The 2017 HMT-WPC Winter Weather Experiment

17 January – 17 February, 2017
Weather Prediction Center
College Park, MD

Findings and Results

Michael Bodner - *NOAA/NWS/WPC, College Park, MD*
Sarah Perfater - *I.M. Systems Group, NOAA/NWS/WPC/HMT*
Benjamin Albright - *Systems Research Group, NOAA/NWS/WPC/HMT, College Park, MD*

Table of Contents

Updated April 14, 2017

1. Introduction

In an effort to support improvements in winter weather forecasts issued by the Weather Prediction Center (WPC) and Weather Forecast Offices (WFOs), the Hydrometeorology Testbed at WPC (HMT-WPC) began hosting an annual Winter Weather Experiment (WWE) in 2011. The experiment brings together members of the operational forecasting, research, and academic communities to address winter weather forecast challenges.

The 2017 WWE continued work that began in the 2016 WWE with the exploration of 1-hour probabilistic snowfall rate guidance as a diagnostic to identify potential-mesoscale snowfall banding. This year's experiment featured improved model parameterization during the Day 1 forecast period in NCEP's Experimental North American Model (Parallel NAMv4). To better ascertain the predictability of these mesoscale snowfall bands, experimental versions of the High-Resolution Rapid Refresh (HRRR) and HRRR Time-Lagged Ensemble (HRRR-TLE) provided by the Earth System Research Laboratory (ESRL), and the National Center for Atmospheric Research (NCAR) Storm-Scale Ensemble were evaluated. A multi-model ensemble of implicit snowfall forecasts was tested this year, as well as an evaluation of a model implicit Probability of Winter Precipitation Forecast (PWPF) in the Day 2-3 time periods.

The 2017 experiment debuted the exploration of issuing both criteria-based winter weather watches and impacts-based winter weather alerts as a national center, compliant with an AFS FY17 Milestone to plan a prototype of a collaborative WPC-WFO Winter Storm Watch issuance. The WPC Watch Collaborator, which uses the WPC PWPF and WFO winter storm watch criteria, was designed to offer the WFO field offices a tool for collaboration among surrounding offices when determining boundaries and issuance times of winter storm watches and was a cornerstone guidance tool. Other tools included disaggregated PWPF into 6-hour forecasts, experimental WPC developed joint probability tools, and the Nash/Cobb Winter Storm Severity Index (WSSI) prototype (http://www.weather.gov/media/btv/wssi/WSSI_PDD.pdf).

To increase exposure to the experimental datasets and concepts being tested in the WWE, a daily forecast briefing was again provided for interested NWS SOOs and WFOs by WWE participants. These briefings attempted to simulate a collaboration concept with center-issued watches and provided an opportunity for WFO forecasters to ask questions and offer feedback.

2. Science and Operations Goals

The goals of the 2017 Winter Weather Experiment were to:

- Explore the Experimental HRRR and HRRR-TLE for winter weather forecasting, including 1-hour snowfall accumulation.
- Evaluate neighborhood versus grid spacing probabilities for forecasting hourly snowfall amounts.
- Evaluate the parallel 3 km NAMv4 for 1-hour snowfall rate forecasts.
- Explore the utility of joint probability tools for winter weather impacts-based forecasting.
- Examine the utility of the WPC Watch Collaborator trend tools.
- Evaluate the utility of a model implicit PWPF methodology.
- Test the utility of issuing both experimental, impacts-based winter weather alerts as well as traditional winter weather watches from a national center perspective.
- Enhance collaboration among NCEP centers, WFOs, and NOAA research labs on winter weather forecast challenges.

Fast findings for science and operations goals can be found in Table 1.

***Table 1**. FY17 Transition Metrics from the 2017 Winter Weather Experiment*

FY17 Transition Metrics

WPC-HMT

- Winter Weather Experiment

Major Tests Conducted	Transitioned to Operations	Recommended for Transition to Operations	Recommended for Further Development & Testing	Rejected for Further Testing	Decision Pending or Deferred on Advancement
WPC Experimental Implicit PWPF			X		
WPC Watch Collaborator Trend Tools		X			
NAMv4 1-Hour Max Hourly Frozen Precipitation Rate		X			
NAMv4/HRRRv3 Hourly Snowfall Accumulation Guidance		X			
HRRR-TLE Probabilistic Snowfall Rate Guidance		X			
WPC Joint Probabilities			X		
Experimental Winter Storm Watches by National Center			X		
Totals	Pending	4	3	0	0

3. Experiment Operations

The experiment was conducted for four weeks beginning January 17, 2017 in the WPC-OPC Collaboration Room at the NOAA Center for Weather and Climate Prediction (NCWCP) in College Park, MD:

Week 1: January 17 – 19, 2017 (Tuesday – Thursday)
Week 2: January 30 – February 3, 2017 (Monday – Friday)
Week 3: February 6 – 10, 2017 (Monday – Friday)
Week 4: February 13 – 17, 2017 (Monday – Friday)

Each morning, experiment participants were paired with a WPC Winter Weather Desk forecaster to form a collaborative forecast team. These forecast teams used a combination of operational and experimental model guidance to create experimental 24 hr deterministic snowfall forecasts across the contiguous United States (CONUS) for weather systems of interest during the **Day 2 (12 – 12 UTC) time period**.   Participants were asked to draw contours of 1, 2, 4, 8, 12, and 20 inch snowfall amounts over the entire country. Participants then created a 24 hr CONUS deterministic ice forecast during the same **Day 2 (12 – 12 UTC) period**, with contours of .01, 0.1, 0.25, and >=0.5” of ice accumulation.

Using these two assessments of winter weather over the CONUS, participants issued experimental winter weather watches for Day 2. The WPC Watch Collaborator thresholds for meeting the WFO snow and ice winter storm warning criteria provided a starting point (Figure 1). Then as a second watch-related exercise, the participants issued impacts-based Winter Alerts based on as experimental joint probabilities, the WSSI tool, and subjective enhancement of available guidance.

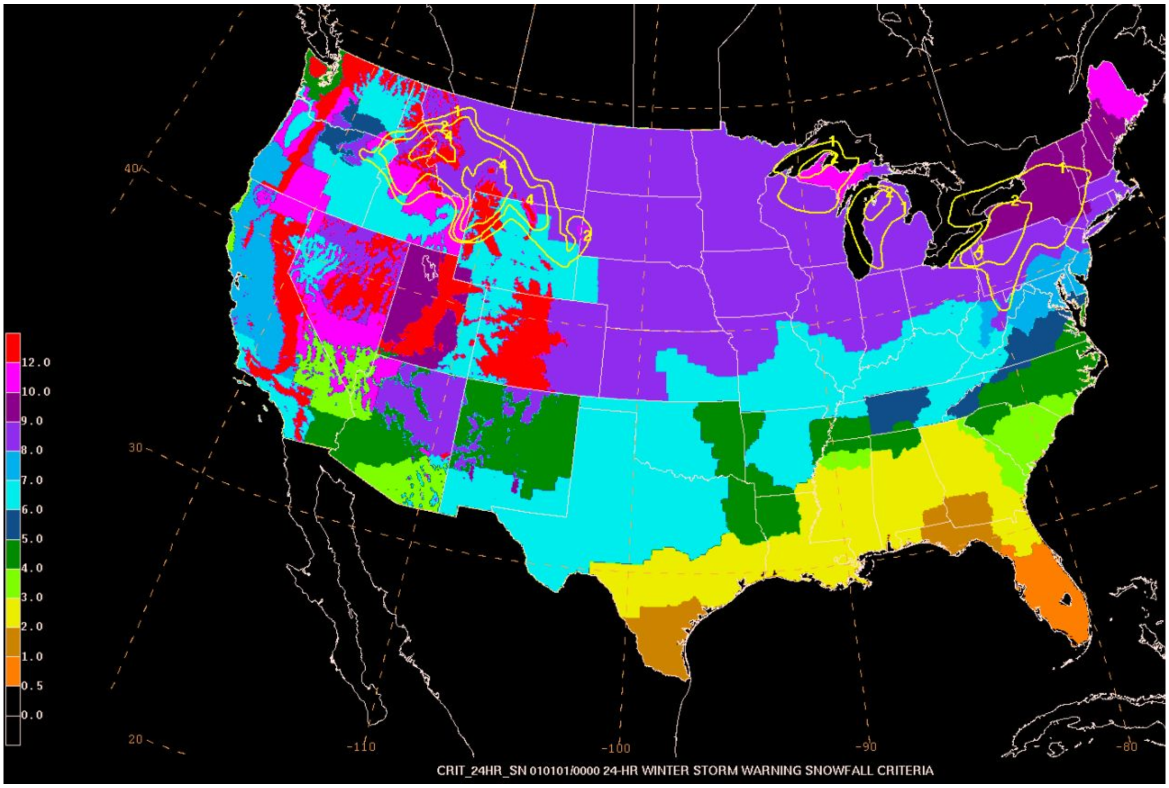


Figure 1. *An example of the participant's deterministic snowfall forecast (yellow contours) overlaid upon the WFO snow winter warning criteria to find areas of exceedance valid February 2, 2017*

During the afternoon, participants examined the latest experimental and operational short-range guidance to assess snowfall for the **Day 1 (18 - 12 UTC) period**. Choosing an active-weather area over a limited domain and a temporal window of 3 hours or more for peak impact, participants issued one or more probabilistic snowfall rate forecasts which included:

1.) The probability of exceeding 0.5", 1", 2" or 3" of snowfall per hour over the forecast period.
2.) An indication of the time period within which the high-impact snowfall rate is expected to occur.

Throughout the forecast process, participants identified specific guidance that supported their forecast rationale. These were saved as images were inserted into a forecast discussion PowerPoint presented during a weather briefing provided to interested local NWS offices each afternoon.

Lastly, participants stepped through daily science questions with the HMT facilitators collecting subjective evaluation and information on the performance of both the experimental forecasts and the experimental model guidance. In addition, at the end of each week participants were asked to provide feedback about the select experimental forecast tools.

Verification

Evaluations of any questions where a 24 hour snowfall analysis was needed (the Day 2 deterministic snowfall forecast, the WPC Deterministic Implicit Blend/5-km Winter Weather Ensemble and PWPFs), utilized version one of National Operational Hydrologic Remote Sensing Center's (NOHRSCv1) two-day quality-controlled 24 hour 1200 UTC NOHRSC snowfall analysis. Data sources for the analysis include all possible observation networks (e.g. COOP and CoCoRaHS) and a spatial interpolation of these observations is performed via a fixed, Barnes 2-pass, method with fixed interpolation parameters. A sample of a NOHRSCv1 analysis is shown in Figure 2.

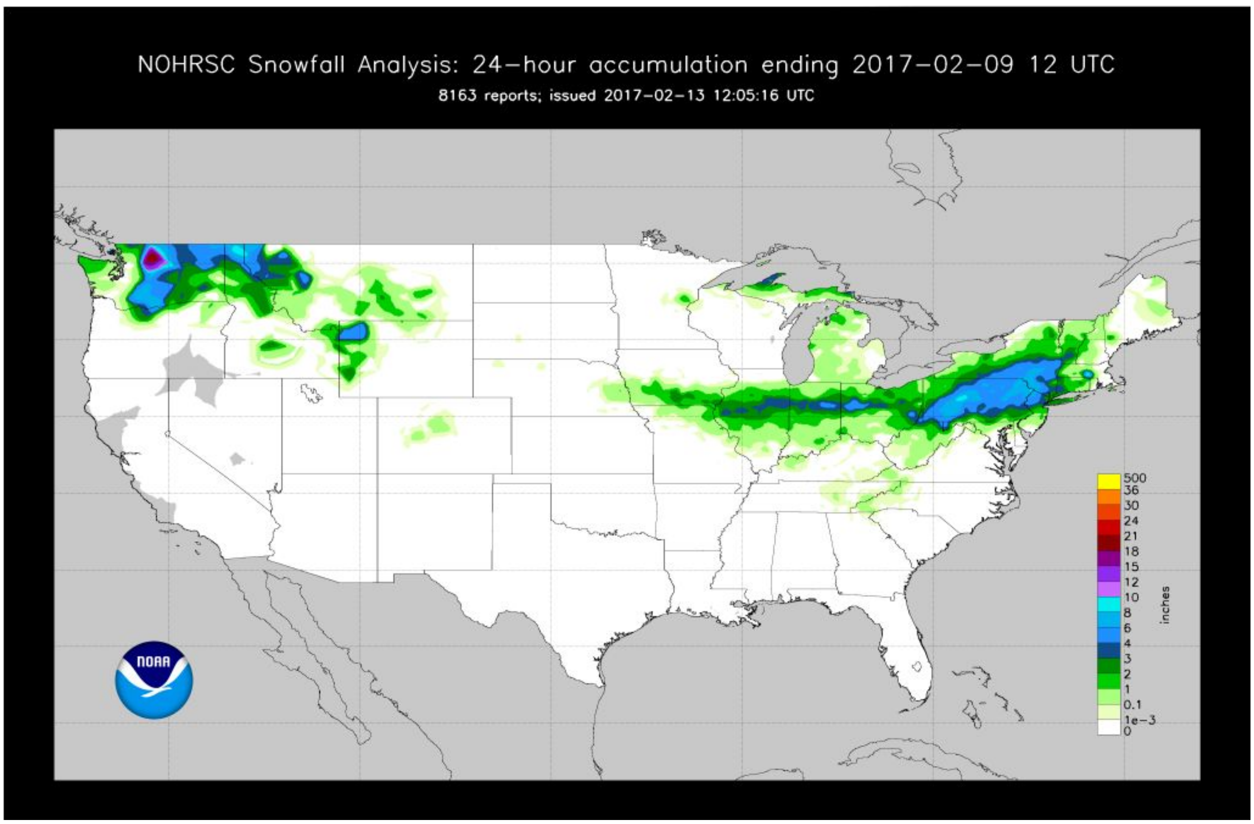


Figure 2. *NOHRSCv1 24-hour snowfall analysis valid at 12Z February 9, 2017. Figure from https://www.nohrsc.noaa.gov/snowfall/.*

An upgrade to the NOHRSC snowfall analysis began running in early January 2017. This developmental version (NOHRSCv2) was only available as a web graphic (available at: http://www.nohrsc.noaa.gov/snowfall_v2/). Although this version is classified as "pre-experimental and non-operational," it showed tremendous promise in capturing snowfall in the Western United States and was used to supplement the NOHRSCv1 in those areas. An example of these differences is shown in Figure 3. Some of the changes in version 2 include improvements to the automatic quality control, the inclusion of a bias-corrected first-guess field based on aggregated HRRR water equivalent snow depth, and the snow-to-liquid ratio now uses an updated version of the Baxter (2005) climatology.

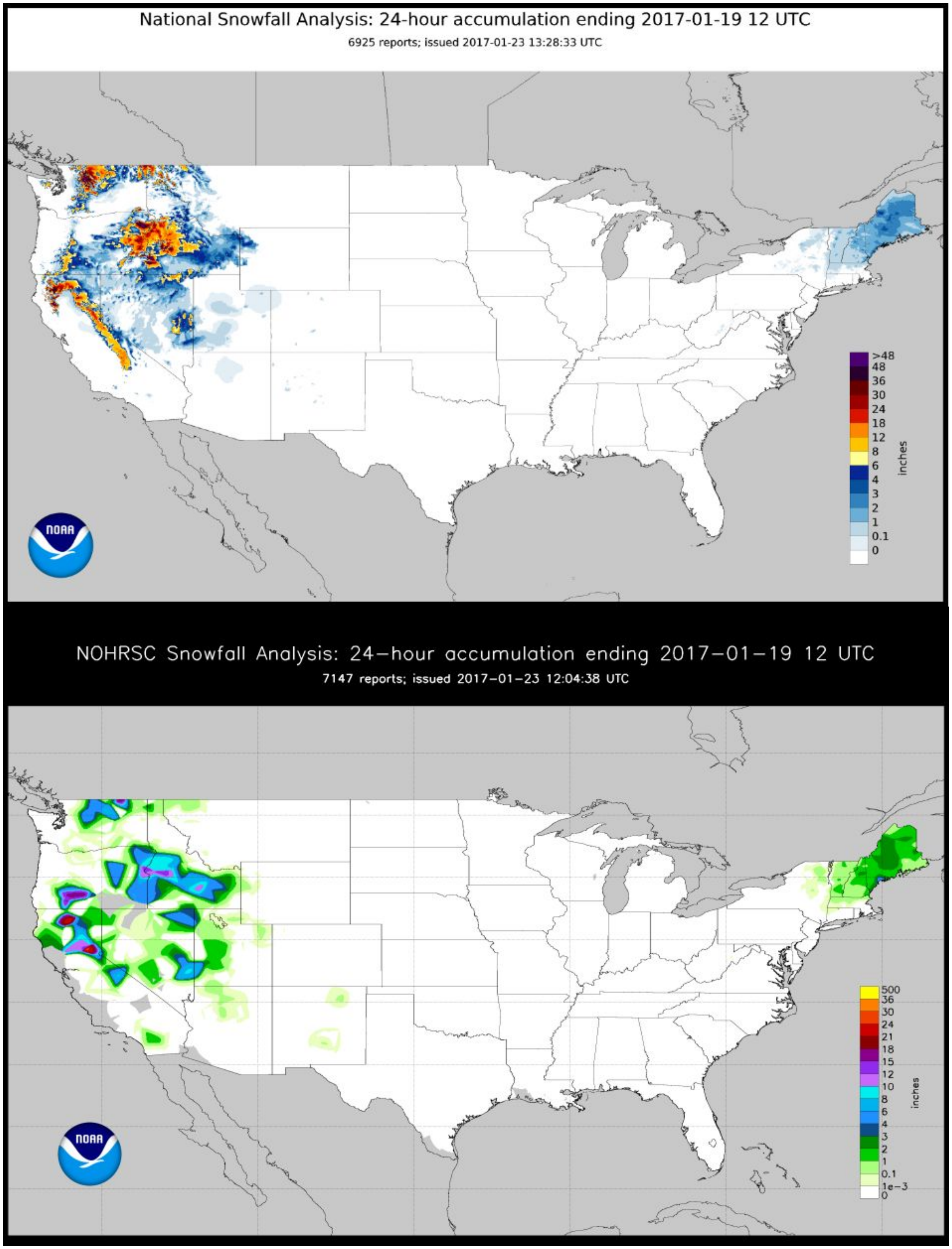


Figure 3. *NOHRSCv2 (top) compared to NOHRSCv1 (bottom) for 24-hour snowfall accumulation valid at 12Z on January 19, 2017.*

A combination of datasets were used in the verification of sub-24 hour snowfall accumulations required for the experimental probabilistic snowfall rate forecast and associated snowfall rate model guidance. The components of the verification methodology are Stage IV hourly precipitation data (Lin and Mitchell, 2005) and hourly initialization fields from the 13 km Rapid Refresh Model (RAP). The precipitation type is derived by employing a WPC algorithm which scans for freezing temperatures at 925 hPa, 850 hPa, and 700 hPa to estimate the depth of the cold layer as well as any melting and re-freezing layers to identify a sleet or freezing rain environment. Two-meter temperatures are used to distinguish freezing rain versus rain. If sleet, freezing rain, or rain are ruled out, snow is identified as the precipitation type where 850 hPa temperatures are less than 0 degrees Celsius (-2 degrees Celsius over higher terrain). For areas where snow is identified during a particular hour, a 10:1 snow-to-liquid ratio is applied to the hourly Stage IV precipitation data. For areas where sleet is determined to be the prevailing precipitation type, a 2:1 SLR is applied. The observed areas of snow and sleet are combined in a mosaic to plot estimated hourly snowfall amounts in 0.25" increments, as shown in Figure 4.

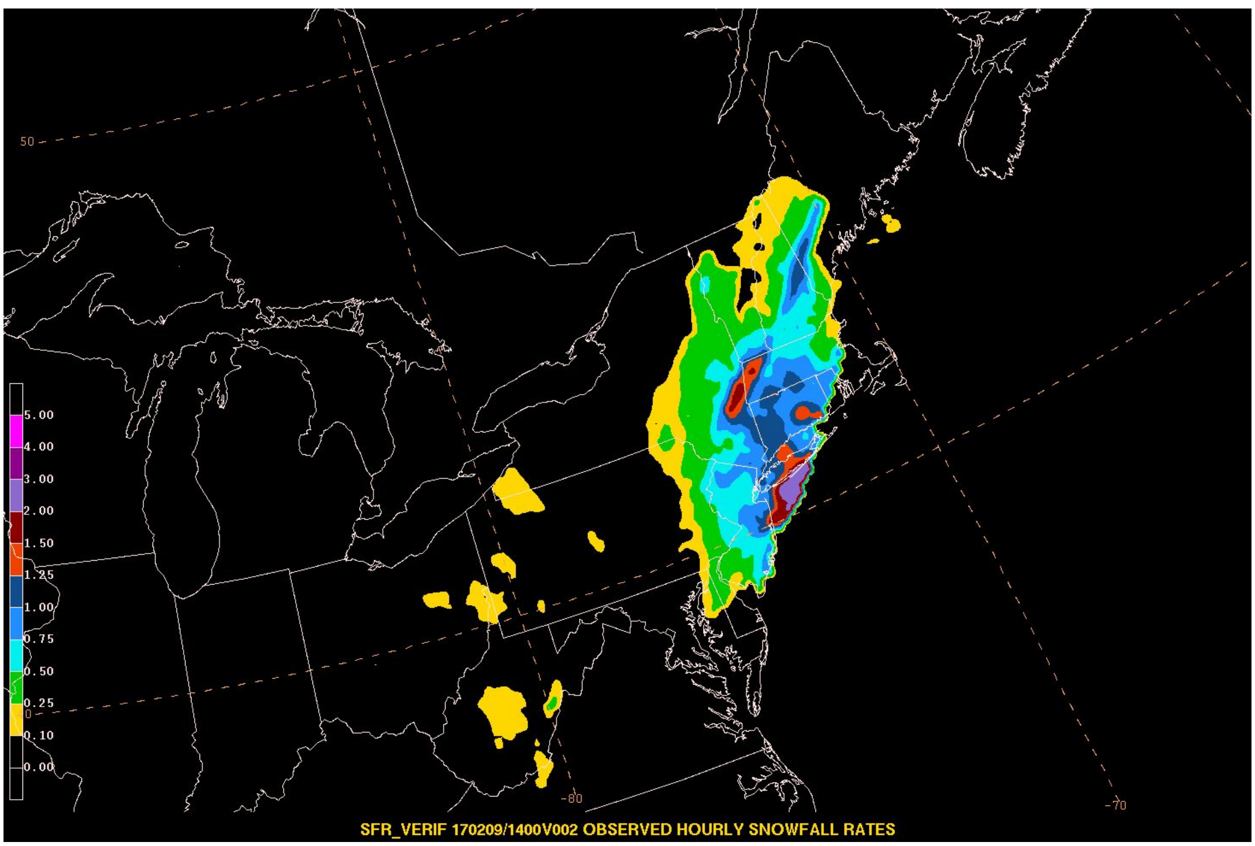


Figure 4. *Hourly snowfall accumulation based on the Stage IV/RAP Analysis valid at 1400Z February 9, 2017.*

The Stage IV/RAP analysis was also used as verification for the deterministic freezing rain forecasts, with METAR reports from stations that were located within the forecast contours used as a supplement. Stage IV quantitative precipitation estimates (QPE) were used to verify quantitative precipitation forecasts (QPF). Finally, precipitation type and precipitation rate from the the Multi-Radar, Multi-Sensor (MRMS) network were used to verify strictly frozen (snow) precipitation rates without a snow-to-liquid ratio. Only where the MRMS precipitation-type showed snow would the precipitation-rates be displayed.

Featured Guidance and Tools for Experimental Forecasts

In addition to the full multi-center suite of deterministic and ensemble guidance available to WPC forecasters, participants considered several different snowfall forecasting data sets and techniques while preparing their Day 2 deterministic snowfall and ice forecasts. Participants had access to the experimental HRRR Time-Lagged Ensemble (HRRR-TLE) system from the Earth System Research Laboratory (ESRL) as well as WPC's operational and experimental PWPF ensemble that is used to derive WPC's probabilistic winter precipitation forecasts. Model implicit snowfall guidance was also available for participants. Table 2 summarizes the model guidance that was available for the short range portion of the experiment, and more information about each dataset is provided below.

Table 2. *Featured Day 1-2 guidance for the 2017 HMT-WPC Winter Weather Experiment. Experimental guidance is shaded.*

Provider	Model	Resolution	Forecast Hours	Notes
EMC	NAM Nest	4 km	60	The operational NAM Nest is a higher resolution nest of the 12 km parent NAM
EMC	SREF (26 members)	16 km (32 km display)	87	Two dynamical cores (ARW and NMMB, 13 members each); Vertical resolution from 35-40 levels, mostly in the PBL; greater diversity in model initial conditions, perturbations, physics
EMC	GEFS (21 members)	55 km	168	Semi-Langrangian; GSI/EnKF hybrid analysis; vertical resolution of 64 levels
WPC	PWPF ensemble (70 members)	20 km	72	Operational PWPF ensemble; membership details follow below; SLR is an average of multiple techniques
EMC	Parallel NAMv4 Nest	3 km	60	3 km CONUS and 3 km Alaska nest
ESRL	HRRR-TLE (high-resolution time-lagged ensemble)	3 km	24	Neighborhood ensembling approach calculated over a 3 km grid of time-lagged HRRRv3 deterministic members
ESRL	HRRRv3	3 km	18 (all hours)	Deterministic Experimental HRRR

			36 (every 3 hours) 48 (00/12Z)	
NCAR	NCAR Storm Scale Ensemble	3 km	48	10-member high resolution, convective allowing model over the CONUS. 00Z only.
WPC	HRAM3E Ensemble	5 km	36	WPC multiplicative downscaling adjustment to hi-res multi-model ensemble
WPC	HRAM3G Ensemble	5 km	36	WPC multiplicative downscaling adjustment to hi-res multi-model ensemble
WPC	HRAM3G/E Winter Ensemble	5 km	36	Same as above but with winter fields including snow accumulation and precipitation type
WPC	Implicit Deterministic Super Blend	20 km	72	Blend of six model implicit snowfall solutions from GFS using Robber SLR, NAM12 rime factor modified Roebber SLR, NAM12 rime factor modified Baxter climatological SLR, the change in snow depth from the GFS, change in snow depth from the NAM, snowfall from 27km ECMWF, and European Center ensemble mean snowfall.
WPC	Implicit PWPF ensemble (72 members)	20 km	72	Experimental PWPF ensemble includes the WPC snowfall forecast, ECENS members, both ARW and NMMB SREF members, GEFS members, & Deterministic Super Blend
WPC	WPC Exp Winter Weather Ensemble	5 km	72	41-member ensemble with snow level grid added to PWPF computation for improvement over complex terrain

<u>**Operational Guidance**</u>

NCEP North American Model Nest (NAM Nest)

The operational NAM Nest is a 4 km nest of the 12-km parent NAM, which covers the CONUS and features full use of the global Ensemble Kalman Filter (EnKF) members as part of its data assimilation system. The nest is available at 00, 06, 12 and 18 UTC.

NCEP Short Range Ensemble Forecast System (SREF)
The SREF is a 26-member, 16 km ensemble consisting of an equal distribution of WRF-ARW, and NMMB members.

WPC PWPF Ensemble
The WPC PWPF Ensemble is a 70-member, 20 km ensemble that is generated internally by WPC and is used extensively in the WPC Winter Weather Desk forecast process. The ensemble membership consists of all 26 SREF members, 10 randomly selected GEFS members, 25 randomly selected ECMWF European ensemble (ECENS) members, and members consisting the latest operational WRF ARW, WRF NMMB, NAM Nest, GFS, GEFS mean, CMC, ECMWF, and ECMWF mean (ECENS) runs. The WPC 24-hour deterministic snow or freezing rain accumulation forecast is also an ensemble member serving as the mode or "most likely" solution, with the full ensemble providing the variance of the distribution. Snowfall from the WPC PWPF Ensemble is calculated using a five member snow-to-liquid ratio (SLR) average. These members include a micro-physics modification of the Roebber SLR (Roebber et al. 2007) applied to the NAM, a micro-physics modification of the Baxter climatological SLR applied to the NAM, the Roebber SLR applied to the GFS, a fixed 11:1 snow-to-liquid ratio, and Baxter SLR climatology.

GEFS
The Global Ensemble Forecast System (GEFS) is a weather forecast model made up of 21 ensemble members (including the GFS). The GEFS attempts to quantify the amount of uncertainty in a forecast by generating an ensemble of multiple forecasts, each minutely different, or perturbed, from the original observations. With global coverage, GEFS is produced four times a day with weather forecasts going out to 16 days. The GEFS is a Semi-Langrangian model with a horizontal resolution of 55 km from 0-168 forecast hours. GEFS has 64 hybrid levels in its vertical resolution to match the GSI/EnKF hybrid analysis system.

<u>Experimental Guidance</u>

NCAR Storm-Scale Ensemble
The 10-member, single-physics NCAR storm-scale ensemble goes out 48 hours which includes a 15-km data assimilation system using Tiedtke cumulus parameterization, Thompson microphysics, a MYJ PBL scheme, and NOAH land-surface model. The data assimilation is a combination of WRF-ARW and the NCAR Data Assimilation Research Testbed (DART) with continuously-cycled ensemble adjustment Kalman filter (EAKF). The ensemble itself offers 3 km grid spacing with initial conditions provided by downscaled members of the 00Z run of the WRF/DART EAKF analyses and perturbed lateral boundary conditions originating from the GFS forecasts.

EMC Parallel NAMv4
Parallel version of NAMv4 Nest using 3-km grid spacing. Maximum hourly snowfall rates are parameterized in the model. Hourly snowfall accumulations are derived by applying a rime

factor modified Baxter climatological SLR to the snowfall rate. The 00Z, 06Z, 12Z, and 18Z cycles will be available out to 60 forecast hours.

EMC/ESRL HRRR-TLE (High-Resolution Rapid Refresh Time-Lagged Ensemble)

This CONUS version of the ensemble will have a 3 km resolution, and use 3-5 time-lagged HRRRv3 initializations to predict the probability of exceeding several snowfall accumulations and accumulated precipitation thresholds. The HRRRv3 uses Thompson physics and a snow algorithm which calculates a variable-density accumulation of all frozen precipitation types, including contributions from snow, graupel and cloud ice. The density of each hydrometer class is calculated as a function of temperature at the lowest model level, with the final density a weighted average of snow, graupel and ice densities based on their relative mixing ratios. This scheme also allows for melting of new snow falling on a warm ground surface. The snow-to-liquid ratio for new snow, prior to any melting, can range from 4:1 to 13:1.

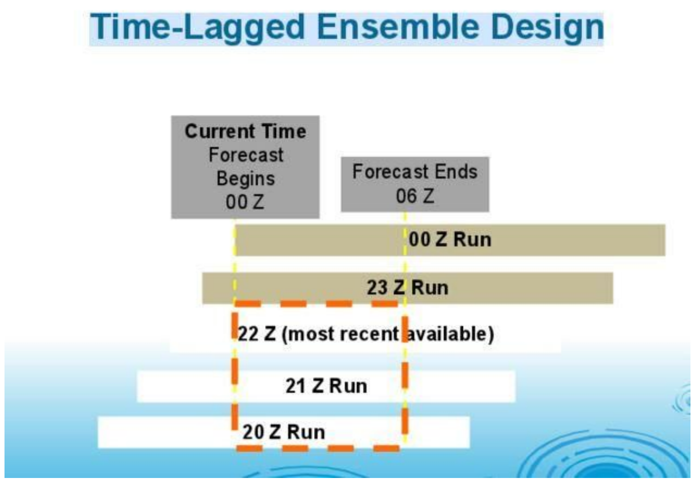


Figure 5. Proposed membership characteristics of the HRRR-TLE for use in the winter weather experiment.

EMC/ESRL Experimental High-Resolution Rapid Refresh (HRRRv3)

The Experimental HRRR (HRRRv3) is a WRF-ARW-based 3 km model initialized with the latest 3-D radar reflectivity using a digital filter initialization (radar-DFI) technique (via the parent 13 km RAP) and is updated hourly. The HRRRv3 uses grid-point statistical interpolation (GSI) hybrid GFS ensemble-variational data assimilation of conventional observations. Building upon the advancements in the operational HRRRv2 at NCEP, HRRRv3 includes assimilation of TAMDAR aircraft observations, refines assimilation of surface observations for improved lower-tropospheric temperature, dewpoint (humidity) winds and cloud base heights and places more weight on the ensemble contribution to the data assimilation. HRRRv3 adds assimilation of lightning flash rates and satellite-derived cloud-top cooling rates as complements to radar reflectivity observations through a similar conversion to specified latent heating rates during a one-hour spin-up period in the model.

HRRRv3 also contains numerous model changes including an update to WRF-ARW version 3.8.1 including the Thompson microphysics, transition to a hybrid sigma-pressure vertical coordinate for improved tropospheric temperature, dewpoint and wind forecasts along with a higher resolution (15 second) land use dataset. Physics enhancements have also been made to the MYNN planetary boundary layer (PBL) scheme and RUC land surface model along with additional refinements to shallow cumulus/sub-grid-scale cloud parameterizations including enhanced interactions with the radiation and microphysics schemes.

WPC HRAM3 Dynamic Downscaling
WPC has implemented several dynamical downscaling methods, and applies them in a multi-model ensemble approach to the QPF solutions from the NAM Nest, HIRESW NMMB, and HIRESW ARW. Both versions of the WPC HRAM being examined in the WWE (HRAM3E and HRAM3G) use a consensus multiplicative downscaling approach. This method computes two downscaling factors at each grid point. The first downscaling factor is a ratio of the HIRES member QPF after remapping to a 20-km coarse resolution then interpolating back to the higher resolution. The second is a ratio of the HIRES QPF to 20-km coarse resolution. The six downscaling factors are averaged to arrive at a consensus for each grid point. A 50% weight for PRISM is applied to HRAM3E, and a 50% weight west of 105 degrees W for HRAM3G.

Table 3. *High-resolution QPF model cycle times utilized for downscaling forecasts projected from the WPC initial cycle times.*

High-Resolution Model	Model cycle time for WPC 00 UTC cycle	Model cycle time for WPC 12 UTC cycle
NAM nest	12	00
HIRESW NMMB	12	00
HIRESW ARW	12	00

HRAM3E and HRAM3G are the result of a multiplicative downscaling adjustment. All of the methods make use of a probability determined by computing the relative frequency of QPF exceeding a WPC-created ensemble bias corrected QPF (ENSQPFBC) in a multi-model ensemble and one minus that probability.

WPC Deterministic Implicit Blend
A super blend of six model implicit snowfall solutions was available in 6 and 24 hour forecast periods during the experiment, and the 24 hour snowfall forecast was evaluated during the daily verification session. Members of the super blend include GFS using Roebber SLR, NAM12 rime factor modified Roebber SLR, NAM12 rime factor modified Baxter climatological SLR, the change in snow depth from the GFS, change in snow depth from the NAM, snowfall from 27km ECMWF, and European Center ensemble mean snowfall.

WPC PWPF Using Model Implicit Snowfall
A model implicit snowfall PWPF is computed using the same WPC snowfall forecast and cumulative binormal distribution methodology as the WPC PWPF. Seventy two member solutions of model implicit snowfall are used as variance to extract the probabilities. The

model implicit PWPF was available during forecast experiment forecast exercises and evaluated in concert with the operational WPC PWPF during verification sessions. The breakdown of members:

30 members from ECENS
5 SREF ARW members - change in snow depth
5 SREF ARW members - snow water equivalent derived snowfall
5 SREF NMMB members - change in snow depth
5 SREF NMMB members - snow water equivalent derived snowfall
10 GEFS members - change in snow depth
10 GEFS members - snow water equivalent derived snowfall Deterministic Super Blend WPC snowfall forecast
1 WPC snowfall forecast
1 Deterministic Super Blend

Total 72 members

WPC Experimental Winter Weather Ensemble

To address concerns regarding the precipitation type algorithm in the operational PWPF, an experimental version of the winter weather ensemble has been constructed. This experimental version performs most processing on the 5 km National Digital Forecast Database (NDFD) grid and introduces a snow level computation to help determine precipitation type.

The snow level is defined as the layer of the atmosphere at which the wet bulb temperature is equal to 0.5 C. This definition is based on the methodology used by Western Region's snow level blender graphical forecast editor (GFE) tool. To find the snow level, the algorithm starts at 500 hPa and searches downward until crossing the 0.5 C wet bulb isosurface. The search is conducted between the 500 and 1000 hPa geopotential height surfaces, with interpolation between available model layers. Using the 5 km RTMA terrain, a snow mask is constructed which identifies grid points above the snow level OR whose precipitation type is snow based on the precipitation type algorithm.

A blended snow-to-liquid (SLR) ratio is used to compute the SLR for each ensemble member. The blended SLR is the arithmetic mean of the following:

Roebber Artificial Neural Network (ANN; Roebber et al. 2007) applied to the GFS
NAM Rime-factor-modified Roebber ANN
NAM Rime-factor-modified Baxter climatological SLR
Baxter et al. (2005) climatology
Straight up 11/1 ratio

For each member, at each grid point that is determined to be snow, the total snowfall is equal to the fraction of QPF determined to be snow multiplied by the SLR.

The current experimental ensemble membership is as shown:

5 SREF ARW (09z; Control, First & Second | negative & positive perturbations)
5 SREF NMMB (09z; Control, First & Second | negative & positive perturbations)
13 GEFS Members (06z)
13 ECMWF Ensemble Members (00z)
1 GFS Deterministic (12z)
1 GEFS Ensemble Mean (06z)
1 NAM Nest (12z Days 1-2; Parent NAM Day 3)
1 ARW High-Res Run (00z; 06z GFS Days 2-3)
1 NMMB High-Res Run (00z; 00z GFS Days 2-3)

___
41 Members Total

Winter Storm Severity Index (WSSI)
The WSSI is made up of a series of sub-component algorithms which use meteorological and non-meteorological data to determine a level of potential societal impact based upon specific characteristics of winter storms. Each of the components produce a 1 to 5 output scale value that equates to the potential impact. The final WSSI value is the maximum value from all the sub-components. The 5 levels are given the following descriptors: Limited, Minor, Moderate, Major and Extreme. The specific sub-components are:

- Snow Load Index
    - Indicates a potential of downed trees/power lines due to the weight of the snow
- Snow Amount Index
    - Indicates a potential of impacts due to the total amount of snow or accumulation rate
- Ice Accumulation
    - Indicates the potential of tree and utility damage as well as transportation difficulties due to combined effects of ice and wind.
- Blowing Snow Index
    - Indicates the potential disruption due to blowing snow
- Flash Freeze Index
    - Indicates the potential impacts of flash freezing during precipitation events

The WSSI builds upon previous work in trying to “categorize” the severity of winter storms. NOAA/NECI has created the Regional Snowfall Index (RSI), though this is only calculated after the fact based upon measured snowfall,and the amount of area covered weighted by population. It does not take into account other impacts from winter storms such as ice, wind and blowing snow. As such, it is of limited operational use to NWS forecasters who may be trying to clearly communicate the combined impact of an upcoming winter storm.

Experimental Tools

Joint Probability Fields

In an attempt to identify multiple impacts probabilistically, joint probabilities were computed using various winter hazards for six hour intervals. Six-hour snow and ice probabilities were derived from the official WPC PWPF. To derive wind and temperature probabilities, the official wind and temperature forecasts from the NDFD were used as a mean and then bias corrected GEFS 10 m-winds/2 m-temperatures and SREF 10 m-winds/2 m-temperatures were used as a variance to create a CDF while using the specified threshold as the integration limit. Finally, the various individual probabilities were then multiplied together to create a variety of joint probabilities. Two examples include: the probability of snow accumulation greater than 2 inches and average wind speed greater than 15 mph in six hours, and the probability of ice accumulation of 0.10 inch and average wind speed greater than 15 mph in six hours. An example of a joint probability field and its component parts are shown in Figure 6.

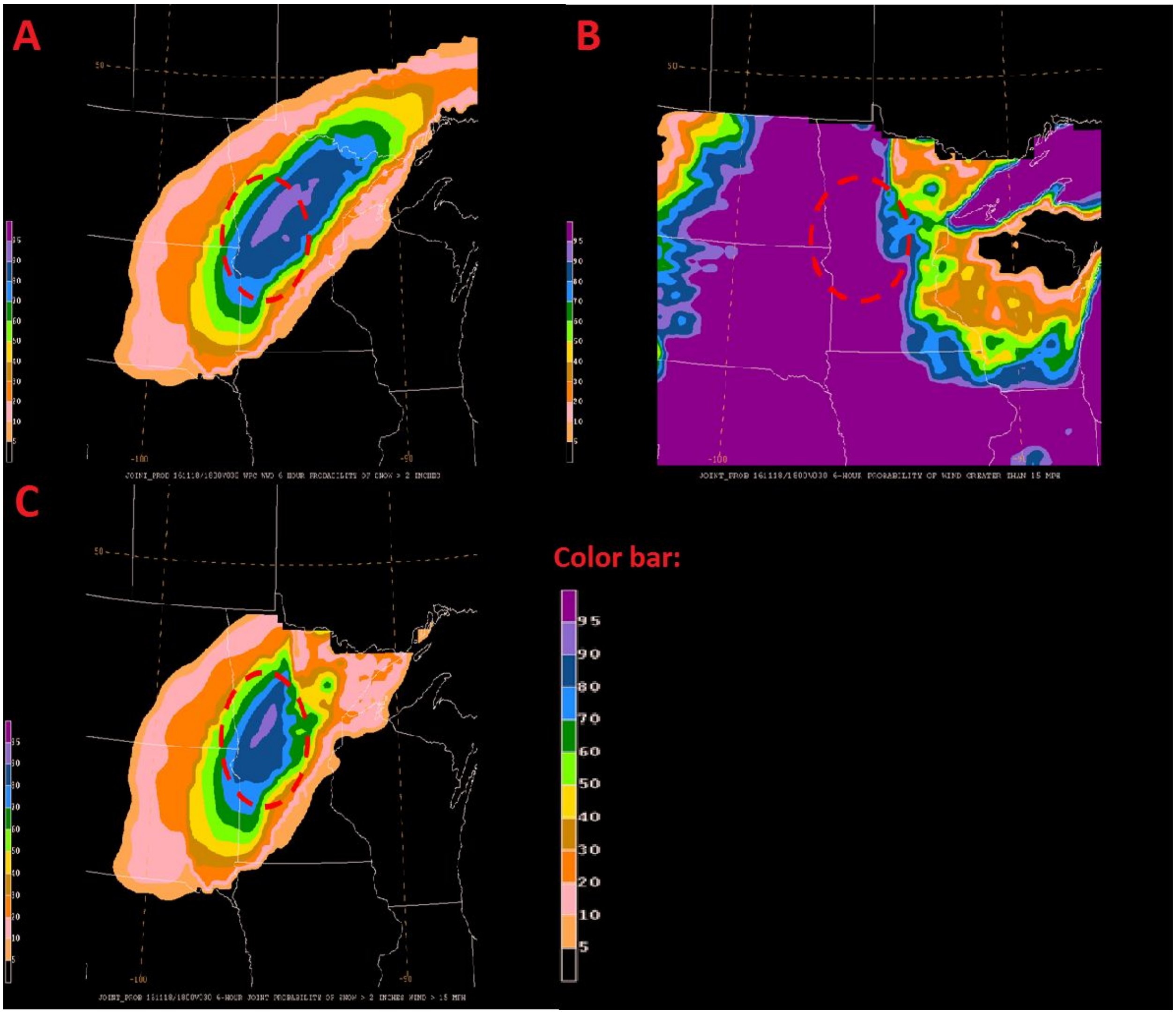


Figure 6. *(A) Probability of greater than 2 inches of snow in six hours; (B) Probability of greater than 15 mph average wind speed over six hours; (C) Joint probability of both greater than 2 inches of snow and greater than 15 mph average wind speed over six hours.*

Watch Collaborator Trend Tools

The WPC Watch Collaborator (WC) predicts the probability of a location exceeding 12- or 24-hour winter storm warning criteria based on the WPC PWPF forecast and the individual weather forecast office (WFO) winter storm warning criteria. The WC is available to all WFOs via an internally hosted WPC webpage. For this year's experiment, three different tools were developed to identify WC trends from cycle to cycle: A count tool which displays how many times the 30 and 50% probabilities were forecast from the WC over three consecutive cycles, a "flip-flop" tool which takes the three most recent WC cycles and displays areas where the probabilities changed either higher or lower between the cycles, and a difference tool which simply subtracts the current WC probabilities from the previous cycle's probabilities and determines the changes from one cycle to the next. An example of all three is shown in Figure 7.

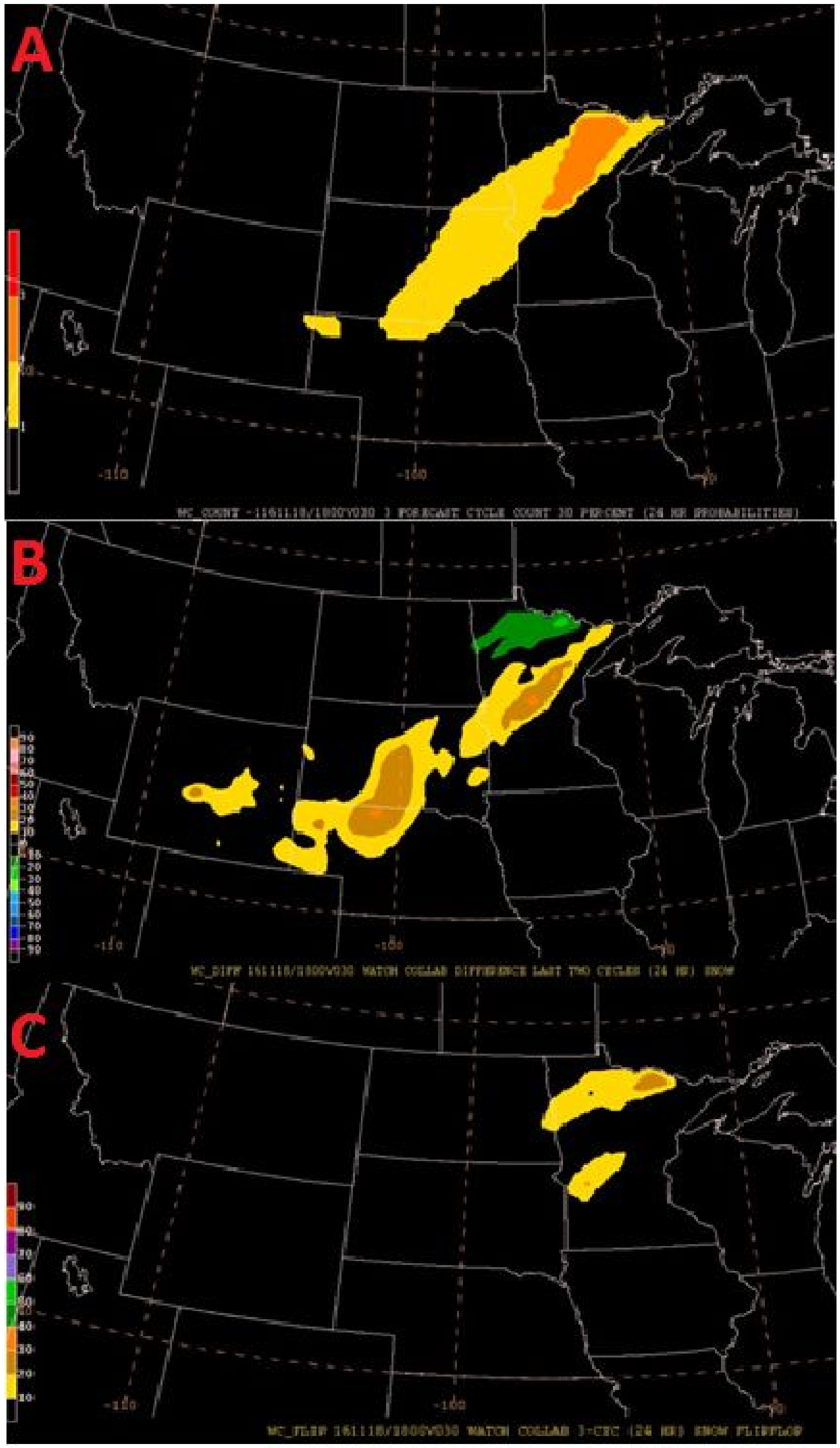


Figure 7. *Examples of the (A) watch collaborator count tool, (B) watch collaborator difference tool, and (C) the watch collaborator flip-flop tool valid at 18Z on November 18, 2016.*

4. Cases

The first week of the 2017 Winter Weather Experiment (January 19-21; when forecasts were valid) was dominated by a sharp 500 hPa trough on the west coast with a corresponding 500 hPa ridge over the eastern third of the country as seen in Figure 8A. This translated to surface temperature anomalies being 6-10 degrees above normal for a large portion of the United States with only some western and southwestern areas recording near normal surface temperature anomalies. The temperature anomalies can be seen in Figure 9A. Two weeks later for the second week of the experiment, the 500 hPa height pattern in Figure 8B, had shifted. A broad trough was located in the east with low heights especially in New England and some slight ridging in the western states. Despite the predominant trough in the east, the surface temperature anomalies in Figure 9B show that temperatures were near normal for much of the Northeast and Great Lakes region with the rest of the country around 2-6 degrees above normal, except for the far Northern Plains. Week three mean 500 hPa heights in Figure 8C were very similar to week 2. A subtle ridge was located in the west, with a trough in the east. Surface temperature anomalies in Figure 9C show much of the country was between 2-10 degrees above normal, with the exception being the state of Maine which had negative temperature anomalies. The 500 hPa mean height field is displayed in Figure 8D for week 4 of the experiment and shows a divergent pattern on the West Coast. The flow splits off the coast of California creating a trough south of Baja, Mexico and a ridge that dominates over much of the U.S. except for Great Lakes and Northeast regions. Surface temperature anomalies in Figure 9D for week 4 again show the majority of the country with positive anomalies except for the southern periphery of the U.S. where anomalies were near normal.

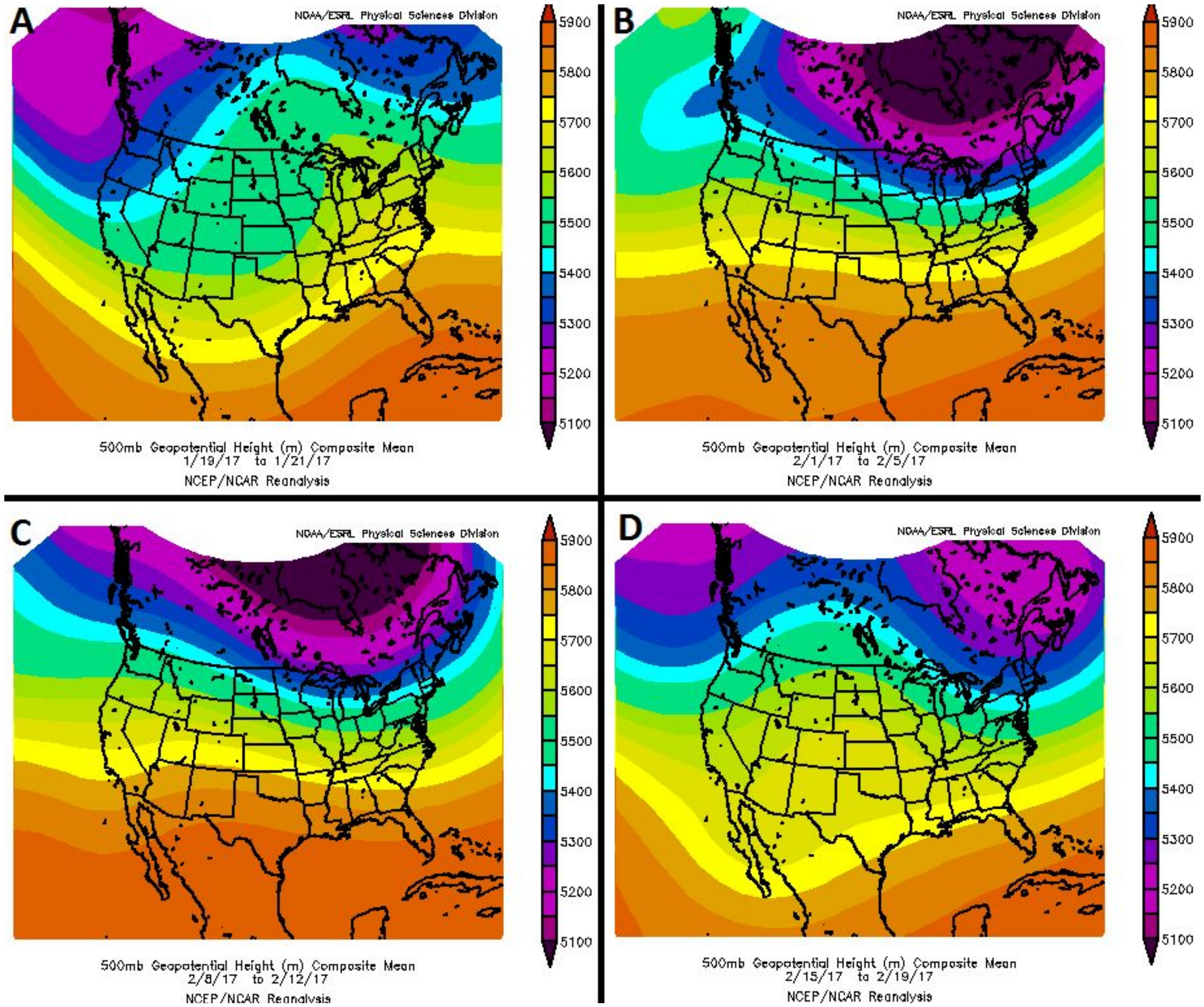


Figure 8. *Composite mean 500 hPa for the forecast valid dates during the four week experiment period: A) January 19-21, B) February 1-5, C) February 8-12, and D) February 15-19, 2017. Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/PSD (http://www.esrl.noaa.gov/psd/data/composites/day/).*

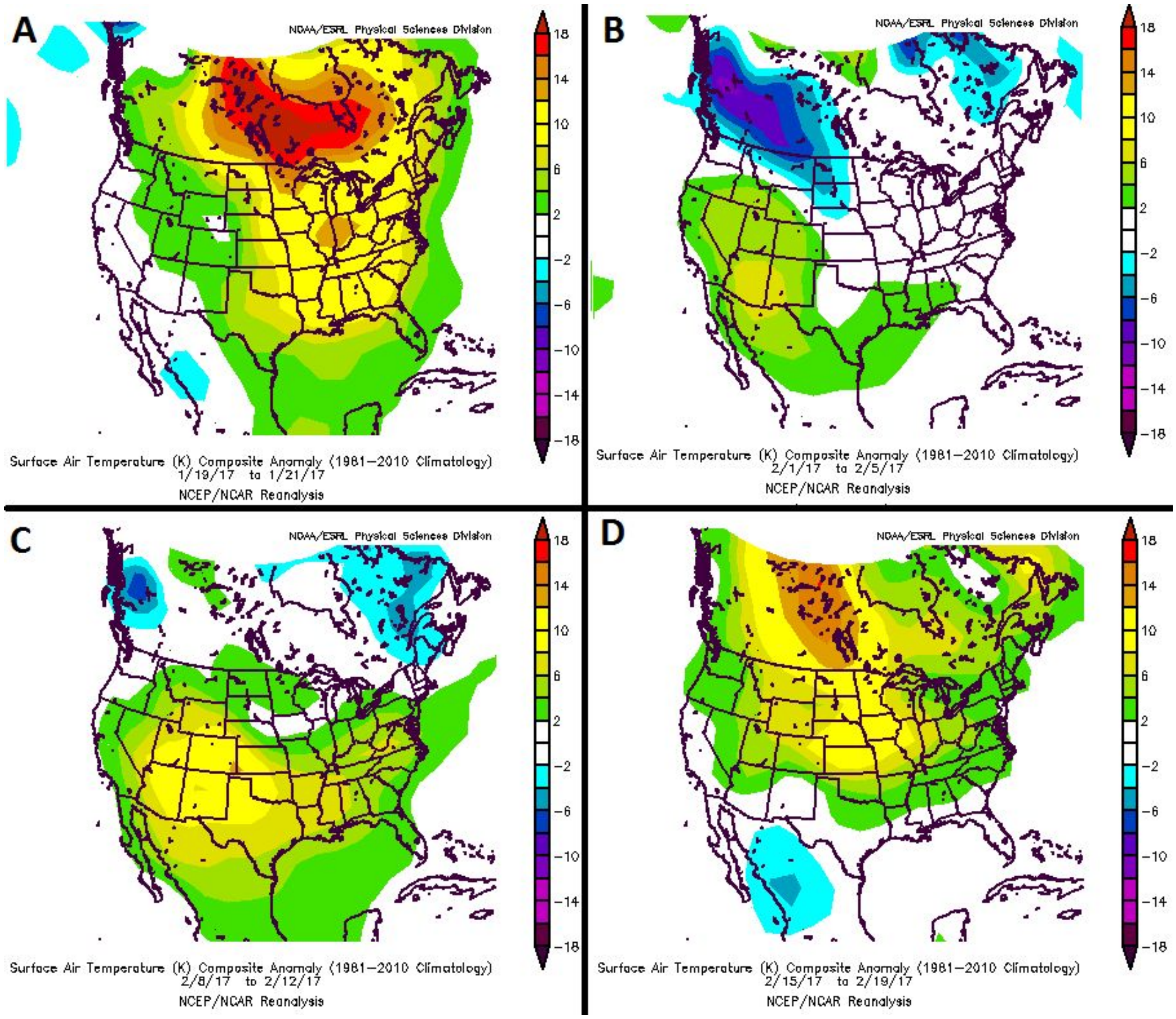


Figure 9. *Composite mean surface temperature (K) anomalies for the forecast valid dates during the four week experiment period: A) January 19-21, B) February 1-5, C) February 8-12, and D) February 15-19, 2017. Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/PSD (http://www.esrl.noaa.gov/psd/data/composites/day/).*

Throughout the four week experiment, the majority of the events were in the Western United States. Table 4 lists each of the days and areas of focus. For the first time in the Winter Weather Experiment, the deterministic snowfall forecast was issued for the entire CONUS rather than picking a limited area of focus. Often there were multiple parts of the country with snow in the forecast. Nine out of eighteen possible forecasts were for snowfall that fell exclusively in the Western United States (Rocky Mountains and west). There were no major storms in the central United States during the four weeks of the experiment. During the first week (forecasts valid January 19-21, 2017), there were major snow events in the Sierra Nevadas in California as well as in the Southwest United States on January 21. Lake effect snow was a forecasting challenge on seven of the eighteen possible forecast days. A major storm affected the Mid-Atlantic and New England during the third week of the experiment on February 9-10, 2017 causing blizzard conditions for coastal New England and Long Island, New York. Finally, a strong storm dropped over a foot of snow in southern Maine and northern New Hampshire on February 16, 2017 during the final week of the experiment.

Table 4. *Forecast valid dates and corresponding forecast area over the eighteen days forecasts were issued during the course of the experiment.*

Forecast Valid Date (12Z)	*Forecast Area*
1/19/2017	West (Sierras); Pacific Northwest
1/20/2017	West (Sierras/AZ/UT/ID)
1/21/2017	West/Southwest (Sierras/NV/UT/AZ/OR)
2/1/2017	West (ID/MT/WY); Great Lakes
2/2/2017	West (WY); Great Lakes; Maine; West Virginia
2/3/2017	West (Sierras, WY); Great Lakes
2/4/2017	West (Sierras, ID, OR, WA); Great Lakes
2/5/2017	West/Pacific Northwest (OR, WA, ID, MT, WY); Michigan
2/8/2017	West (Sierras, UT, CO, WY); Upper Great Lakes; New England
2/9/2017	West/Pacific Northwest (WA, ID, WY, CO); Mid-Atlantic (PA, OH, WV, MD)
2/10/2017	West (Sierras, ID, WA, WY); Mid-Atlantic; New England; West Virginia
2/11/2017	West (Sierras, OR, WA, WY, UT, ID); New England (NY, MA, VT)
2/12/2017	West (CO); New England (ME)
2/15/2017	Great Lakes (MI, NY)
2/16/2017	West (WA); New England (NY, VT, NH, ME)
2/17/2017	West (Sierras, WA, ID); New England (NY, VT, NH, ME)
2/18/2017	West (Sierras, NV, ID)
2/19/2017	West/Southwest (Sierras, AZ, UT, CO, ID, NV)

5. Hourly Snowfall Guidance

Deterministic Guidance

The 3km NAMv4 offered two different hourly snowfall parameters to the 2017 WWE. The first was a maximum hourly frozen precipitation rate to assess the potential for heavy snowfall rates based on vertical motion and precipitation efficiency in the model. The second was a derived hourly snowfall accumulation taking the aforementioned hourly frozen precipitation rate and applying riming data from the model microphysics to adjust the Baxter climatological snow-to-liquid ratio (Figure 10 right).

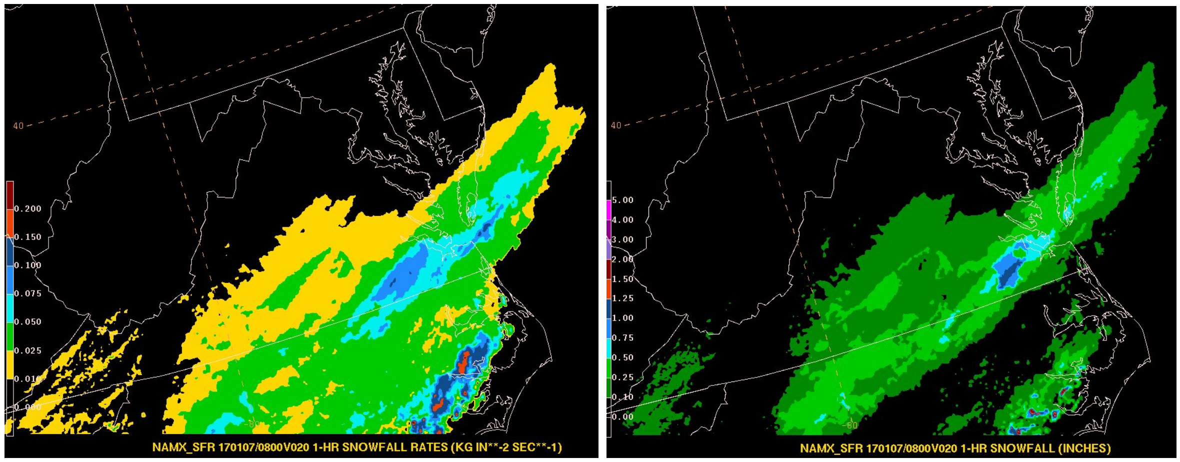


Figure 10. *(left) is an example of the Max Hourly Frozen Precipitation field with no SLR and (right) is an example of the NAMv4 1-hour snowfall accumulation with modified Baxter SLR valid January 7, 2017*

Generally, both NAMv4 fields were evaluated as being too light as a snowfall forecast when compared to observed data. However, the participants indicated that the hourly snowfall fields provided useful guidance when determining spatial extent and timing of the experimental probabilistic snowfall rate forecasts. The NAMv4 maximum 1-hour frozen precipitation rate (Figure 10 right) helped identify areas where more intense snowfall and sleet bands were forecast.

Also featured was the HRRRv3 1-hour snowfall accumulation (Figure 11b) which comes directly from the model microphysics. Different from the NAMv4, however, is a basic 10:1 SLR is applied to frozen QPF as it interacts with the model land surface parameterization. Participants commented frequently throughout the experiment that although the spatial extent of this field had the general idea of the event occurring, magnitude of the snowfall was too underdone to be useful. The light amounts generated from the HRRRv3 would have misled forecasters regarding the intensity of the snowfall events. The HRRRv3 was consistently lighter in amounts than the NAMv4.

The NCAR Ensemble provided more dispersive guidance to the hourly snowfall rates with its 1-hour mean hourly snowfall accumulation (Figure 11d). Overall, the multi-member ensemble did well by improving the general spatial extent and increasing the magnitude of snowfall accumulations, although was geographically offset at times (most often northward, as in the Figure 11 example). Ensembling was an improvement upon the deterministic products at capturing the magnitude and intensity of the snow events.

Comments: In slightly fewer than half of the 14 events evaluated, participants reported that the snowfall from the NCAR ensemble closely resembled verification. Despite this improvement, there were cases in which the ensemble struggled with intensity.

Suggestions: Provide a probability-matched mean to improve the snowfall accumulation output. It was also suggested that a mean for snowfall accumulation may not have forecaster usability due to the brute-force averaging not producing a physically consistent structure to the field.

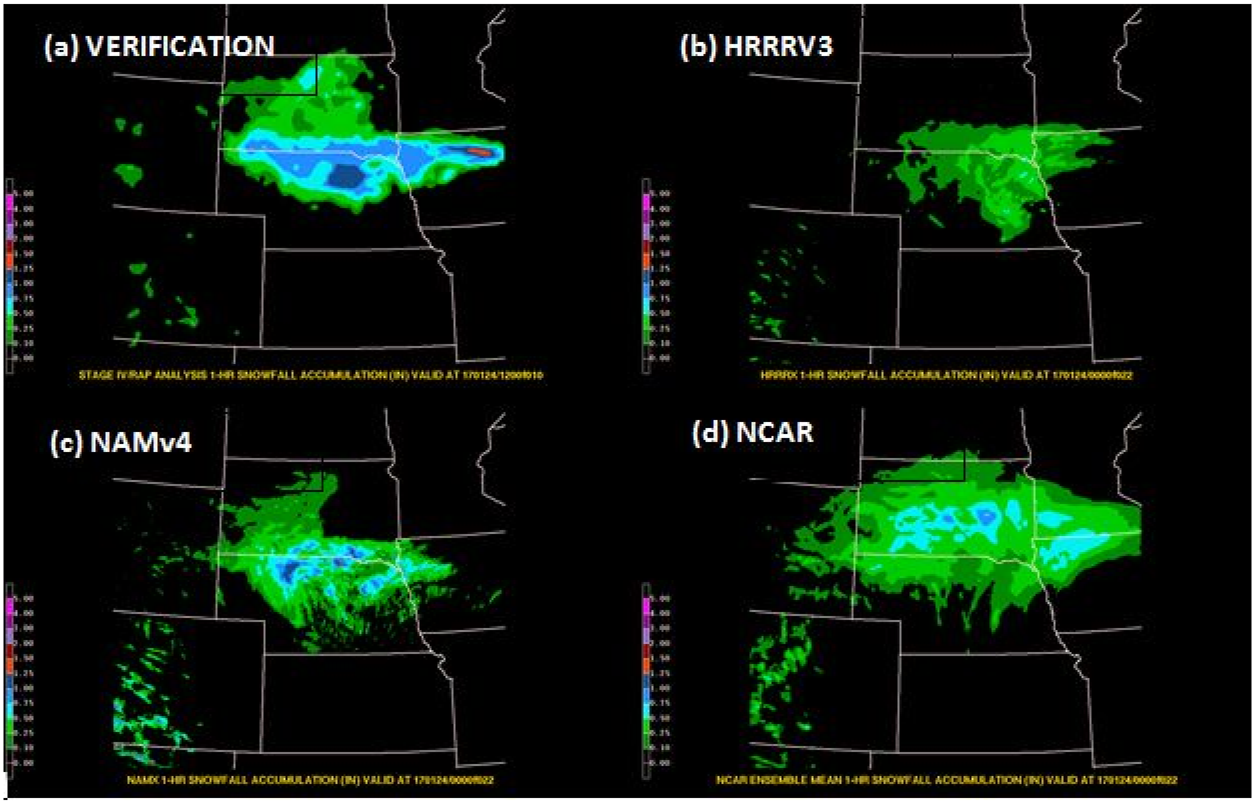


Figure 11. *1-hour snowfall valid Jan. 24, 2017. a - Verification (Stage IV /RAP Analysis ; b - HRRRv3 1-hour snowfall accumulation, c- NAMv4 1-hour snowfall accumulation , d- NCAR Ensemble 1-hour mean snowfall.*

The NAMv4 1-hour snowfall accumulation and NCAR Ensemble hourly mean snowfall accumulation received the same average participant score of 5.5 out of 10. However, as depicted in Figures 12 and 13, the distribution of the scoring was quite different. The NAMv4 scores were more variable from event to event and obtained high scores (7 or above) more often than the NCAR Ensemble. The smaller median distribution of the NCAR Ensemble likely indicates the consistency of its performance throughout the experiment.

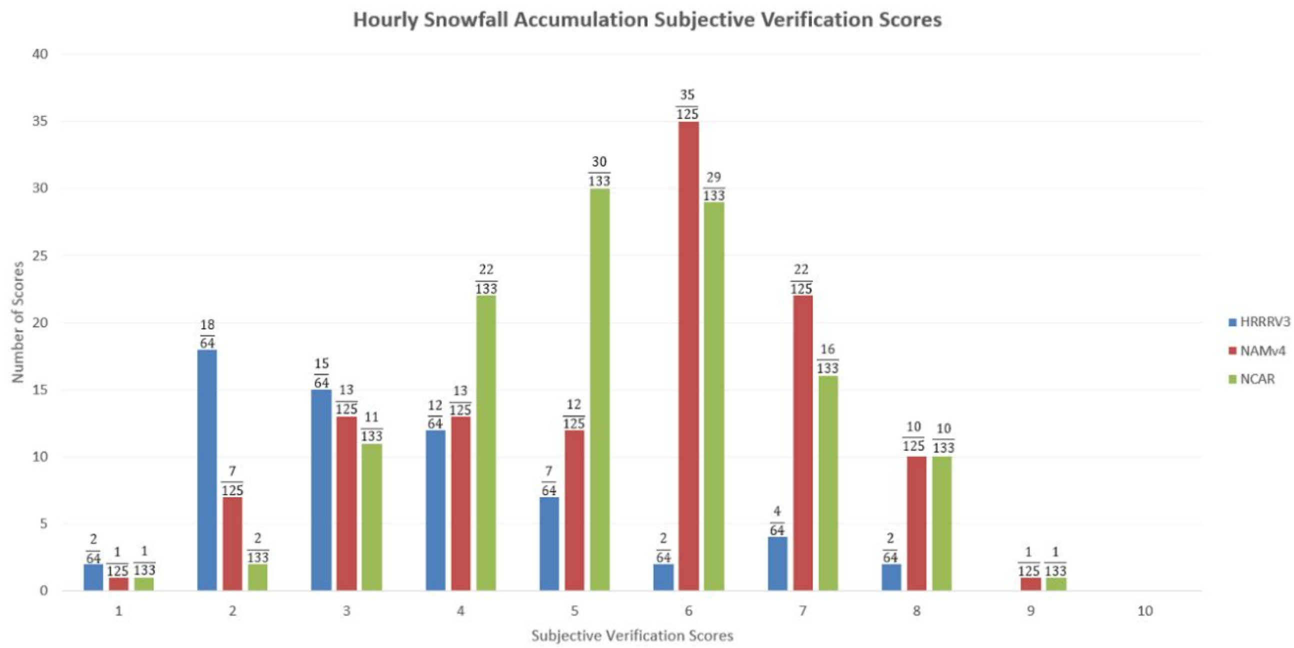


***Figure 12**. The 2017 WWE distribution of subjective verification scores for the hourly accumulated snowfall guidance from the HRRRv3, NAMv4, and NCAR Ensemble. The products were scored each day they were available on a scale of 1 (very poor) to 10 (very good).*

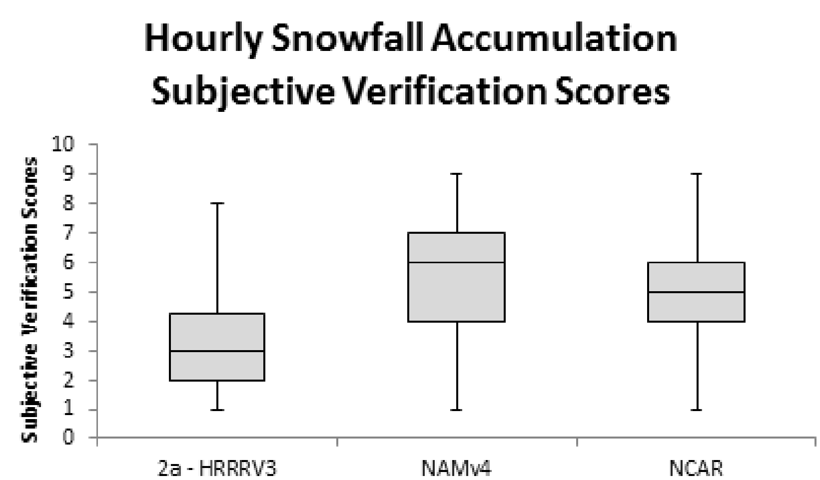


***Figure 13**. The box plot for the 1-hour snowfall accumulation guidance from the HRRRv3, NAMv4, and NCAR Ensemble showing the median dispersion of participant scores.*

The HRRRv3 1-hour snowfall accumulation rate received an average score of 3.6 out of 10. The low score is reflective of its consistent low bias, and the significantly fewer cases participants had to evaluate due to outages and missing data.

Probabilistic Guidance

In addition to deterministic one hour snowfall accumulations that were examined each day, one hour probabilistic snowfall rates from the NCAR Ensemble and the HRRR-TLE were evaluated subjectively on a 1-10 scale based on the deterministic Stage IV/RAP analysis used for verification. The HRRR-TLE provided 40 km neighborhood probability of half inch and one inch snowfall rates per hour. The NCAR Ensemble also provided a 40 km neighborhood probability at the same thresholds but additionally included grid point-based probabilities. It should be noted that for verification, the NCAR Ensemble was only available once a day and initialized at 00Z, and because the HRRR-TLE only forecasts out 24 hours, it was necessary to use the 12Z cycle in order to cover the entire 18-12Z forecast period.

All of the subjective scores for the two different models and probability displays at each threshold were very similar, with mean scores over the entire experiment ranging from a 4.8 to a 5.6 out of a possible 10. Figure 14 is the box plot showing statistics of the subjective verification scores for each model at each threshold. Figure 15 shows the distribution of scores throughout the experiment for the half inch threshold and Figure 16 shows the distribution for the one inch threshold. The median for all fields was either a 5 (NCAR ensemble, both probabilities/thresholds) or a 6 (HRRR-TLE both thresholds) and the standard deviation of the subjective scores was between 1.6 (NCAR Ensemble 1” neighborhood probability) and 1.9 (NCAR Ensemble 0.5” neighborhood probability). Additional HRRR-TLE 1-hour snowfall objective verification performed by NCAR/Development Training Branch are in **Appendix A**.

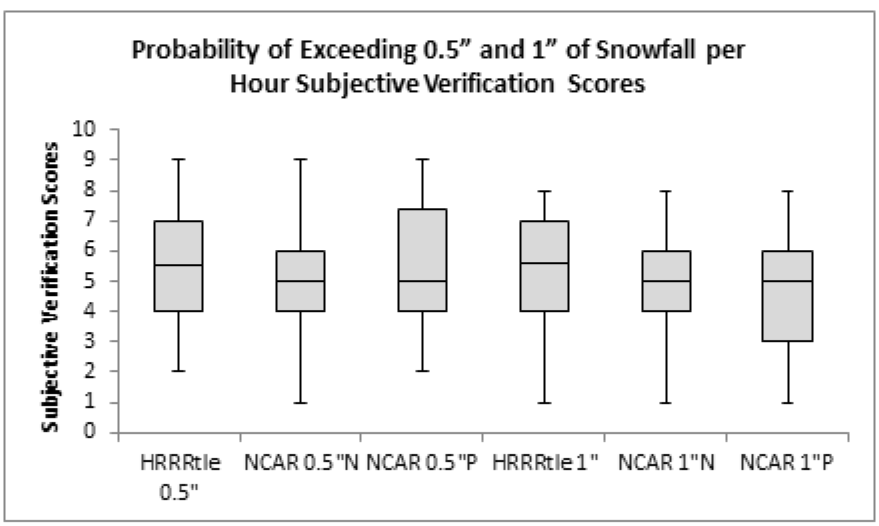


Figure 14. *Box plot of the subjective scores for the HRRR-TLE, NCAR Neighborhood, and NCAR Point probabilities of exceeding 0.5” and 1” of snowfall per hour.*

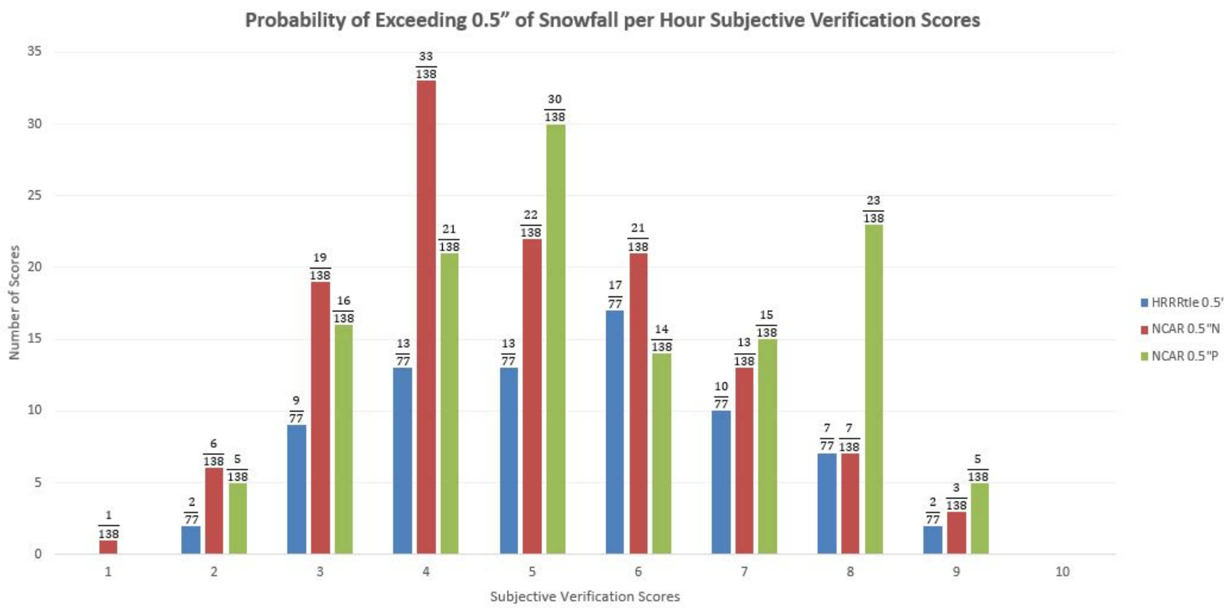


Figure 15. *Distribution of scores for the HRRR-TLE, NCAR Neighborhood, and NCAR Point probabilities of exceeding 0.5" of snowfall per hour.*

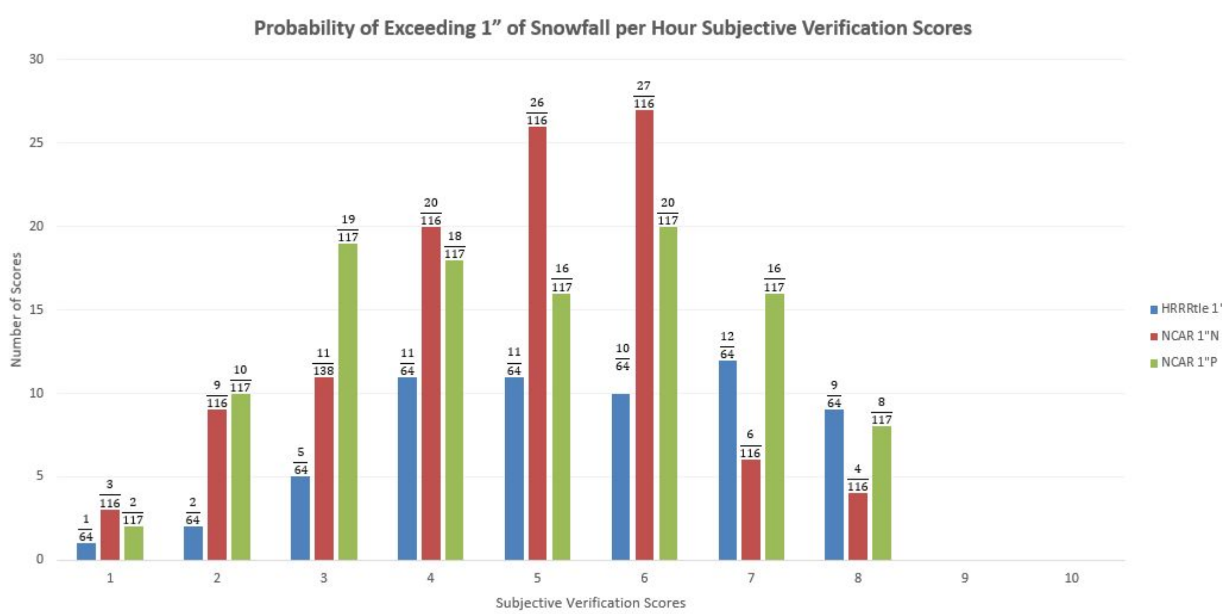


Figure 16. *Distribution of scores for the HRRR-TLE, NCAR Neighborhood, and NCAR Point probabilities of exceeding 1" of snowfall per hour.*

Although the subjective scores were very similar between the two models at both thresholds, participants had valuable feedback each day on how each model and probability display performed for the variety of different cases examined during the experiment. Figure 17 shows an example of a more synoptically-forced event over Maine at the one inch threshold valid at

00Z on February 10, 2017. Figure 18 shows an example out in the Western United States at the half inch threshold valid at 02Z on January 20, 2017.

Comments: Participants frequently discussed throughout the experiment the preferred display of probabilistic snowfall information. Neighborhood probabilistic techniques are more common in severe weather forecasting and forecasters are more accustomed to viewing them in that context. For smaller scale lake-effect events, snow events in the mountains like what is shown in Figure 18, or banded structures in larger scale systems, many participants preferred the point probabilities, citing that the point probabilities more closely resembled the spatial distribution of the snowfall event and the neighborhood probabilities would often be too broad leading to "false-alarms." However, some of the model developer participants felt probabilistic guidance should not strive to accurately portray the exact shape of the weather phenomenon, but merely serve to show the probabilistic chance of occurrence.

Suggestions: A compromise between the two for testing future displays, in which the neighborhood probabilities are still present but possibly combined with a point-based probability from a control member of the ensemble.

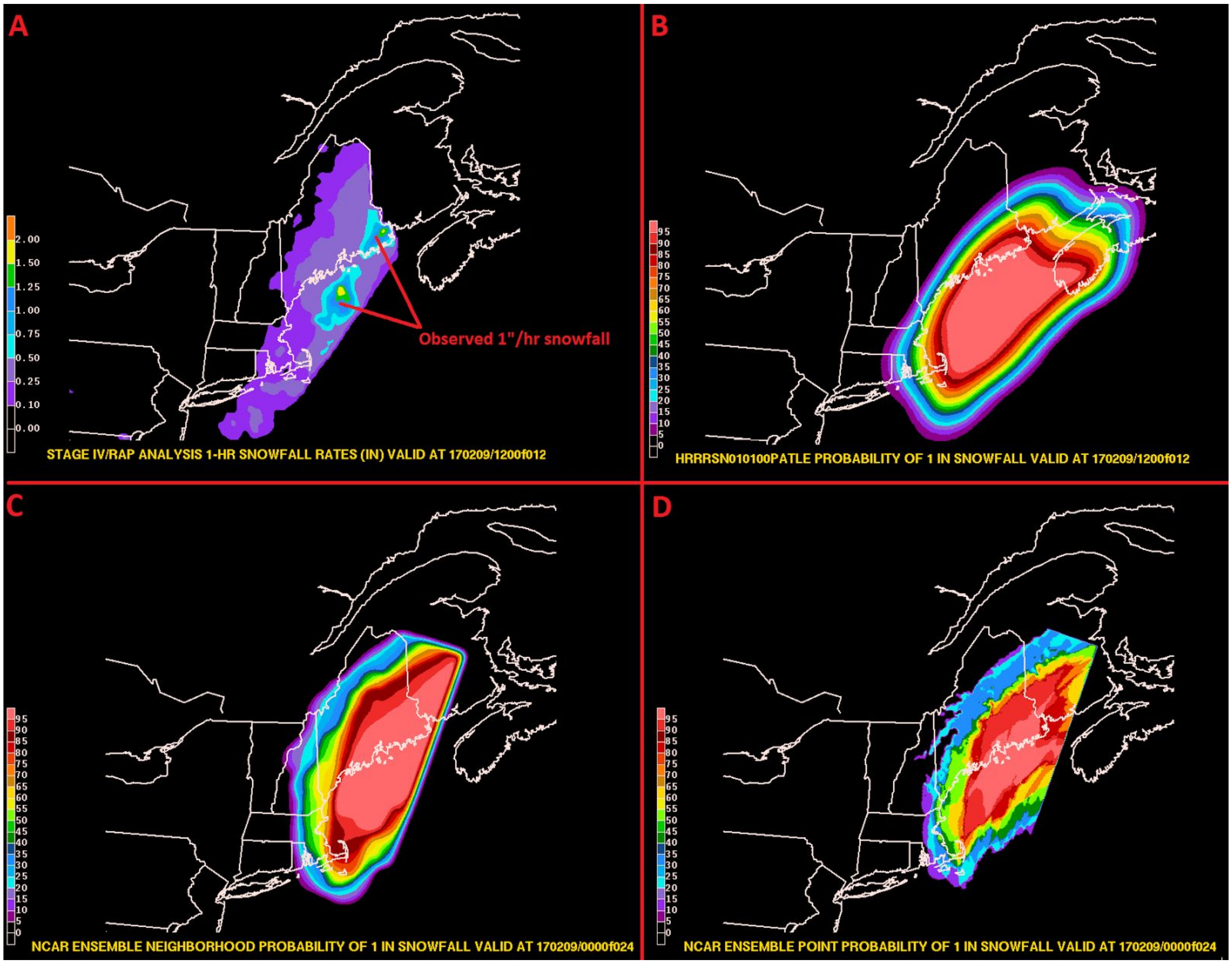


Figure 17. *(A) Stage IV/RAP analysis observed snowfall used for verification; (B) HRRR-TLE neighborhood probabilities for one inch snowfall rates; (C) NCAR Ensemble neighborhood probabilities for one inch snowfall rates; (D) NCAR Ensemble point probabilities for one inch snowfall rates; All valid at 00Z February 10, 2017.*

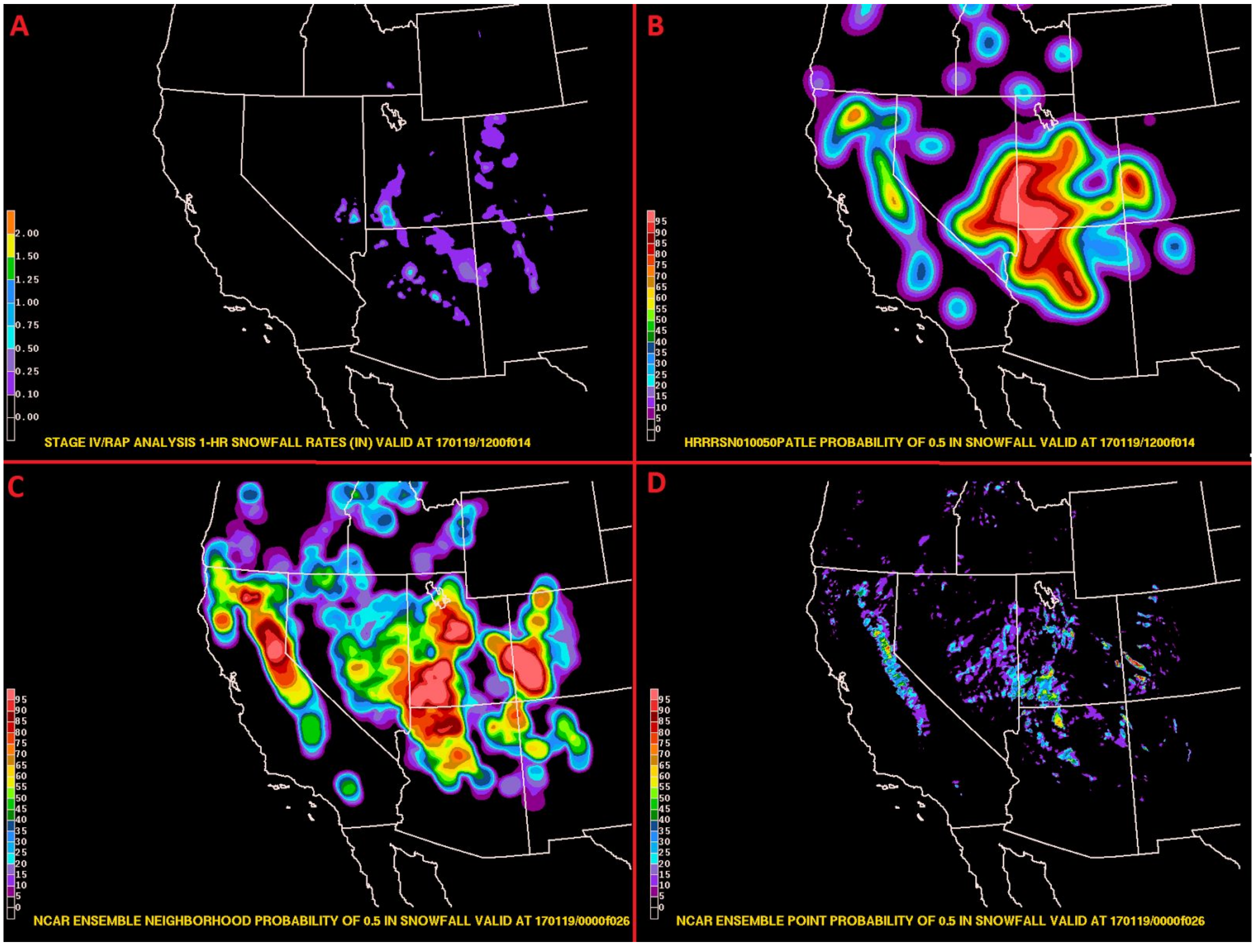


Figure 18. *(A) Stage IV/RAP analysis observed snowfall used for verification; (B) HRRR-TLE neighborhood probabilities for one inch snowfall rates; (C) NCAR Ensemble neighborhood probabilities for one inch snowfall rates; (D) NCAR Ensemble point probabilities for half-inch snowfall rates; All valid at 02Z January 20, 2017. NOTE: Stage IV hourly data not provided by California-Nevada and Northwest River Forecast Offices.*

Snowfall Rate Algorithm

The remotely-sensed Snowfall Rate Algorithm (SFR) was formally evaluated in the 2017 WWE to establish its utility when verifying the intensity of snowfall events. The SFR detects snowfall and calculates its intensity using the passes of 5 satellites and an NWP data screening as a last step before providing the data to operations after each pass. Figure 19 shows an example of the SFR algorithm compared against the Stage IV/RAP Analysis to discern snow for the same time step. In this case, the two verification resources closely resemble each other. However, the inability of the SFR to resolve precipitation near coasts is noted and limited its use in this case during which heavy snowfall fell on Long Island and along the coastline in a short period of time.

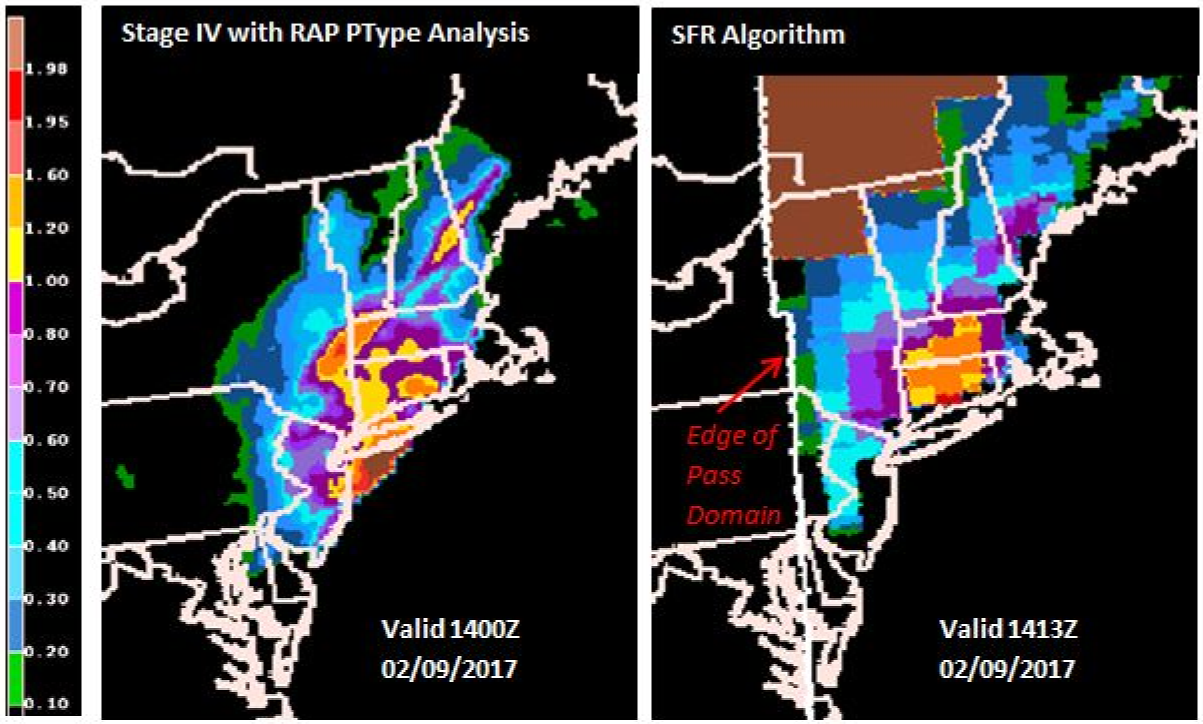


Figure 19. *Snowfall rate (inches) data from the Stage IV QPE+RAP precipitation type analysis (left) compared with the SFR Algorithm (right) valid 14Z and 1413Z respectively on February 09, 2017. Both verification data sets have a 10:1 SLR applied.*

Despite the SFR not being reliable below 7 degrees Fahrenheit at the surface and the bounding upper rate limit of 2" of snowfall per hour, the anticipation is that the remotely-sensed SFR can provide snowfall rate information in areas where the radar and other ground observations may be lacking. The SFR can also supplement areas over which the River Forecast Centers in the western CONUS (California-Nevada and Northwest) do not provide hourly Stage IV QPE.
In the following example from February 23rd, 2017, the experimental NOHRSC Version 2 (http://www.nohrsc.noaa.gov/snowfall_v2/) in Figure 20 shows that accumulating snowfall occurred in the area of southern Idaho, northern Nevada, and southeastern Oregon.

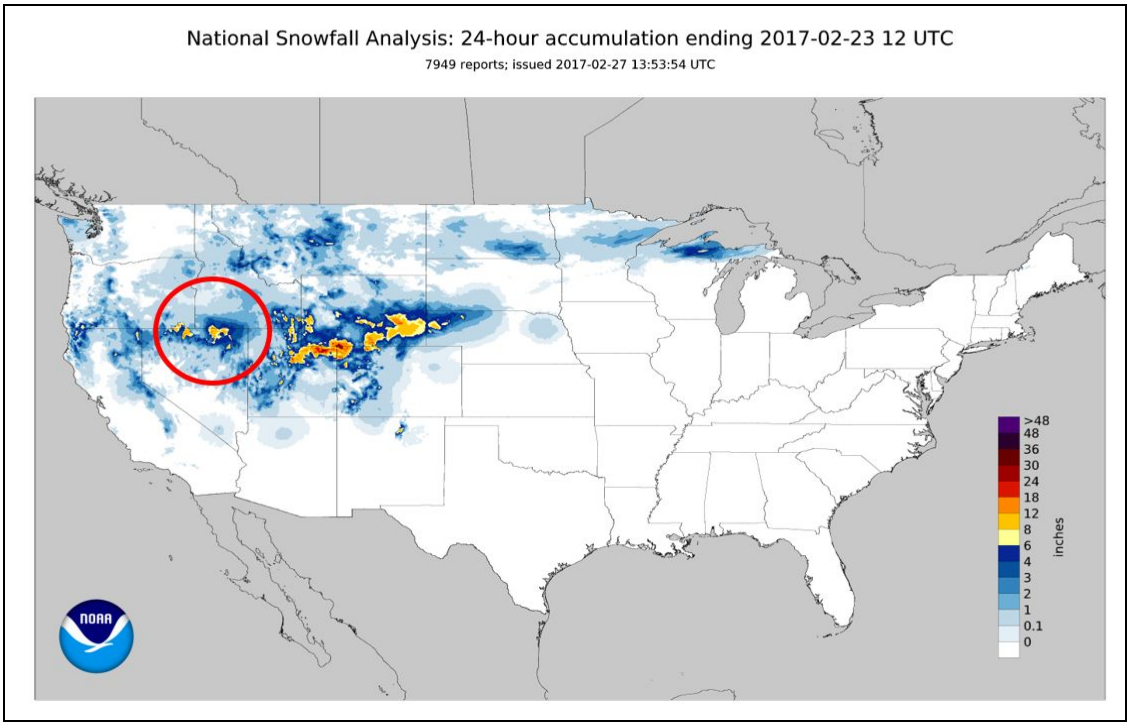


Figure 20*. NOHRSC Version-2 24-hour snowfall analysis valid 12Z February 23, 2017*

The series of images in Figure 21 show 7 hours of snowfall that contributed to the snowfall analysis, both from the Stage IV/RAP Analysis on the left and the SFR algorithm observations on the right. The areas circled in red show a region where the hourly Stage IV QPE data is not provided by the RFC over southern Idaho, northern Nevada, and southeastern Oregon, the SFR algorithm returns snowfall signatures.

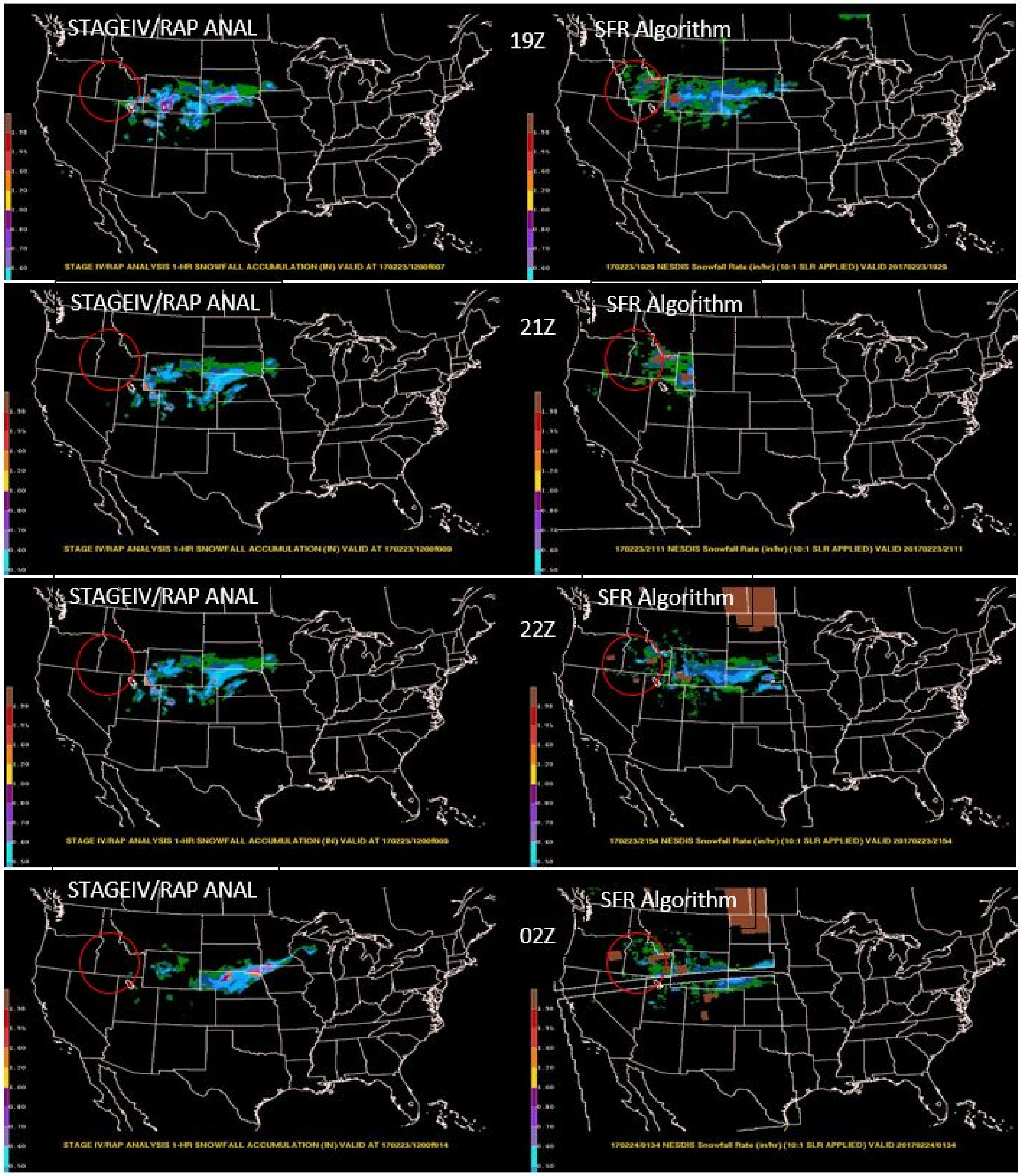


Figure 21. *Stage IV+RAP Analysis on left, SFR algorithm on right from 19Z on January 23 to 02Z on January 24, 2017*

Given the comparisons of these two verification resources, participants overall were satisfied with the performance of the SFR algorithm giving an average score of 5.6 out of 10. As shown in the scoring distribution (Figure 22), the score was highly event-dependent and pass-dependent. Participants agreed it provided supplemental data and an advantage where radar may be lacking. Not every event was structurally well-captured by the SFR, most notably along the edges of the pass due to the “limb effect” from the sweep of the imager. The SFR most often demonstrated a low bias, but some events too heavy in intensity. Most Great Lakes events were completely missed by the SFR. a known weakness in the algorithm due to the microwave retrievals unable to resolve shallow convection that often generate the lake effect snows.

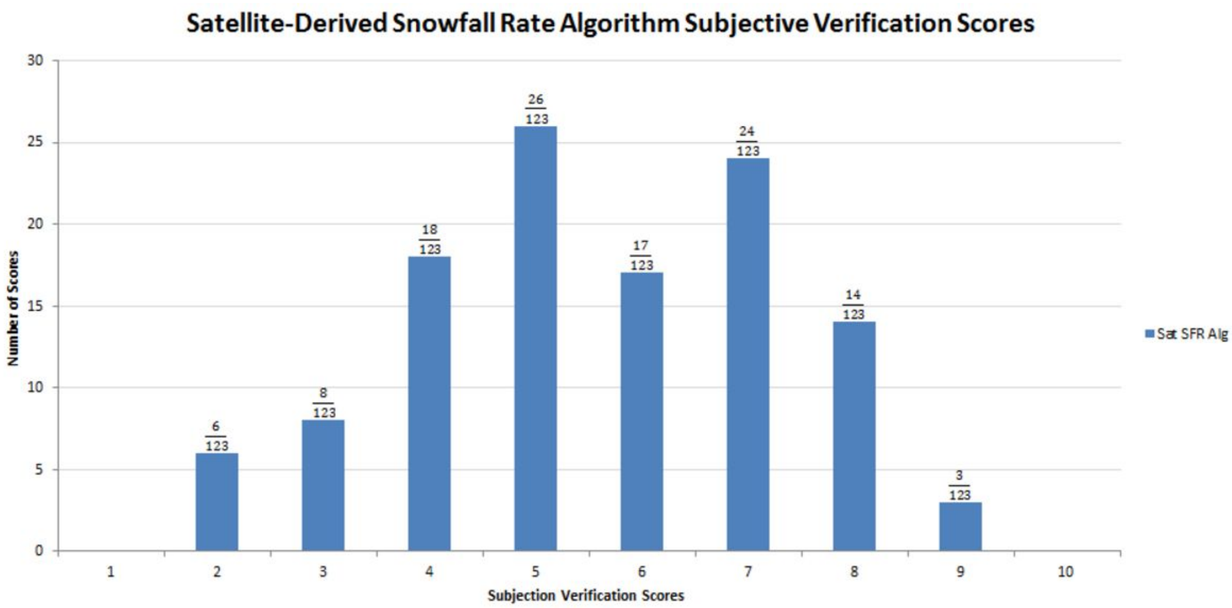


Figure 22. *The 2017 WWE distribution of subjective verification scores for the SFR algorithm. The SFR was scored each day it was available on a scale of 1 (very poor) to 10 (very good).*

These inconsistencies in its performance degrade forecaster trust. To restore confidence, it is suggested that the SFR algorithm be held against more ground truth observations to ensure possible verification resource errors were not misrepresenting the snowfall accumulation. The unreliability at temperatures below 7 degrees Fahrenheit nor the upper limit of 2" of snowfall per hour were not of great importance to the participants. Only one event over the experiment had snowfall rates achieving 3" per hour.

Probabilistic Snowfall Rate Forecast

Participants were asked to use a suite of experimental guidance to produce a probabilistic snowfall rate forecast. Each day a threshold of 0.5, 1, or 2 inches per hour was chosen and then probability contours of 25, 50, and 75% were drawn over a select area to indicate the likelihood of the chosen snowfall rate. In addition to the probability contours, forecasters were asked to specify a time range during which the activity was most likely to occur.

Participants rated each probabilistic snowfall rate forecast subjectively each day on a scale of 1-10, where 1 was very poor and 10 was very good. These ratings were based on how well both the probability contours captured the observed snowfall rates (aerial coverage) and if the observed rates that occurred during the specified time period (snowfall intensity). Participants were asked to forecast for the period of 18-12Z over the Day 1 period. During the course of the experiment, the average length chosen for the probabilistic snowfall rate forecasts was seven hours. Figure 23 displays the box plot of the subjective verification scores for the probabilistic rate forecast and Figure 24 shows the distribution of scores over the entire experiment. The mean score was a 6.5 out of 10 with a median of 5.0 and standard deviation of 1.7. An example of a rate forecast valid 18Z January 30 - 06Z January 31, 2017 for the probability of a half inch of snow in an hour is shown in Figure 25 with verification at 22Z January 30, 2017 overlaid. At 22Z the half inch of snow in light blue is entirely within the 75% probability contour. Overall, this particular forecast received an average score of an 8 out of 10 when all of the hours were considered.

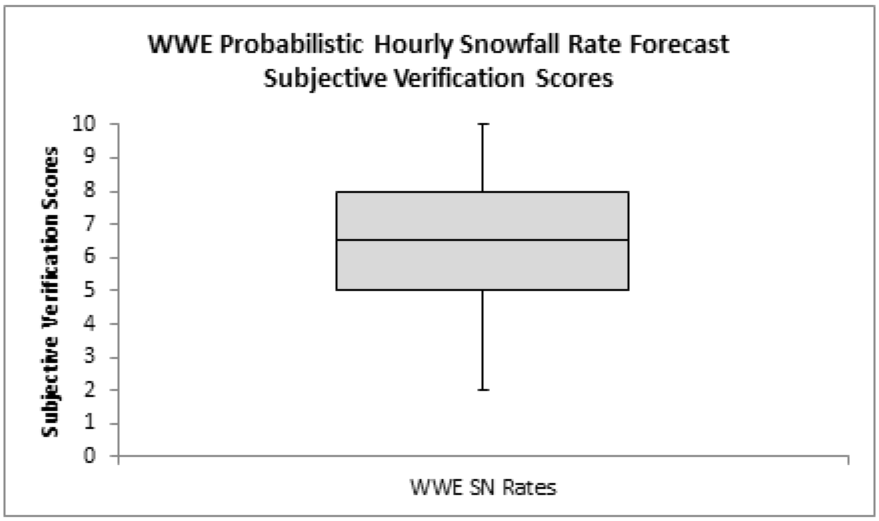


Figure 23. *Box plot of the subjective scores for the probabilistic hourly snowfall rate forecasts issued by participants during the WWE.*

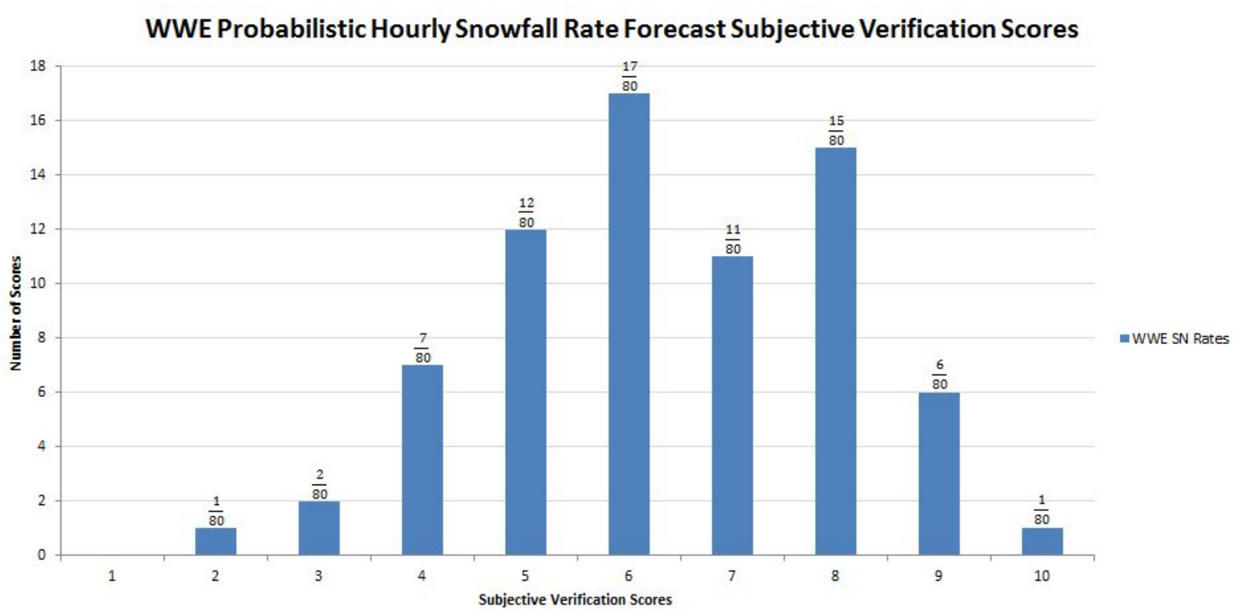


***Figure 24**. Distribution of scores for the WWE probabilistic hourly snowfall rate forecast throughout the entire experiment.*

A major challenge this year was verifying the forecasts. With many events, and thus many hourly rate forecasts in the Western United States, challenges arose in verifying the hourly snowfall. The Northwest and California-Nevada River Forecast Centers do not provide one hour Stage IV data, thus the Stage IV/RAP analysis also did not provide data in those regions. Stage IV data also has difficulty in other areas of the West where there is complex terrain. The Stage IV/RAP analysis verification uses a 10:1 snow-to-liquid ratio (SLR) that was called into question for some events where participants thought the SLR may be higher due to colder temperatures. Moving forward, it is recommended that the Stage IV/RAP analysis SLR computation be improved, as well as finding a better or alternate source for verifying one hour snowfall in the Western U.S.

<u>*Comments*</u>: Regarding the actual forecasts, participants frequently commented that the probabilities were too high after comparing with the verification. There were several examples when participants felt the forecast was too confident given what happened or that the probabilities were too high for the given threshold chosen for the forecast.

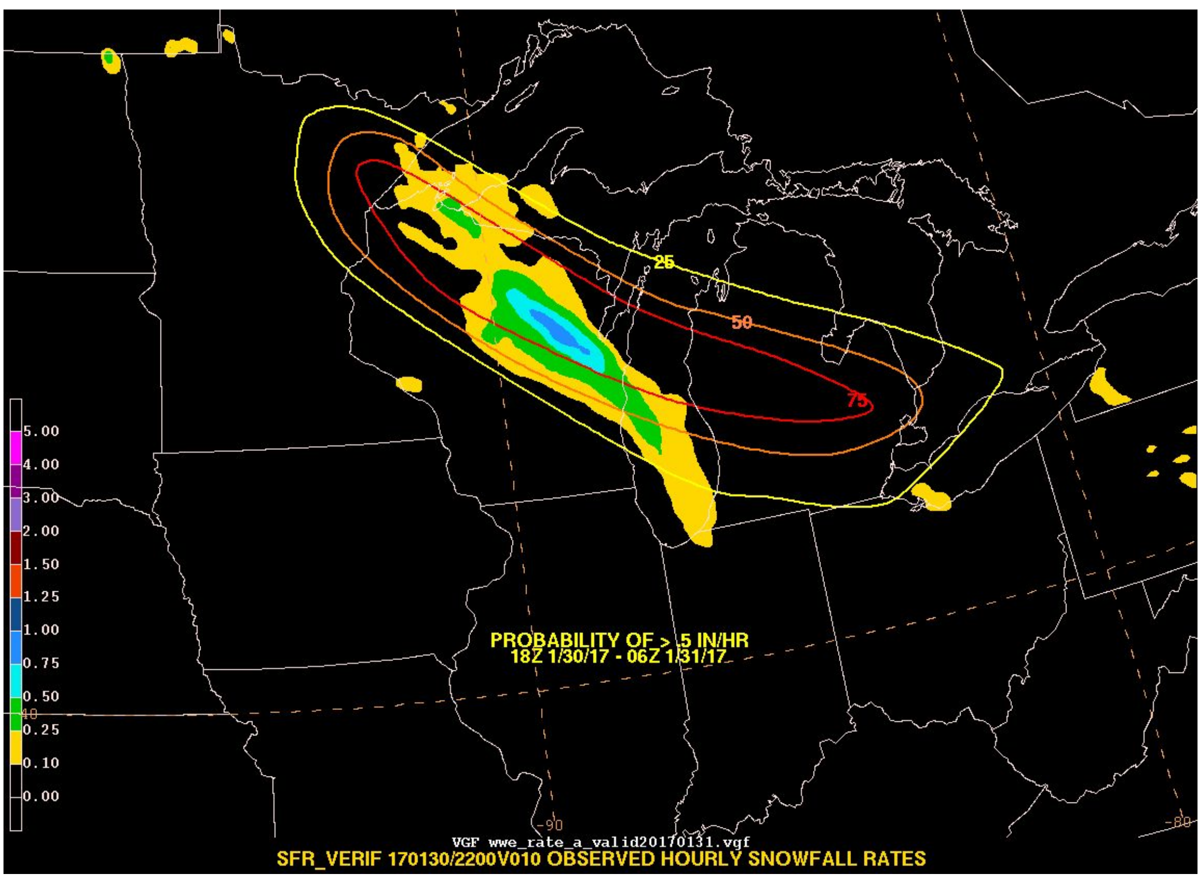


***Figure 25.** Probabilistic hourly snowfall rate forecast (contours) for 0.5 inch of snowfall between 18Z January 30 - 06Z January 31, 2017 with Stage IV/RAP Analysis verification (fill) valid at 22Z January 30, 2017.*

6. Experimental Day 2 Snowfall Forecast Models

During this year's experiment, three ensembles developed by WPC were tested and used in the creation of the Day 2 deterministic snowfall forecast. Two of these ensembles featured exclusively implicit snowfall directly from their respective members to explore benefits and drawbacks when compared to the more traditional component method using QPF, precipitation type estimate, and snow-to-liquid ratios (SLR). The three models tested were the WPC Experimental Implicit PWPF, the WPC Deterministic Implicit Blend, and the WPC Experimental Winter Weather Ensemble.

WPC Experimental Implicit PWPF

Operationally, WPC's PWPF consists of 70 members including the GFS, SREF, CMC, ECMWF, and three high resolution convection allowing models (CAMs). A 5-member average SLR is applied to the WPC deterministic snowfall forecast, which is the mode of the ensemble, and each individual member to a create a probabilistic snowfall forecast. In contrast, WPC's Experimental Implicit PWPF consists of 72 members exclusively using model implicit snowfall as well as the same WPC deterministic snowfall forecast as the mode. See the experimental data/models section for more details on exact membership. The daily subjective evaluation sought comments from the participants on the spatial and quantitative differences between the two different PWPFs.

Day 2 (48-hour) forecasts of 24-hour snowfall probabilities of greater than 2 and 6 inches were evaluated for both the operational and experimental PWPF each day during the experiment and verified based on NOHRSCv1 and NOHRSCv2 data. Figure 26 shows a box plot of the subjective scores for both thresholds and both versions of the PWPF. The mean score for the experimental PWPF was slightly higher for both thresholds (6.3 out of 10) than the operational PWPF (6.0 out of 10) while each had similar standard deviations of 1.2-1.3 at the 2 inch threshold and 1.4-1.5 at the 6 inch threshold. In general, participants scored both versions more favorably at the lower 2 inch threshold than the 6 inch threshold. Figure 27 shows the distribution of scores for all versions and threshold over the entire experiment.

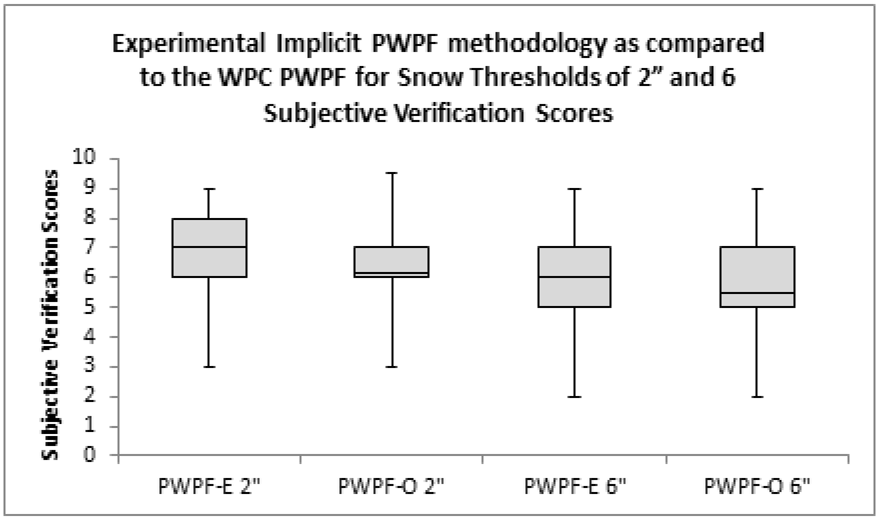


Figure 26. *Box plot of the subjective scores for the PWPF Implicit Experimental and the Operational PWPF at the 2 inch and 6 inch thresholds.*

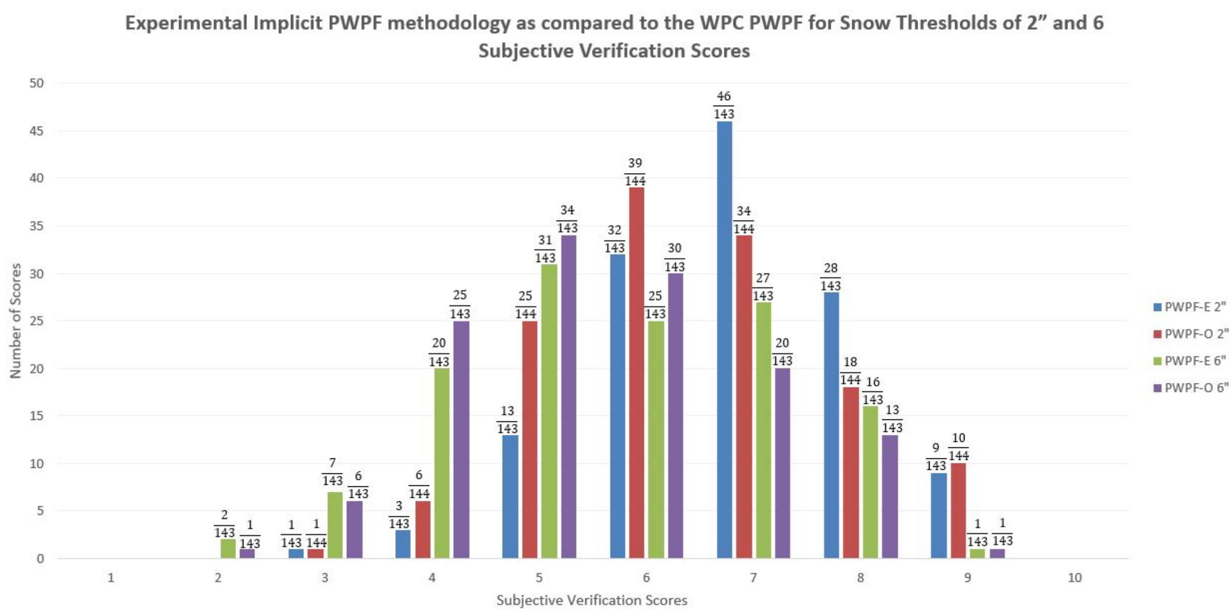


Figure 27. *Distribution of scores throughout the entire experiment for the PWPF Implicit Experimental and Operational PWPF at the 2 inch and 6 inch thresholds.*

Figure 28 shows an example of both the Experimental Implicit PWPF (B and C) and the Operational PWPF (D and E) and both thresholds valid at 12Z on February 10, 2017. Overall comments from the participants were quite positive for both versions. Participants indicated that the experimental implicit PWPF probabilities better represented the spatial distribution of snowfall as compared to the operational version in some cases. Participants frequently commented that both versions produced probabilities that were too low when compared to verification. This occurred most often in Western areas, for some lake effect snow events, and for the 6 inch threshold in general. However, for several more marginal events it was noted that the experimental version often produced low probabilities when the operational version did not produce any. Based on the positive feedback and performance, it is recommended that the WPC Experimental Implicit PWPF undergo further testing and development.

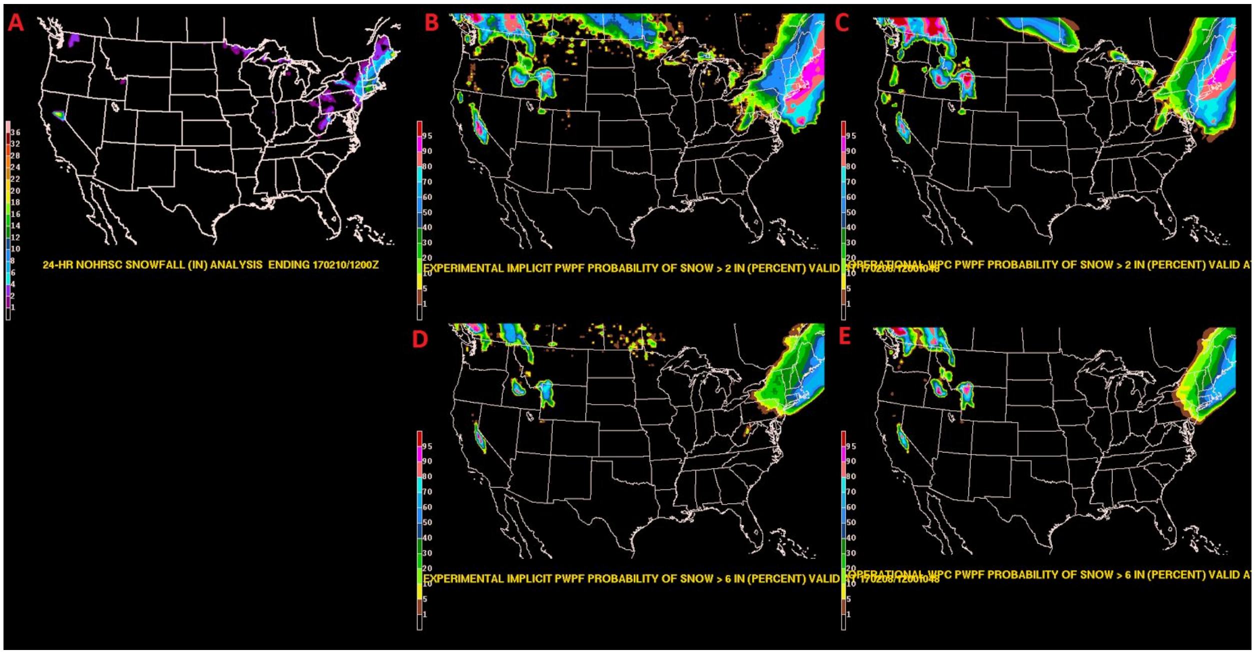


Figure 28. *(A) NOHRSCv1 24 hour snowfall accumulation; (B) Experimental Implicit PWPF > 2 inches of snow; (C) Operational PWPF > 2 inches of snow; (D) Experimental Implicit PWPF > 6 inches of snow; (E) Operational PWPF > 6 inches of snow all valid at 12Z on February 10, 2017. NOTE: NOHRSCv2 (not shown) did show 24 hour snowfall greater than 6 inches in many mountain ranges in the West including the Sierra Nevadas in California, Western Wyoming, Idaho, Oregon, and Western Montana.*

WPC Deterministic Implicit Blend

The WPC Deterministic Implicit Blend is made up of six different model implicit snowfall solutions and its Day 2 (48 hour), 24 hour deterministic snowfall forecast was evaluated during the experiment. The use of a blend of implicit snowfall was motivated in part to try and reduce inflated snowfall amounts in mixed precipitation or transition areas. The forecasts were again verified using NOHRSCv1 and NOHRSCv2. Please refer to the model guidance section for exact details on the model implicit snowfall solutions. Figure 29 shows the box plot representing the subjective ratings for the Implicit Blend. The mean score was a 4.8 out of 10 and standard deviation of 1.5. Figure 30 shows a distribution of the individual scores throughout the experiment. The lowest individual score received was a 2 out of 10 (10/146) and highest was a 7 out of 10 (20/146).

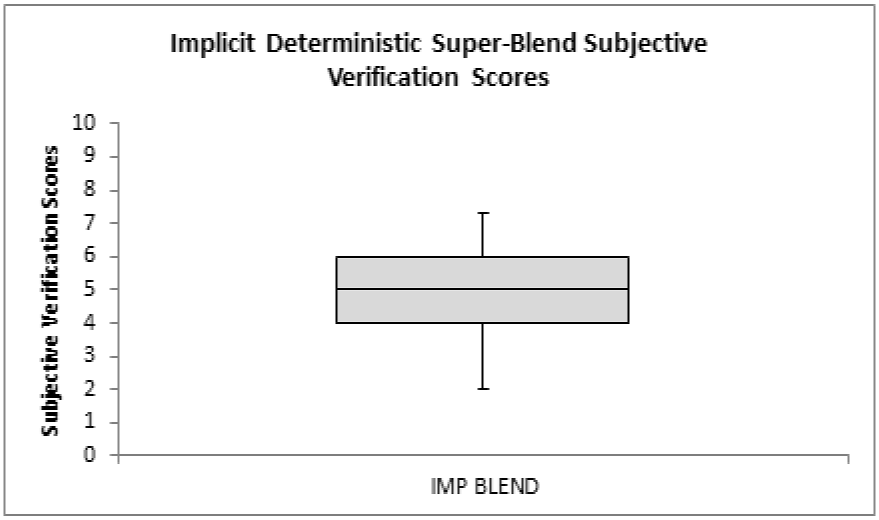


Figure 29. *Box plot of the subjective scores for the Deterministic Implicit Blend.*

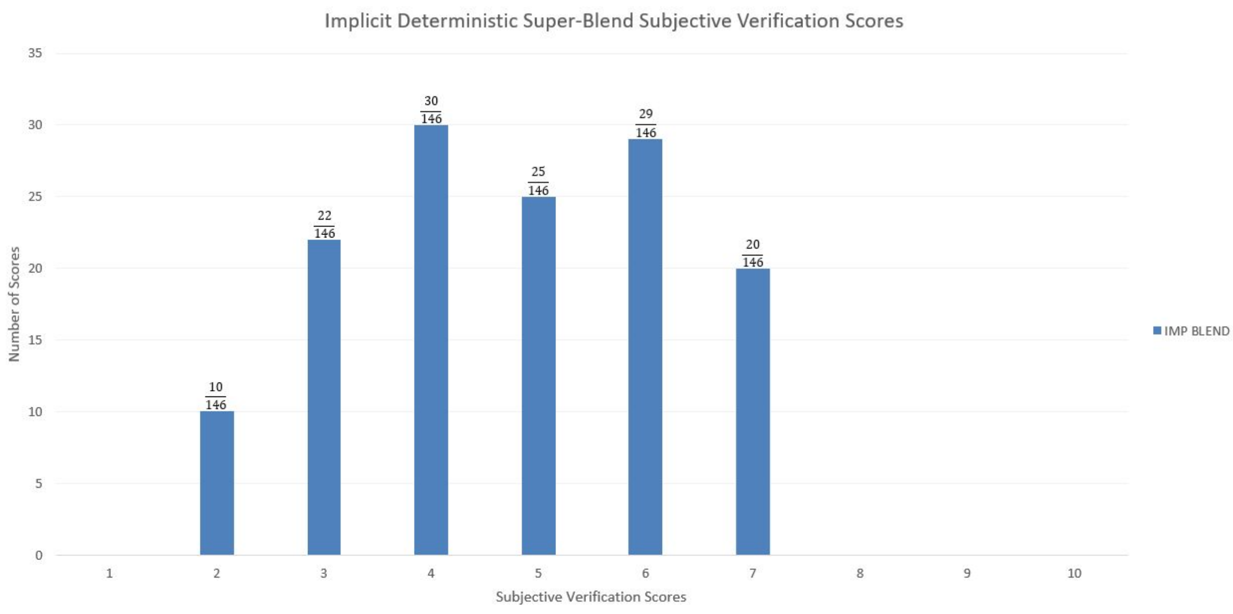


Figure 30. *Distribution of scores throughout the entire experiment for the Deterministic Implicit Blend.*

<u>*Comments*</u>: The overwhelming feedback almost every day was that the Deterministic Implicit Blend consistently under forecasted the snowfall amounts. This was especially evident in the Western regions as well as along the Great Lakes. Despite its strong synoptic forcings, the implicit blend still presented a low bias for the storm on February 9-10, 2017 in the Northeast. Figure 31 shows an example from February 10, 2017 valid at 12Z compared to the NOHRSCv1 verification. The participants, however, did comment favorably on the spatial extent of the snowfall for this event. This model was often used as a first guess for the Day 2 deterministic forecast because it performed well with the spatial coverage of the snowfall.

Suggestions: Despite the overwhelming feedback concerning the low bias, it is still recommended that further development of the implicit blend continue as it proved to be a good first guess field and to see if the magnitude of snowfall can be improved.

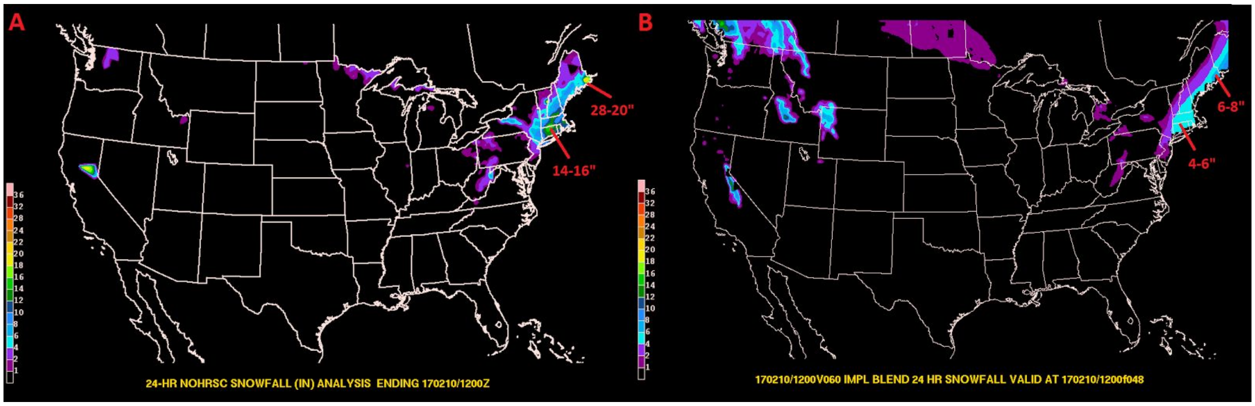


Figure 31. *(A) NOHRSCv1 24 hour snowfall accumulations and (B) WPC Implicit Blend 24 hour snowfall accumulations, both valid at 12Z on February 10, 2017. NOTE: NOHRSCv2 (not shown) did show 24 hour snowfall in California, Idaho, Washington, and Wyoming.*

WPC Experimental Winter Weather Ensemble

The WPC Experimental Winter Weather Ensemble (WWENS) differs from the previous two as it does not use implicit snowfall techniques but rather utilizes a blended SLR to compute snowfall from each of its 41 ensemble members. The WWENS does most of its processing on a 5 km National Digital Forecast Database (NDFD) grid and applies a snow level computation to try to improve snow prediction out west. More details on the membership and snow level computation can be found in the experimental model guidance section.

Participants were asked to comment on the performance of the Winter Weather Ensemble, especially in the intermountain west and areas of complex terrain. Figure 32 shows the box plot for the WPC WWENS subjective scores. The mean score was a 5.9 out of 10 with a standard deviation of 1.5. Figure 33 shows a distribution of the individual scores throughout the experiment. The lowest individual score received was a 2 out of 10 (3/146) and highest was a 9 out of 10 (1/146).

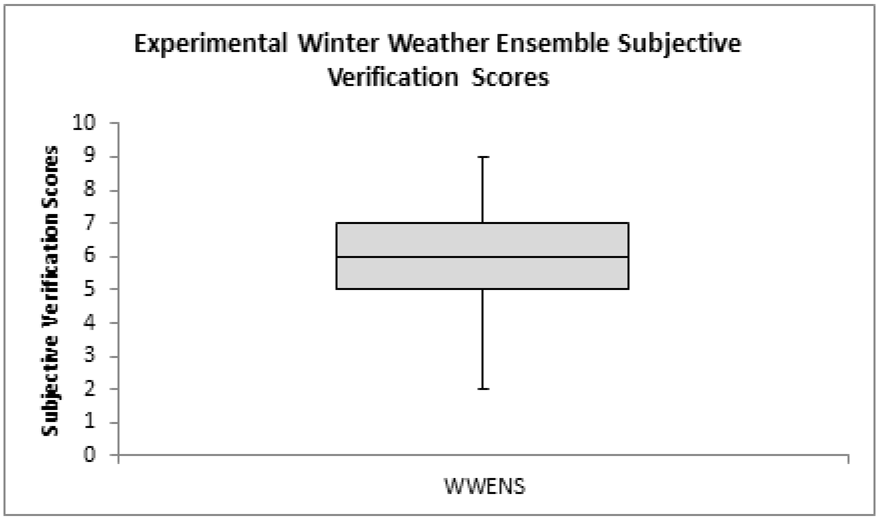


Figure 32. *Box plot of the subjective scores for the WPC Experimental Winter Weather Ensemble.*

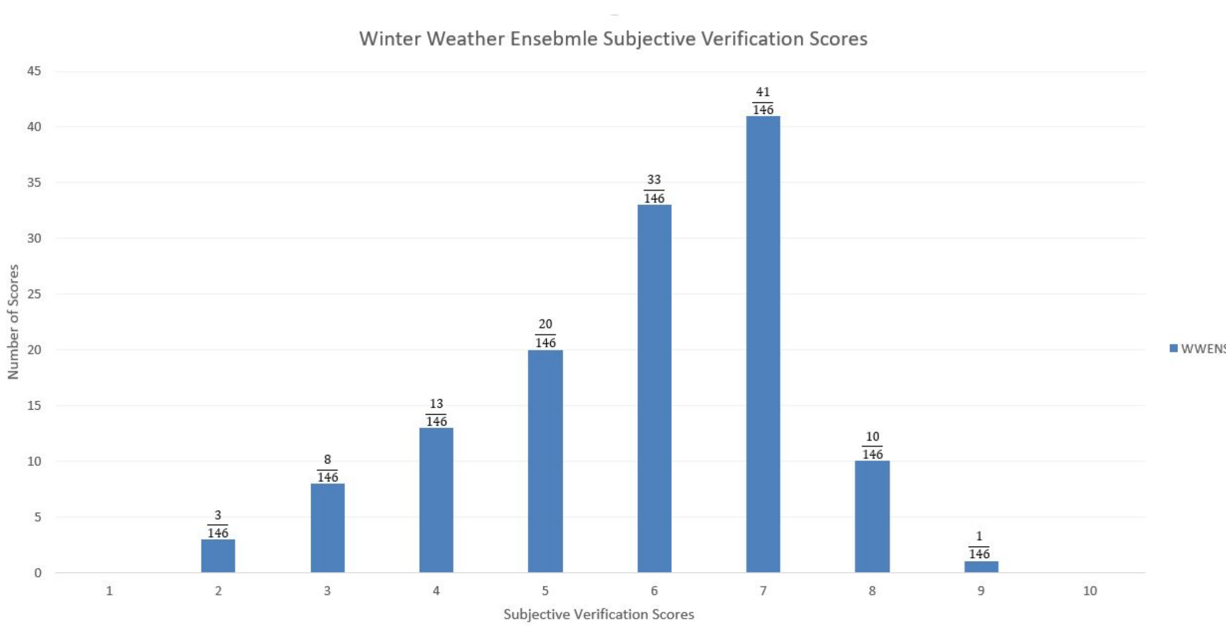


Figure 33. *Distribution of scores throughout the entire experiment for the WPC Experimental Winter Weather Ensemble.*

<u>Comments</u>: Participants were very enthusiastic about the resolution the WWENS provided for a Day 2 forecast as well as the amounts, particularly in the mountains in the Western U.S. There were several cases where amounts were underdone however overall feedback on amounts and spatial coverage was positive. Figure 34 shows an example of the WWENS and NOHRSCv2 for verification valid at 12Z January 21, 2017. Outside of the Western U.S., the WWENS struggled at times with amounts being too light for events in the Northeast and around the Great Lakes.

Due to the large number of Western events during this year's experiment, this ensemble was relied on heavily during the actual forecasting process. Although not evaluated formally in subjective verification, WWENS PWPF was also provided and used as guidance whenever there was an event over the west as the participants relied heavily on the detail from this ensemble.

Suggestions: Because of the tremendous promise the WWENS has shown in forecasting snowfall and freezing rain over the higher terrain in the Western U.S., further development and testing is recommended.

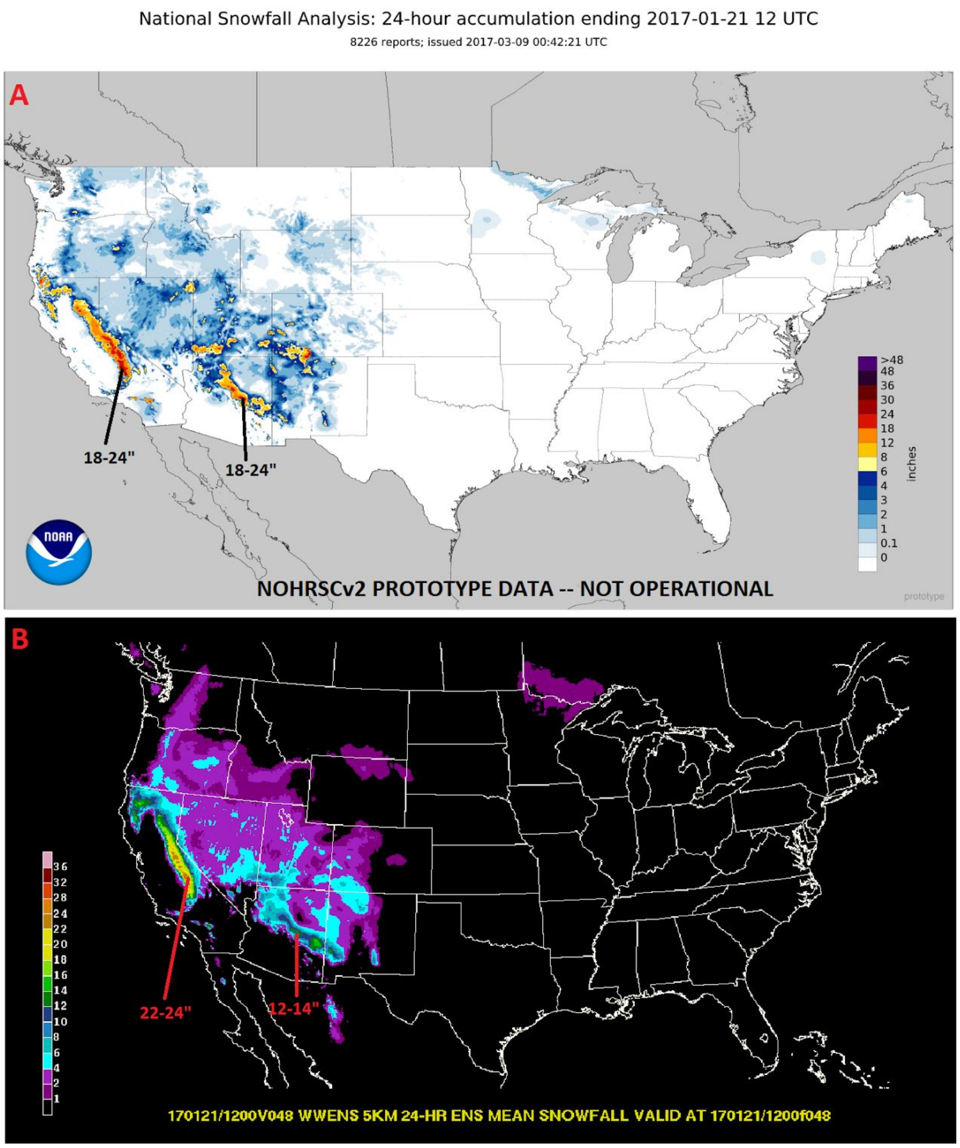


Figure 34. *(A) NOHRSCv2 24 hour snowfall accumulations and (B) WPC Winter Weather Ensemble 24 hour snowfall forecast both valid at 12Z on January 21, 2017.*

7. Day 2 WWE Deterministic Forecasts

Participants created two deterministic forecasts for 24 hour snowfall and freezing rain accumulations each day that covered the entire CONUS region and were valid for the Day 2 period from 12Z to 12Z. A total of eighteen deterministic snowfall forecasts were created during the experiment. NOHRSCv1 and the prototype NOHRSCv2 snowfall analyses were used to verify the forecasts. Figure 35A shows an example of the Day 2 snow forecast and associated verification (Figure 35B) valid from 12Z February 9, 2017 to 12Z February 10, 2017.

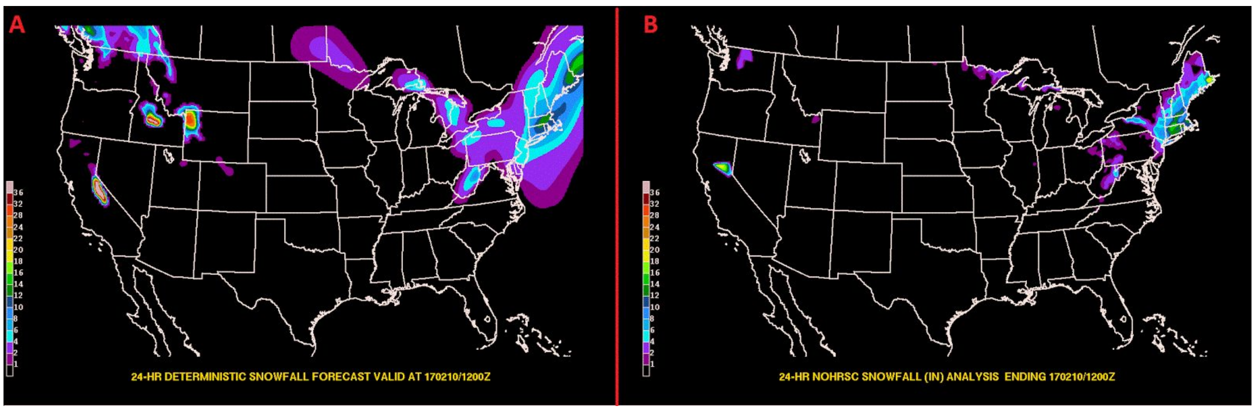


Figure 35. *(A) Day 2 forecast for 24 hour snowfall; (B) NOHRSCv1 24 hour snowfall analysis; both valid at 12Z February 10, 2017.*

Figure 36 shows the box plot for the deterministic snowfall forecast subjective scores and Figure 37 displays the distribution of the scores throughout the experiment. The mean and median score was a 5.9 out of 10 with a standard deviation of 1.3. The highest individual score was a 9 out of 10 (1/121) and the lowest individual score was a 3 out of 10 (4/121). The most common criticism regarding the skill of the forecasts was that the forecasted amounts were often underdone when compared to the analysis. Comments regarding the forecast being underdone were recorded for sixteen of the eighteen forecasts. The most common geographical region where participants felt the forecasts were often underdone was the Western United States in the mountainous regions. Many participants felt, however, that the forecasts often captured the spatial extent of the areas that did receive snowfall quite well for Day 2 forecasts.

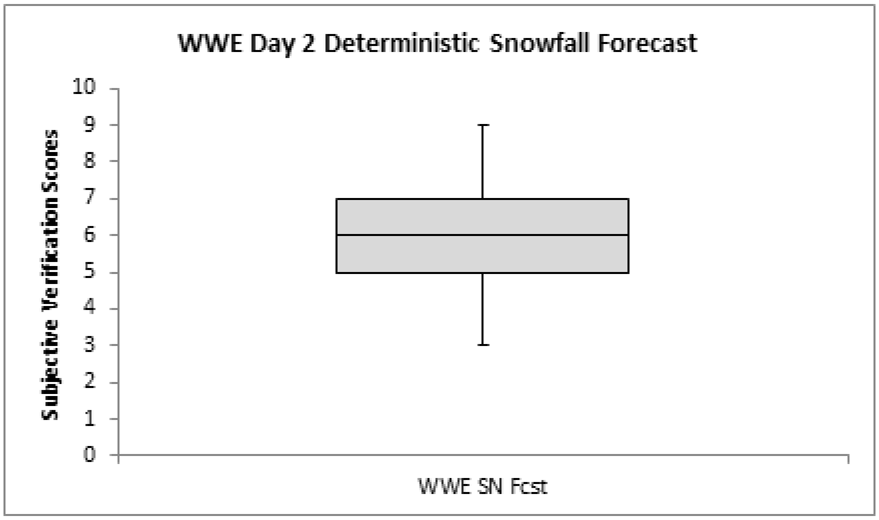


Figure 36. *Box plot of the subjective scores for the WWE Day 2 deterministic snowfall forecast.*

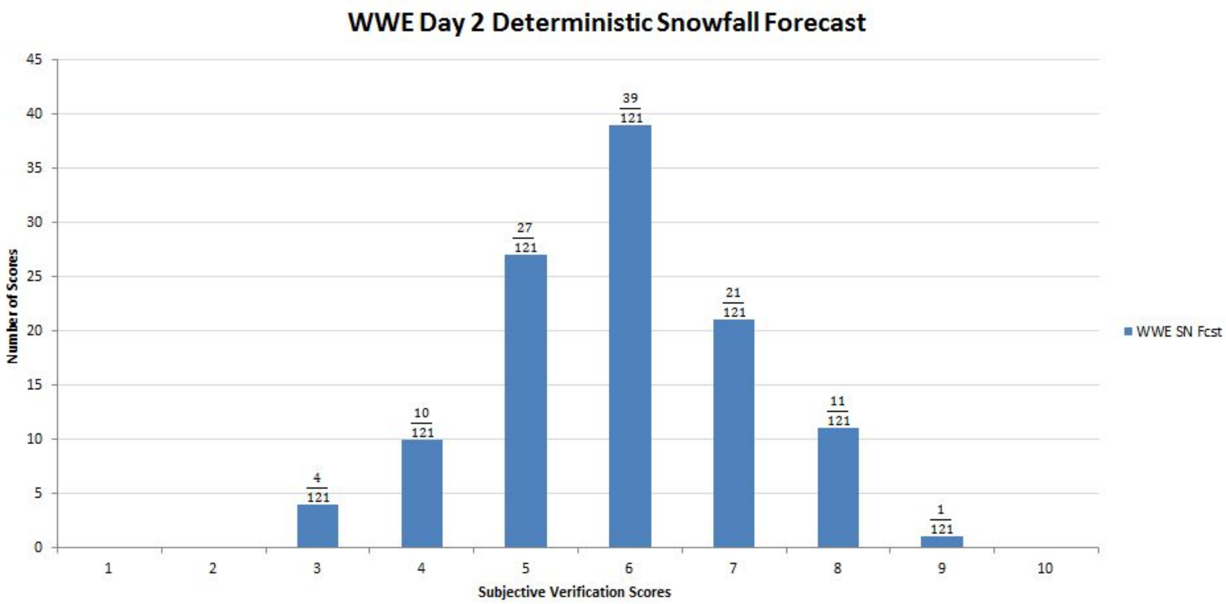


Figure 37. *Distribution of scores for the WWE Day 2 deterministic snowfall forecast over the entire experiment.*

A deterministic freezing rain accumulation forecast was made each day if there were accumulations over 0.01 inch expected. Out of the eighteen possible forecast days, twelve freezing rain forecasts were issued and out of the twelve, eight forecasts had areas in the Pacific Northwest. To verify, the Stage IV/RAP analysis was used and supplemented with METARs from stations within the forecast contours in an attempt to discern if freezing rain was reported and how much precipitation fell over the valid forecast period. Figure 38 shows an example of a Day 2 freezing rain forecast issued during the experiment.

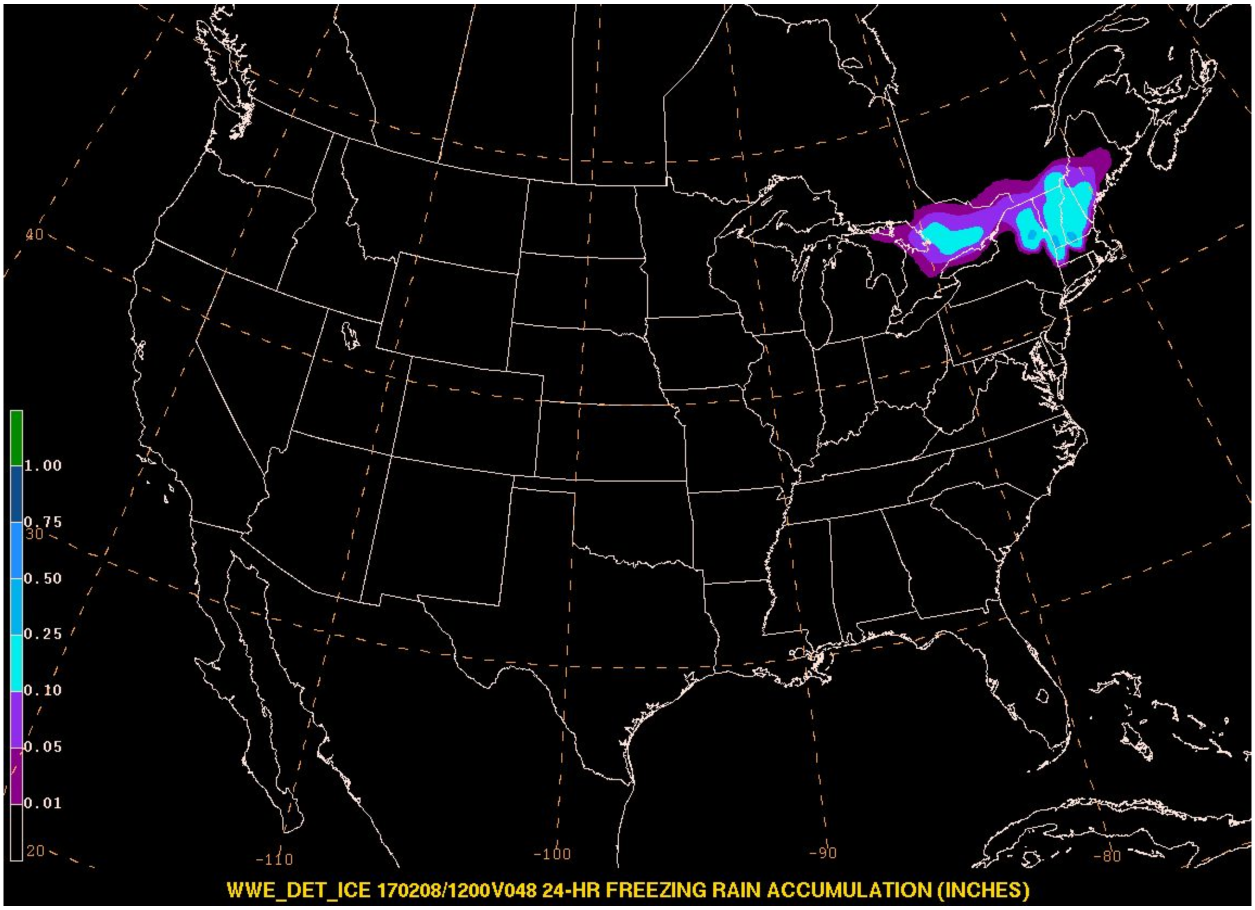


Figure 38. *Day 2 24 hour freezing rain accumulation forecast valid at 12Z on February 8, 2017.*

Figure 39 shows the box plot for the deterministic freezing rain accumulation forecast subjective scores and Figure 40 displays the distribution of the scores throughout the experiment. The mean score was 6.0 out of 10 and the median score was also a 6 out of 10, with a standard deviation of 2. The highest individual score was an 9 out of 10 (4/59) and the lowest individual score was a 1 out of 10 given just once (1/59). The majority of the comments focused on the need for an improvement in verification of freezing rain. The Stage IV/RAP analysis was not reliable in depicting freezing rain in the Pacific Northwest where many of the METARs did report freezing rain. However, due to station distribution and density, there were often too few stations in the forecast areas in the Pacific Northwest to really give strong confidence that the forecast was successful.

<u>Comments</u>: The participants often felt the forecasted areas spatially were generally good for a Day 2 forecast, however it was often hard to know how well the amounts were forecast.

<u>Suggestions</u>: A better, reliable freezing rain verification source is needed in the future in order to really diagnose performance of freezing rain forecasts.

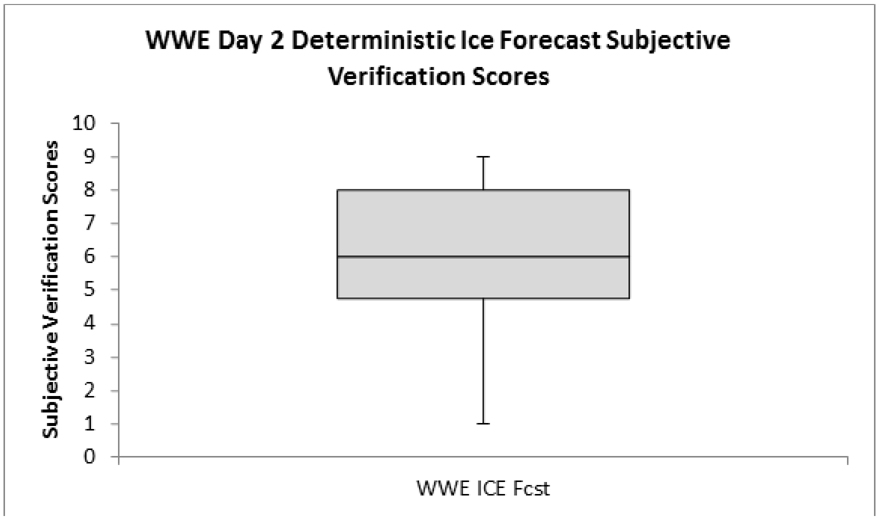


Figure 39. *Box plot of the subjective scores for the WWE Day 2 deterministic ice forecast.*

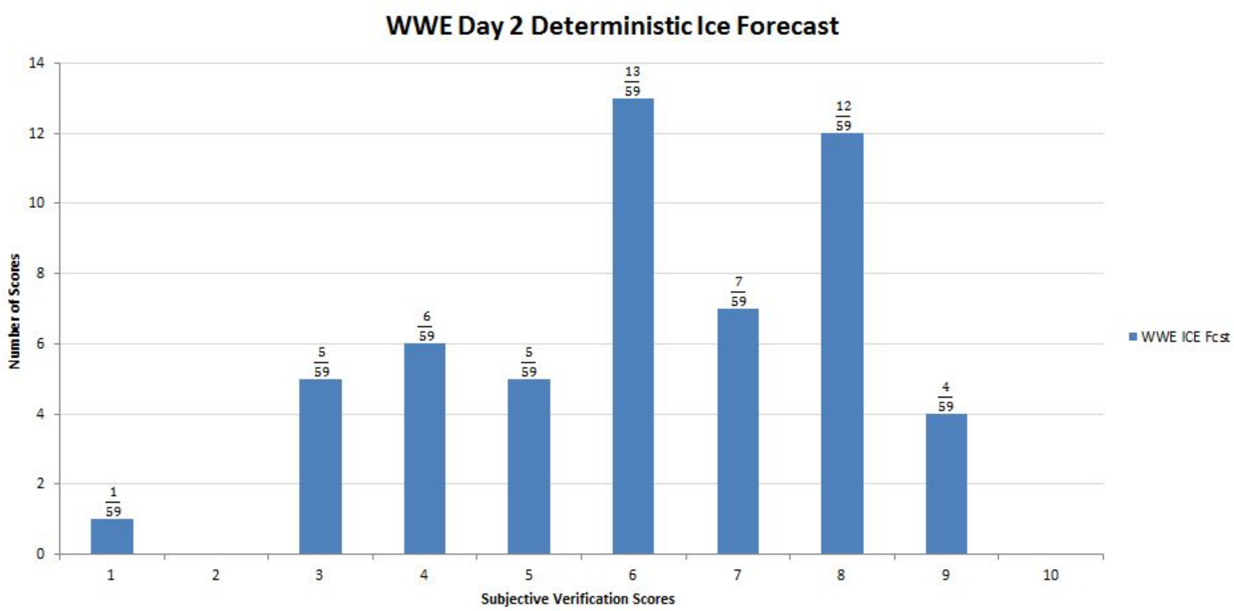


Figure 40. *Distribution of scores for the WWE Day 2 deterministic ice forecast over the entire experiment.*

8. Experimental Watches

To address the FY17 AFS Milestones, issuance of Winter Weather Watches from a national center perspective was put into practice in the 2017 WWE. Two different watch types were created each day. The first considered areas where the standard snowfall or ice accumulation forecast exceeded NWS WFO warning criteria. The second was an innovative impacts-based "Winter Alert" which allowed participants the flexibility to issue within a shorter temporal scale (as small as a 6-hour period) and based on more than just exceeding criteria. Participants could also consider the impacts of multiple hazards, time of day, or other concerns that may encourage a forecaster to issue an alert.

Criteria-Based Watches

Watches are currently issued by regional WFOs 2 or 3 days in advance of a winter weather system that will exceed warning criteria. As a national center, WPC would need to account for all winter watch potential over the CONUS. The WPC Watch Collaborator (WC) assists this effort by providing a full suite of probabilistic output that displays areas where the WPC PWPF will exceed both the 12- and 24-hour WFO warning criteria (Figure 41).

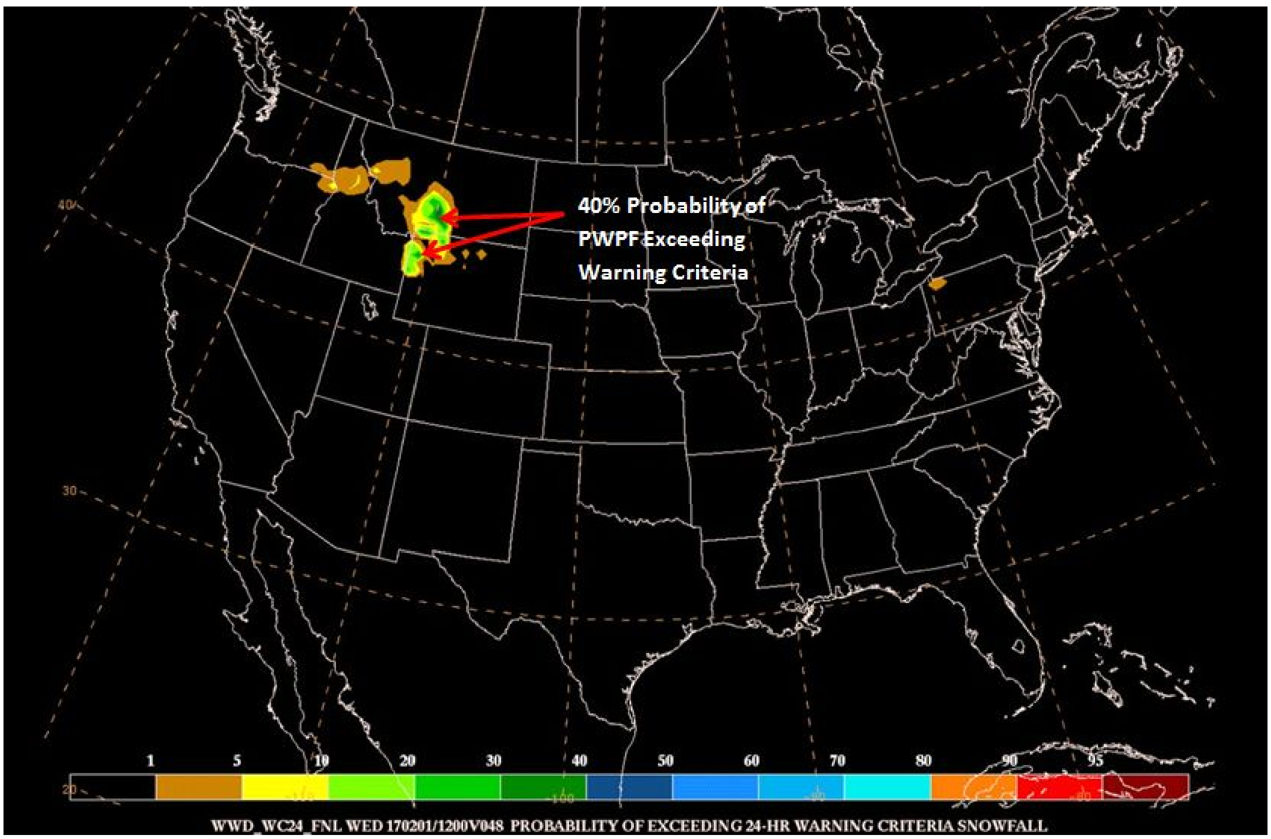


***Figure 41**. Watch Collaborator probabilities of PWPF exceeding 24-hour snowfall warning criteria valid 12Z February 1, 2017.*

Watch Collaborator Trend Tools

As mentioned in the experimental tools section, three different Watch Collaborator trend tools were available to participants. In Figure 42, we can see how each tool indicated different trend information of the WC. Again, the count tool displayed geographical areas of agreement among the most recent 3 watch collaborator cycles using yellow (1 run out of 3), orange (2 runs out of 3) and red (all 3 runs). The difference tool displayed areas that differed between the 2 most recent WC cycles. And the flip-flop tool displayed the areas over which there were notable changes in continuity over the last 3 cycles of the WC.

Figure 42. *Examples of the three Watch Collaborator trend tools available in the 2017 WWE.*

<u>*Comments*</u>: The WWE participants were surveyed on the utility of the WC trend tools at the end of their week. Overall, the participants agreed that the trend tools increased forecaster confidence in the WC output and the assessment of evolving winter storms. There is great utility in using the trend tools to collaborate with neighboring WFOs on adjustments to watch and warning issuances by offering a quick understanding of trends on a larger scale. These tools could aid in the reduction of inconsistencies in the watches CONUS-wide. Affinity for the count and difference tools outweighed the flip-flop tool.

<u>*Suggestions*</u>: Caveats to using the trend tools included the accuracy of the underlying guidance itself, (WPC's PWPF), caution against weighting the trends too heavily without checking against other guidance, and noting that although this is a great visual, verbal collaboration among WFOs remains vital to the process.

Using the Watch Collaborator, the trend tools, the experimental Day 2 deterministic snowfall forecast, the WFO warning criteria map, and any relevant experimental guidance, the participants drew contours around regions where they issued Winter Storm Watches valid 12Z-12Z on Day 2 (Figure 43).

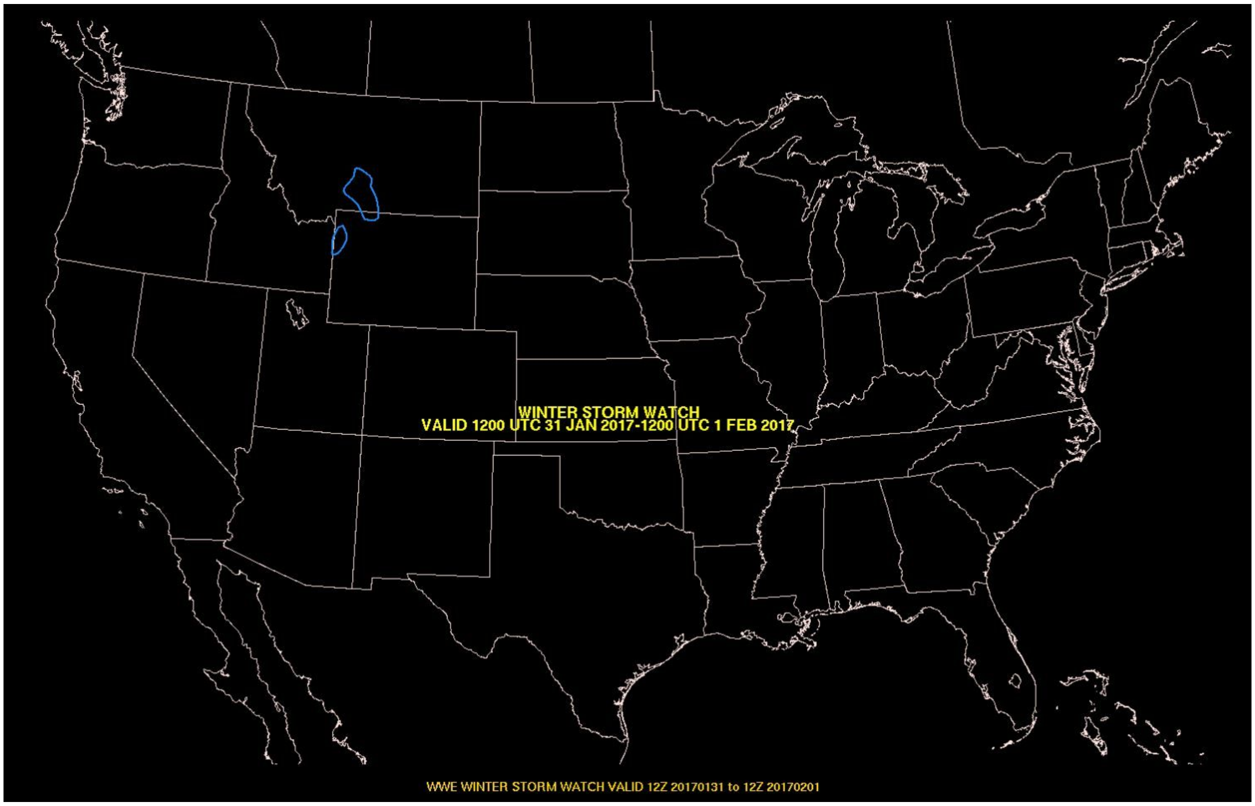


Figure 43. *Experimental Watches issued for the Day 2 period valid 12Z February 1, 2017.*

Watches were verified by identifying regions where NOHRSCv1/v2 indicated snowfall exceeded WFO warning criteria as well as METAR and other observational reports for ice. Participants were generally satisfied with the results of the experimentally watched areas. As was the trend of the experimental guidance being often underdone, noted deficiencies centered around being too conservative with the spatial extent of the watch areas or not issuing where events were marginal.

The criteria-based watches received an average score of 6.2 out of 10 by the participants. Figure 44 shows the distribution of the subjective verification scores of the criteria-based watches. The box plot in Figure 45 shows that although there was one watch that scored a perfect 10 and one that scored very poorly, the majority of scores were a 6 or 7.

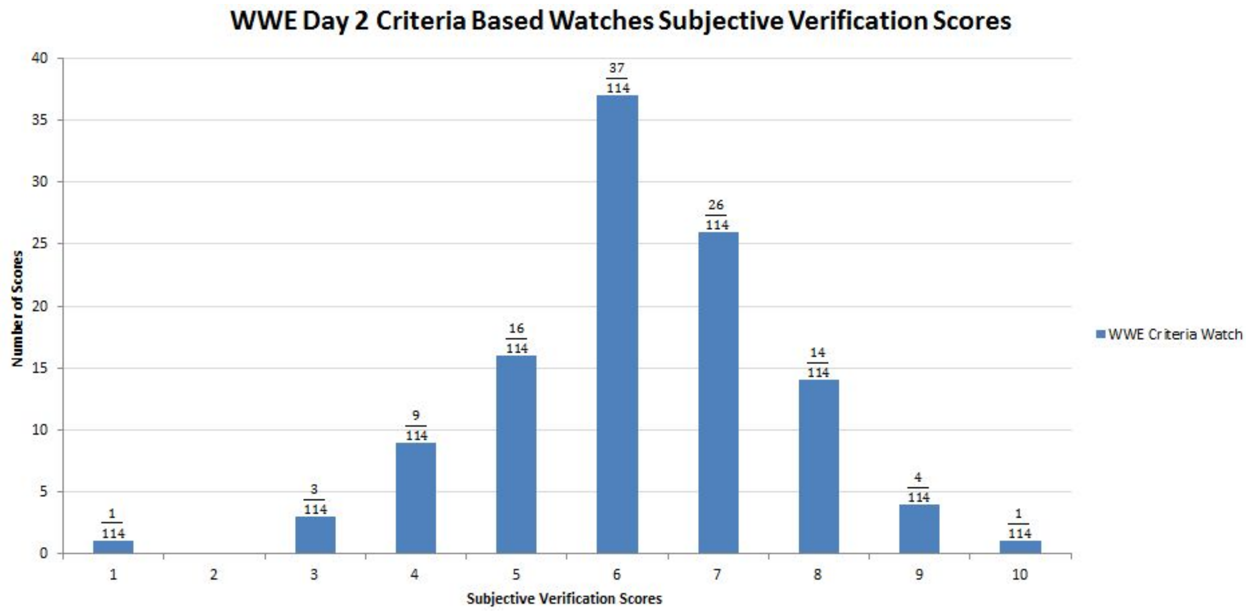


Figure 44. *The 2017 WWE distribution of subjective verification scores for the criteria-based winter weather watches. The watches were scored each day it was available on a scale of 1 (very poor) to 10 (very good).*

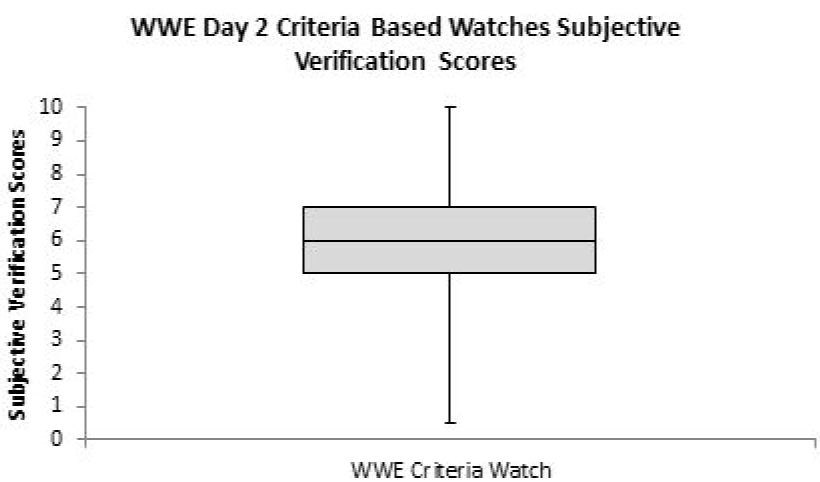


Figure 45*. The box plot of the subjective scores for the winter weather watches.*

Impacts-Based Winter Alerts

The Impacts-Based Winter Alerts examined fields that went beyond simply exceeding warning criteria. Participants were given the flexibility to issue an alert for as short as 6 hours rather than the full 24 hours on Day 2. Joint probabilistic guidance and other tools were used to determine areas where multiple hazards may impact regions. Factors such as time of day, population affected, and major roadways were also considered. For these next sections, the guidance and resulting experimental Winter Weather Alerts product correspond with events that were forecast for February 1-2, 2017.

Joint Probabilities

In response to comments and suggestions derived from the evaluation of the Watch Collaborator in the 2016 WWE and an increasing need for more impact-based guidance, WPC created a variety of joint probability tools for assessment in the 2017 WWE. Using NDFD grids and GEFS model data to compute probabilities of temperatures and wind, and snow and ice from the WPC PWPF, individual probabilities were combined multiplicatively to produce the joint probabilities (Figure 46). Please see the model guidance section of this document for more details.

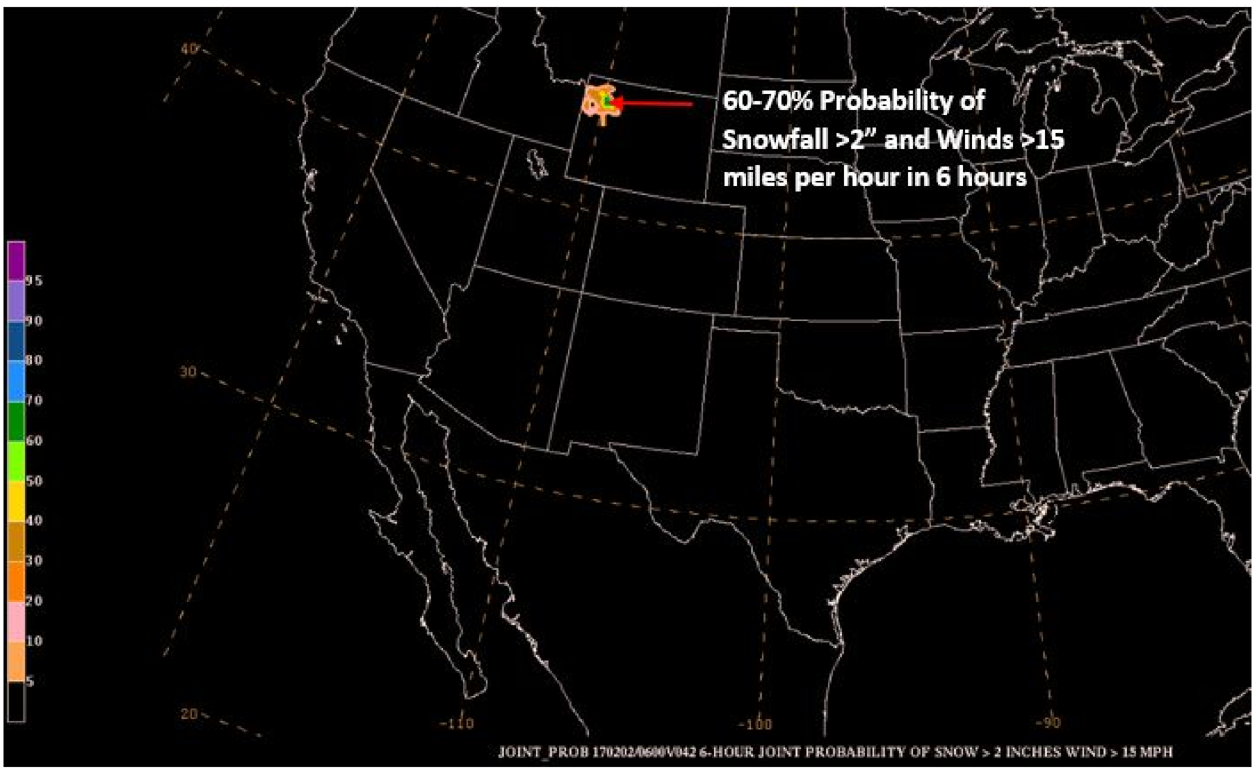


Figure 46*. An example of the joint probability of getting snowfall greater than 2" AND winds greater than 15 miles per hour in a 6-hour period, valid 06Z on February 2, 2017*.

<u>Comments:</u> Most participants agreed there is significant value in having joint probabilistic fields available when building a forecast.

<u>Suggestions</u>: Experiment with broader NWP data, and ensemble data with a neighborhood technique applied to bring up the resulting probabilities. Training would also be very important to the integration of joint probabilities in the field, as forecasters interpreting the resulting low joint probabilities would need to adjust their thinking as a more moderate risk than a low probability would imply. A high-resolution ensemble with added snow level calculation would

be preferred out west in the shorter range for the best assessment of impacts. The design of a tool in which forecasters could pick their own hazards and thresholds to get a resulting probability was also suggested.

It is the consensus of the HMT staff and model developers to move forward with further development of NWP joint probabilistic guidance and make that guidance available to the national centers and to the field. Some envision these as first-guess fields when creating a forecast. As a tool with combined fields relative to potential impacts, joint probabilities are useful for quickly evaluating the risk for multiple hazards which is valuable to the impacts-based probabilistic forecasting to which the National Weather Service is evolving.

Winter Storm Severity Index

The Winter Storm Severity Index (WSSI) was developed at WFO Burlington to show the combination of NDFD forecast information and additional datasets to assist NWS operational forecasters in maintaining situational awareness of the possible significance of weather related impacts based upon the current official forecast. The impact severity graphic (Figure 47) was designed to help communicate a general level of societal impacts and their spatial distribution, assuming minimal preparation activities are done, due to winter storms.

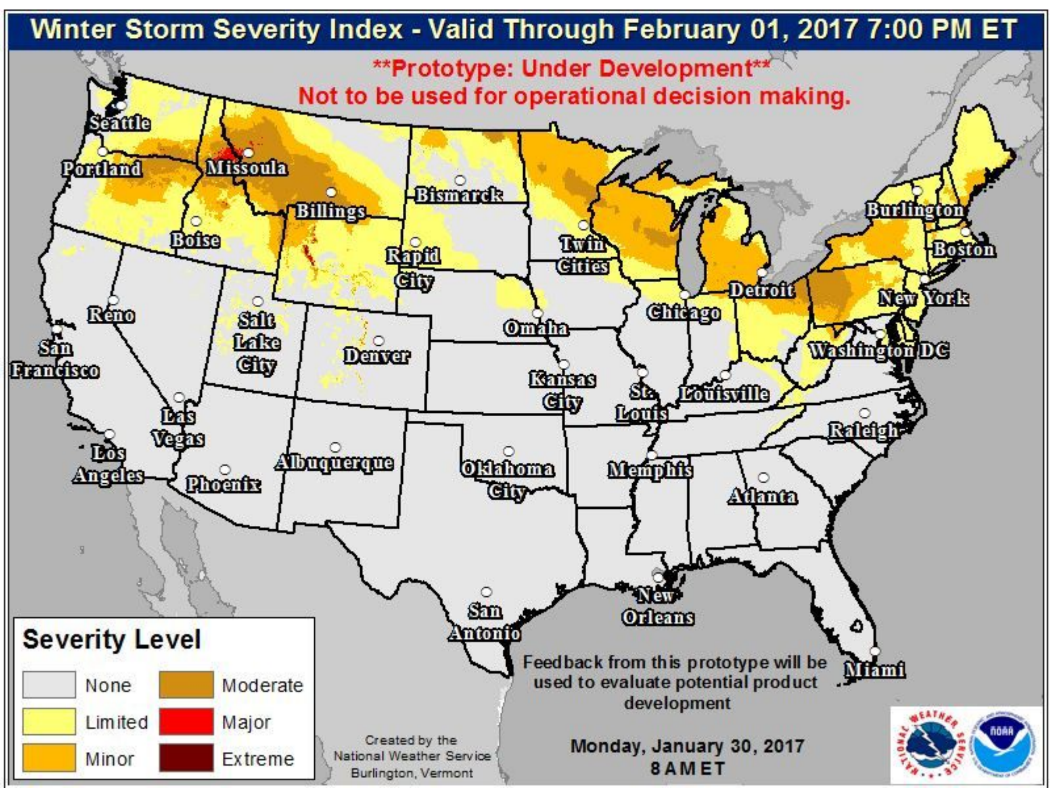


Figure 47. *An example of the WSSI cumulative 72-hour summary graphic displaying winter weather impact severity levels valid Tuesday, February 1, 2017.*

The WSSI is made up of a series of sub-components (Figure 48) which use meteorological and non-meteorological data to determine a level of potential societal impact based upon specific characteristics of winter storms. Each of the components produce a 1 to 5 output scale value that equates to the potential impact. The final WSSI value is the maximum value from all the sub-components. The 5 levels are given the following descriptors: Limited, Minor, Moderate, Major and Extreme. The specific sub-components are:

- Snow Load Index (potential of downed trees/power lines due to the weight of the snow)
- Snow Amount Index (potential of impacts due to the total amount of snow or accumulation rate)
- Ice Accumulation (potential of tree and utility damage as well as transportation difficulties due to combined effects of ice and wind)
- Blowing Snow Index (potential disruption due to blowing snow)
- Flash Freeze Index (potential impacts of flash freezing during precipitation events)

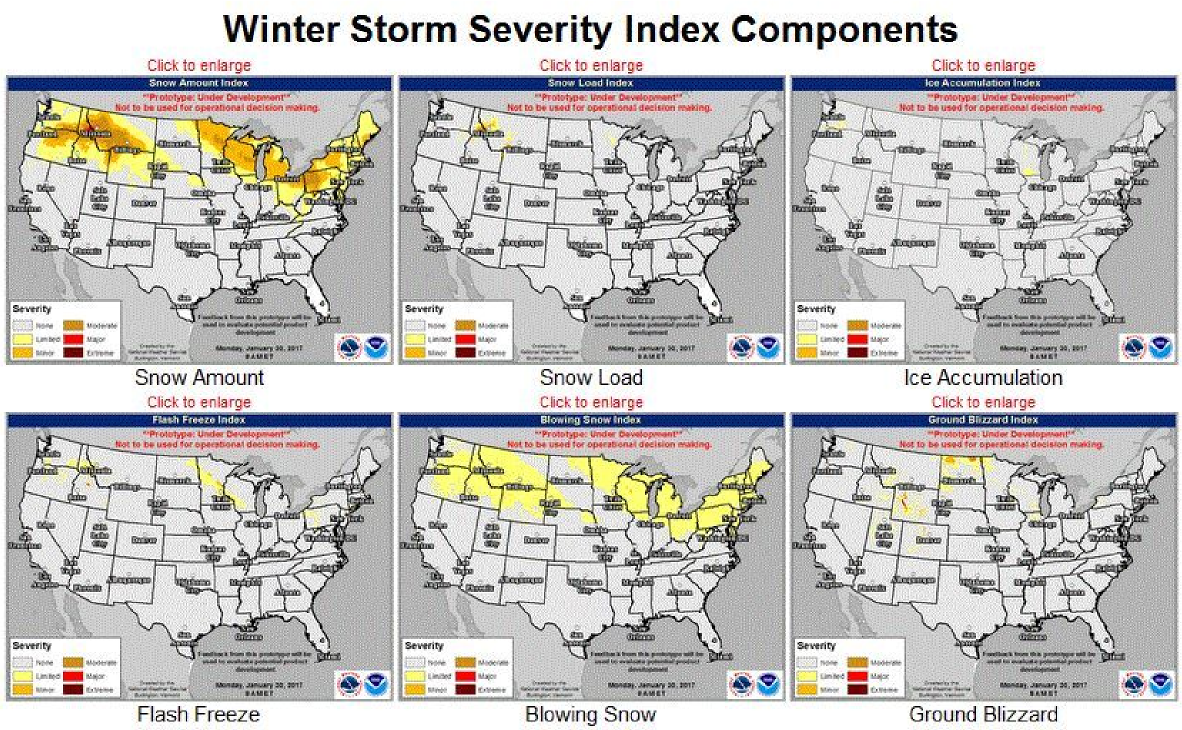


Figure 48. *An example of the WSSI breakdown of winter weather impacts and their relative severity levels valid Monday, January 30, 2017. These components contribute to the summary.*

During the 2017 WWE, the WSSI was used to aid in the assessment of potential winter weather impacts for the issuance of the Winter Weather Alerts. Participants appreciated the mission and presentation of the WSSI, but noted the need for a 24 hour forecasts to maximize its application to the Day 2 Winter Alerts activity as opposed to the current 72-hour cumulative forecast. It was useful as a last check to compare the alert assessment with the WSSI expected impacts. It also fostered conversation of impacts-based forecasting and communication with the public. Most are in agreement that a tool like the WSSI is needed and would have utility when collaborating with emergency management and the public.

The participants were asked to comment on the mapping technology, color choices, clarity of information offered, and scientific support of the WSSI at the end of their week in the WWE. Opinions were sought through a survey question about its utility and how it may be improved.

Comments: Most survey responses echoed that the idea of the WSSI is valued and needed to improve impacts-based forecasting and delivering decision support services (DSS) to the public. Output could be applied to materials used when creating DSS packages. However, there are many areas where the WSSI misses the mark. Common noted deficiencies centered around the cumulative spatial and temporal scale being too limited and the lack of confidence information or predictability.

Suggestions: Smaller time steps in 6, 12 and 24 hours are desired as well as broadening input to include Numerical Weather Prediction (NWP) and not just NDFD. Probabilistic output derived from ensembles to convey uncertainty would improve the WSSI and provide more meaningful information.

Comments: Regarding presentation, comments included concerns about the color scale not being broad enough to distinguish between severity levels. Some felt "minor" and "limited" were not necessary and focus should be toward "moderate, major, and extreme" only. Others also felt the summary graphic did not supply enough information and the breakdown of the different impacts was more valuable. Even better would be providing a probabilistic likelihood of the individual hazards having an impact. Three participants desired a Wind Chill calculation (a temperature component) in addition to the winter hazards available in the WSSI.

Suggestions: The WSSI is a solid first step toward offering an automated tool for public consumption taking us into the realm of impact-based forecasting and weather event communication. To maximize its potential would require NWP model infusion to build predictability, probabilistic output to convey uncertainty, GIS and Python development to improve the interface including better labeling and metadata, social science to improve the color scale and strength of messaging, and further testing and experimentation of the science and end-user application.

Winter Alerts

Joint probabilities, experimental guidance, operational guidance, and the WSSI contributed to the resulting Winter Alerts product (Figure 49).

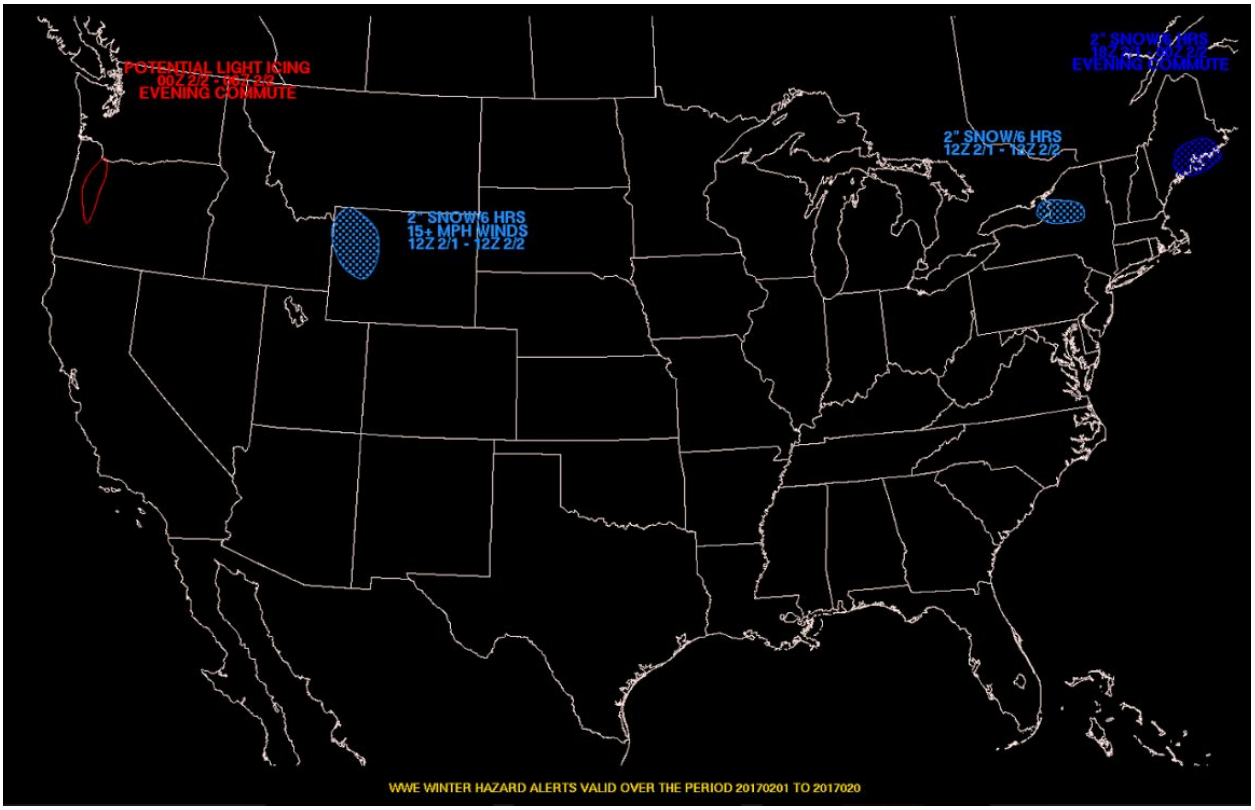


Figure 49. *Experimental Winter Weather Alerts issued for the Day 2 period valid 12Z February 2, 2017. Shaded light blue areas indicate heavy snow and wind in 6 hours. The dark blue indicated heavy snow and wind during the evening commute. And the red area indicated light icing on the roadways during the evening commute.*

Verification for this activity was limited and challenging. Although the majority of forecaster participants prefer this type of alerting over watches for impacts-based messaging, the alerts scored lower for under-prediction and having too low of confidence that an event will occur. The winter alerts received an average rating of 6.5 out of 10 (Figure 50), which was slightly better than the watches.

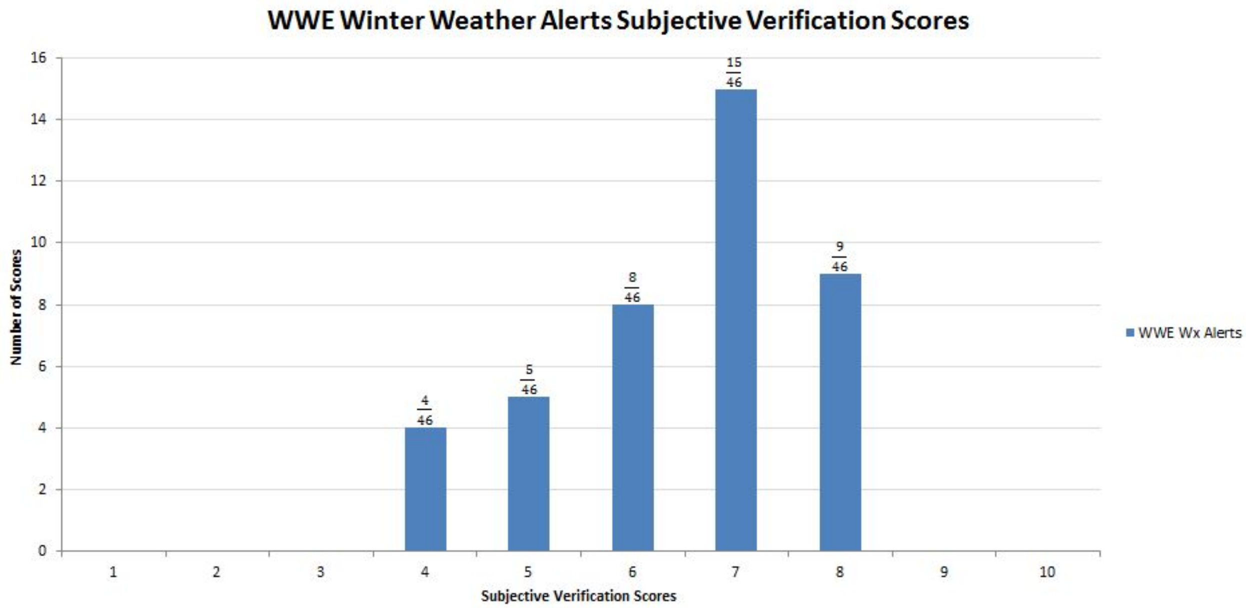


Figure 50. *The 2017 WWE distribution of subjective verification scores for the winter weather alerts. The alerts were scored each day it was available on a scale of 1 (very poor) to 10 (very good).*

Similar to the watch exercise, where verifiable data was available, participants felt areas should have been broadened spatially and that confidence in the potential impacts increased. It was also suggested that joint probabilities and localized winter alerts draw from the developing NWS Impacts-Based Catalog.

General Feedback on Center-Based Winter Watches:

Participants were polled at the end of the experiment week to provide an opportunity to share their thoughts on the criteria-based watches, the impacts-based winter weather alerts, and the ability of a National Center to issue them. Roughly 70% of the responses favored the impact-based winter alerts paradigm as it would give the public detailed, actionable information and would not be constrained to the current watch time scales and directives. The participants overwhelmingly agree that this is the future of the NWS delivering impact-based decision support services to emergency management, transportation agencies, and the public. Reasonably high confidence in specific impacts two or three days out should be communicated to the public in a preparatory format without waiting for the appropriate warning window.

However, in context of watches being issued by a national center, thinking shifted to the need to have a hybrid product that included both exceedance criteria and additional hazards or impacts. Criteria-based Winter Weather Watches are objective and can be automated. It is strongly suggested that the national center would need an automated starting part, such as the Watch Collaborator, to identify winter hazards CONUS-wide. Winter Weather Alerts are more subjective, demanding local knowledge and temporal input that cannot easily be automated nor captured by a national center without deep collaboration with WFOs. Thus, a Winter

Weather Alert paradigm may defeat the goal of reducing WFO time dedicated to issuing watches and allow them to focus on more imminent hazards and warnings. Other concerns were the dedication and availability of WPC forecasters to be able to issue over the CONUS and collaborate with each affected WFO (staffing increases suggested, as well as dividing responsibility into western CONUS and eastern CONUS). Thresholds would need to be determined regarding the minimum areal extent and level of geographic detail included in a national center winter storm watch.

The majority of participants felt Winter Watches issued by a national center could be accomplished with tools such as an enhanced Watch Collaborator, joint probabilities, the Impacts-Based Catalog, improved WFO collaboration, and increased WPC staff. This conclusion is in recognition of the need for better alignment with watch responsibilities already held by the Storm Prediction Center and the National Hurricane Center. Other participants felt the requirement for detail and local knowledge would be too great for a national center and watch responsibilities should stay with the WFOs.

9. The Re-evaluation of the HRAM Ensemble QPF

An error in the calculation of the HRAM downscaled QPF used during the 2016 Winter Weather Experiment resulted in the desire to re-evaluate the different dynamic downscaling techniques offered in the HRAM3E and HRAM3G QPF fields this year. The High Resolution Window NMMB was used two times in the ensemble averaging alongside the NAM CONUS Nest ensembling, thus the advertised High Resolution Window ARW member was excluded during the period of the 2016 WWE. A correction was made in April 2016, past the conclusion of the experiment. For the 2017 WWE, the HRAM3E and HRAM3G were both subjectively evaluated and objectively compared to Stage IV QPE (Figure 51).

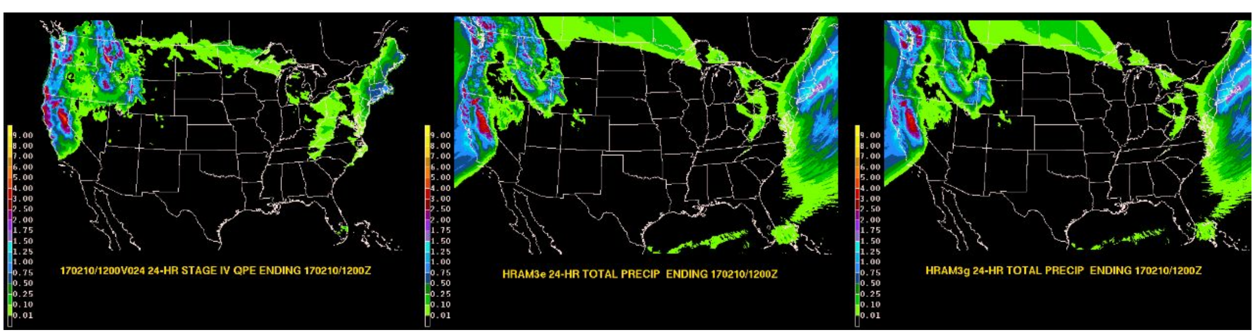


***Figure 51**. An example of the 24 hour Stage IV QPE (left) compared with the HRAM3E (middle) and HRAM3G (right) during subjective evaluation of the forecast valid 12Z February 10, 2017.*

Although the two models differ in dynamic downscaling techniques, their resulting QPF fields very closely resemble each other on a run-to-run basis. Comments regarding the QPF were fairly consistent on several points with few events varying:

1. The capturing of precipitation amounts, spatial extent and detail of terrain over the intermountain west and Pacific northwest were impressive overall
2. The Gulf Coast and Southeast were often underdone, scouring out too much of the lighter precipitation and bringing down the QPF maxima
3. The Northeast and Central Plains were well-captured for most events
4. Details of the QPF spatial extent and intensity often struggled over the Great Lakes and was not particularly helpful in identifying lake effect snow banding

The participants scored the HRAM3E and HRAM3G an average of 6.8 and 6.2 out of 10, respectively. Figure 52 shows the distribution the of the participant scores.

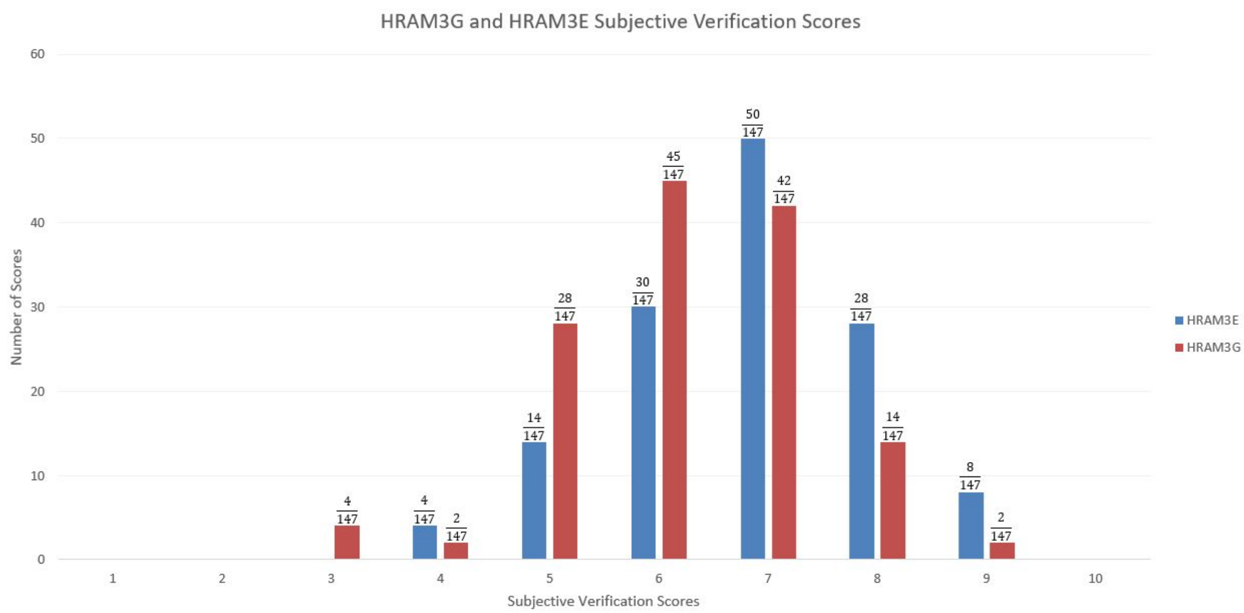


Figure 52. *The 2017 WWE distribution of subjective verification scores for the HRAM3E and HRAM3G. The ensembles were scored each day on a scale of 1 (very poor) to 10 (very good).*

Although the averages were very close, evident in the dispersion of the participant scores shown in Figure 53 is that the HRAM3E was a superior performer, having more frequent high scores and no scores below 4 out of 10.

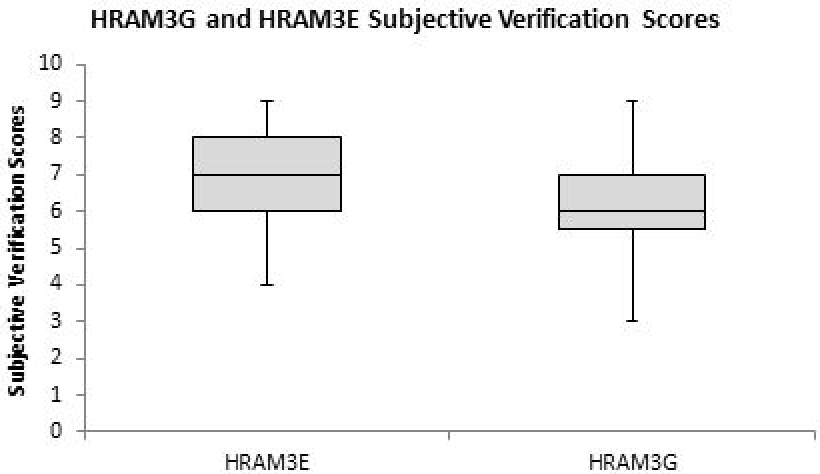


***Figure 53**. The box-and-whisker plot for the HRAM3E and HRAM3G showing the median dispersion of participant scores.*

In an attempt to offer participants objective verification to supplement the subjective, the Developmental Testbed Center's (DTC) Method for Object-Based Diagnostic Evaluations (MODE) was run to identify forecast QPF objects at different thresholds and match to observation objects. Figures 54 and 55 demonstrate the MODE verification of the HRAM3G and 3E for the 1" QPF threshold on the same valid date as the example above. The green shading is the Stage IV QPE and the red outline is the model. The interest value is a measure of how similar the QPF objects are to the QPE objects and ranges between 0 and 1. An interest of 1 means that the forecast and observation objects have very similar attributes. For this event, both HRAMs were overdone in the Northeast, but closely represented the verifying analysis in most areas of the west.

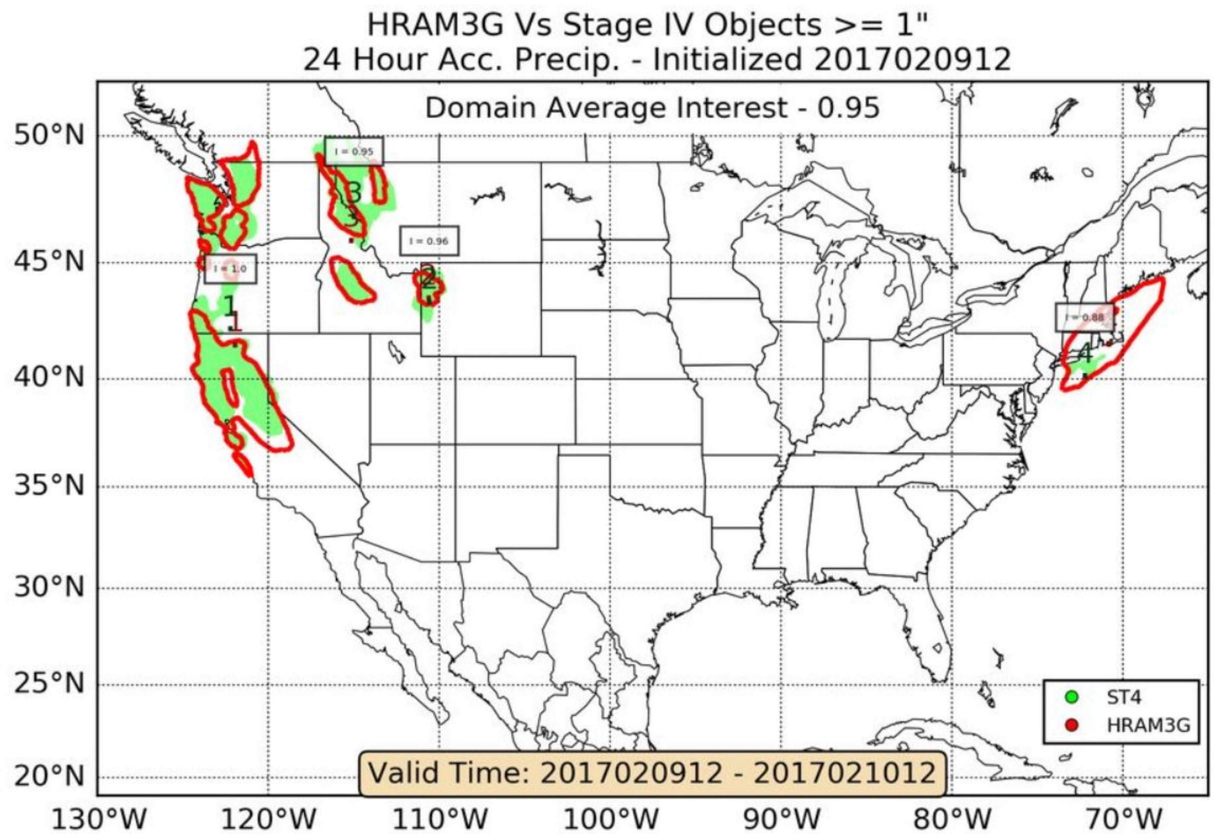


Figure 54*. MODE verification of the HRAM3G vs Stage IV for QPF objects of greater than or equal to 1" valid February 10, 2017.*

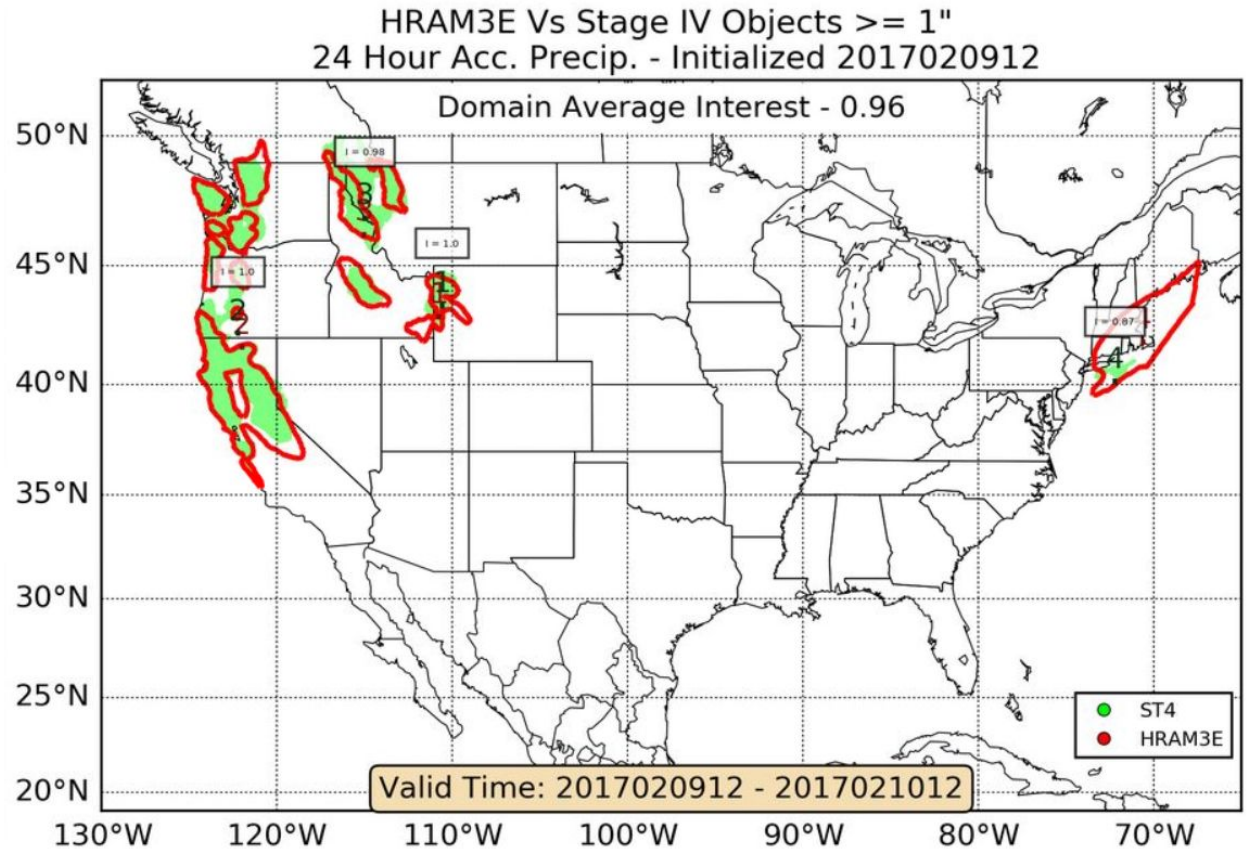


***Figure 55**. MODE verification of the HRAM3E vs Stage IV for QPF objects of greater than or equal to 1" valid February 10, 2017.*

10. Summary and Research-to-Operations Recommendations

The HMT conducted the 7th annual Winter Weather Experiment from January 19-February 19, 2017 at the NCWCP building in College Park, MD. This experiment brought together WFO forecasters, academic researchers, and numerical weather prediction modelers to explore new model data sets, forecast tools, and forecasting techniques with the goal of improving winter weather prediction both at WPC and in the field. Each day, the participants deterministically forecasted snow and ice for Day 2 (12Z-12Z) over the CONUS and probabilistically forecasted snowfall rates throughout Day 1 (18Z-12Z) over a chosen limited domain. Based on the Day 2 forecasts, the Watch Collaborator and experimental guidance were applied to the issuing Winter Storm Watches and Winter Weather Alerts. Participants also subjectively evaluated the experimental forecasts and guidance each day.

The hourly snowfall accumulation fields from the NAMv4 and HRRRv3 were determined to be beneficial to the forecast process and it is recommended that these fields be made available to the WPC and WFO forecasters to be used in regular operations. WPC-HMT will work with the Storm Prediction Center (SPC) to ensure transfer to Heavy Snowfall Mesoscale Discussion operations, and the WFO IT officers to ensure these fields can be extracted for visualization in AWIPS.

Likewise, the probabilistic hourly snowfall fields from the HRRR-TLE provided valuable information when forecasting snowfall and snow intensity. It is recommended that the neighborhood filter size be altered to better represent snowfall, and to consider exploring a point probability visualization to capture more terrain, convective banding, and lake effect streamer detail. These new representations of the HRRR-TLE probabilistic snowfall data should then be further evaluated.

The WPC Implicit PWPF showed significant potential compared to the operational PWPF in the 2017 WWE particularly with increasing terrain detail and spatial distribution of the snowfall over the CONUS. Participants would have preferred the PWPF produce higher probabilities in many cases. It is recommended that WPC continue development and testing of the Implicit PWPF.

The Watch Collaborator Trend Tools were useful for establishing forecast trends and tailoring Winter Storm Watch considerations. Establishing areas over which the probabilities are increasing or decreasing over model runs enables deeper collaboration with neighboring WFOs during Winter Storm Watch decision- making. It is recommended that the Watch Collaborator Trend Tools be transitioned into operations by migrating them to the Watch Collaborator web page currently accessible through the WPC Winter Weather Desk internal site designed for WFOs.

The concept of joint probabilistic guidance was of great interest and value to participants for a quick assessment for the potential of multiple hazards. The WPC joint probabilities were a first attempt at providing the guidance. The HMT staff recommends improving the forecaster derived probabilities and adding more multi-variable fields, and working with model developer partners to extract assorted winter weather joint probabilities from NWP (for example, the HRRR-TLE now offers a probability of "blizzard conditions" by combining multiple fields). This joint probabilistic guidance will work to improve forecaster decision support services by focusing on winter weather impacts.

Winter Storm Watches and Winter Weather Alerts from a national center perspective were successfully tested for the first time, as encouraged by the FY17 AFS Milestone. Feedback was collected by WWE participants and processed. It is the consensus that with better predictive tools and increased staffing, a national center could issue WInter Storm Watches. It is recommended that if widely supported, testing should continue in a year-round capacity parallel with operations on the WPC Winter Weather Desk.

Acknowledgements

We would like to acknowledge the vital facilitation and technical development provided by Joshua Kastman, as well as his production of graphs and box plots for this document. Thank you also to Jim Nelson for the recruitment of participants for the 2017 WWE.

References

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

Hamill, T. M., G.T. Bates , J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble forecast dataset. *Bull. Amer. Meteor. Soc.* **94**, 1553–1565.

Manikin, G. S., 2005: An overview of precipitation type forecasting using NAM and SREF data. *21st Conference on Weather Analysis and Forecasting/17th Conference on Numerical Weather Prediction*, Washington, D.C., 8A.6.

Roebber, P. J., S. L. Bruening, D. M. Schultz, and J. V. Cortinas, 2003: Improving snowfall forecasts by diagnosing snow density. *Wea. Forecasting*, **18**, 264-287.

Roebber, P.J., Melissa R. Butt, Sarah J. Reinke, and Thomas J. Grafenauer, 2007: Real-Time Forecasting of Snowfall Using a Neural Network. Wea. Forecasting, 22, 676–684.

Von Storch, H., and F.W. Zwiers 1999. *Statistical Analysis in Climate Research*. Cambridge University Press, 496 pp.

Wilks, D.S., 2006. *Statistical Methods in the Atmospheric Sciences, 2nd Edition*. International Geophysics, 648 pp.

APPENDIX A

HRRR-TLE Retrospective Preliminary Objective Evaluation

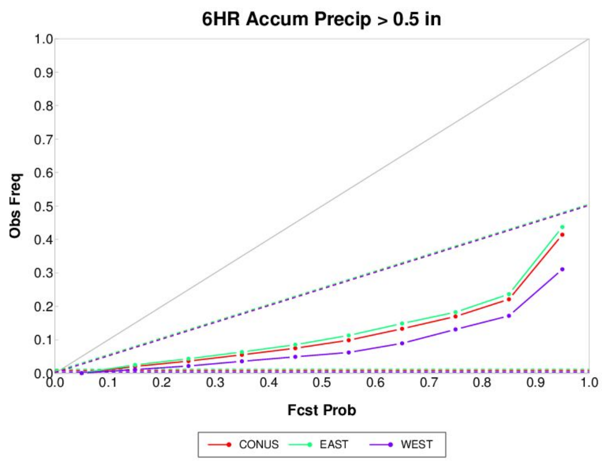
The High Resolution Rapid Refresh – Time Lagged Ensemble (HRRR-TLE) was evaluated retrospectively using Stage IV accumulated precipitation and the Weather Prediction Center (WPC) snowfall analysis. Available fields from March 2016-May 2016, September 2016-December 2016, and January 2017-February 2017 were evaluated. The Winter Weather Experiment (WWE) occurred between January-February 2017 and was evaluated separately to determine if the HRRR-TLE performed similarly during this period. Table 1 indicates the probability of 1,3, and 6-hr accumulated precipitation exceeding several thresholds were evaluated. Additionally, the probability of 1-hr snowfall exceeding several thresholds were also evaluated. Three regions were evaluated using the National Center for Environmental Prediction (NCEP) masking regions for the Continental US (CONUS), the Eastern US (EAST) and the Western US (WEST).

Table 1. HRRR-TLE fields and thresholds evaluated

Field	Thresholds (inches)	Analysis
1hr Accum. Precip.	0.5, 1, 2	Stage IV
3hr Accum. Precip.	0.5, 1, 2, 3	Stage IV
6hr Accum. Precip.	0.5, 1, 2, 3, 6	Stage IV
1hr Snowfall	0.5, 1, 2, 3, 6	Stage IV (using 10:1 ratio)
1hr Snowfall	0.5, 1, 2, 3, 6	WPC Snowfall Analysis

The Model Evaluation Tools (MET) were used to compute scores for the probabilistic forecasts. Reliability diagrams are used in this section to evaluate how well predicted probabilities of an event correspond to the observed frequency of the event. Forecasts that lie along the 1-to-1 line (grey line that runs from lower left corner to upper right) are considered perfectly reliable. If the curve lies below the line, this indicates the probabilities tend to be too high. Points above the line indicate probabilities tend to be too low.

Analysis shown in Figures 1-3 all indicate that the HRRR-TLE tends to over-predict probabilities for the sample we evaluated. However, due to sample size limitations, this analysis should be considered preliminary being many of the events being predicted are limited in temporal and spatial extent. The 6 hour accumulated precipitation fields were evaluated over the entire dataset and then over the 2017 WWE time period. Figure 1 shows that the HRRR-TLE tends to over-predict probabilities more in the WEST (purple) than in the EAST (green) with the CONUS (red) reliability diagram, with the largest over prediction in the low probability events. During the 2017 WWE, the HRRR-TLE tended to be more reliable in for the high probability events than during the previous year. A similar trend of over-prediction, especially of lower probability events, is present in Figure 2. This figure shows the 1hr snowfall accumulation field evaluated against both the Stage IV 1-hr precipitation analysis and the WPC snowfall analysis. A ratio of 10:1 was applied to the Stage IV precipitation field to perform the evaluation. In both panels of Figure 2, the probabilities of 1hr snowfall exceeding 1 inch for the EAST are much more reliable than those in the West. The difference in using one analysis versus another is evident by comparing the two East region curves. If Stage IV is used, the predictions in the EAST all appear to be above the no-skill line (dashed line), while if the WPC snow analysis is used, only the high probability events show skill. Figure 3 shows the reliability diagrams for 1hr snowfall exceeding 3 inches. Similar to Figure 1, the entire period is provided in the left panel while the results just prior, during and after the WWE are in the right. Once again, the HRRR-TLE shows marked improvement in reliability over the EAST region but showed a little less reliability during the 2017 WWE. This is likely due to a very small sample size for this type of event.



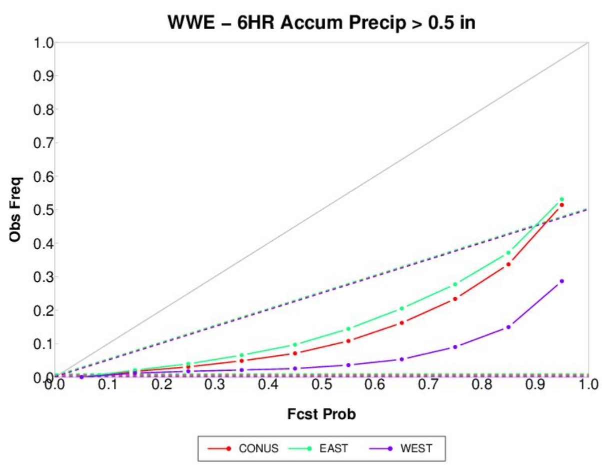
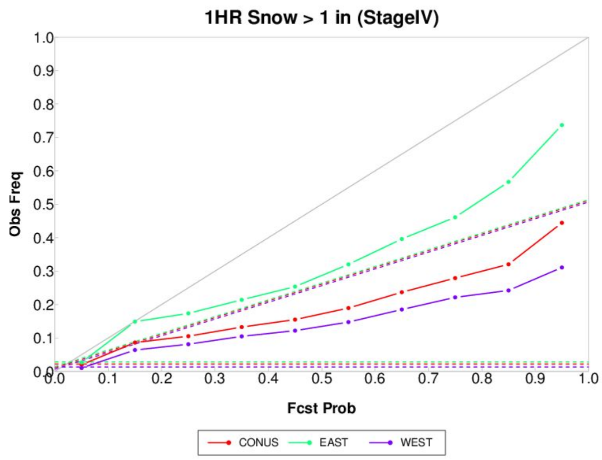


Figure 1. Reliability Diagrams for 6hr accumulated precipitation for the entire evaluation period (left) and during the 2017 Winter Weather Experiment (right).



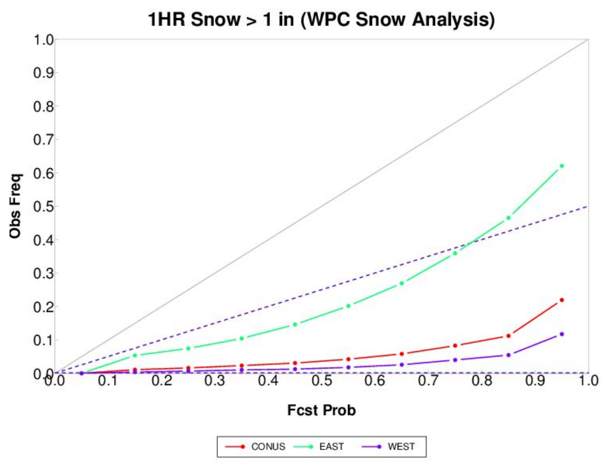
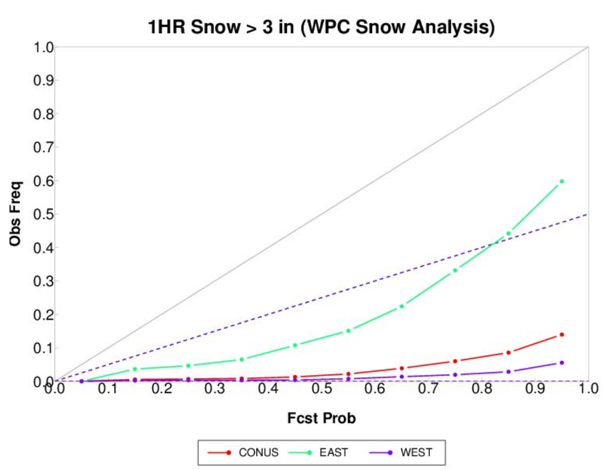


Figure 2. Reliability Diagrams for 1hr snowfall for the entire evaluation period evaluated against Stage IV (left) using a 10:1 ratio and WPC Snow Analysis (right).



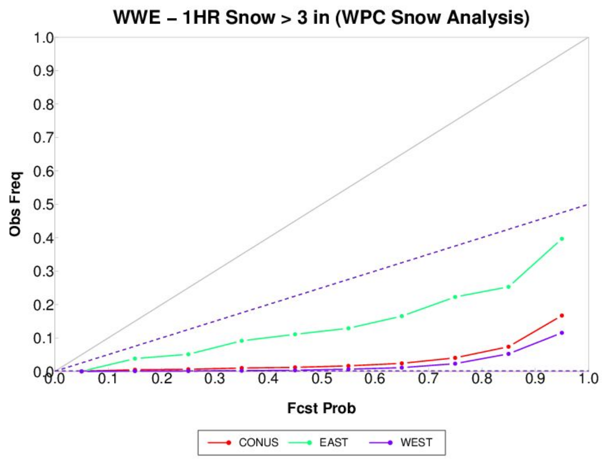


Figure 3. Reliability Diagrams for 1hr snowfall > 3 inches for the entire evaluation period (left) and during the 2017 Winter Weather Experiment (right) using the WPC Snow Analysis.