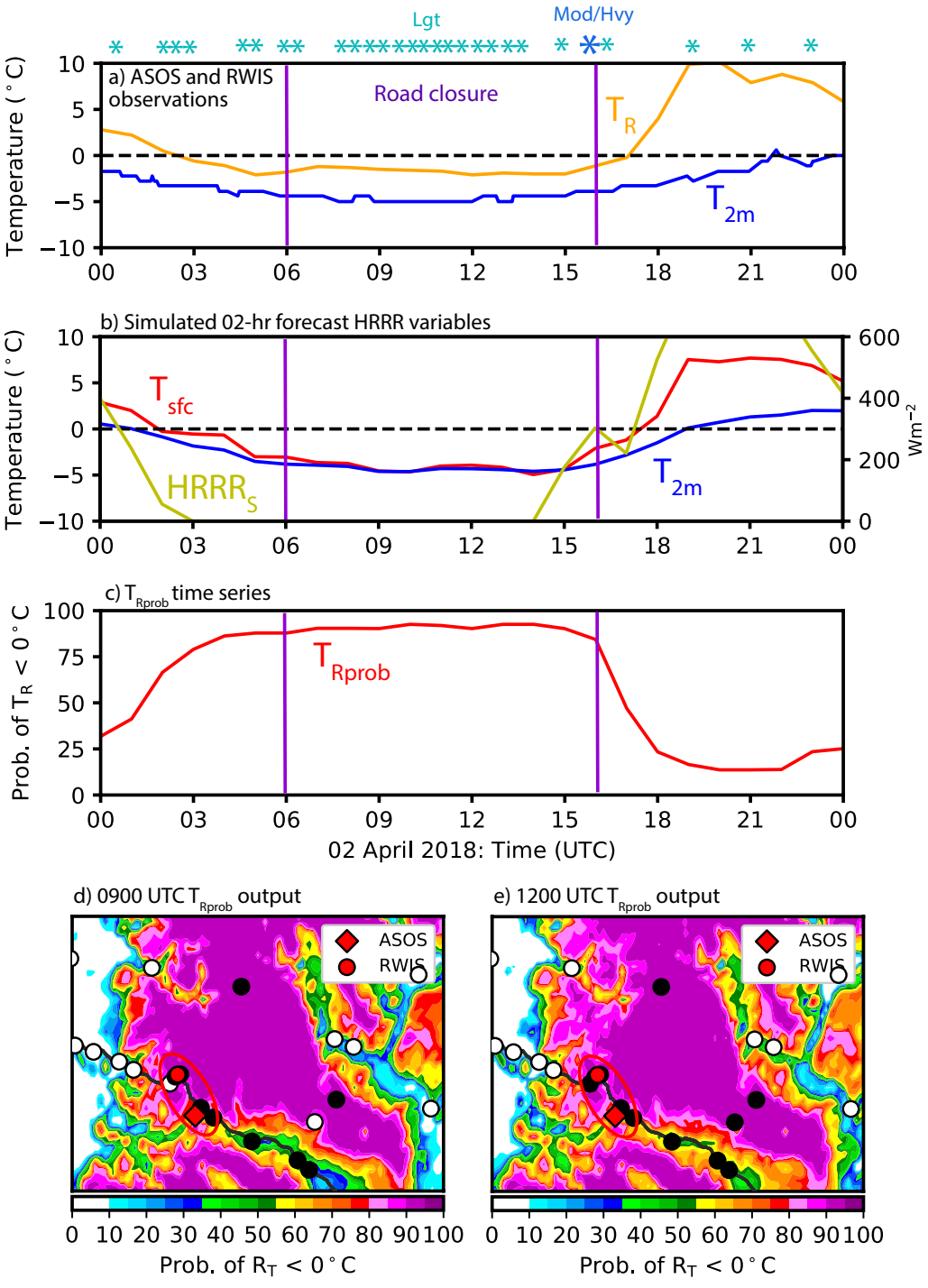


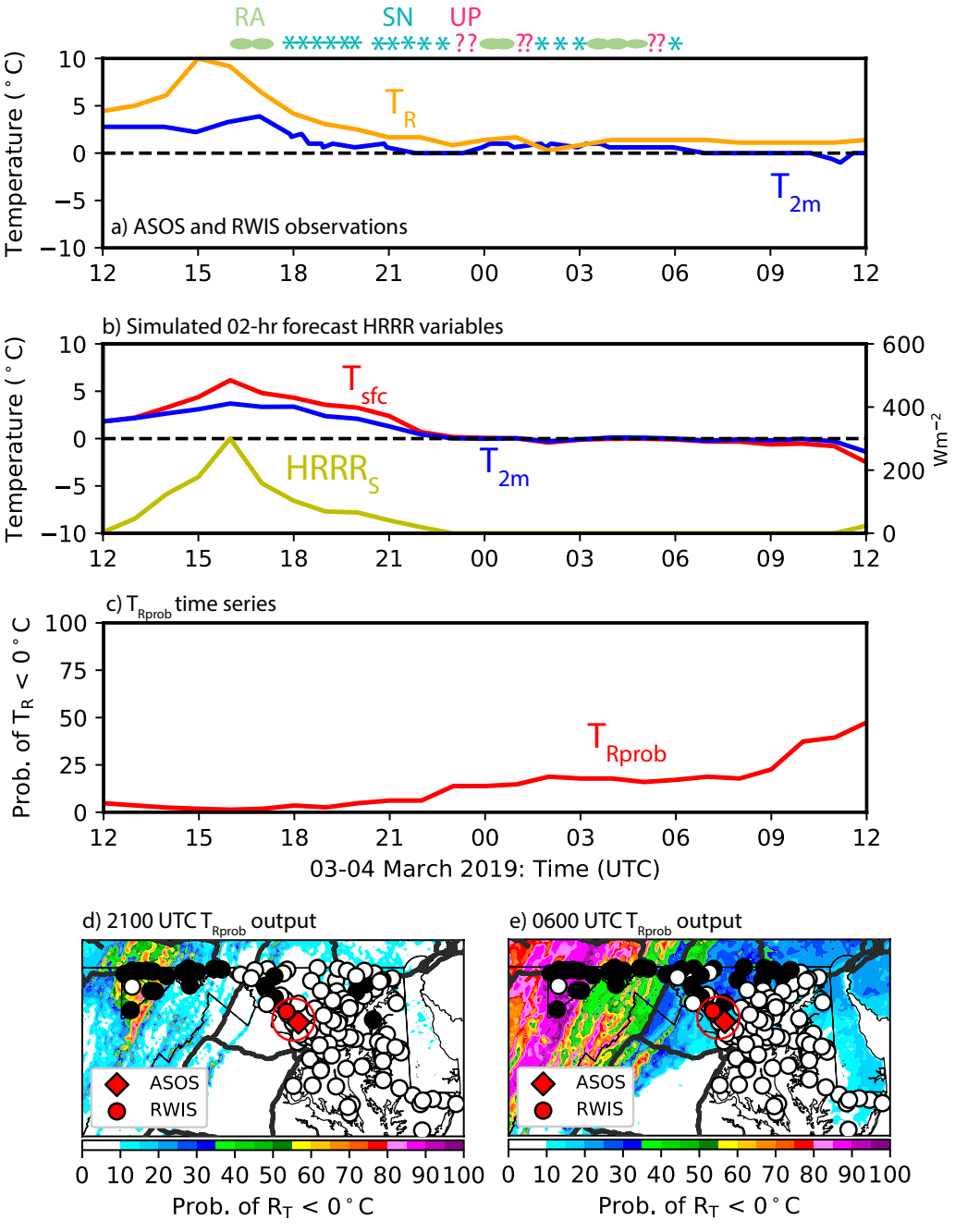
FIG. 4. Analysis from the 04 February 2018 Missouri multi-car pile up showing time series of (a) observed T_{2m} and precipitation type from the KLBO ASOS and T_R from the MO009 RWIS sites (locations of these are indicated in panel d), (b) the HRRR 02-hr forecast T_{2m}, T_{sfc}, and insolation ($HRRR_s$), (c) the T_{Rprob} output, and (d,e) plan views of T_{Rprob} at 2000 and 2300 UTC, respectively. RWIS observations are overlaid in (d,e). Sites indicated as black (white) have subfreezing (above-freezing) T_R.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-19-0159.1.



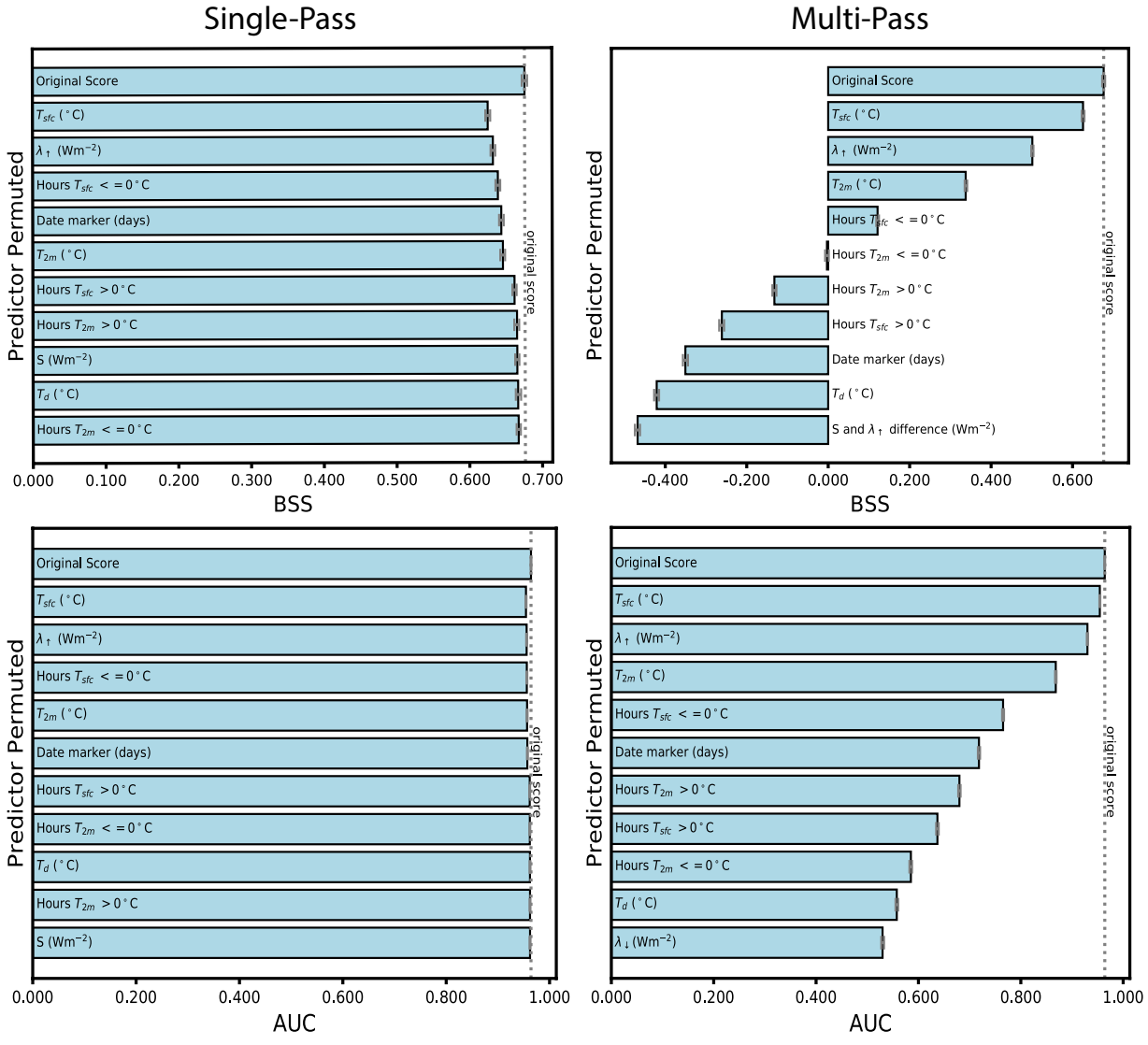
747 FIG. 5. Analysis from the 01-02 April 2018 Washington snow in complex terrain case-study showing time
748 series of (a) observed T_{2m} and precipitation type from the KSMP ASOS and T_R from the TFRAN RWIS sites
749 (locations of these are indicated in panel d), (b) the HRRR 02-hr forecast T_{2m}, T_{sfc}, and insolation ($HRRR_s$), (c)
750 the T_{Rprob} output, and (d,e) plan views of T_{Rprob} at 0900 and 1200 UTC, respectively. RWIS observations are
751 overlaid in (d,e). Sites indicated as black (white) have subfreezing (above-freezing) T_R.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-19-0159.1.



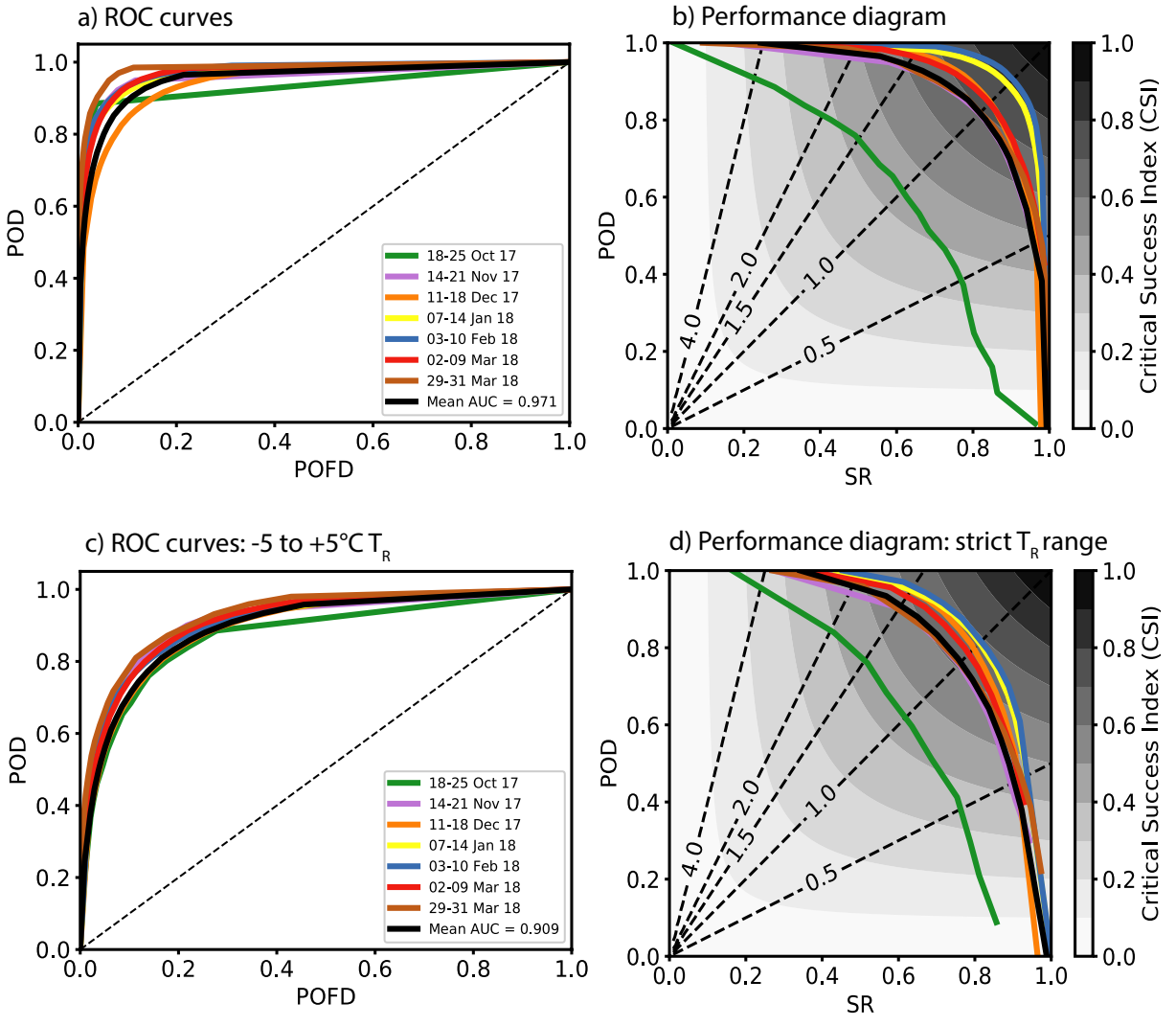
752 FIG. 6. Analysis from the 03-04 March 2019 Maryland transition season case-study showing time series of
753 (a) observed T_{2m} and precipitation type from the KGAI ASOS and T_R from the MD056 RWIS sites (locations
754 of these are indicated in panel d), (b) the HRRR 02-hr forecast T_{2m}, T_{sfc}, and insolation ($HRRR_s$), (c) the T_{Rprob}
755 output, and (d,e) plan views of T_{Rprob} at 2100 and 0600 UTC, respectively. RWIS observations are overlaid in
756 (d,e). Sites indicated as black (white) have subfreezing (above-freezing) T_R.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-19-0159.1.



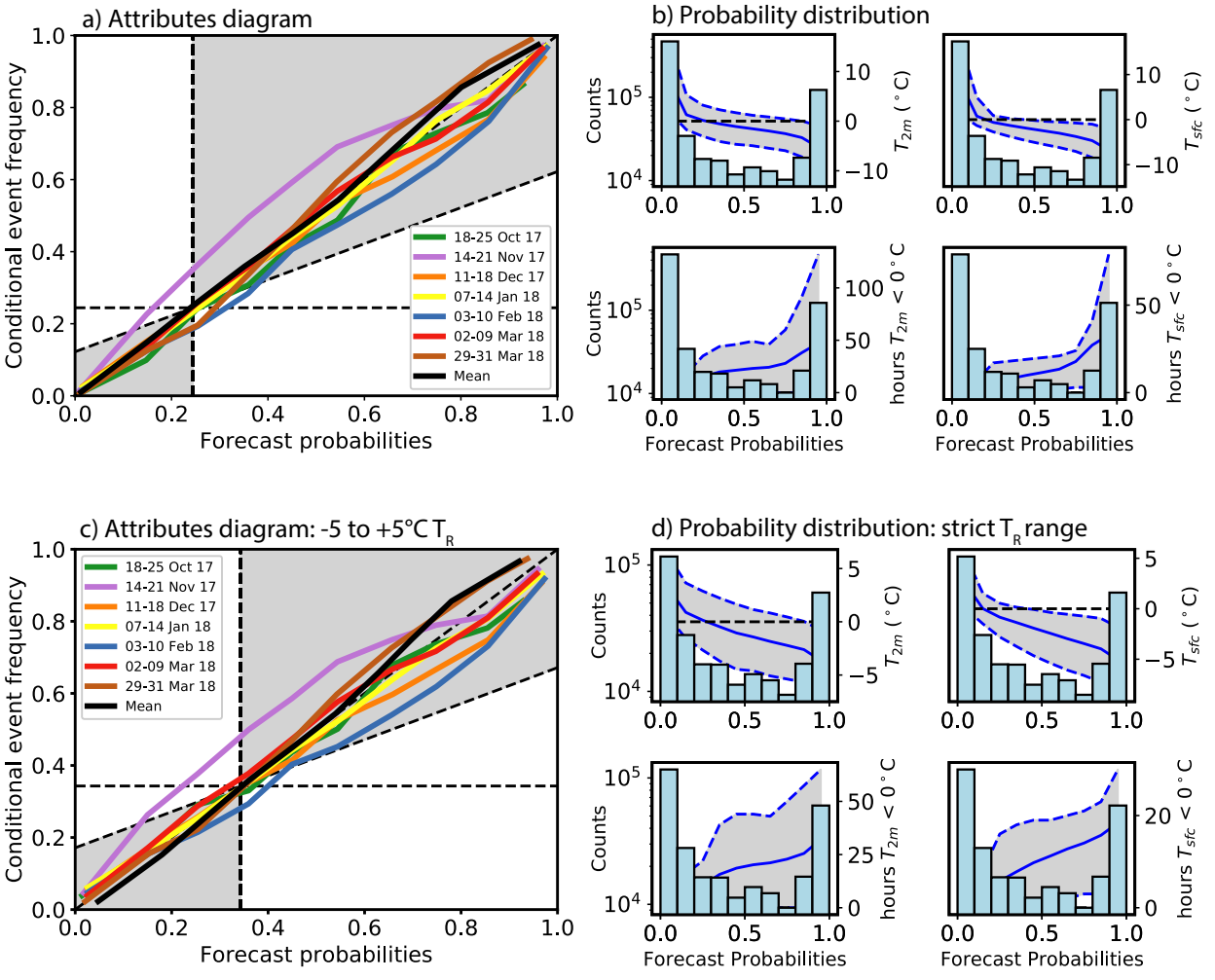
757 FIG. 7. Feature importance plots determined by the two methods of permutation. The single-pass (multi-pass)
758 method is shown on the left (right) column, and the the BSS (AUC) skill metric is shown on the top (bottom)
759 row.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.



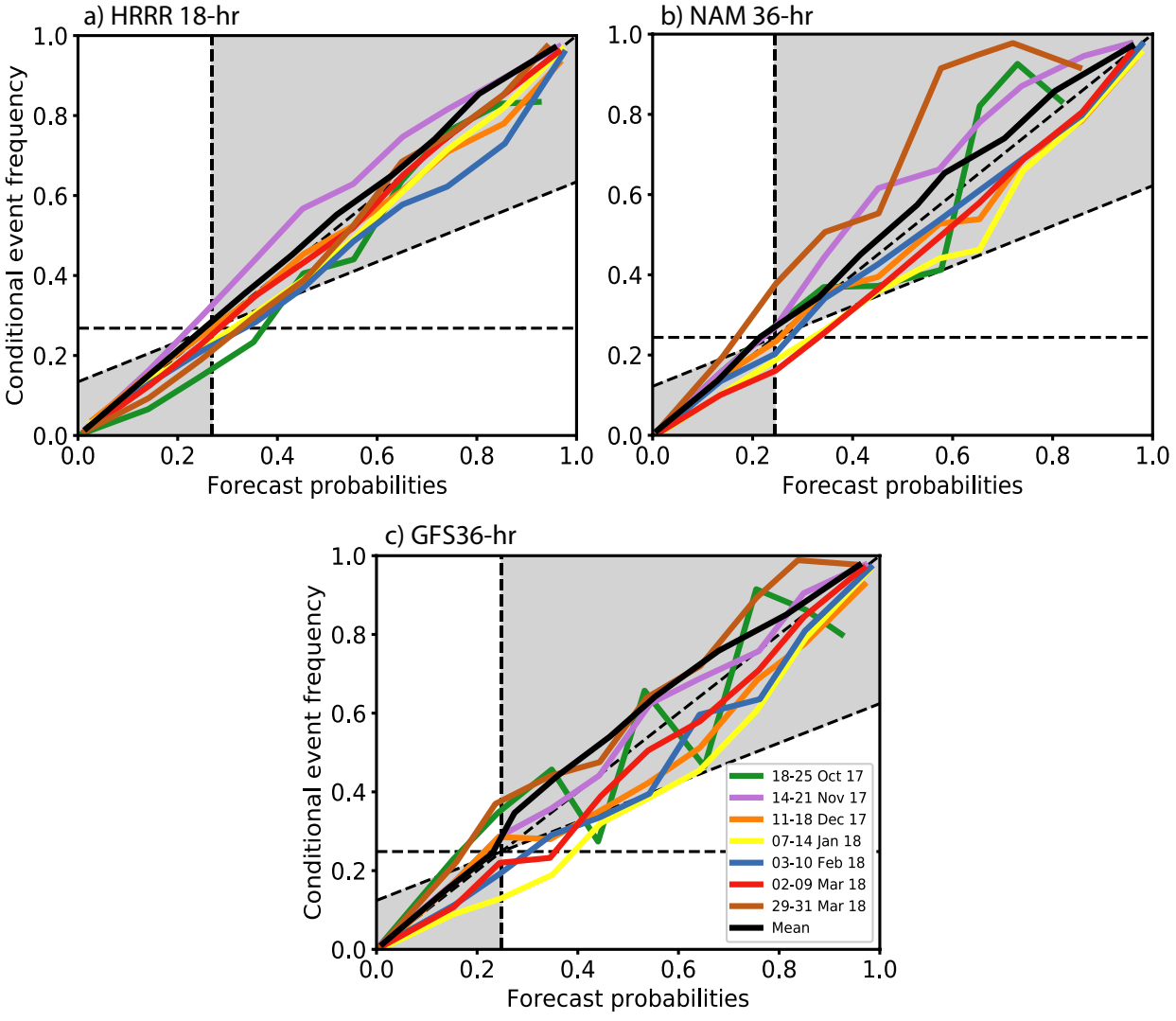
760 FIG. 8. a) Receiver operator characteristic (ROC) curve for each of the 7 testing weeks. Each color represents
761 a testing week as denoted in the inset of panel (a). The solid black line represents the mean of the entire testing
762 set. Dashed line represents the no-skill line. b) Performance diagram with each individual testing week shown
763 along with the overall mean. Contoured dashed lines represent the frequency bias, whereas the colored contours
764 represent the critical success index (CSI) values. c) Similar to (a) except for -5 to +5°C T_R range only. d) Similar
765 to (b) except for -5 to +5°C T_R range only.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-19-0159.1.



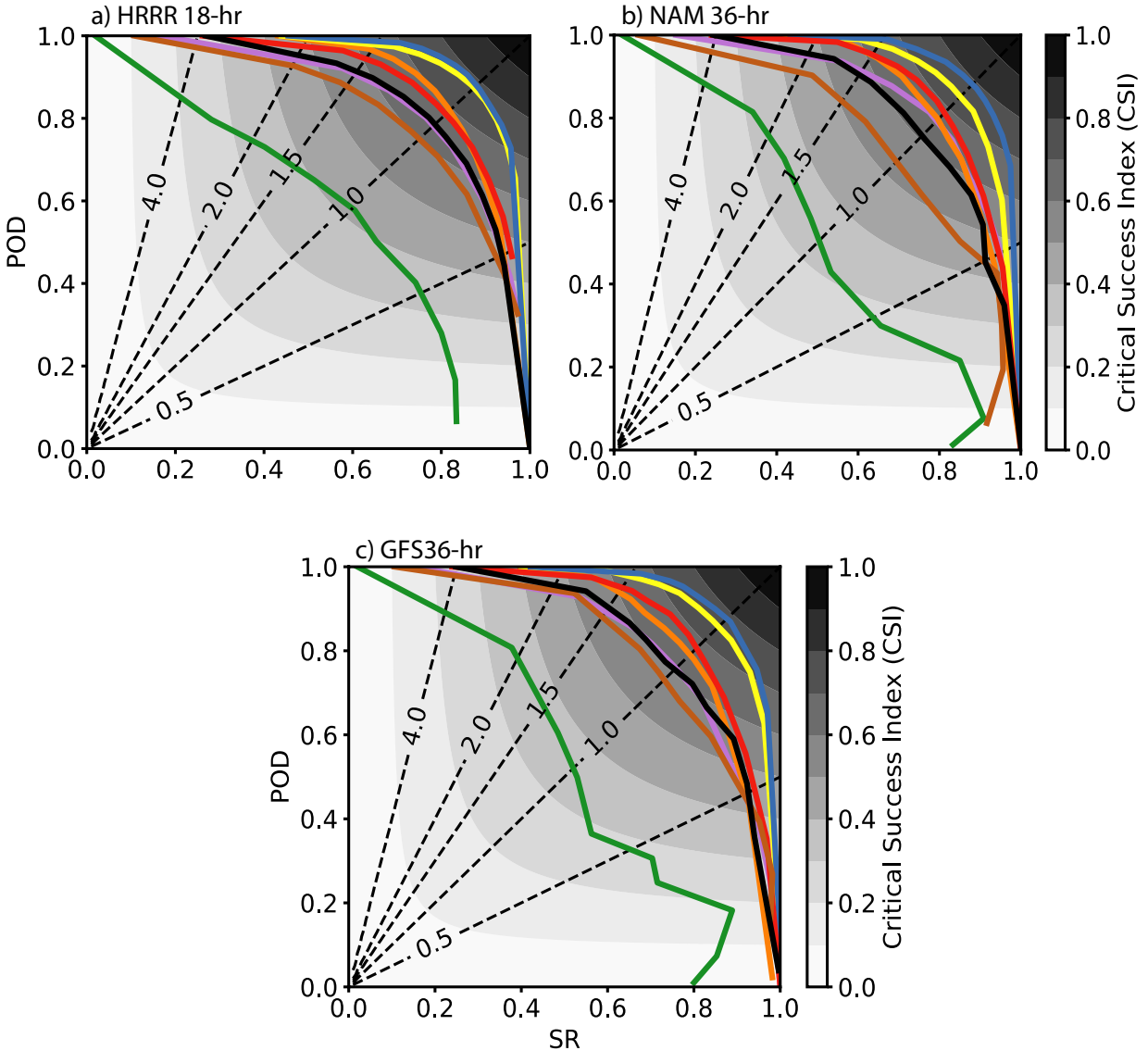
766 FIG. 9. (a) Attributes diagram for the testing set. The (1:1) diagonal dashed line represents the perfect
767 reliability curve, the vertical dashed line represents the climatology line, the horizontal dashed line is the no-
768 resolution line, and the second quasi-horizontal line that intersects with the perfect reliability line is the no-skill
769 line. The shaded region corresponds to a skillful forecast (i.e., greater than climatology). (b) Histogram of
770 forecast probabilities. The height of each probability bin corresponds to the number of counts in each bin. Solid
771 blue line represents the mean 2-m temperature (top left), surface temperature (top right), number of hours the
772 2-m temperature is subfreezing (bottom left), and number of hours the SFC temperature is subfreezing (bottom
773 right). Dashed lines correspond to 10th and 90th percentiles of each variable. c) Similar to (a) except for -5 to
774 +5°C T_R range only. d) Similar to (b) except for -5 to +5°C T_R range only.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-19-0159.1.



775 FIG. 10. As in figure 9a, except for each forecast model. (a) HRRR 18-hr forecast (b) NAM 36-hr forecast,
776 and (c) GFS 36-hr forecast.

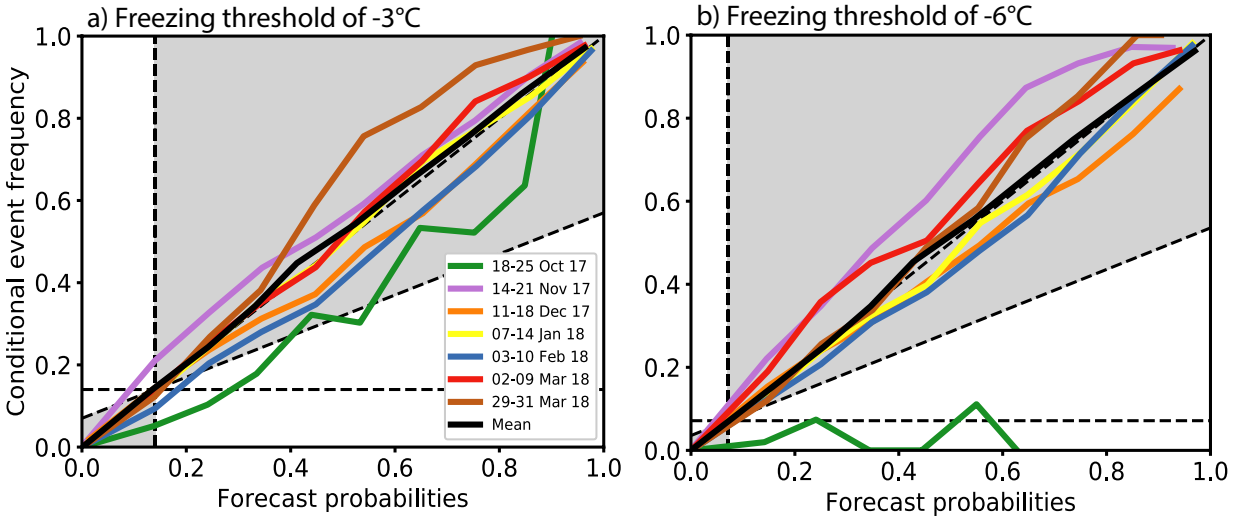
Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.



777 FIG. 11. As in figure 8b, except for each forecast model. (a) HRRR 18-hr forecast (b) NAM 36-hr forecast,

778 and (c) GFS 36-hr forecast.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0159.1.



779 FIG. 12. As in figure 9a, except for (a) perturbed freezing threshold of -3°C and (b) perturbed freezing
780 threshold of -6°C.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-19-0159.1.