SPRING FORECASTING EXPERIMENT 2026

Conducted by the

NOAA/HAZARDOUS WEATHER TESTBED

HWT Facility – National Weather Center
27 April – 29 May 2026
https://hwt.nssl.noaa.gov/sfe/2026/

Program Overview and Operations Plan

17 April 2026

Adam Clark[2,4], Israel Jirak[1], Thomas Galarneau[2,4], Tim Supinie[1], Kent Knopfmeier[2,3], David Harrison[1,3], Jake Vancil[1,3], Miranda Silcott[2,3], David Jahn[1,3], Chris Karstens[1], Andy Wade[1,3], Kimberly Hoogewind[2,3], Corey Potvin[2,4], Nusrat Yussouf[2,3], Sean Ernst[1,3,5], Joey Picca[1,3], Michael Baldwin[1,3], Jorge Duarte[2,3], Joshua Martin[2,3], and Brian Matilla[2,3]

(1) NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma
(2) NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma
(3) Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma
(4) School of Meteorology, University of Oklahoma, Norman, Oklahoma
(5) Institute for Public Policy Research and Analysis, University of Oklahoma, Norman, Oklahoma

Table of Contents

1. Introduction....4
2. Overview of Experimental Products and Models....5
a) The 2026 Community Leveraged Unified Ensemble (CLUE).... 5
b) High Resolution Ensemble Forecast (HREFv3) System.... 9
c) WoFS configurations....9
i. WoFS VLAB.... 10
ii. WoFS MPAS (NSSL).... 12
iii. WoFSCast (NSSL).... 12
d) Artificial Intelligence Weather Prediction (AIWP) Models.... 12
i. GraphCast (Brightband).... 14
ii. Aurora (Brightband).... 14
iii. PanguWx (Brightband).... 14
iv. AIFS (ECMWF).... 15
v. AIGFS (Project EAGLE).... 15
vi. AIGEFS (Project EAGLE).... 15
vii. Nested-EAGLE (Project EAGLE).... 16
viii. WeatherNext 2.... 16
ix. AIFS Ensemble (ECMWF).... 17
x. HRRRCast Ensemble.... 17
xi. Stormscope (Nvidia).... 18
e) Calibrated Forecast Products.... 18
i. GEFS reforecast MLP (SPC/CSU).... 18
ii. GEFS operational MLP (SPC/NSSL).... 19
iii. NSF NCAR AIWP.... 20
iv. OU AIWP.... 20
v. SUPER-EC.... 21
vi. StormNet.... 21
vii. REFS/GEFS & HREF/GEFS.... 22
iix. Loken HREF & Vancil REFS.... 22
ix. Nadocast HREF & Nadocast REFS.... 23
x. STP Cal Inflow for HREF and REFS.... 23
3. SFE 2026 Core Interests and Daily Activities....24
a) Formal Evaluation Activities.... 25
(C)alibrated Guidance.... 26
C1. Medium-Range Severe ML Guidance.... 26
C2. Day 1 00Z Tornado ML Guidance.... 26
C3. Day 1 00Z Hail ML Guidance.... 26
C4. Day 1 00Z Wind ML Guidance.... 26
CAM (E)nsembles.... 27

E1. CLUE: Day 1 Ensemble Flagships..... 27
E2. CLUE: Day 2 Ensemble Flagships..... 27
E3. CLUE: Short-Term Ensembles.....27
(D)eterministic CAMs..... 28
D1. CLUE: 00Z Day 1 Deterministic Flagships..... 28
D2. CLUE: 12Z Day 2 Deterministic Flagships..... 28
D3. CLUE: 00Z Day 3 Deterministic Flagships..... 28
D4. CLUE: Day 1 MPAS ICs..... 28
D5. CLUE: Data Assimilation Comparisons..... 29
(A)nalyses..... 29
A1. Mesoscale Analysis Background..... 29
A2. Upscaled Mesoscale Analysis Background..... 29
A3. Storm Scale Analysis..... 29
(A)rtificial (I)ntelligence Evaluations..... 30
AI1. Global AIWP Models: GFS ICs..... 30
AI2. Global AIWP Models: EC ICs..... 30
AI3. Global AIWP Models: IC Comparison..... 30
AI4. Global AIWP Ensembles..... 30
AI5. WoFS Deterministic Comparisons..... 31
AI6. WoFS Ensemble Comparisons..... 31
AI7. 00Z Day 1 HRRRCast vs. HRRR & REFS..... 31
AI8. Short-Term AIWP..... 31
AI9. MD Peak Intensity Bin Guidance..... 32
(O)utlook Evaluations..... 32
O1. Day 1/2/3/4 Outlooks..... 32
O2. 4-h Day 0 Outlooks..... 32
O3. SPC Impacts System: Day 1 Outlook Tornado, Hail, and Wind Counts. 32
b) Forecast Products and Activities..... 33
Appendix A: Organizational structure of the NOAA/Hazardous Weather Testbed..35
Appendix B Mandatory 2026 CLUE Fields.....38
References.....39





The NOAA Hazardous Weather Testbed (photo credit: James Murnan, NSSL)

1. Introduction

Each spring, the NOAA/NWS Storm Prediction Center (SPC) and NOAA/OAR National Severe Storms Laboratory (NSSL) conduct the NOAA/Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE). This experiment aims to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather. The primary goals of the SFE are to accelerate the transfer of promising new tools from research to operations, inspire new initiatives for operationally relevant research, and document performance characteristics of state-of-the art operational and experimental storm-scale modeling systems with 1- to 3-km grid-spacing.

The 2026 SFE will be conducted 27 April – 29 May. This will be the fourth hybrid experiment with both in-person and virtual participation. SFE 2026 will feature morning and afternoon forecasting activities, as well as next-day model evaluations. As in previous years, a suite of new and improved experimental guidance contributed by our large group of collaborators will be central to these forecasting and model evaluation activities. Several of these contributions comprise an ensemble framework called the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018). The 2026 CLUE is constructed by using common model specifications (e.g., grid-spacing, model version, domain size, post-processing, etc.) wherever possible, so that the simulations contributed by each group can be used in carefully designed controlled experiments. This design will once again allow us to conduct several experiments geared toward identifying optimal configuration strategies for deterministic convection-allowing models (CAMs) and CAM ensembles. The 2026 CLUE includes 11 members. The SFE 2026 will also continue testing of the Warn-on-Forecast System (WoFS), which produces 18-member, 3-km grid-spacing forecasts, and will be used for the 10th year to issue very short lead-time products. As a first step toward operational transition of WoFS, the NWS will be conducting the WoFS runs on VLab resources and working with SFE coordinators on daily domain placement (WoFS VLAB, hereafter).

With implementation of the Rapid Refresh Forecast System (RRFS) and RRFS Ensemble Forecast System (REFS) expected in late July 2026, a major point of emphasis will once again be evaluating these systems relative to the operational systems they are designed to replace. Additionally, with the recent announcement that the Model for Prediction Across Scales (MPAS) will be the future dynamical core for global and regional UFS applications, several MPAS evaluations will be conducted, including MPAS-based RRFSv2 prototypes contributed by the NOAA/Global Systems Laboratory (GSL) and a new MPAS-based version of WoFS from NSSL (WoFS MPAS). Finally, SFE 2026 will expand upon evaluations of global and regional Artificial Intelligence Weather Prediction (AIWP) models from the previous two years. Highlights include first-time evaluations of GSL's HRRRCast Ensemble, next-day evaluations of an improved version of NSSL's WoFSCast, which will also be used in an afternoon forecasting activity, first-time examinations of 0-3 h forecasts from a data-driven model contributed by NVIDIA, and several global AIWP models including the most recent version of Google DeepMind's global ensemble called WeatherNext 2.

This document summarizes the core interests of SFE 2026 with information on experiment operations. The organizational structure of the HWT and information on various forecast tools and diagnostics can also be found in this document. The remainder of the operations plan is organized as follows: Section 2 provides details on models and products being tested during SFE 2026 and Section 3 describes the core interests and new concepts being introduced for SFE 2026. Details on the SFE forecasting and more general information on NOAA's HWT are found in appendices.

2. Overview of Experimental Products and Models

Daily model evaluation activities will occur Tuesday through Friday from 9:00 – 11:00am (CDT) focusing on various CLUE subsets and other models, guidance, and products. The 2026 CLUE includes deterministic and ensemble forecasts using the most recent versions of the Finite Volume Cubed-Sphere Model (FV3) and MPAS. In addition to the CLUE, the 3-km grid-spacing High-Resolution Ensemble Forecast system version 3 (HREFv3) and the High-Resolution Rapid Refresh version 4 (HRRRv4) will be examined as the operational modeling baselines. The rest of this section provides further details on each modeling system utilized in SFE 2026.

a) The 2026 Community Leveraged Unified Ensemble (CLUE)

The CLUE is a carefully designed ensemble with members contributed by NOAA units: NSSL, GSL, and the Environmental Modeling Center (EMC); and a research group at the National Aeronautics and Space Administration (NASA). All CLUE members have convection-allowing resolution over a CONUS domain, except the RRFS and REFS, which cover North America. CLUE members have 3-km grid-spacing, except NASA FV3 (2.2-km). Depending on the CLUE subset, forecast lengths range from 48 to 72 h. Table 1 summarizes all 2026 CLUE contributions. Subsequent tables provide details on members in each subset, as well as ensembles comprising different combinations of members that will be evaluated to test different configuration strategies.

Table 1 Summary of the five unique subsets that comprise the 2026 CLUE.

CLUE Subset	# of mems	IC/LBC perts	Mixed Physics	Data Assimilation	Dynamical Core	Agency	Init. Times (UTC)	Forecast length (h)	Domain
RRFS	1	none	no	Hybrid 3DEnVar	FV3	EMC	00,06,12,18	84	N. America
REFS	5	EnKF	yes	Hybrid 3DEnVar	FV3	EMC	00,06,12,18	60	N. America
NSSL MPAS	2	none	no	HRRR or RRFS ICs	MPAS	NSSL	00,06,12,18	48 or 84	CONUS
GSL MPAS	2	none	no	RRFSv1 or Hybrid 3DEnVar	MPAS	GSL	00,06,12,18	36-84	CONUS
NASA FV3	1	none	no	GEOS-DA	FV3	NASA	00,12	72	CONUS

Table 2 Specifications for the Rapid Refresh Forecast System (RRFS). The RRFS is initialized from a hybrid 3DEnVar analysis and is the control member of the RRFS Ensemble Forecast System (REFS). The ensemble component of the 3DEnVar uses the RRFS Data Assimilation System (RDAS) ensemble Kalman filter. The RDAS uses a wide variety of conventional observations along with radar reflectivity and satellite radiance data. It also includes a nonvariational cloud analysis. For gravity wave drag, the small scale and turbulence orographic form drag options are used. RRFS forecasts are initialized from 00, 06, 12, and 18 UTC with forecasts to 84 h.

Members: RRFS	**ICs**	**LBCs**	**Micro-physics**	**PBL/SFC**	**LSM**	**Radiation**	**Cumulus**	**Dynamical Core**
RRFS	RRFS hybrid 3DEnVar	GFS	Thompson	MYNN/MYNN	RUC	RRTMG	saSAS	FV3

Table 3 Specifications for the RRFS Ensemble Forecast System (REFS). REFS model forecasts are initialized from 00, 06, 12, and 18 UTC with forecasts to 60 h. Schemes marked with an asterisk (*) indicate stochastically perturbed parameterizations (SPP) and those marked with a hashtag (#) include fixed parameter perturbations.

Members: REFS	**ICs**	**LBCs**	**Micro-physics**	**PBL/SFC**	**LSM**	**Radiation**	**Cumulus**	**Dynamical Core**
REFS01	RRFS enkf1	GEFS m1	Thompson*	TKE-EDMF/GFS	RUC*	RRTMG*	GF-deep*+sh	FV3
REFS02	RRFS enkf2	GEFS m2	Thompson*	MYNN*/ MYNN*	RUC*	RRTMG*	saSAS deep	FV3
REFS03	RRFS enkf3	GEFS m3	NSSL#	MYNN*/ MYNN*	RUC*	RRTMG*	GF deep	FV3
REFS04	RRFS enkf4	GEFS m4	NSSL#	TKE-EDMF/GFS	RUC*	RRTMG*	GF-deep*+sh	FV3
REFS05	RRFS enkf5	GEFS m5	NSSL#	MYNN*/MYNN*	RUC*	RRTMG*	saSAS deep	FV3

Table 4 Specifications for the NSSL MPAS CLUE members. These members use 3-km grid-spacing covering the CONUS and are driven by the HRRR or RRFS. The last two letters of each member denote the ICs and microphysics [HT = HRRR-TEMPO (Li et al. 2024), RN = RRFS-NSSL 2-moment (Mansell 2010)]. NSSL MPAS HT runs are initialized from 00, 06, 12, and 18 UTC with forecasts to 48 h. The NSSL MPAS RN is initialized from 00 and 12 UTC with forecasts to 84 h.

Members: NSSL MPAS	**ICs**	**LBCs**	**Micro-physics**	**PBL**	**LSM**	**Radiation**	**Dynamical Core**
NSSL MPAS HT	HRRR	HRRR	TEMPO	MYNN	RUC	RRTMG	MPAS
NSSL MPAS RN	RRFS	RRFS	NSSL (2-mom)	MYNN	RUC	RRTMG	MPAS

Table 5 Specifications for the GSL MPAS CLUE members. These forecasts use 3-km horizontal cell spacing with 60 vertical levels across the CONUS. The PBL parameterization is an updated version of MYNN versus what is available in the NCAR-maintained MPAS release. The aerosol-aware Thompson-Eidhammer Microphysics Parameterization for Operations, or TEMPO, is used to parameterize microphysical processes and includes a 2-moment graupel representation with predicted density. Initial and lateral boundary conditions for aerosols for both forecasts are provided by MERRA-2 global climatological data.

Members: GSL MPAS	**ICs**	**LBCs**	**Micro-physics**	**PBL**	**LSM**	**Radiation**	**Dynamical Core**
GSL MPAS RT	RRFS	RRFS	TEMPO	MYNN	RUC	RRTMG	MPAS
GSL MPAS JEDI	JEDI Hybrid 3DEnVar	GFS	TEMPO	MYNN	RUC	RRTMG	MPAS

Table 6 Specifications for the NASA-FV3 CLUE member. The NASA-FV3 is also known as the NASA GEOS model and will run an FV3-based stretched global grid. The target resolution is a c2160 grid with 137 vertical levels, the stretching will produce a 2-km domain over CONUS with the coarsest global resolution of 12-km over the Indian Ocean. We will be running this case in a replay mode using an incremental analysis update (IAU) to our GEOS-FP 12-km production data assimilation system. The IAU approach permits our higher resolution model to evolve dynamically with time and avoids having to cold start forecasts each day. The NASA FV3 model will produce 72 h forecasts initialized at 0000 and 1200 UTC daily. Updates for SFE 2025 include retuned turbulence and convection, updated GFDL-MP cloud microphysics, new radar reflectivity calculations using the Thompson scheme calculation that includes brightbanding, and updates to the diffusion parameters in FV3.

Members: NASA FV3	**ICs**	**LBCs**	**Micro-physics**	**PBL**	**LSM**	**Radiation**	**Dynamical Core**
NASA FV3	GEOS-FP	None	GEOS-GFDL	Lock-Louis & UW	Nasa Catchment	RRTMGP	FV3

The configuration of the 2026 CLUE will allow for several unique experiments that have been designed to examine issues immediately relevant to the design of a NCEP/EMC operational CAM-based ensemble. Some of the major themes are listed below:

Operational ensemble membership (REFS vs. HREFv3 vs. SSEOv2): With operational implementation of REFS planned for late July 2026, ensemble comparisons of REFS vs. HREFv3 will be made at Day 1 & 2 lead times for 0000 and 1200 UTC initializations. An alternative configuration of REFS designed by SPC (SPC REFS) will also be included in the ensemble comparisons. SPC REFS includes more HRRR and RRFS control members compared to REFS. Finally, an ensemble leveraging all available sources of model diversity will be configured using HRRR, MPAS, and FV3 membership. This ensemble, the SPC Storm Scale Ensemble of Opportunity version 2 (SSEOv2), will be examined to assess whether it provides superior performance relative to REFS, and could thus be used while the MPAS-based REFSv2 is being developed. See Table 7 for configuration details.

Table 7 Ensemble members comprising REFS, SPC REFS, and SSEOv2 based at 1200 UTC. The members are shaded by model core and initialization time. The RRFS/REFS members (FV3) are shaded red and red/orange, the HRRR members (WRF-ARW) are shaded gray, and the MPAS members are shaded light orange and yellow. More recent runs have darker shades while older runs use lighter shades.

	REFS		SPC REFS		SSEOv2	
#	Member	Init. Time	Member	Init. Time	Member	Init. Time
1	RRFS	12Z	RRFS	12Z	RRFS	12Z
2	REFS01	12Z	REFS01	12Z	HRRR	12Z
3	REFS02	12Z	REFS02	12Z	RRFS	06Z
4	REFS03	12Z	REFS03	12Z	HRRR	06Z
5	REFS04	12Z	REFS05	12Z	MPAS GSL	06Z
6	REFS05	12Z	HRRR	12Z	MPAS NSSL	06Z
7	HRRR	12Z	RRFS	06Z	RRFS	00Z
8	RRFS	06Z	HRRR	06Z	HRRR	00Z
9	REFS01	06Z	RRFS	00Z	MPAS GSL	00Z
10	REFS02	06Z	HRRR	00Z	MPAS NSSL	00Z
11	REFS03	06Z				
12	REFS04	06Z				
13	REFS05	06Z				
14	HRRR	06Z				

Data assimilation comparisons: Comparisons will be made during the first 12 h of RRFS, HRRR, and GSL MPAS JEDI forecasts to evaluate the effectiveness of the data assimilation strategies in each system. While RRFS and HRRR, which use the Community Gridpoint Statistical Interpolation (GSI) software (Hu et al. 2018), have been evaluated for several years, this marks the first time that a fully-cycled, JEDI-based (Guerrette et al. 2023) convective-scale data assimilation system has been tested.

RRFSv2 Development Systems: NOAA/GSL will be providing two MPAS configurations (Table 5), which are being tested to form the foundation for RRFSv2. These configurations contain the most up-to-date physics suites tuned for MPAS. GSL MPAS RT is initialized from RRFS, while GSL MPAS JEDI uses its own fully-cycled JEDI-based data assimilation system. Comparisons will be made to each other and additional deterministic CAMs such as HRRR, RRFSv1, NSSL MPAS HT, and NASA FV3 at a range of lead times.

3D-RTMA Background and Storm-Scale Analyses: An hourly version of 3D-RTMA that uses the HRRR for the background first guess (3D-RTMA HRRR) will be compared to the surface objective analysis HRRR (sfcOA HRRR), which is created by performing a simple 2-pass Barnes analysis on surface observations with the HRRR analysis as the first-guess and direct use of the HRRR analysis for the atmospheric state above the surface. Analyses from WoFS VLAB and WoFS MPAS will also be incorporated into these comparisons. Upscaled versions of the analyses to 40-km will also be examined and compared with SPC's RAP-based surface objective analysis (sfcOA). Finally,

15-minute WoFS VLAB forecasts of hourly maximum 80-m winds, UH, and updraft speed will be compared to Multi-Radar, Multi-Sensor (MRMS) products to gauge whether these 15-minute WoFS VLAB forecasts are a viable proxy for observed hazards.

To ensure consistent post-processing, visualization, and verification, post-processing is standardized as much as possible, so that a consistent set of model output fields are output on the same grid. For the 2026 CLUE, all groups output fields to the 3-km CONUS grid used for the operational HRRR. For FV3 and MPAS the Unified Post-Processor software (UPP; https://www.epic.noaa.gov/unifiedpost-processor) is used and a minimum set of 49 output fields is provided at hourly intervals. This list of mandatory CLUE fields is provided in Appendix B and includes fields that are relevant to a broad range of forecast needs, including aviation, severe weather, and precipitation.

b) High Resolution Ensemble Forecast (HREFv3) System

HREFv3 is a 10-member CAM ensemble that was implemented on 11 May 2021. The design of HREFv3 originated from the SSEOv1, which demonstrated skill for six years in the HWT and SPC prior to operational implementation as the HREF in 2017. In HREFv3, the HRW NMMB simulations have been replaced with HRW FV3 and HRRRv3 has been upgraded to HRRRv4. HREFv3 specifications are listed in Table 8.

Table 8 Model specifications for HREFv3.

HREFv3	**ICs**	**LBCs**	**Micro-physics**	**PBL**	**dx (km)**	**Vertical Levels**	**HREF hours**
HRRRv4	HRRRDAS	RAP -1h	Thompson	MYNN	3.0	50	0 - 48
HRRRv4 -6h	HRRRDAS	RAP -1h	Thompson	MYNN	3.0	50	0 - 42
HRW ARW	RAP	GFS -6h	WSM6	YSU	3.2	50	0 - 48
HWR ARW -12h	RAP	GFS -6h	WSM6	YSU	3.2	50	0 - 36
HRW FV3	GFS	GFS -6h	GFDL	EDMF	3	50	0 - 60
HRW FV3 -12h	GFS	GFS -6h	GFDL	EDMF	3	50	0 - 48
HRW NSSL	NAM	NAM -6h	WSM6	MYJ	3.2	40	0 - 48
HRW NSSL -12h	NAM	NAM -6h	WSM6	MYJ	3.2	40	0 - 36
NAM CONUS Nest	NAM	NAM	Ferrier-Aligo	MYJ	3.0	60	0 - 60
NAM CONUS Nest -12h	NAM	NAM	Ferrier-Aligo	MYJ	3.0	60	0 - 48

c) WoFS configurations

The NWS is running a demonstration version of WoFS version 1 on the NOAA Virtual Laboratory (WoFS VLAB), and these runs will coincide with the daily SFE domain

in 2026. NSSL is also running two next-generation WoFS configurations in real-time, WoFS MPAS and WoFSCast. The three WoFS configurations are described below.

i. WoFS VLAB

WoFS VLAB is a rapidly-updating 36-member, 3-km grid-spacing WRF-ARW-based ensemble data assimilation and forecast system. WoFS VLAB forecasts are initialized every 30 minutes and produce very short-range (0-6/0-3 hour for runs launched at top/bottom of the hour) probabilistic forecasts of individual thunderstorms and their associated hazardous weather phenomena, such as hail, high winds, heavy rainfall, and supercell thunderstorm rotation. The 900-km x 900-km daily WoFS VLAB domain will target the primary region where severe weather is anticipated, using the SPC Day 1 Convective Outlook as a guide. WoFS VLAB is capable of running over two different regions simultaneously. A second domain will only be implemented when there are two distinct regions where severe weather is expected (e.g., the Plains and the East Coast), or when there is a very large single area for which two domains are needed to cover the entire risk area.

WoFS VLAB is based on the cloud-based Warn-on-Forecast System (cb-WoFS; Martin et al. 2025) which uses current technologies in containerization and cloud computing. The entire WoFS VLAB application was built on top of multiple Platform-as-a-Service and Infrastructure-as-a-Service technologies on the Azure platform and the WRF model itself was rebuilt to run in containers optimized for HPC. With the WoFS VLAB interface, administrators can easily configure the domain and dynamically create an HPC infrastructure for the run, and upon completion, tear it down, thereby reducing costs by only paying for used resources. Another benefit is that as Azure continues to add new, updated computer core types from chip manufacturers, these options are passed down to Azure customers, giving WoFS VLAB operators the choice of running on the latest technologies. All parts of WoFS VLAB have been rebuilt for scalability: the containerized WRF can be executed on any node, the post-processing is built on high performance queues and containerized, so any number of post-processing jobs can run concurrently.

The initial conditions for WoFS VLAB are provided by the High-Resolution Rapid Refresh Data Assimilation System (HRRRDAS) using the nearest-in-time 1-hour HRRRDAS forecast provided by NCEP Central Operations. Currently WoFS VLAB can be started four times a day (15, 21, 03, or 09 UTC). For instance, if WoFS VLAB is scheduled to begin data assimilation cycling at 1500 UTC, a 1-h forecast from the 1400 UTC, 36-member, hourly-cycled HRRRDAS analysis provides the ICs for WoFS VLAB. Boundary conditions are from the nearest-in-time 48-h deterministic HRRR forecast (e.g., the 12, 18, 00, or 06 UTC run) where perturbations from the previous GEFS (e.g., 06, 12, 18, or 00 UTC, respectively) are added to that HRRR forecast. The GEFS perturbations are scaled such that the ensemble spread at the lateral boundaries is similar to that provided from 2018-2021 by the experimental HRRR ensemble. Table 9 provides a summary of the model specifications for WoFS VLAB, and Figure 1 shows an example of a SPC Day 1 convective outlook and corresponding WoFS VLAB domain

with WSR-88D radars used for data assimilation overlaid. Further details on WoFS VLAB are included below.

The 36-member WoFS VLAB cycles its data assimilation every 15 minutes by GSI-EnKF assimilation of MRMS radar reflectivity and radial velocity data, cloud water path retrievals and clear-sky radiances from the GOES-19 imager, and Oklahoma Mesonet observations (when available). Conventional (i.e., prepbufr) observations are also assimilated at 15 minutes past each hour. All WoFS VLAB ensemble members use the NSSL 2-moment microphysics parameterization and the RUC land-surface model; however, the PBL and radiation physics options are varied amongst the ensemble members to increase ensemble spread, given the fact that the EnKF may underrepresent model physics errors. 6-h (3-h) forecasts are initialized and launched from the first 18 members from the real-time WoFS VLAB analyses on each hour (half-hour). The first 6-h forecast will be launched 2 hours after the initialization time (17, 23, 05, or 11 UTC). The final forecasts are launched at either 12 or 15 hours after initialization. These forecasts will be viewable using the web-based WoFS VLAB Forecast Viewer (https://cbwofs.vlab.noaa.gov).

Table 9 WoFS VLAB configuration.

	WoFS VLAB
Model Version	WRF-ARW v3.9+
Grid Dimensions	300 x 300 x 50
Grid Spacing	3 km
EnKF cycling	36-mem. w/ GSI-ENKF every 15 min
Observations	- prepbufr conventional observations - Oklahoma Mesonet (when available) - MRMS reflectivity ≥ 15 dBZ; radar 'zeroes'; radial velocity - GOES-16 cloud-water path & clear sky radiances
Radiation LW/SW	Dudhia/RRTM, RRTMG/RRTMG
Microphysics	NSSL 2-moment
PBL	YSU, MYJ, or MYNN
LSM	RUC (Smirnova)

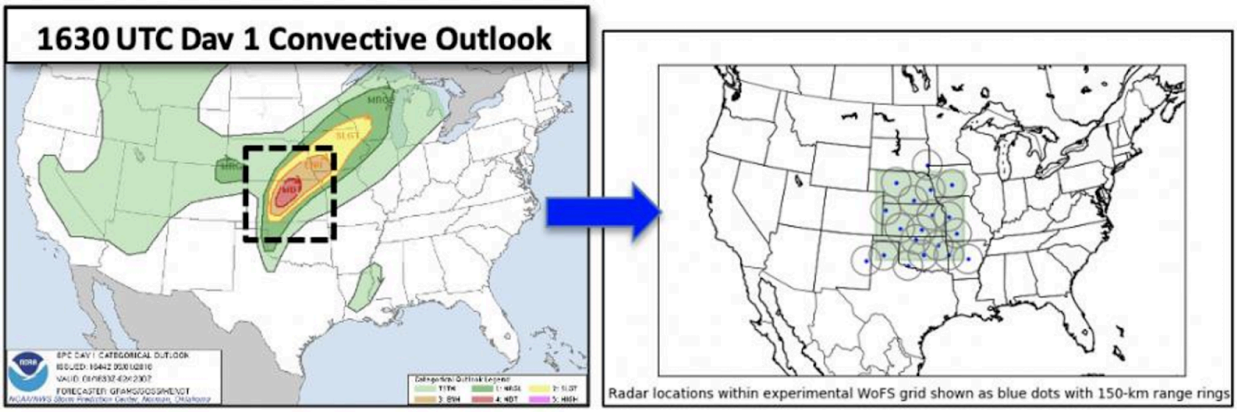


Figure 1 SPC 1630 UTC issued Day 1 convective outlook (left) and corresponding WoFS VLAB grid (right).

ii. WoFS MPAS (NSSL)

A prototype WoFS MPAS system will be evaluated to assess its performance relative to WoFS VLAB, utilizing the same domain locations for direct comparison. The WoFS MPAS configuration is built on the GSL MPAS V8.2.2 codebase, including NSSL additions and fixes, using the Ensemble Adjustment Kalman Filter from NCAR's Data Assimilation Research Testbed. The ensemble consists of 36 members with 3-km horizontal cell spacing and a single-physics setup featuring the MYNN-EDMF planetary boundary layer scheme and NSSL two-moment microphysics. The system is deployed on the Azure cloud platform.

The data assimilation system assimilates MRMS radar reflectivity and radial velocity, GOES-19 cloud water path retrievals, Oklahoma Mesonet observations when available, and conventional prepbufr observations. The ensemble is cycled every 15 minutes from 1500 UTC to 0300 UTC the following day. It is initialized at 1500 UTC using 1-hour forecasts from the 1400 UTC HRRRDAS, while lateral boundary conditions between analyses are generated by adding scaled perturbations from 0600 UTC GEFS forecasts to the 1200 UTC HRRR forecast.

Forecasts are generated using the first 18 ensemble members. Specifically, 6-hour forecasts are launched hourly while 3-hour forecasts are launched on the half-hour. The first 6-hour forecast is launched two hours after the initialization time, such as at 1700 UTC. The final forecasts are launched at either 12 or 15 hours after the initial start time. All output will be viewable using the web-based WoFS MPAS Forecast Viewer at https://cbwofs.nssl.noaa.gov/mpas.

iii. WoFSCast (NSSL)

An improved version of WoFSCast will be evaluated to assess its performance relative to WoFS VLAB. WoFSCast is a Graphcast-based AIWP model that emulates WoFS VLAB forecasts. The current WoFSCast model was trained on real-time WRF-based, WoFS forecasts from Springs 2019–2020. Initial and boundary conditions for WoFSCast will be provided by WoFS VLAB. WoFSCast requires only a minute to produce a 6-h, 18-member forecast on a small number of A100 GPUs. Forecasts will be initialized hourly beginning an hour after the first WoFS VLAB forecast initialization. All output will be viewable using the web-based WoFSCast Forecast Viewer at https://cbwofs.nssl.noaa.gov/wofscast. Note that the WoFSCast web viewer includes fewer products than the other web viewers due to WoFSCast's relatively small set of state variables (U, V, W, 2–5-km UH, θ, Φ, 10-min rainfall, 2-m T, composite reflectivity).

d) Artificial Intelligence Weather Prediction (AIWP) Models

Artificial Intelligence Weather Prediction (AIWP) models have rapidly advanced in the last three years, primarily emerging from the private sector for global weather prediction. This research area is potentially revolutionary, as AIWP models often demonstrate skill measures that exceed those of highly-regarded systems like the

ECMWF's Integrated Forecast System (IFS). Crucially, these AIWP models can produce forecasts in seconds, offering computational resource savings orders of magnitude greater than traditional NWP systems. While initial development focused on the global domain and training using vast, multi-year reanalysis datasets like ERA5, government agencies like NOAA are now increasingly running, training, and testing these algorithms, expanding the focus to include storm-scale applications. The SFE 2026 will expand upon evaluations of global and regional AIWP models, building on the work from the previous two years. The model evaluation activities will feature the first-time testing of GSL's HRRRCast Ensemble, alongside initial examinations of 0–3 hour forecasts from a 3-km grid-spacing AIWP model contributed by Nvidia. Additionally, an improved version of NSSL's WoFSCast will be assessed in next-day evaluations and utilized in real-time during an afternoon forecasting activity. These evaluations will also include several global AIWP models, such as the most recent iteration of Google DeepMind's global ensemble, WeatherNext 2 (https://deepmind.google/science/weathernext), and models produced for Project EAGLE (Experimental AI Global and Limited-area Ensemble forecast system; https://www.epic.noaa.gov/ai/eagle-overview/), which is a joint effort between NOAA OAR, NWS, and EPIC (Earth Prediction Innovation Center; https://epic.noaa.gov/about-epic/). Finally, several global AIWP models provided by Brightband (https://www.brightband.com/) will be examined. Information on each AIWP model that will be evaluated in SFE 2026 is summarized in Table 10 below, followed by more detailed descriptions of each AIWP model.

Table 10 Specifications for AIWP models contributed to SFE 2026.

Algorithm	**Contributor**	**Training data**	**Initialization data**	**Ensemble or deterministic**	**# of mems**	**grid-spacing**	**Forecast length**
GraphCast	Brightband	ERA5	GFS & IFS	Deterministic	1	0.25°	360 h
Aurora	Brightband	ERA5 & GFS	GFS & IFS	Deterministic	1	0.25°	360 h
PanguWx	Brightband	ERA5	GFS & IFS	Deterministic	1	0.25°	360 h
AIFS	ECMWF	ERA5 & IFS analyses	IFS	Deterministic	1	0.25°	360 h
AIFS-Ensemble	ECMWF	ERA5 & IFS analyses	IFS Ens	Ensemble	51	0.25°	360 h
AIGFS	Project Eagle	GDAS	GFS	Deterministic	1	0.25°	384 h
AIGEFS	Project Eagle	GDAS	GEFS	Ensemble	31	0.25°	384 h
Eagle Nest	Project Eagle	GFS & HRRR	GFS & HRRR	Deterministic	1	6-km	360 h
WeatherNext 2	Google DeepMind	ERA5	IFS	Ensemble	65	0.25°	360 h
HRRRCast	GSL	HRRR+GFS	HRRR	Ensemble	9	3-km	18-48 h
Nvidia Stormscope	Nvidia	GOES-19 & MRMS	GOES-19 & MRMS	Deterministic	1	3-km	3 h
WoFSCast	NSSL	WoFS VLAB	WoFS VLAB	Ensemble	18	3-km	6 h

i. GraphCast (Brightband)

GraphCast is a machine-learning algorithm developed by Google DeepMind that is trained directly from ERA5 data (Lam et al. 2022). GraphCast predicts hundreds of weather variables over 10 days using 0.25° resolution and produces forecasts in under one minute. Objective verification found that GraphCast significantly outperformed the IFS on 90% of 1380 verification targets. The code for GraphCast is available publicly at https://github.com/deepmind/graphcast.. Two versions of GraphCast configured by Brightband, one with GFS initial conditions and the other with IFS, will be evaluated during SFE 2026. Specifically, Brightband runs the `GraphCast_operational` model, which resolves outputs on 13 pressure levels; this model was pre-trained on ERA5 data spanning 1979 to 2017, and then was further fine-tuned on HRES forecasts spanning 2016-2021. The model is not initialized with precipitation inputs, but these become available at the first output step (6-hour forecast increment). The GFS-initialized forecasts were not further fine-tuned on either GFS forecasts or GDAS analyses.

ii. Aurora (Brightband)

Aurora is a large-scale foundation model developed by Microsoft for the Earth system trained on millions of hours of diverse data, and can be fine-tuned for diverse applications at only modest computational costs (Bodnar et al. 2024). The code for Aurora is available publicly at https://github.com/microsoft/aurora. Objective verification found that Aurora outperforms the IFS HRES for all lead times up to 10 days. Versions of Aurora configured by Brightband with IFS and GFS initial conditions will be evaluated during SFE 2026. Brightband runs Aurora with the open source weights fine-tuned to target IFS HRES T0. The GFS-initialized forecasts were not further fine-tuned on either GFS forecasts or GDAS analyses.

iii. PanguWx (Brightband)

Pangu-Weather is a deep learning-based system trained using 43 years of ERA5 data and was developed by Huawei Cloud (China) (Bi et al. 2022). The forecasts are produced with 0.25° resolution. At time ranges of 1 h to 1 week, Pangu-Weather was found to outperform the IFS in terms of RMSE and anomaly correlation coefficient (ACC) for fields like geopotential, specific humidity, wind speed, and temperature. Pangu-Weather is applied by designing a 3D Earth Specific Transformer architecture that formulates the pressure level information into cubic data, and applying a hierarchical temporal aggregation algorithm to alleviate cumulative forecast errors. The code is publicly available at https://github.com/198808xc/Pangu-Weather. Two versions of Pangu-Weather configured by Brightband, one with GFS initial conditions and the other with IFS, will be evaluated during SFE 2026. Brightband runs Pangu-Weather with its hierarchical temporal aggregation algorithm. The GFS-initialized forecasts were not further fine-tuned on either GFS forecasts or GDAS analyses

iv. AIFS (ECMWF)

AIFS is a machine learning–based global forecasting system developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Lang et al. 2024). The model is trained on ERA5 reanalysis data together with ECMWF operational NWP analyses. AIFS is built using a graph neural network (GNN) encoder–decoder architecture combined with a sliding-window transformer processor, enabling efficient learning of spatial–temporal relationships in atmospheric data. The system features a flexible and modular design and supports multiple levels of parallelism to allow training at high spatial resolution. AIFS forecast skill has been evaluated against both NWP analyses and direct observations, demonstrating high accuracy for upper-air fields, surface variables, and tropical cyclone track prediction.

v. AIGFS (Project EAGLE)

Project EAGLE's deterministic configuration AIGFS (also called Global-EAGLE-Solo) is a global AI-based forecasting system developed jointly by NOAA OAR, EPIC, and NOAA NWS. The model is initialized from GFS initial conditions and serves as a complement to NOAA's physics-based GFS. The first version is based on GraphCast, which has been retrained using NOAA's Global Data Assimilation System (GDAS) analyses as both inputs and targets. AIGFS produces global forecasts at 0.25° (~28 km) grid-spacing with 13 pressure levels, generating 6-hourly output up to 16 days at twice-daily cycles (00Z and 12Z). Forecast variables include key atmospheric and surface fields such as geopotential height, temperature, wind, humidity, vertical velocity, and precipitation.

Project EAGLE is designed as a near–real-time testing and demonstration framework for AI weather models, enabling evaluation against standard NOAA verification metrics used for operational systems. Forecast products are disseminated through the NOAA Open Data Dissemination Program (NODD), supporting community access and intercomparison with traditional numerical weather prediction systems. More information on Project Eagle can be found at: https://www.epic.noaa.gov/ai/eagle-overview/.

vi. AIGEFS (Project EAGLE)

The ensemble configuration of Project EAGLE, AIGEFS, extends the deterministic system to a multi-member framework designed to quantify forecast uncertainty and improve probabilistic guidance. The ensemble is generated by perturbing the initial conditions derived from the Global Data Assimilation System, producing multiple realizations of the atmospheric state that evolve forward using the same underlying AI model architecture. This approach is analogous to traditional numerical weather prediction ensembles but leverages the efficiency of machine learning to produce large ensembles at relatively low computational cost.

AIGEFS produces global forecasts at 0.25° (~28 km) grid-spacing with multiple members, providing 6-hourly output out to medium-range lead times. Ensemble output supports probabilistic products such as spread, exceedance probabilities, and percentile-based forecasts for key atmospheric variables including geopotential height, temperature, wind, and precipitation. The system is designed to be directly comparable to NOAA's operational ensemble systems, enabling evaluation using standard ensemble verification metrics such as spread–skill relationships and probabilistic skill scores.

vii. Nested-EAGLE (Project EAGLE)

Project EAGLE's nested configuration (Nested-EAGLE) is a global forecasting system featuring a high resolution (6-km) nested domain over CONUS. This framework follows the approach pioneered by Nipen et al., (2024) at Met Norway. Nested-EAGLE was jointly developed by NOAA OAR, EPIC, and NOAA NWS. The model is trained on a decade of 0.25-degree GFS data and 6-km HRRR data (conservatively regridded from its native 3-km resolution). During near-real time inference, it is initialized with both GFS and HRRR initial conditions. Nested-EAGLE was trained using the Anemoi framework from ECMWF and employs a encoder-processor-decoder architecture with a graph encoder and decoder and sliding-window transformer processor. Nested-EAGLE produces forecasts with 6-hourly timesteps out to 15-days and can be initialized 4 times daily. Output resolution is 0.25-degree globally and 6-km over CONUS. Forecast variables include key surface and atmospheric variables at 12 levels (geopotential, u, v, w, t, q, 80m winds, 6h precipitation). More information on Project Eagle can be found at: https://www.epic.noaa.gov/ai/eagle-overview/.

viii. WeatherNext 2

WeatherNext 2 is a machine learning–based global forecasting system developed to extend the capabilities of its predecessor through improved accuracy, stability, and computational efficiency. The model is trained on multi-decadal ERA5 reanalysis dataset fields together with operational analysis data from the ECMWF Integrated Forecasting System (IFS), enabling it to learn consistent atmospheric dynamics across a wide range of spatial and temporal scales. WeatherNext 2 employs a hybrid architecture that combines transformer-based sequence modeling with neural operator components to better capture both long-range temporal dependencies and fine-scale spatial structure (Alet et al. 2025). Compared to the previous version, it demonstrates reduced forecast error growth, improved representation of extreme events, and increased skill in medium-range predictions, particularly for upper-air fields and tropical cyclone tracks. The system is trained at approximately 0.25° horizontal grid spacing with multiple vertical levels and produces forecasts out to 10–15 days with sub-daily temporal resolution. Initialization is performed using IFS operational analyses, ensuring consistency with real-time atmospheric states. Designed with a modular and distributed training framework, WeatherNext 2 supports efficient scaling to high-resolution configurations and rapid inference. More information can be found at:

https://developers.google.com/weathernext/guides/models. A version of the model configured with IFS initial conditions will be evaluated during SFE 2026.

ix. AIFS Ensemble (ECMWF)

AIFS ENS v1 is a machine learning–based probabilistic global forecasting system developed by the ECMWF to extend the deterministic AIFS into an ensemble prediction framework. The system is trained using a multi-stage procedure based on the Continuous Ranked Probability Score (CRPS), beginning with single-step learning on ERA5 reanalysis dataset (1979–2017), followed by progressive rollout training and fine-tuning on operational IFS analyses (2016–2023) to improve temporal consistency and probabilistic calibration. AIFS ENS generates 51 ensemble members (50 perturbed plus one control) by sampling from a learned stochastic distribution, with each member initialized from perturbed initial conditions provided by the IFS ensemble data assimilation system. Forecasts are produced four times daily on a ~0.25° (~31 km) global grid with 6-hourly output to 15 days, enabling efficient representation of forecast uncertainty alongside competitive medium-range skill. Compared to traditional physics-based ensemble systems, AIFS ENS demonstrates improved accuracy for upper-air and surface variables while achieving orders-of-magnitude gains in computational efficiency. The system is designed as a fully data-driven ensemble model in which uncertainty is intrinsically represented, rather than introduced through explicit model perturbations, and it will serve as a complementary component to ECMWF's operational forecasting suite.

x. HRRRCast Ensemble

HRRRCast is a machine-learning based convection-allowing model (CAM) trained to emulate the High Resolution Rapid Refresh (HRRR) at its native 3km resolution. Details about the HRRRCast architecture for version 1 can be found in Abdi et. al (2026). The latest version, HRRRCast v3, is trained on approximately 9 years of HRRR data and outputs a broad set of variables (~164) at 20 pressure levels, which is unlike many other AI models. Most of the variables are treated as prognostic (i.e. are used as input by the model and also predicted as outputs). The HRRRCast 9-member ensemble introduces stochasticity through diffusion-based sampling, generating multiple plausible realizations. Forecasts are currently initialized from the HRRR analyses and produced at hourly cadence for short-range lead times (18h to 48h), targeting convective-scale hazards such as precipitation and reflectivity. HRRRCast v3 is being prepared for integration into NOAA's project EAGLE. HRRRCast is a collaborative effort of several laboratories including GSL, EMC, CIRES, CIRA, NSSL, PSL and WPO. HRRRCast will be evaluated during the 2026 SFE for the first time particularly focusing on composite reflectivity.

HRRRCast code: https://github.com/NOAA-GSL/HRRRCast-live/tree/master
HRRRCast data:

https://noaa-gsl-experimental-pds.s3.amazonaws.com/index.html#HRRRCast
Visualization: https://sites.gsl.noaa.gov/desi
Verification: https://gsl.noaa.gov/mats
HRRRCast v1 paper:
https://journals.ametsoc.org/view/journals/aies/5/2/AIES-D-25-0061.1.xml

xi. Stormscope (Nvidia)

Stormscope is a data-driven weather forecasting model that directly predicts future satellite radiance and radar observations based on a generative diffusion transformer architecture. Stormscope generates ensemble forecasts of geostationary satellite radiance observations and ground radar data at a spatial resolution of 3km and temporal resolution of 10 min over the Continental United States domain. The model grid is identical to the Lambert Conformal Conic grid used by the High-Resolution Rapid Refresh model. The model can be initialized as frequently as GOES-19 data and MRMS composite reflectivity data are updated. Details about the Stormscope architecture can be found in the preprint linked below. The model has been updated since the preprint was published. The updates include increased resolution from 6km to 3km. The forecast fields generated by the model are GOES-19 ABI channels at wavelengths 0.47µm (Blue), 0.64µm (Red), 0.86µm (Veggie), 3.90µm, 6.19µm, 6.95µm, 7.34µm, 10.35µm and MRMS composite and base reflectivity. Forecasts from the nowcast version, which only forecasts observed geostationary and radar fields without any additional NWP model conditioning, will be evaluated. The optimal skill window for the model is 0-3 hours. Experimentation with additional modalities continues and model updates will occur as improvements are discovered. The model inference code and weights are open-source and can be downloaded through earth2studio (link below).

Learning Accurate Storm-Scale Evolution from Observations:
https://arxiv.org/abs/2601.17268
Earth2Studio open model inference code: https://github.com/NVIDIA/earth2studio

e) Calibrated Forecast Products

i. GEFS reforecast MLP (SPC/CSU)

Similar to previous SFEs, the GEFS Machine Learning Probabilities (Hill et al. 2020, 2023; hereafter, GEFS Reforecast MLP) forecasts severe weather hazards through the application of random forests (RFs). The GEFS Reforecast MLP RFs are trained with about 9 years of daily 0000 UTC initializations from the FV3-based Global Ensemble Forecast System reforecast dataset (FV3-GEFS/R) along with severe weather reports. For consistency with SPC outlooks as well as SFE activities, RFs are trained separately for individual hazards in the day 1-3 timeframes, such that separate forecasts are issued for each hazard type. Then, for days 4-7, forecasts are issued for any hazard type.

Predictors from the FV3-GEFS/R correspond to parameters expected to be related to severe weather occurrence, including bulk wind shear, convective available potential energy, low-level wind and thermodynamics, as well as derived quantities like lifting condensation level. To be consistent across variables and times, all predictors are gridded to a 0.5-degree grid for preprocessing. Severe weather reports (i.e., storm data) are similarly gridded over the training period, where each point is labeled a 0, 1, or 2 for the occurrence of no severe report, a severe report, and a significant severe report. For every gridded event of severe weather across the contiguous United States, predictors are selected around the training point with spatiotemporal dimensions to capture any pre-existing dynamical model biases from the FV3-GEFS/R, which allows the RFs to learn predictor biases during training. Spatially, predictors are gathered within a latitudinal and longitudinal radius (set to 3 in these models) around the training point so each grid point represents a separate predictor. Temporally, this procedure is followed at each model output time over the forecast window; the FV3-GEFS/R has 3-hourly output through day 10. For example, during the day-1 period, predictors are gathered 3-hourly from forecast hour 12 through hour 36, totaling nine predictor times. The predictor assembly results in approximately 6,500 predictors for each training point in which to build the RFs.

ii. GEFS operational MLP (SPC/NSSL)

NSSL has formulated a similar RF model using archived operational GEFS data that provides probabilities of any severe weather at lead times of 1 to 15 days (Clark et al. 2025). The global GEFS fields are subset over a 0.5° by 0.5° grid covering the CONUS. In addition to the variables available within the GEFS forecasts, the pressure level data are used to derive additional diagnostics and indices commonly used for severe weather forecasting like bulk shear and the significant tornado parameter. After domain subsetting, the fields are interpolated to a coarser, 81-km grid that tightly encompasses the CONUS. After interpolation, only a set of 1,385 masked points covering CONUS land areas are used to train the RF. The fields are extracted at 3-hourly intervals from forecast hours 12 to 225 (Days 1-10) and 6-hourly intervals from forecast hours 228 to 372 (Days 11-15). In Table 16, the GEFS fields used for predictors in the new RF algorithm (GEFS Operational MLP, hereafter) are listed, with the fields that required additional post-processing marked with an asterisk.

There are several notable differences between GEFS Operational MLP and GEFS Reforecast MLP. First, GEFS Reforecast MLP uses 12 different variables as predictors and no additional diagnostics are computed from the pressure level data (aside from bulk vertical wind shear), while GEFS-ops RF uses 18 predictors and does include severe weather diagnostics computed from pressure level data. Both algorithms use the GEFS output at 3-hourly intervals, but GEFS Reforecast MLP uses 9 times per day: 12, 15, 18, 21, 00, 03, 06, 09, and 12 UTC, while GEFS Operational MLP uses 8 times: 12, 15, 18, 21, 00, 03, 06, and 09 UTC. Second, GEFS Reforecast MLP uses a higher resolution grid of about 55-km (0.5° x 0.5°), while GEFS-ops RF uses the 81-km NCEP 211 grid. Additionally, GEFS Reforecast MLP uses predictors at the grid-point

being considered, as well as all points within a 7 x 7 point box surrounding the point. GEFS Operational MLP only uses predictors at the grid-point being considered. This means that for each point, GEFS Reforecast MLP uses: 49 surrounding points x 12 fields x 9 output times = 5292 predictors, while GEFS Operational MLP uses 1 point x 18 fields x 8 output times + 1 latitude coordinate + 1 longitude coordinate = 146 predictors for Days 1-10, and 74 predictors for Days 11-15 (since those lead times only contain 6-hourly GEFS output resulting in 4 output times per day). Third, for training, GEFS Operational MLP uses the ensemble mean of all 31 GEFS members, while GEFS Reforecast MLP is trained on the ensemble median of 5 GEFS reforecast members. Fourth, GEFS Reforecast MLP performs training over 4 distinct regions of the CONUS and stitches them together for a CONUS-wide forecast, while GEFS Operational MLP trains over the entire CONUS. Finally, for forecast input, GEFS Reforecast MLP uses the median of the first 21 GEFS members, while GEFS Operational MLP uses the mean of all 31 GEFS members.

iii. NSF NCAR AIWP

Day 1–8 AI-NWP-based convective hazard forecasts are generated based on output from two 52-member ensembles generated with two different AI NWP emulators (PanguWeather and FengWu) and run on NSF NCAR supercomputing resources. This 104-member multi-model ensemble is initialized daily at 00 UTC using the 51-member 00 UTC ECMWF ensemble, plus an additional member initialized with the ECMWF HRES. Hazard forecasts are then generated for each member with a decoder-only transformer, with forecasts averaged over the ensemble and used for evaluation. Each transformer, one for each model, was trained to take Pangu/FenguWu output and generate probabilities for the occurrence of ≥ 1 convective weather report (tornado, hail, or wind) within 24-hr and 40-km of a point. The transformers were trained with 5 years of 10-day deterministic forecasts from 2018–2023 initialized with the 00 UTC ECMWF HRES. These forecasts are called “AIWP Mean IFS” for the next-day evaluation activity. Hazard forecasts are also being generated with the Weathernext2 64-member ensemble mean forecast, produced in real-time by Google DeepMind. Post-processing for hazard prediction is done at NSF NCAR using the same transformer-based approach described above, training on 2022–2024 WeatherNext2 mean forecasts. These forecasts are called “WxNext2 Mean” for the next-day evaluations. Real-time forecasts using the Pangu, FengWu, and WeatherNext2, and GEFS forecasts are available here: https://www2.mmm.ucar.edu/projects/ncar_ensemble/ainwp/. More details about the specific decoder-only transformer architecture used for these forecasts can be found in Hua et al. (2026).

iv. OU AIWP

The RF architecture of the FV3-GEFS/R is replicated with inputs from multiple AIWP models, including GraphCast, PanguWeather, and FourCastNet. RF models are trained to probabilistically predict individual daily weather hazards from 1–8 days (White

and Hill 2026). Models are trained separately for each day, hazard (i.e., tornado, wind, and hail), and AIWP model. The inputs to the RFs correspond to environmental variables available from 0000 UTC GFS-initialized AIWP model output related to severe weather forecasting, and the targets are severe weather reports. Because of how the AIWP models are trained, namely with ERA5 data through 2021, the RF training period is limited to 2022-2023. To make real-time forecasts with the RFs, GFS and IFS initial conditions are used to generate AIWP to feed the RF predictions. Forecasts are generated from both 0000 and 1200 UTC-initialized AIWP models at all lead times. To further exploit the RF-based architecture, the AIWP inputs are assembled in a "small ensemble" and the median of three AIWP outputs are used to train another set of RFs, referred to as "combined" RFs. This approach mimics that of the FV3-GEFS/R and minimizes the model-specific input biases from individually trained AIWP RFs. Real-time forecasts are available here: https://weather.ou.edu/~whitee/aimlp.html. Additional details about the RFs, the model training process, and performance characteristics can be found in White and Hill (2026).

v. SUPER-EC

The Severe U-Net Prediction Engine Realtime - ECMWF (SUPER-EC) is a neural network-based deep-learning model that forecasts total severe coverage out to Day 10 using the operational ECMWF as input to the model. The model was trained on 15 years of ERA5 Reanalysis data and Practically Perfect storm reports from 2008–2023 and applies that training to the deterministic ECMWF forecast in a perfect-prognosis approach. The model runs twice daily (00z and 12z cycles) and includes a 3-day mean time-lagged ensemble (TLE) for the 00z run. This evaluation will focus on the TLE version of SUPER-EC.

vi. StormNet

StormNet is a multi-timescale, multi-modal severe weather prediction model developed by OpenSnow. StormNet outputs CONUS probability fields 4x per day for tornadoes, large hail, damaging wind, and lightning. For the SFE 2026, StormNet's long-range tornado, hail, and wind probability fields will be combined into a 1x daily 'total severe' probability field, initialized at 0000 UTC. StormNet's long-range component (forecasts for days 3 through 14) uses the AIFS-Ensemble as the NWP baseline. StormNet makes bias and resolution adjustments to the AIFS-ENS members, then post-processes the adjusted members to generate a 50-member probability field. The final output is the mean probability of all 50 members. The model architecture is a custom deep-learning architecture built specifically for severe weather prediction, taking inspiration from transformers, graph neural networks, and convolutional neural networks. The labels used in training StormNet come from the NCEI Storm Events Database, masked to 25 miles per event.

vii. REFS/GEFS & HREF/GEFS

Calibrated probabilities for tornadoes, severe hail, and damaging winds valid over a 24-h time window corresponding to a convective day (i.e., 1200 – 1200 UTC) are produced for two CAM ensembles (i.e., HREF and SPC REFS) using the following procedure. At every grid-point for the valid forecast hour, two probabilities are paired:

1. Maximum neighborhood probability of HREF or SPC REFS storm attribute variable(s). For all three convective hazards, UH ≥ 99.85th percentile for each member is used over all 4-h periods valid within the previous 24-h period. In addition, severe wind probabilities are composed of two additional storm attributes, including the calibrated thunder probabilities and the neighborhood probability of 10- m AGL Wind Speeds ≥ 30 kt, which are masked by the aforementioned UH probabilities exceeding 5%.

2. Maximum probability of GEFS environmental field(s) meeting a threshold over 3-h periods valid within the previous 24-h period. These fields include the Significant Tornado Parameter (STP), Most-Unstable CAPE (MUCAPE), and Effective Bulk Shear.

The historical frequency of a hazard report occurring within 25 miles of that grid point and within the 24-h period for that forecast pair of probabilities is substituted as the 24-h calibrated hazard probability. The HREF/GEFS calibrated severe guidance is the current operational standard. The REFS/GEFS version evaluated here has a much more limited training sample (April-October 2025) given the limited REFS archive.

iix. Loken HREF & Vancil REFS

Automated “first-guess” Day 1 (12-36 h lead times; 1200 UTC – 1200 UTC) and Day 2 (24-48 h lead times; 1200 UTC – 1200 UTC) hazard probabilities are created from random forests (RFs) trained on HREF or SPC REFS fields and observed SPC storm reports. Separate RFs predict the probability of tornadoes, severe hail, and severe wind. These predictions are made on an 80-km CONUS grid, but are bilinearly interpolated to the native HREF grid for the SFE. The tornado-, severe hail-, and severe wind-predicting RFs use the same set of predictors for a given lead time. These predictors are derived from simulated HREFv3 variables. The procedure for creating the predictors is as follows:

1. Aggregate HREF or SPC REFS fields in time by computing a maximum or minimum over the forecast period.
2. Upscale the temporally-aggregated fields to an 80 km grid.
3. Compute the ensemble mean for each field on the 80 km grid.
4. Spatially smooth the daily maximum 2-5 km updraft helicity (UH2-5km) forecasts from each HREF or SPC REFS member using a 2-dimensional Gaussian kernel density function with = 90 km.
5. The final predictors include ensemble mean temporally-aggregated fields, spatially-smoothed individual-member UH2-5km, and 0-1 km relative vorticity

from the two HRRR members. Predictors are taken from the point of prediction as well as the closest eight additional 80 km grid points. Latitude and longitude at the point of prediction are also included.

The overall procedure for creating RFs is similar to that described in Loken et al. (2020). Key differences from past SFEs include the removal of most raw environmental field predictors (e.g., 2-m temperature) and the addition of: period-minimum storm and index predictors, individual-member smoothed UH2-5km, skewness of UH2-5km at the time of maximum (or minimum) UH2-5km, and 0-1km relative vorticity from the HRRR members.

ix. Nadocast HREF & Nadocast REFS

Nadocast is a machine-learning system, initially focused on tornadoes, that aims to produce timely, calibrated, severe weather probabilities on the Day 1 time scale (2-35 hours). Probabilities are generated by gradient-boosted decision trees trained on 10,000+ storm and storm-adjacent hours of HREF outputs. Nadocast performs extensive feature engineering: each grid point from the HREF hourly output is supplemented by adding spatial blurs of various radii, spatial gradients, parameters from 1 h future and 1 h past, summary statistics over a 3-h window, and additional information such as climatology and an estimate of recent convective forcing. To provide rotational invariance, winds at each grid point are rotated relative to an estimate of the 500m-5000m shear vector. The result is over 10,000 features per grid point per hour, upon which the decision trees operate to produce hourly probabilities. To capture uncertainty at longer lead times, different models are trained for short- (2-13hr), medium- (13-24hr), and longer-range (24-35hr) forecasts. Hourly probabilities are pooled into day-long guidance on a 15km grid and rescaled to follow the historical characteristics of SPC thresholds. A preliminary objective comparison (n=260 days) suggests tornado-prediction performance that, on average, matches or slightly exceeds SPC 6Z Day 1 guidance.

x. STP Cal Inflow for HREF and REFS

Calibrated tornado probabilities over a 24-hr period (1200 – 1200 UTC) are produced based on a correlation of STP value to tornado frequency. For a model grid point with a right-moving, rotating storm (identified such that UH exceeds the 99.985th percentile of the respective HREF or REFS member's climatology), the storm environment is depicted using the 50th percentile of a distribution of STP values that is populated from points within the storm's inflow region. The inflow area is defined as a quadrant region of 40-km radius that is centrally oriented along the direction of the environmental wind at 1 km AGL. The maximum STP value for each point over the 24-hour period is used to assign a probability based on the climatological frequency of tornado occurrence correlated to STP for right-moving supercells. Grid point tornado probabilities are calculated separately for each ensemble member and a grid point mean

is taken across the (HREF or REFS) ensemble. Finally, a Gaussian smoother is applied with σ = 50 km.

3. SFE 2026 Core Interests and Daily Activities

2026 SFE activities will occur from 9am-4pm CDT on Mondays, 8:30am-4pm CDT Tuesday-Thursday, and 8:30am-12pm CDT Fridays. Tuesday-Friday there is an optional 8-8:30am period for map analysis, data loading, and networking. Each day will have a lunch break from 12:30-2pm CDT. On Wednesdays there will be an optional science discussion from 1:15-2pm. Tables 11-13 provide a daily schedule for Monday, Tuesday-Thursday, and Friday, respectively. Further details are provided in subsequent sections.

Table 11 Monday schedule.

Time (CDT)	
9:00 AM – 9:45 AM	**Welcome and Introductions** *Hybrid All* (Israel Jirak & Participants)
9:45 AM – 10:30 AM	**HWT SFE Scientific Objectives and Goals** *Hybrid All* (Israel Jirak & Adam Clark)
10:30 AM – 10:45 AM	***Break*** Fill out IRB Consent Form, Program CACs
10:45 AM – 11:00 AM	**Conditional Intensity Forecasting Overview** *Hybrid All* (Israel Jirak)
11:00 AM – 11:15 AM	**Weather Briefing** *Hybrid All* (David Imy)
11:15 AM – 12:30 PM	**Group Forecasting Activity** (Coverage and Conditional Intensity Outlooks) *In-Person R2O* (Day 1); *In-Person Innovation* (Days 3 & 4); *Virtual* (Day 2)
12:30 PM – 2:00 PM	***Lunch/Break***
2:00 PM – 2:15 PM	**Update on Today's Weather** *Hybrid All* (SPC Forecaster/Israel Jirak)
2:15 PM – 3:15 PM	**Individual Forecasting Activity** (0-4h Outlooks and Training) *IP R2O* (0-4h;20-00Z); *IP Inno* (WoFS MPAS); *Virtual 1* (WoFSCast); *Virtual 2* (WoFS VLAB)
3:15 PM – 4:00 PM	**Individual Forecasting Activity Continued** (0-4h Updates and MDs) *IP R2O* (0-4h;21-01Z); *IP Inno* (WoFS MPAS); *Virtual 1* (WoFSCast); *Virtual 2* (WoFS VLAB)

Table 12 Tuesday - Thursday schedule.

Time (CDT)	
8:00 AM – 8:30 AM	*(Optional) Map Analysis, Data Loading, and Networking* *In-Person (Optional)*
8:30 AM – 9:00 AM	**Overview of Yesterday's Severe Weather** *Hybrid All* (David Imy)
9:00 AM – 10:30 AM	**Model & Outlook Evaluation** (Orientation, Surveys, and Discussion) *Hybrid Groups* (Group 1; Group 2; Group 3)
10:30 AM – 10:45 AM	***Break***
10:45 AM – 11:00 AM	***Evaluation Highlights*** *Hybrid All* (Group 1; Group 2; Group 3)
11:00 AM – 11:15 AM	**Weather Briefing** *Hybrid All* (David Imy)
11:15 AM – 12:30 PM	**Group Forecasting Activity** (Coverage and Conditional Intensity Outlooks) *In-Person R2O* (Day 1); *In-Person Innovation* (Days 3 & 4); *Virtual* (Day 2)
12:30 PM – 2:00 PM	***Lunch/Break*** ***Science Discussion (Wednesdays @ 1:15)***
2:00 PM – 2:15 PM	**Update on Today's Weather** *Hybrid All* (SPC Forecaster/Israel Jirak)
2:15 PM – 3:15 PM	**Individual Forecasting Activity** (0-4h Outlooks and MDs) *IP R2O* (0-4h;20-00Z); *IP Inno* (WoFS MPAS); *Virtual 1* (WoFSCast); *Virtual 2* (WoFS VLAB)
3:15 PM – 4:00 PM	**Individual Forecasting Activity Continued** (0-4h Updates and MDs) *IP R2O* (0-4h;21-01Z); *IP Inno* (WoFS MPAS); *Virtual 1* (WoFSCast); *Virtual 2* (WoFS VLAB)

Table 13 Friday schedule

Time (CDT)	
8:00 AM – 8:30 AM	*(Optional) Map Analysis, Data Loading, and Networking* *In-Person (Optional)*
8:30 AM – 9:00 AM	**Overview of Yesterday's Severe Weather** *Hybrid All* (David Imy)
9:00 AM – 10:30 AM	**Model & Outlook Evaluation** (Orientation, Surveys, and Discussion) *Hybrid Groups* (Group 1; Group 2; Group 3)
10:30 AM – 10:45 AM	***Break***
10:45 AM – 11:00 AM	***Evaluation Highlights*** *Hybrid All* (Group 1; Group 2; Group 3)
11:00 AM – 12:00 PM	**Weekly Wrap-up and Discussion** *Hybrid All* (Israel Jirak)

a) Formal Evaluation Activities

SFE 2025 will feature one period of formal evaluation from 9-11:00am CDT Tuesday-Friday. The evaluations will be done in three hybrid groups (i.e., each group will have in-person and virtual participants) and involve comparisons of different ensemble diagnostics, CLUE ensemble subsets, and other products and guidance.

Participants will be split into Groups 1, 2, & 3, which will each conduct a separate set of evaluations. In each group, for each set of evaluations, a short tutorial will be presented and then participants will conduct the evaluations independently while facilitators remain available for questions. Following each set of evaluations, there will be a short discussion period during which participants can discuss noteworthy aspects of the evaluations, evaluation philosophy, questions, or any other topics related to the evaluations. The evaluations will end at 10:30am, followed by a 15-minute break, and from 10:45-11:00am each evaluation group will have 5 minutes to discuss highlights from their group with all participants. The evaluations are categorized as "CAM (E)nsembles", "(D)eterministic CAMs", "(A)nalyses", "(C)alibrated Guidance", "(O)utlooks", and "(A)rtificial (I)ntelligence". The letter in parentheses combined with a number is used to label the individual evaluations in each category (e.g., E1 refers to the first CAM Ensemble evaluation). Each evaluation group will conduct a mix of evaluations from each category. On Fridays, there will be a weekly wrap-up discussion, including aggregate objective-verification statistics. The evaluations in each category are summarized below:

(C)alibrated Guidance

C1. Medium-Range Severe ML Guidance

Eight different sets of extended range, ML-derived total severe probabilities for Day 3-7 lead times are subjectively rated. These methods include: (1) GEFS Reforecast MLP, (2) GEFS Operational MLP, (3) AIWP Mean IFS, (4) AIWP Median GFS, (5) AIWP Median IFS, (6) WxNext2 Mean, (7) SUPER-EC TLE, and (8) StormNet.

C2. Day 1 00Z Tornado ML Guidance

Two versions of four different methods for deriving Day 1 tornado guidance are subjectively rated. One version of each method is calibrated with HREF and the other with REFS. These methods include: Nadocast, GEFS, Loken/Vancil, and STP Calibrated Inflow.

C3. Day 1 00Z Hail ML Guidance

Two versions of three different methods for deriving Day 1 hail guidance are subjectively rated. One version of each method is calibrated with HREF and the other with REFS. These methods include: Nadocast, GEFS, and Loken/Vancil.

C4. Day 1 00Z Wind ML Guidance

Two versions of three different methods for deriving Day 1 wind guidance are subjectively rated. One version of each method is calibrated with HREF and the other with REFS. These methods include: Nadocast, GEFS, and Loken/Vancil.

Primary Science Question(s): What are the strengths and weaknesses of the various calibrated hazard guidance, and what are the best approaches and techniques to develop calibrated hazard probabilities? Can reasonable calibrated Day 1 hazard guidance be generated from a limited training sample of REFS data?

CAM (E)nsembles

E1. CLUE: Day 1 Ensemble Flagships

This evaluation will feature in-depth examinations of severe storm-attribute and environmental fields from 0000-UTC initializations of HREF, REFS, SPC REFS, and SSEOv2 for Day 1 lead times. These comparisons will serve to unearth ways in which the currently operational CAM ensemble (i.e., HREF) differs from the candidate to replace it (i.e., REFS), and whether the REFS improves upon or degrades forecasts of the HREF for fields relevant to forecasting severe weather. Additionally, the SSEOv2 will be examined to assess whether it provides superior performance relative to REFS, and could thus be used while the MPAS-based REFSv2 is being developed.

Primary Science Question(s): How do probabilistic forecasts of REFS compare to those of the HREF at Day 1 lead times (e.g., spread and skill)? Are there systematic shortcomings or advantages of REFS? Does the SPC REFS improve upon the proposed full 14-member REFS? Does an ensemble with three sources of model diversity, SSEOv2, provide an advantage relative to HREF or REFS configurations?

E2. CLUE: Day 2 Ensemble Flagships

This evaluation is similar to the Day 1 Ensemble Flagships but for Day 2 lead times and 1200 UTC initializations.

Primary Science Question(s): How do probabilistic forecasts of REFS compare to those of the HREF at Day 2 lead times (e.g., spread and skill)? Are there systematic shortcomings or advantages of REFS? Does the SPC REFS improve upon the proposed full 14-member REFS? Does an ensemble with three sources of model diversity, SSEOv2, provide an advantage relative to HREF or REFS configurations?

E3. CLUE: Short-Term Ensembles

To gauge the performance of WoFS VLAB and measure performance relative to operationally available hourly systems over the CONUS, WoFS VLAB forecasts initialized at 2100 and 0000 UTC will be compared to corresponding HRRR/RRFS time-lagged (HRRR/RRFS-TL) forecasts.

Primary Science Question(s): How does WoFS VLAB performance compare to short-term ensembles composed of operational hourly CAMs?

(D)eterministic CAMs

D1. CLUE: 00Z Day 1 Deterministic Flagships

This activity will focus on rating the primary deterministic CAMs provided by several SFE collaborators – NSSL (NSSL MPAS HT), EMC (RRFS), NASA (NASA FV3), and GSL (GSL MPAS RT) – based on their skill and utility for severe weather forecasting at Day 1 lead times. These runs will be compared to the operational HRRR, which was developed by GSL. Particular attention will be given to simulated storm structure, convective evolution, and location/coverage of storms. Storm surrogate fields, like hourly maximum updraft helicity, will also be examined to gauge their utility for forecasting severe storms.

Primary Science Question(s): How do various deterministic CAMs compare to the operational standard for convective forecasting (i.e., HRRRv4) at Day 1 lead times?

D2. CLUE: 12Z Day 2 Deterministic Flagships

The same models from D1 are evaluated but for 1200 UTC initializations and Day 2 lead times.

Primary Science Question(s): How do various deterministic CAMs compare to the operational standard for convective forecasting (i.e., HRRRv4) at Day 2 lead times?

D3. CLUE: 00Z Day 3 Deterministic Flagships

The following 0000 UTC initialized models are evaluated for Day 3 lead times: RRFS, GSL MPAS RT, NSSL MPAS RN, NASA FV3, and EAGLE Nest.

Primary Science Questions(s): What strategies for CAM configurations perform the best at Day 3 lead times, and what are their forecast characteristics for severe weather forecasting applications?

D4. CLUE: Day 1 MPAS ICs

Pairs of similarly configured MPAS configurations initialized at 0000 and 1200 UTC are evaluated for Day 1 lead times. GSL MPAS RT uses RRFS ICs and LBCs, while GSL MPAS JEDI uses its own fully-cycled, JEDI-based data assimilation system.

Primary Science Question(s): What are the impacts of using the next-generation, JEDI-based data assimilation system with MPAS for convective forecasting applications?

D5. CLUE: Data Assimilation Comparisons

The HRRR, RRFS, and GSL MPAS JEDI are examined in the first six (6) hours of the forecast period for 0000 UTC initializations to evaluate the impact of their data assimilation.

Primary Science Question(s): How do the data assimilation strategies in HRRR, RRFS, and GSL MPAS JEDI impact short term convective weather forecasts?

(A)nalyses

A1. Mesoscale Analysis Background

Hourly versions of the 3D-RTMA using HRRR as the background (3D-RTMA HRRR) will be compared to a 3-km grid-spacing version of the sfcOA that uses HRRR forecasts as the background (sfcOA HRRR) and applies a simple 2-pass Barnes objective analysis to incorporate the latest surface observations. The goal is to assess the utility of these analysis systems for situational awareness and short-term forecasting for convective-weather scenarios. Analyses from WoFS VLAB and WoFS MPAS will also be incorporated into these comparisons.

Primary Science Question(s): What are the optimal methods for producing quality mesoscale analyses for convective forecasting applications? What are differences in analysis quality between WoFS VLAB and WoFS MPAS?

A2. Upscaled Mesoscale Analysis Background

3D-RTMA HRRR and sfcOA HRRR will be upscaled to a 40-km grid and compared to the 40-km grid-spacing sfcOA that uses the RAP as the background.

Primary Science Question(s): What are the optimal methods for producing quality mesoscale analyses for convective forecasting applications? Is the 40-km upscaled version sufficient for SPC operations? What about WFO operations?

A3. Storm Scale Analysis

WoFS VLAB-based “analyses” (actually 15-minute maximum forecasts) of 80-m wind are compared to preliminary local storm reports, including gust measurements and estimates. Additionally, similar WoFS VLAB-based, 15-minute maximum 2-5 km AGL UH and updraft speed are compared to MRMS Mid-Level Rotation Tracks (MLRT) and MRMS MESH, respectively.

Primary Science Question(s): Can a high resolution, rapidly updating ensemble DA system serve as a verification source for severe winds, mesocyclone tracks, and hail?

(A)rtificial (I)ntelligence Evaluations

AI1. Global AIWP Models: GFS ICs

The AIWP models GraphCast, Pangu-Weather, and Aurora from Brightband, and operational AIGFS starting with GFS initial conditions will be assessed, compared, and subjectively rated alongside the GFS. The target lead time will be 7 days (i.e. forecast hours 156-180) and all 5 times (12, 18, 00, 06, and 12Z) that fall within the convective day will be considered. Participants will primarily consider the 500-mb height-wind patterns, but will use other available fields (e.g., 850 mb heights/winds, 2-m temperatures, etc.) to supplement their ratings. Finally, the Day 7 QPFs in the NWP emulators that have QPF available will also be subjectively rated alongside the GFS.

Primary Science Question(s): How do forecasts from AIWP models initialized from the GFS compare to traditional NWP forecasts? Is there value in the AIWP models for extended range severe weather forecasting applications?

AI2. Global AIWP Models: EC ICs

Following the same format as AI1, The AIWP models GraphCast, Aurora, and Pangu-Weather from Brightband, and AIFS from ECMWF starting with EC initial conditions will be assessed, compared, and subjective rated alongside the ECMWF.

Primary Science Question(s): How do forecasts from AIWP models initialized from the EC compare to traditional NWP forecasts? Is there value in the AIWP models for extended range severe weather forecasting applications?

AI3. Global AIWP Models: IC Comparison

GraphCast versions with both GFS and EC initial conditions are compared to GFS and EC analyses, respectively, at the target lead time of 7 days.

Primary Science Question(s): How sensitive are the NWP emulators to initial conditions, and do either of the ICs result in better performance?

AI4. Global AIWP Ensembles

The global AIWP ensembles WeatherNext 2, operational AIGEFS, and AIFS ensemble from Brightband are compared to traditional NWP ensembles from the GFS and IFS, including ensemble mean and spread fields of 500 mb height and 2-m temperature.

Primary Science Question(s): How do global AIWP ensembles compare to traditional NWP ensembles in terms of the mean pattern and overall spread?

AI5. WoFS Deterministic Comparisons

This evaluation examines the quality of deterministic reflectivity forecasts from 0000 UTC initializations of WoFS VLAB, MPAS WoFS and WoFSCast.

Primary Science Question(s): How do storm-scale AIWP forecasts from WoFSCast, which are designed to emulate WoFS VLAB, compare to WoFS VLAB itself? How does WoFS MPAS compare to WoFS VLAB?

AI6. WoFS Ensemble Comparisons

This evaluation will examine the quality of ensemble reflectivity forecasts (neighborhood probability of 40 dBZ exceedance) from 0000 UTC initializations of WoFS VLAB, WoFS MPAS, and WoFSCast.

Primary Science Question(s): How do storm-scale ensemble AIWP forecasts from WoFSCast, which are designed to emulate WoFS VLAB, compare to WoFS VLAB itself? How does MPAS WoFS compare to VLAB WoFS?

AI7. 00Z Day 1 HRRRCast vs. HRRR & REFS

This evaluation will examine the quality of deterministic and probabilistic reflectivity forecasts from 0000 UTC initializations of HRRRCast, and compare them with the HRRR and REFS, respectively.

Primary Science Question(s): What are the performance characteristics of deterministic and ensemble forecasts from HRRRCast, and how do they compare with operational baselines?

AI8. Short-Term AIWP

Deterministic and ensemble simulated composite reflectivity forecasts from 0000 UTC initializations of WoFSCast and NVIDIA Stormscope are compared to MRMS at 10-minute intervals at 0-3 h lead times.

Primary Science Question(s): What are the performance characteristics for short-term severe weather forecasting applications for NVIDIA Stormscope at 0-3 h lead times and how do they compare to WoFSCast?

AI9. MD Peak Intensity Bin Guidance

A suite of random forest (RF) models have been developed using sfcOA and RAP inputs to provide an objective first guess on peak-intensity bins (PIBs) in SPC MCDs. These models are trained independently from each other and use a selected set of parameter field statistics collected from the area and vicinity of the MCD. These RF models output a deterministic value of the given hazard intensity which is then converted into a PIB based on the distribution of tree predictions in the forest. This suite of models will be compared and evaluated against the current Outlook-based first-guess implementation of PIB predictions as well as the SPC forecaster's prediction.

Primary Science Question(s): How well does the RF first-guess PIB compare to the current Outlook-based first guess and the official forecast? Are there scenarios when the RF models excel? struggle?

(O)utlook Evaluations

O1. Day 1/2/3/4 Outlooks

The experimental Day 1-3 outlooks for tornado, wind, and hail, and Day 4 outlook for total severe produced by SFE teams are subjectively rated and compared.

Primary Science Question(s): How does the skill for tornado, hail, and wind severe outlooks vary with increasing lead time? How skillful are the Day 4 total severe outlooks and was CAM guidance useful at this lead time?

O2. 4-h Day 0 Outlooks

The 4-h Day 0 outlooks for tornado, wind, and hail valid 20-00Z and 21-01Z are compared across three groups: an operational forecaster with no data restrictions, a consensus of participants focused on WoFS VLAB output, and a consensus of participants focused on RRFS/HRRR-TL output.

Primary Science Question(s): How skillful are current CAM ensembles at identifying higher resolution probabilistic corridors of significant severe weather in the 0-4 h timeframe? What role should the human forecaster play in the process?

O3. SPC Impacts System: Day 1 Outlook Tornado, Hail, and Wind Counts

SPC maintains an internal analytics system for predicting the number of tornadoes expected on a particular forecast day, given human-generated coverage and intensity forecasts as initial inputs. Recently, the SPC has added internal prediction of final hail/wind local storm report counts to complement these tornado predictions. Each of these predictions (tornado, hail, and wind) are produced for each Day 1 convective

outlook (other than the 0100 UTC issuance) only if there are non-zero coverage probabilities for the respective hazard. All three hazard predictions are based on the historical nature of the frequency of these hazards relative to SPC's convective outlooks. Historical diversity in outcomes (collected from validation data) provide the background distributions to add uncertainty to these forecasts.

Primary Science Question(s): Would this information be helpful in communicating the potential severe weather impacts on a given day? What is the best way to visualize this information? Is the statistical output consistent and aligned with my expectation of potential severe weather for the day?

b) Forecast Products and Activities

There will be two periods of experimental forecast activities during SFE 2026. The first will occur from 11:00am – 12:30pm CDT and will focus on generating probabilistic outlooks for individual hazards, as well as more precise information on the intensity of specific hazards. Participants will be split into three groups: (1) In-Person R2O, (2) In-Person Innovation, and (3) Virtual. As the naming convention suggests, in-person participants will be split into the R2O and Innovation groups, while remote participants will be in the Virtual group. The In-Person R2O group will issue products for Day 1, the Virtual group will issue products for Day 2, and the In-Person Innovation group will issue products for Days 3 & 4. The experimental forecasts will cover a limited-area domain typically covering the primary severe threat area with a center-point based on existing SPC outlooks and/or where interesting convective forecast challenges are expected. The Day 3 & 4 forecast is the only exception to the smaller domain, and will instead cover a full CONUS domain. Also, the Day 4 outlooks will only cover total severe (i.e., no individual hazards or conditional intensity forecasts).

In all groups, the morning forecasts will be done collectively. The individual hazard forecasts will mimic the SPC operational Day 1 & 2 Convective Outlooks by producing individual probabilistic coverage forecasts of large hail, damaging wind, and tornadoes within 25 miles (40 km) of a point. The Day 1 outlooks will cover the period 1800 UTC to 1200 UTC the next day, while the Days 2, 3, & 4 outlooks will cover 1200 – 1200 UTC periods. Additionally, for experimental outlooks covering Days 1, 2, & 3, conditional intensity forecasts of tornado, wind, and hail will be issued, in which areas are delineated with reports that are expected to follow intensity distributions defined by conditional intensity groups (see more information below). These conditional intensity forecasts are similar to those issued during SFEs 2019-2025, and became operational 3 March 2026. When generating Day 1 Convective Outlooks, SPC forecasters currently draw probabilities that represent the chance of each hazard occurring within 25 miles of a point. Previously, forecasters could also delineate "hatched" areas, which represented regions with a 10% chance or greater of significant severe weather (EF-2 or greater tornadoes, winds ≥ 65 kts, or hail ≥ 2-in.) within 25 miles of a point. Research by the SPC has shown that a conditional-intensity framework can be used to better forecast and communicate the severe-weather risk. These results have been used to identify four

conditional intensity groups (CIG) that can be forecast via examination of the atmospheric environment: CIG 0, CIG 1, CIG 2, and CIG 3. In plain language, CIG 0 refers to a typical severe weather day, where significant severe weather is unlikely, CIG 1 areas indicate where significant severe weather is possible, CIG 2 areas indicate where high impact significant severe weather is expected, and CIG 3 represents intensity on historic severe weather days. All groups will have access to all available operational and experimental guidance products for issuing their outlooks.

The second period of experimental forecasting activities will occur during the 2-4pm CDT time period. From 2-2:15pm CDT, a weather briefing led by SPC will be conducted for all participants during which an update on current weather will be given. During the 2:15-3:15pm CDT time period on the R2O Desk, some In-Person participants will generate 4-h probabilistic tornado, hail, and wind outlooks valid 2000 – 0000 UTC using WoFS VLAB and other available observations and CAM guidance within the SFE Drawing Tool while others will focus on RRFS/HRRR-TL output. After a brief discussion about the difference in CAM ensemble guidance, each In-Person participant will update the 4-h probabilistic tornado, hail, and wind outlooks for the period 2100 – 0100 UTC.

Also during the 2:15-4pm CDT time period, virtual and the remaining in-person participants will split into three groups – in-person participants will be in one group, while the virtual participants will be split into two groups, for an activity to evaluate the subjective performance of the three different versions of WoFS. The first group will have access to WoFS VLAB, the second group to WoFS MPAS, and the third group to WoFSCast. The version of WoFS evaluated with the in-person group will rotate each week. All participants will stay in the same group each week to maintain a control group (WoFS VLAB) for forecast performance comparison of the other two versions (WoFS MPAS and WoFSCast) to the control group. On the first day of each week, all participants will complete a short training on all versions of WoFS from 2:15-2:45 PM CDT. Then, from 2:45-3:45 PM CDT, participants will create their own MDs using their specific version of WoFS via the drawing tools in Google Slides. From 3:45-4:00 PM CDT, all participants will complete a survey on their subjective assessment of the system's performance. After the first day, all participants will create their own MDs from 2:15-3:15 PM CDT in their respective groups. This will be followed by completing the survey from 3:15-3:40 PM CDT. The activity will conclude with a combined, large group discussion where all participants (in-person and virtual) will share their MDs and discuss the overall trends they are noticing about each version of WoFS.

These WoF activities are the tenth year WoFS has been tested in the SFE to explore the potential utility of WoFS products for issuing guidance between the watch and warning time scales (i.e. 0.5 to 6-h lead times). These activities explore ways of seamlessly merging probabilistic severe weather outlooks with probabilistic severe weather warnings as part of NOAA's Warn-on-Forecast (WoF; Stensrud et al. 2009) and Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) initiatives. These efforts also support the transition to higher temporal resolution forecasts at the SPC.

Appendix A: Organizational structure of the NOAA/Hazardous Weather Testbed

NOAA's Hazardous Weather Testbed (HWT) is a facility jointly managed by the National Severe Storms Laboratory (NSSL), the Storm Prediction Center (SPC), and the NWS Oklahoma City/Norman Weather Forecast Office (OUN) within the National Weather Center building on the University of Oklahoma South Research Campus. The HWT is designed to accelerate the transition of promising new meteorological insights and technologies into advances in forecasting and warning for hazardous mesoscale weather events throughout the United States. The HWT facilities are situated between the operations rooms of the SPC and OUN. The proximity to operational facilities, and access to data and workstations replicating those used operationally within the SPC, creates a unique environment supporting collaboration between researchers and operational forecasters on topics of mutual interest.

The HWT organizational structure was modified in 2025 to reflect the evolution of activities within the HWT over the last decade (Fig. A1). Formerly, the Experimental Forecast Program (EFP) had focused on predicting hazardous mesoscale weather events on time scales ranging from hours to a week in advance, and on spatial domains ranging from several counties to the CONUS, whereas the Experimental Warning Program (EWP) had concerned with detecting and predicting mesoscale and smaller weather hazards on time scales of minutes to a few hours, and on spatial domains from several counties to fractions of counties. With the growth of activities linked to the Warn-on-Forecast System (WoFS) program (e.g.; Heinselman et al. 2024) and Probabilistic Hazard Information (PHI) project (e.g.; Karstens et al. 2018), activities have increasingly focused on goals outside of the traditional "forecast" and "warning" timescales. Furthermore, there has been a recent growth in the number of HWT activities focused on the forecast scale but run separately from the SFE, which had been the sole constituent of the formal EFP. To better reflect the modern operations of the HWT, the formal EFP and EWP programs were discontinued and the HWT structure was modified to an activity-based structure. The HWT Executive Officer (XO) is responsible for coordinating with each activity to ensure the successful execution of HWT goals across the spectrum of activities and science foci represented. The XO also coordinates with the executive leadership of the HWT, inclusive of the Board of Directors and Operational Managers, to facilitate long-term strategic planning and address any challenges that may arise within the HWT.

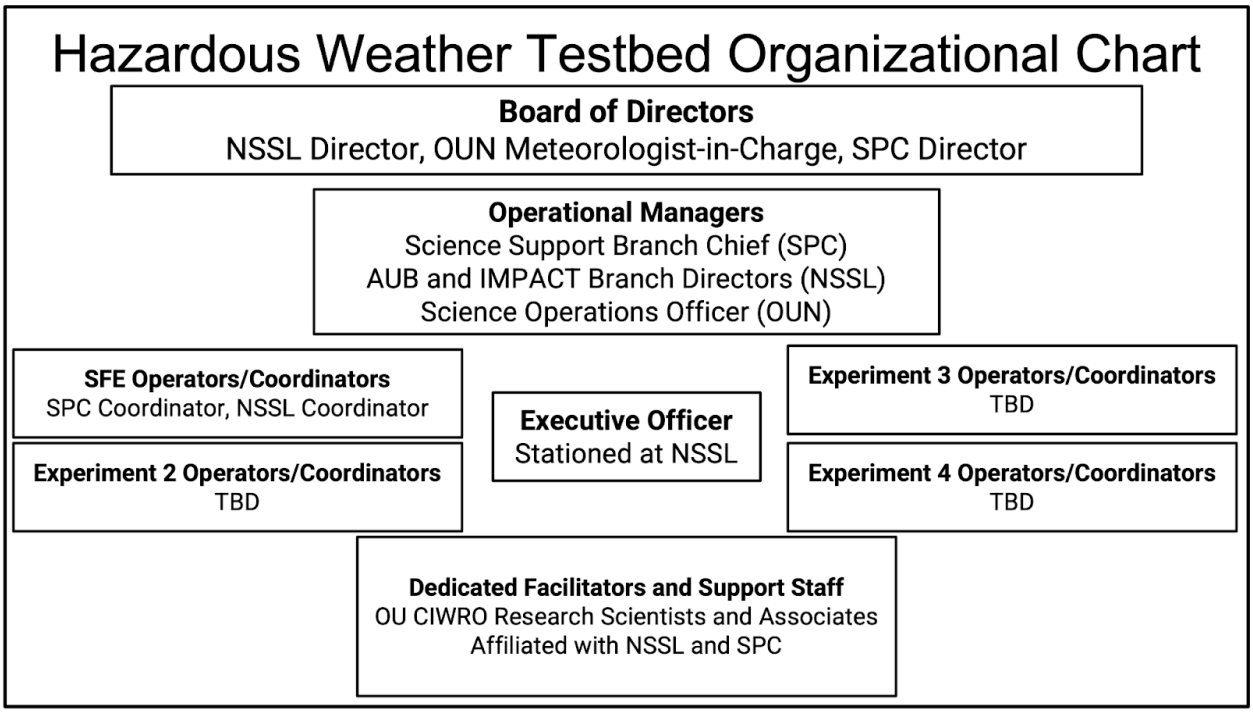


Figure A1 The organizational structure of the NOAA Hazardous Weather Testbed (HWT) as of June 2025.

Rapid science and technology infusion for the advancement of operational forecasting requires direct, focused interactions between research scientists, numerical model developers, information technology and communication specialists, and operational forecasters. The HWT provides a unique setting to facilitate such interactions and allows participants to better understand the scientific, technical, and operational challenges associated with the prediction and detection of hazardous weather events. The HWT allows participating organizations to:

- Refine and optimize emerging operational forecast and warning tools for rapid integration into operations,
- Educate forecasters on the scientifically correct use of newly emerging tools and to familiarize them with the latest research related to forecasting and warning operations,
- Educate research scientists on the operational needs and constraints that must be met by any new tools (e.g., robustness, timeliness, accuracy, and universality), and
- Motivate other collaborative and individual research projects that are directly relevant to forecast and warning improvement.

For more information about the HWT, see https://hwt.nssl.noaa.gov/. Detailed historical background about the EFP Spring Experiments, including scientific and operational motivation for the intensive examination of high resolution NWP model applications for convective weather forecasting, and the unique collaborative interactions

that occur within the HWT between the research and operational communities, are found in Kain et al. (2003), Weiss et al. (2010 – see http://www.spc.noaa.gov/publications/weiss/hwt-2010.pdf), Clark et al. (2012; 2018; 2020; 2021; 2022; 2023), and Gallo et al. (2017).

Appendix B Mandatory 2026 CLUE Fields

1. Mean Sea Level Pressure	26. CIN (most unstable)
2. Composite reflectivity	27. CAPE (mixed layer)
3. Reflectivity at -10 C	28. CIN (mixed layer)
4. Maximum surface wind gust	29. 0-3 km AGL storm relative helicity
5. hrly-max upward motion 100-1000 hPa	30. 0-1 km AGL storm relative helicity
6. hrly-max downward motion 100-1000 hPa	31. 2-5 km AGL UH (instantaneous)
7. Reflectivity at 1-km AGL	32. Echo Top Height
8. Hrly-max reflectivity at 1-km	33. 300 hPa Height
9. Hrly-max reflectivity at -10 C	34. 300 hPa u-wind
10. Hrly-max 2-5 km AGL UH	35. 300 hPa v-wind
11. Hrly-min 2-5 km AGL UH	36. 300 hPa temperature
12. Hrly-max 0-3 km AGL UH	37. 500 hPa Height
13. Hrly-min 0-3 km AGL UH	38. 500 hPa u-wind
14. Surface Pressure	39. 500 hPa v-wind
15. Surface Height	40. 500 hPa temperature
16. 2-m temperature	41. 700 hPa Height
17. 2-m dewpoint	42. 700 hPa u-wind
18. 2-m relative humidity	43. 700 hPa v-wind
19. 10-m u-wind	44. 700 hPa temperature
20. 10-m v-wind	45. 850 hPa Height
21. Hrly-max 10-m Wind Speed	46. 850 hPa u-wind
22. Surface total precipitation (run total)	47. 850 hPa v-wind
23. CAPE (surface parcel)	48. 850 hPa temperature
24. CIN (surface parcel)	49. 850 hPa specific humidity
25. CAPE (most unstable)	

References

Abdi, D., I. Jankov, P. Madden, V. Vargas, T. A. Smith, S. Frolov, M. Flora, and C. Potvin, 2026: HRRRCast: A Data-Driven Emulator for Regional Weather Forecasting at Convection-Allowing Scales. Artif. Intell. Earth Syst., 5, 250061, https://doi.org/10.1175/AIES-D-25-0061.1.

Alet, F., I. Price, A. El-Kadi, D. Masters, S. Markou, T. R. Andersson, J. Stott, R. Lam, M. Willson, A. Sanchez-Gonzalez, and P. W. Battaglia, 2025: Skillful joint probabilistic weather forecasting from marginals. arXiv preprint, arXiv:2506.10772, https://doi.org/10.48550/arXiv.2506.10772.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2022: Pangu-Weather: A 3D High-resolution Model for Fast and Accurate Global Weather Forecast. https://doi.org/10.48550/arXiv.2211.02556.

Bodnar, C., and Coauthors, 2024: A Foundation Model for the Earth System. https://arxiv.org/pdf/2405.13063.

Clark, A. J., and Coauthors, 2012: An Overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74.

Clark, A.J. and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, https://doi.org/10.1175/BAMS-D-16-0309.1.

Clark, R., P. T. Marsh, R. S. Schneider, and S. A. Erickson, 2019: Using NOAA/NWS Storm Prediction Center Forecasts to Estimate Potential Societal Impacts from Severe Weather. *35th Conf. on Environmental Information Processing Technologies*, Amer. Meteor. Soc., 2.1 (https://ams.confex.com/ams/2019Annual/webprogram/Paper352385.html).

Clark, A. J., and Coauthors, 2020: A real-time, simulated forecasting experiment for advancing the prediction of hazardous convective weather. *Bull. Amer. Meteor. Soc.*, **0**, (https://doi.org/10.1175/BAMS-D-19-0298.1).

Clark, A. J., and Coauthors, 2021: A Real-Time, Virtual Spring Forecasting Experiment to Advance Severe Weather Prediction. *Bull. Amer. Meteor. Soc.*, E814-816. https://doi.org/10.1175/BAMS-D-20-0268.1.

Clark, A. J., and Coauthors, 2022: The Second Real-Time, Virtual Spring Forecasting Experiment to Advance Severe Weather Prediction. *Bull. Amer. Meteor. Soc.*, E1114-1116. https://doi.org/10.1175/BAMS-D-21-0239.1.

Clark, A. J., and Coauthors, 2023: The Third Real-Time, Virtual Spring Forecasting Experiment to Advance Severe Weather Prediction Capabilities. *Bull. Amer. Meteor. Soc.*, E456-458. https://doi.org/10.1175/BAMS-D-20-0268.1.

Clark, A. J., A. Hill, K. Hoogewind, E. Loken, and M. Hosek, 2025: Extended range machine-learning severe weather guidance based on the operational GEFS. *Wea. Forecasting*, **40**, 1325-1343.

ECMWF, 2020: IFS Documentation CY47R1–Part V: Ensemble prediction system. ECMWF IFS Doc. 5, 23 pp., https://doi.org/10.21957/d7e3hrb.

Flora, M. L., and C. K. Potvin, 2025: WoFSCast: A Machine Learning Model for Predicting Thunderstorms at Watch-to-Warning Scales. *Geo. Res. Lett.* https://doi.org/10.22541/essoar.172574503.30734251/v1.

Gallo, B.T., and Coauthors, 2017: Breaking New Ground in Severe Weather Prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Heinselman, P. L., and Coauthors, 2024: Warn-on-Forecast System: From vision to reality. *Wea. Forecasting*, **39**, 75–95, https://doi.org/10.1175/WAF-D-23-0147.1.

Hersbach, H., and Coauthors, 2020: The ERA5 Global Reanalysis. *QJRMS*, **146**, 1999-2049. https://doi.org/10.1002/qj.3803.

Guerrette, J. J., Z. Liu, C. Snyder, and Coauthors, 2023: Data assimilation for the Model for Prediction Across Scales–Atmosphere with the Joint Effort for Data assimilation Integration (JEDI-MPAS 2.0.0-beta): Ensemble of 3DEnVar assimilations. *Geosci. Model Dev*., **16**, 7123–7142, https://doi.org/10.5194/gmd-16-7123-2023.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting Severe Weather with Random Forests. *Mon. Wea. Rev*.,**148** (5), 2135–2161, doi:10.1175/MWR-D-19-0344.1.

Hu, M., G. Ge, C. Zhou, D. Stark, H. Shao, K. Newman, J. Beck, and X. Zhang, 2018: Grid-point Statistical Interpolation (GSI) user's guide version 3.7. Developmental Testbed Center, 147 pp., http://www.dtcenter.org/com-GSI/users/.

docs/index.php.

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between Forecasters and Research Scientists at the NSSL and SPC: The Spring Program. *Bull. Amer. Meteor. Soc*., **12**, 1797-1806.

Karstens, C. D., and Coauthors, 2018: Development of a human–machine mix for forecasting severe convective events. *Wea. Forecasting*, **33**, 715–737, https://doi.org/10.1175/WAF-D-17-0188.1.

Lam, R., and Coauthors, 2022: GraphCast: Learning skillful medium-range global weather forecasting. https://doi.org/10.48550/arXiv.2212.12794.

Lang, S., M. Alexe, M. Chantry, J. Dramsch, F. Pinault, B. Raoult, M. C. A. Clare, C. Lessig, M. Maier-Gerber, L. Magnusson, Z. Ben Bouallègue, A. Prieto Nemesio, P. D. Dueben, A. Brown, F. Pappenberger, and F. Rabier, 2024: AIFS—ECMWF's data-driven forecasting system. *J. Adv. Model. Earth Syst*., in review, https://doi.org/10.48550/arXiv.2406.01465.

Li, H., G. A. Grell, R. Ahmadov, L. Zhang, S. Sun, J. Schnell, and N. Wang, 2024: A simple and realistic aerosol emission approach for use in the Thompson–Eidhammer microphysics scheme in the NOAA UFS Weather Model. *Geosci. Model Dev.*, 17, 607–619, https://doi.org/10.5194/gmd-17-607-2024.

Mansell, E. R., 2010: On sedimentation and advection in multimoment bulk microphysics. J. *Atmos. Sci.*, **67**, 3084-3094.

Martin, J. J., A. J. Clark, N. Yussouf, L. Wicker, P. Heinselman, K. Knopfmeier, B. C. Matilla, P. C. Burke, and S. Adili, 2024: Cb-WoFS: Migrating the Warn-on-Forecast System to the Cloud, *Bull. Amer. Meteor. Soc*., **105**, E1962-E1971.

Price, I., and Coauthors, 2025: Probabilistic weather forecasting with machine learning. *Nature*, **637**, 84-90.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. D. Karsten, G. J. Stumpf, and t. M. Smith, 2018: FACETs: A Proposed Next-Generation Paradigm for High-Impact Weather Forecasting. *Bull. Amer. Meteor. Soc*., **99**, 2025-2043.

Stensrud, D. J., and Coauthors, 2009: Convective-Scale Warn-on-Forecast System. *Bull. Amer. Meteor. Soc*., **90**, 1487–1499.

Weiss, S. J., A. J. Clark, I. L. Jirak, C. J. Melick, C. Siewert, R. A. Sobash, P. T. Marsh, A. R. Dean, J. S. Kain, M. C. Coniglio, M. Xue, F. Kong, K. W. Thomas, J. Du, D. R. Novak, F. Barthold, M. J. Bodner, J. J. Levit, C. B. Entwistle, R. S. Schneider, and T. L. Jensen, 2010: An Overview of the 2010 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. *25th Conf. on Severe Local Storms*, Amer. Meteor. Soc., 7B.1.