SPRING FORECASTING EXPERIMENT 2025

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA/HAZARDOUS WEATHER TESTBED

HWT Facility – National Weather Center
28 April – 30 May 2025
https://hwt.nssl.noaa.gov/sfe/2025/

Program Overview and Operations Plan

16 April 2025

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], Eric Loken[2,3], Andy Wade[1,3], Jeffrey Milne[1,3], Kimberly Hoogewind[1,3], Sean Ernst [1,3,5], Joey Picca[1,3], Michael Baldwin[1,3], Montgomery Flora[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

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

1. Introduction

Each spring, the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT), organized by the Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL), conducts a collaborative experiment to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather. The primary goals of the HWT are to accelerate the transfer of promising new tools from research to operations, to inspire new initiatives for operationally relevant research, and to identify and document sensitivities and the performance of state-of-the art experimental convection-allowing (1- to 3-km grid-spacing) modeling systems.

The 2025 HWT Spring Forecasting Experiment (SFE 2025), a cornerstone of the EFP, will be conducted 28 April – 30 May. This will be the third hybrid experiment with both in-person and virtual participation. SFE 2025 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 CAM guidance contributed by our large group of collaborators will be central to these forecasting and model evaluation activities. These contributions comprise an ensemble framework called the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018). The 2025 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 CAMs and CAM ensembles. The 2025 CLUE includes 23 members. The SFE 2025 will also continue testing of the Warn-on-Forecast System (WoFS, hereafter), which produces 18-member, 3-km grid-spacing forecasts, and will be used for the 9th year to issue very short lead-time products. As a first step toward operational transition of WoFS, the NWS Office of Science and Technology Integration (OSTI) will be conducting the WoFS runs and working with SFE coordinators on daily domain placement.

With plans for operational implementation of the Rapid Refresh Forecast System (RRFS) and RRFS Ensemble Forecast System (REFS) in 2026, a major point of emphasis will be evaluating these systems relative to the operational systems they are designed to replace. Additionally, SFE 2025 will include more experimental configurations of the Model for Prediction Across Scales (MPAS), including several configurations run by NOAA’s Global Systems Laboratory (GSL) and NSSL, and an extended-range MPAS ensemble run by the National Center for Atmospheric Research (NCAR). Finally, SFE 2025 will expand the evaluations of AI NWP emulators that were first conducted last year. Specifically, a WoFS emulator called WoFS-Cast will be examined along with both deterministic and ensemble global AI NWP emulators.

This document summarizes the core interests of SFE 2025 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 model and products being tested during SFE 2025 and Section 3 describes the core interests and new concepts being introduced for SFE 2025. A list of daily participants, 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 2025 CLUE includes deterministic and ensemble forecasts using the most recent versions of the Finite Volume Cubed-Sphere Model (FV3), the Advanced Research Weather Research and Forecasting (WRF-ARW) model, and MPAS. In addition to the CLUE, the operational 3-km grid-spacing High-Resolution Ensemble Forecast system version 3 (HREFv3), individual HREFv3 members, 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 2025.

a) The 2025 Community Leveraged Unified Ensemble (CLUE)

The CLUE is a carefully designed ensemble with members contributed by NOAA units: NSSL, GSL, Environmental Modeling Center (EMC), and the Geophysical Fluid Dynamics Laboratory (GFDL); and research groups at the National Aeronautics and Space Administration (NASA) and the National Center for Atmospheric Research (NCAR). All CLUE members cover a CONUS domain with convection-allowing resolution, except the RRFS and REFS, which cover North America, and the NCAR-MPASgbl, which covers the globe with a 3-km mesh. CLUE members have 3-km grid-spacing, except NASA FV3 (2.2-km) and GSL-MPAS3.5 (3.5-km). Depending on the CLUE subset, forecast lengths range from 36 to 132 h. Table 1 summarizes all 2025 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 12 unique subsets that comprise the 2025 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	3	none	no	HRRR or RRFS ICs	MPAS	NSSL	00, 12	48 or 84	CONUS
GSL-MPAS-RRFS	1	none	no	RRFS ICs	MPAS	GSL	00, 06, 12, 18	18 or 60	CONUS
GSL-MPAS3.5	1	none	no	RRFS ICs	MPAS	GSL	00	36	CONUS
GFDL-FV3	1	none	no	GFS cold start	FV3	GFDL	00	126	CONUS
NASA-FV3	1	none	no	GEOS-DA	FV3	NASA	00, 12	72	CONUS
NCAR-MPAS	8	GEFS	no	GEFS cold start	MPAS	NCAR	00	132	CONUS
NCAR-MPASgbl	1	none	no	GFS cold start	MPAS	NCAR	00	60	Global
NCAR-MPASctl	1	none	no	GFS cold start	MPAS	NCAR	00	60	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 forecasts are initialized from 00, 06, 12, and 18 UTC with forecasts to 60 h. Schemes marked with an asterisk (*) include 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 ("HN" = HRRR-NSSL (Mansell 2010), "HT" = HRRR-Thompson, and "RT" = RRFS-Thompson). All NSSL-MPAS runs are initialized from 00 and 12 UTC; the NSSL-MPAS-HN and NSSL-MPAS-HT have forecast lengths of 48 h, while NSSL-MPAS-RT runs to 84 h.

Member: NSSL-MPAS	**ICs**	**LBCs**	**Microphysics**	**PBL**	**LSM**	**Radiation**	**Dynamical Core**
NSSL-MPAS-HN	HRRR	HRRR	NSSL	MYNN	RUC	RRTMG	MPAS
NSSL-MPAS-HT	HRRR	HRRR	Thompson	MYNN	RUC	RRTMG	MPAS
NSSL-MPAS-RT	RRFS	RRFS	Thompson	MYNN	RUC	RRTMG	MPAS

Table 5 Specifications for the GSL-MPAS-RRFSA CLUE member. These forecasts use 3-km horizontal grid spacing with 60 vertical levels across the CONUS. Note that the PBL parameterization is an updated version of MYNN versus what is available in the NCAR-maintained MPAS-Atmosphere 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 are provided by RRFS analyses and forecasts.

Member: GSL-MPAS-RRFSA	**ICs**	**LBCs**	**Microphysics**	**PBL**	**LSM**	**Radiation**	**Dynamical Core**
GSL-MPAS-RRFS	RRFS	RRFS	TEMPO	MYNN	RUC	RRTMG	MPAS

Table 6 Specifications for the GSL-MPAS3.5 CLUE member. These forecasts are identical to GSL-MPAS-RRFS except for having 3.5-km horizontal grid spacing across the CONUS. Model output is post-processed to the same 3-km CONUS grid as the other GSL MPAS forecasts, however.

Member: GSL-MPAS3.5	**ICs**	**LBCs**	**Microphysics**	**PBL**	**LSM**	**Radiation**	**Dynamical Core**
GSL-MPAS3.5	RRFS	RRFS	TEMPO	MYNN	RUC	RRTMG	MPAS

Table 7 Specifications for the GFDL FV3 CLUE member. GFDL's C-SHiELD (Harris et al., 2019) is an FV3-based model that uses a 13-km global grid and a 3-km CONUS nest, coupled to a modified form of the GFS Physics. C-SHiELD uses version 3 of the GFDL In-line Microphysics (Zhou et al. 2022) and the EMC/UW TKE-EDMF PBL scheme (Han and Bretherton 2019). On the CONUS nest the Noah-MP LSM is used; the global domain uses the GFS Noah LSM. Initialization is cold start from regridded GFS real-time analyses. GFDL will provide simulations run daily at 00Z out to 126 hours to demonstrate the potential for medium-range prediction of convective-scale events. For more info see: http://www.gfdl.noaa.gov/shield.

Member: GFDL FV3	**ICs**	**LBCs**	**Microphysics**	**PBL**	**LSM**	**Radiation**	**Dynamical Core**
GFDL-FV3	GFS	n/a	GFDL	TKE-EDMF	NOAH-MP	RRTMG	FV3

Table 8 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.

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

Table 9 Specifications for the NCAR-MPAS ensemble members. All 8 ensemble members use NCAR's MPAS model and identical physics, with ensemble diversity solely provided by ICs. These runs use a global 13-km grid-spacing domain with a refined 3-km grid-spacing mesh over the CONUS. Initialization is cold-start from members 1–8 of real-time GEFS ICs. Simulations run daily at 00Z out to 132 hours.

Members: NCAR-MPAS	**ICs**	**LBCs**	**Micro-physics**	**PBL**	**LSM**	**Radiation**	**Cumulus**	**Dynamical Core**
NCAR-MPAS01	GEFS m1	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS
NCAR-MPAS02	GEFS m2	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS
NCAR-MPAS03	GEFS m3	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS
NCAR-MPAS04	GEFS m4	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS
NCAR-MPAS05	GEFS m5	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS
NCAR-MPAS06	GEFS m6	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS
NCAR-MPAS07	GEFS m7	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS
NCAR-MPAS08	GEFS m8	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS

Table 10 Specifications for the NCAR-MPASgbl member. This member uses uniform 3-km grid-spacing covering the entire globe with forecasts to 60 h and is initialized from the GFS. Other aspects of the configuration are identical to the NCAR-MPAS ensemble members.

Member: NCAR-MPASgbl	**ICs**	**LBCs**	**Micro-physics**	**PBL**	**LSM**	**Radiation**	**Cumulus**	**Dynamical Core**
NCAR-MPASgbl	GFS	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS

Table 11 Specifications for the NCAR-MPASctl member. This member is identical to the NCAR-MPASgbl member, but with 13-km grid-spacing covering the globe and a 3-km refined mesh over the CONUS.

Member: NCAR-MPASctl	ICs	LBCs	Micro-physics	PBL	LSM	Radiation	Cumulus	Dynamical Core
NCAR-MPASctl	GFS	n/a	Thompson	MYNN	NOAH	RRTMG	Scale-aware New Tiedtke	MPAS

The configuration of the 2025 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:

RRFS vs. Operational CAMs and REFS vs. HREF: With plans for operational implementation of RRFS in 2026, a critical evaluation activity for SFE 2025 will involve comparing RRFS to the operational CAMs that EMC plans to retire once RRFS is implemented, which includes the NAM Nest, HRW ARW, HRW NSSL, and HRW FV3. Comparisons will be made at Day 1 lead times for 0000 and 1200 UTC initializations. Similarly, ensemble comparisons of REFS vs. HREF will be made at Day 1 & 2 lead times for 0000 and 12000 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. See table 14 for configuration details. Finally, comparisons will be made during the first 12 h of RRFS and HRRR forecasts to evaluate the effectiveness of the data assimilation strategies in each system.

Table 12 Ensemble members comprising two versions of REFS based at 1200 UTC.

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

Medium-Range CAM Ensembles: NCAR will be providing an 8-member, 0000-UTC initialized MPAS ensemble with forecasts to 5 days (Table 11). Severe weather forecasts derived from machine learning and neighborhood maximum ensemble probabilities (NMEPs) will be examined and compared to other methods for generating extended-range severe weather probabilities that are based on global NWP ensembles.

RRFSv2 Development Systems: NOAA/GSL will be providing two MPAS configurations (Tables 5-6), which are being tested to form the foundation for RRFSv2. These configurations contain the most up-to-date physics suites tuned for MPAS. Comparisons will be made to other deterministic CAMs such as HRRR, RRFSv1, and NSSL-MPAS-RT. GSL-MPAS3.5 is identically configured to GSL-MPAS-RRFS, except it uses 3.5-km grid-spacing. MPAS is believed to have higher effective resolution than WRF because of its unstructured grid and numerics; thus, this resolution sensitivity test will examine whether 3.5-km grid-spacing could meet performance requirements while saving computational time relative to 3-km runs.

Global CAM vs. Global with Refined Mesh CAM: In a first-of-its-kind test, NCAR will provide 0000 UTC, GFS-initialized, 60-h forecasts from a global MPAS configuration with uniform 3-km grid-spacing over the entire globe (Table 12; NCAR-MPASgbl). Comparisons will be made to an identical configuration that uses a 13-km grid-spacing global mesh with refinement to 3-km grid-spacing over the CONUS (Table 13; NCAR-MPASctl).

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. Versions of the analyses upscaled to 40-km will also be examined and compared with SPC's RAP-based surface objective analysis (sfcOA). Finally, 15-minute WoFS 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 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 2025 CLUE, all groups output fields to the 3-km CONUS grid used for the operational HRRR. For WRF-ARW, FV3, and MPAS the Unified Post-Processor software (UPP; https://www.epic.noaa.gov/unified-post-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 C 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 11 May 2021. The design of HREFv3 originated from the SSEO, 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 13.

Table 13 Model specifications for HREFv3.

HREFv3	ICs	LBCs	Microphysics	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
HRW 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) NSSL cloud-based Warn-on-Forecast Experiments

Cb-WoFS is a rapidly-updating 36-member, 3-km grid-spacing WRF-ARW-based ensemble data assimilation and forecast system. The cb-WoFS forecasts are initialized every 30 minutes and used to produce very short-range (0-6/0-3 hour at top/bottom of the hour) probabilistic forecasts of individual thunderstorms and their associated hazardous weather phenomena such as supercell hail, high winds, flash flooding, and supercell thunderstorm rotation. The 900-km x 900-km daily cb-WoFS domain will target the primary region where severe weather is anticipated, using the SPC Day 1 Convective Outlook as a guide. Cb-WoFS is capable of running over two different regions. 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.

The cloud-based Warn-on-Forecast System (cb-WoFS; Martin et al. 2025) uses current technologies in containerization and cloud computing. The entire WoFS application was built on top of multiple Platform-as-a-Service and Infrastucture-as-a-Service technologies on the Azure platform and the WRF model itself rebuilt to run in containers optimized for HPC. With the cb-WoFS 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 cb-WoFS operators the choice of running on the latest technologies. All parts of WoFS 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 cb-WoFS 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 the WoFS can be started four times a day (15, 21, 03, or 09 UTC). For instance, if WoFS 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 cb-WoFS. 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 16 provides a summary of the model specifications for cb-WoFS, and Figure 1 shows an example of a SPC Day 1 convective outlook and corresponding cb-WoFS domain with WSR-88D radars used for data assimilation overlaid. Further details on the cb-WoFS are included below.

The 36-member cb-WoFS 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 cb-WoFS 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 cb-WoFS 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 cb-WoFS Forecast Viewer (https://cbwofs.nssl.noaa.gov).

Table 14 cb-WoFS configuration.

	WoFS
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)

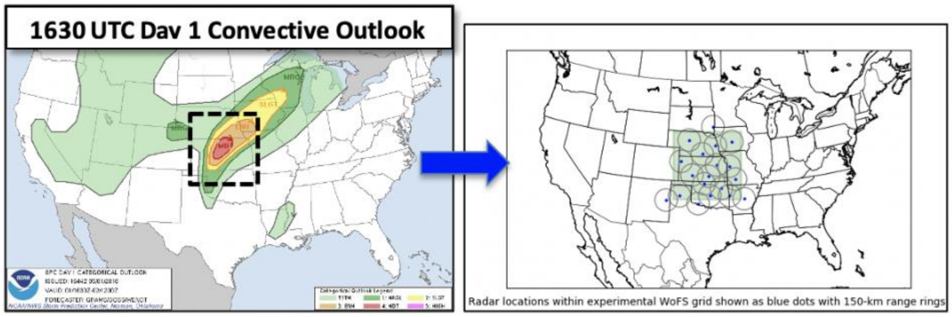

1630 UTC Day 1 Convective Outlook
Radar locations within experimental WoFS grid shown as blue dots with 150-km range rings

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

d) AI NWP Emulators

In the last three years, fully AI-based models (i.e., AI NWP emulators) have been developed by the private sector for global weather prediction. This area of research is advancing rapidly and has the potential to be revolutionary for weather prediction since skill measures of the NWP emulators commonly exceed those of the ECMWF's Integrated Forecast System (IFS; ECMWF 2020), the world's most skillful global NWP system. Furthermore, the NWP emulators can produce forecasts in seconds, orders of magnitude faster and with fewer computational resources than traditional NWP systems. The algorithms are trained using large, global, multi-year reanalysis datasets like ERA5 (Hersbach et al. 2020). Several of these algorithms have been made public, and government agencies are beginning to run and train the models themselves. While objective skill measures have been impressive, these NWP emulators have had only limited testing for real-time operational forecasting applications. Thus, during SFE 2025, we will evaluate several of the publicly available algorithms that were trained using ERA5 data. The AI-based NWP emulators are being run experimentally at the Cooperative Institute for Research in the Atmosphere (CIRA) by providing both GFS and IFS initial conditions to these AI models. The CIRA forecasts can be viewed at: https://aiweather.cira.colostate.edu/. Ensemble forecast data from Google Deep Mind's WeatherNext Gen system is also being provided through a public repository.

Additionally, CIWRO/NSSL has developed a convective scale AI NWP Emulator for WoFS, which is known as WoFSCast (Flora and Potvin 2025). Information on each AI NWP Emulator that will be evaluated during SFE 2025 is contained below.

i. Pangu-Weather

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 CIRA, one with GFS initial conditions and the other with IFS, will be evaluated during SFE 2025.

ii. GraphCast

GraphCast is a machine-learning algorithm developed by Google 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. GraphCast is pre-trained with ERA5 reanalysis data. Two versions of GraphCast configured by CIRA, one with GFS initial conditions and the other with IFS, will be evaluated during SFE 2025. Another version of GraphCast run by EMC will also be evaluated. The EMC version is fine-tuned with NCEP's GDAS data as inputs and ERA5 data as ground truth to calculate new weights in creating global forecasts

iii. Aurora

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. Two versions of Aurora configured by CIRA, one with GFS initial conditions and the other with IFS, will be evaluated during SFE 2025.

iv. WeatherNext Gen

WeatherNext Gen is a probabilistic weather model developed by Google DeepMind that generates global 15-day, 64-member ensemble forecasts at 0.25-degree resolution, which have been shown to outperform the ECMWF ensemble (Price et al. 2025). Generation of a single 15-day WeatherNext Gen forecast takes about 8 minutes on a cloud TPUv5 device, and an ensemble can be generated in parallel. WeatherNext Gen uses a conditional diffusion model, a generative ML method capable of modeling the probability distribution of complex data and generating new samples. It is trained on 40 years of ERA5 data. The forecasts are provided through a Google DeepMind repository.

v. WoFSCast

WoFSCast is an AI emulator of the NSSL Warn-on-Forecast System (WoFS; Flora and Potvin 2025). Refactored from Google's GraphCast, but for limited area domain modeling, WoFSCast predicts a combination of 3D and 2D variables at high spatiotemporal resolution (3-km grid spacing and 10-min timesteps). With a single NVIDIA A100, WoFSCast can produce 18-member, 6-h forecasts in under 2 minutes. WoFSCast is trained from WoFS forecasts and at inference time uses the WoFS analysis and 10-min forecast as initial conditions and WoFS forecasts for boundary conditions. Objective verification found that WoFSCast performs similarly to WoFS compared to MRMS out to 6 hrs. A public version of the code base is available at https://github.com/NOAA-National-Severe-Storms-Laboratory/frdd-wofs-cast.

e) Calibrated Forecast Products

i. GEFS-based, ML-derived Hazard Probabilities (credit: A. Hill)

Similar to previous SFEs, the GEFS Machine Learning Probabilities (Hill et al. 2020; 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 (e.g., Figure 2). 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; all predictors are listed in Table 15. 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 preexisting 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.

Table 15 Short-hand notation (left) and long description (right) of predictor variables used to train GEFS Reforecast MLP severe weather RFs. Derived variables from FV3-GEFS/R output are denoted with an asterisk (*).

Predictor Acronym	**Predictor Description**
APCP	3-hourly accumulated precipitation
CAPE	Convective available potential energy
CIN	Convective inhibition
U10	10 m latitudinal wind speed
V10	10 m longitudinal wind speed
T2M	2 m temperature
Q2M	2 m specific humidity
MSLP	Mean sea level pressure
PWAT	Precipitable water
UV10	10 m wind speed
SRH03	0 - 3km storm relative helicity
SHEAR850*	0 - 850 hPa bulk wind shear
SHEAR500*	0 - 500 hPa bulk wind shear
ZLCL*	Height of lifting condensation level
RH2M*	2 m relative humidity

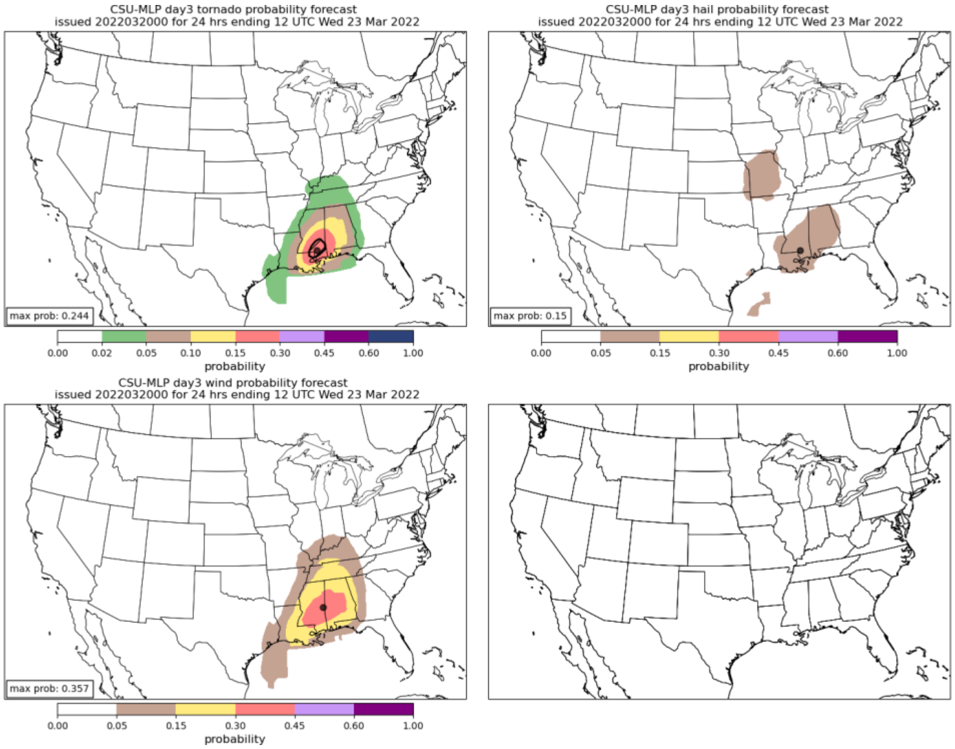

CSU-MLP day3 tornado probability forecast
issued 2022032000 for 24 hrs ending 12 UTC Wed 23 Mar 2022
max prob: 0.244
0.00
0.02
0.05
0.10
0.15
0.30
0.45
0.60
1.00
probability
CSU-MLP day3 hail probability forecast
issued 2022032000 for 24 hrs ending 12 UTC Wed 23 Mar 2022
max prob: 0.15
0.00
0.05
0.15
0.30
0.45
0.60
1.00
probability
CSU-MLP day3 wind probability forecast
issued 2022032000 for 24 hrs ending 12 UTC Wed 23 Mar 2022
max prob: 0.357
0.00
0.05
0.15
0.30
0.45
0.60
1.00
probability

Figure 2 Probabilistic day-3 forecasts of (upper left) tornado, (upper right) hail, and (bottom left) wind hazards valid 1200 - 1200 UTC ending 23 March 2022. Hatched contours represent a 10% probability of significant severe hazards.

ii. NSSL GEFS-based, ML-derived Hazard Probabilities (credit: A. Clark)

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. Table 17 summarizes the main differences between the algorithms.

Table 16 GEFS-based predictors used in GEFS-ops RF.

GEFS Operational MLP Predictors		
(1) Bulk Shear (0-1 km AGL)*	(8) Surface-based lifting condensation level (LCL) height	(15) u-wind (10-m)
(2) Bulk Shear (0-3 km AGL)*	(9) Significant tornado parameter (STP)*	(16) v-wind (10-m)
(3) Bulk Shear (0-6 km AGL)*	(10) Mean-sea-level pressure	(17) Wind magnitude (10-m)
(4) Surface-based convective available potential energy (CAPE)	(11) Precipitable water	(18) Most unstable CAPE*
(5) Surface-based convective inhibition	(12) Specific Humidity (2-m)	(19) Latitude
(6) Storm relative helicity (0-3 km)	(13) Temperature (2-m)	(20) Longitude
(7) Lape Rate (700-500 mb)*	(14) Precipitation (3-h accumulation)	

Table 17 Summary of differences between GEFS Operational MLP and GEFS Reforecast MLP.

	GEFS Operational MLP	**GEFS Reforecast MLP**
Grid-spacing	81-km (interpolated)	55-km (0.5° x 0.5°)
Training	Ensemble mean of 31 GEFS operational members	Ensemble median of 5 GEFS reforecast members
Lead time	Days 1-15	Days 1-8
Products	Total Severe	Total Severe (Days 1-8), Hazard probs & Sig Severe (Days 1-3)
Predictors	18	12
Forecast input	Mean of all 31 GEFS members	Median of first 21 GEFS members
Regional training?	No (CONUS land points only)	yes; 4 distinct regions over the CONUS
Neighboring points used for predictors?	no	yes; 7 x 7 point surrounding box
Latitude/longitude coordinate used for predictors?	yes	no

iii. NSF NCAR ML-derived MPAS-based convective hazard probabilities (R. Sobash)

For the SFE 2025, gridded machine learning-based probabilistic convective hazard guidance is being generated using neural networks (NNs) and the medium-range real-time MPAS-based ensemble forecasts generated at NSF NCAR. Two modifications were made to the system based on evaluations of the C-SHiELD-based ML hazard forecasting system in the 2024 SFE. First, the MPAS system natively outputs 24-hour probabilities (valid 12 UTC – 12 UTC) for Days 1–5. Second, the MPAS ensemble mean is used for training and inference, rather than individual members.

More specifically, NNs were trained (Table 18) using the 80 sets of 0000 UTC-initialized MPAS ensemble mean forecasts during the 2023 and 2024 SFEs. Features include a set of 23 diagnostics (Table 19) that were upscaled onto an 80-km grid and an additional set of “neighborhood” features constructed by taking the non-static predictors and computing means (for the environmental predictors) and maxima (for the explicit predictors) in space over 3x3 and 5x5 arrays of 80-km grid boxes. The hourly upscaled MPAS output was further aggregated in time to reduce the feature set by taking the mean or maximum of each field within three-hour intervals.

Each grid box was labeled as a “hit” if a severe weather report occurred within a 24-hr period (1200 – 1200 UTC) and 40-km of the grid box center point. The NNs were designed to output six independent probabilities: probability of hail, wind, tornado, significant hail, significant wind, or any storm report. Slight changes were made to the NN training settings, including the addition of regularization, compared to 2024 based on hyperparameter experiments. We trained 10 individual NNs, with the final output probabilities computed using an average of the 10 networks, based on the work of Sobash and Ahijevych (2024).

Table 18 Settings used to construct and train the NNs. Ten NNs with different initial weights were trained separately with their output probabilities averaged together.

Neural network hyperparameter	**Value**
Number of hidden layers	1
Number of neurons in hidden layer	16
Dropout rate	0.1
Learning rate	0.001
Number of training epochs	30
Hidden layer activation function	Rectified Linear Unit
Output layer activation function	Sigmoid
Optimizer	Adam
Loss function	Binary Cross-entropy
Batch size	1024
Regularization	0.01
Batch normalization	On

Table 19 The 23 base predictors used to train the NNs. The mean of the environmental fields, and the maximum of the explicit fields, within each 80-km grid box, was used as input into the NNs. Neighborhood predictors were also constructed by taking larger spatial and temporal means and maximums of the environmental and explicit fields as described in the text.

Base Predictor	**Type**
Forecast Day, Local Solar Hour, Latitude, Longitude, Day of Year	Static
SBCAPE, SBCIN, MUCAPE, MLLCL	Environment
2-m Temperature & 2-m Dewpoint Temperature	Environment
0-6 km & 0-1 km AGL bulk shear	Environment
0-1 km AGL & 0-3 km AGL Storm-relative helicity	Environment
Fixed-layer significant tornado parameter	Environment
Product of MUCAPE and 0-6 km AGL bulk shear	Environment
700 hPa – 500 hPa lapse rate	Environment
Hourly-maximum 2–5 km AGL UH (positive & negative)	Explicit
Hourly-maximum 0–1 km & 0–3 km AGL UH	Explicit
Hourly-maximum 1 km AGL relative vorticity	Explicit
Hourly-maximum updraft & downdraft speed below 400 hPa	Explicit
Hourly-maximum column-integrated graupel	Explicit
Hourly-maximum 10-m wind speed	Explicit
Column-maximum reflectivity	Explicit

iv. NSF NCAR ML-derived GEFS-based convective hazard probabilities (R. Sobash)

In a similar way to the MPAS ML hazard probabilities, we have generated GEFS-based probabilities for the 2025 SFE. The GEFS-based system was designed to be nearly identical to the MPAS-based system, to facilitate inter-comparisons. Three major differences exist. First, the feature set is restricted to 19 base predictors (Table 20) that are nearly identical to those used in the operational GEFS random forest-based system (e.g., Hill et al. 2023). Second, the training dataset is much larger than the MPAS forecast dataset, covering March – June of 2021–2024 (i.e., 16 months of GEFS 0000 UTC initializations). Finally, 24-hour probabilities are output for Days 1–8, rather than Days 1–5 with MPAS. Similar to Hill et al. (2023), we use the 3-hourly GEFS output fields as predictors into the NNs. Other than the number of input features, the NN architecture is identical to the MPAS-based system (Table 18).

Table 20 As in Table 21, but for the 16 base predictors used to train the GEFS-based ML NNs.

Base Predictor	Type
Forecast Day, Local Solar Hour, Latitude, Longitude, Day of Year	Static
MLCAPE, MLCIN	Environment
2-m Temperature & 2-m Specific Humidity	Environment
Surface–500 hPa and Surface–850 hPa bulk shear	Environment
10-m wind speed and components (zonal and meridional)	Environment
Precipitable water	Environment
3-hourly Accumulated Precipitation	Environment

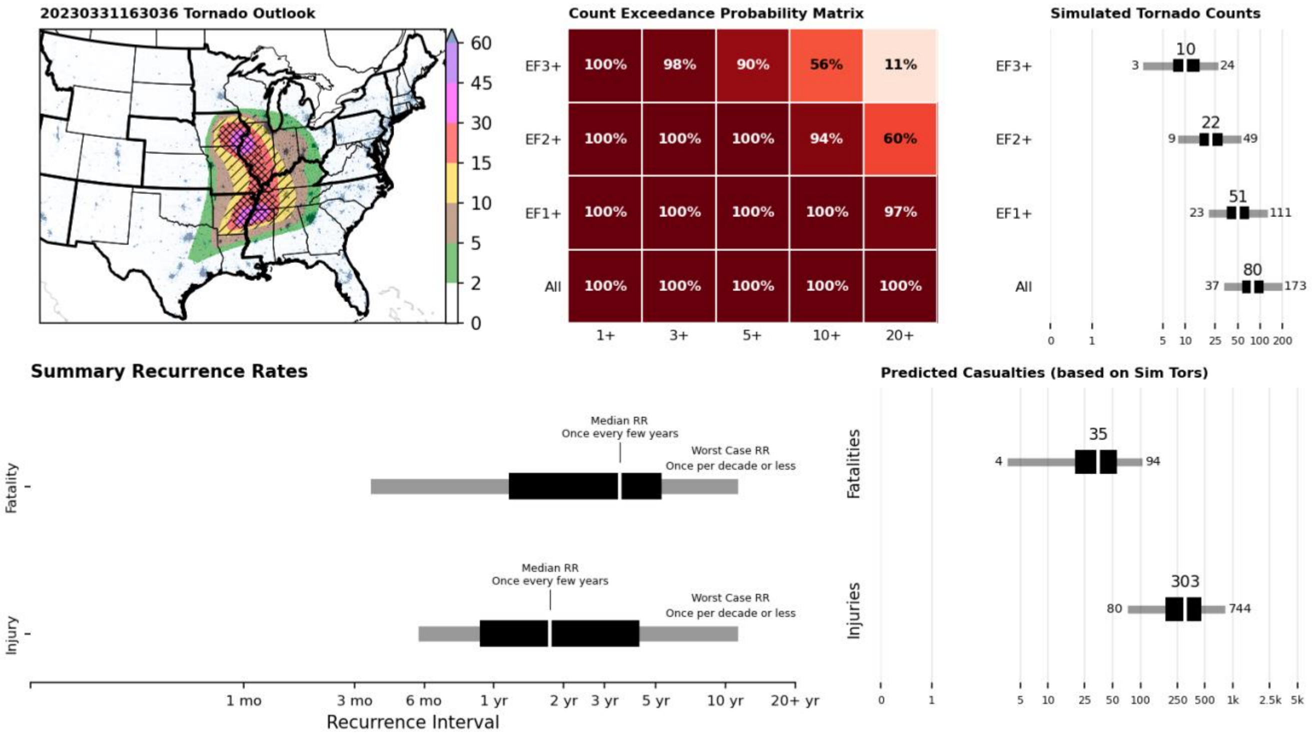
f) SPC Impacts System

SPC maintains an internal analytics system (e.g., Clark et al 2019) for predicting the number of tornadoes, their characteristics, and their potential impacts to society, using the SPC's Day 1 tornado forecast (both coverage and conditional intensity forecasts) as the initial input. From this input, the system runs a series of Monte Carlo simulations (currently set to n=10000 simulations) that draw from a number of historical distributions (tornado frequency per unit area, tornado rating, tornado duration, path width) to produce many possible realizations of a tornado day.

Within a simulation/realization, individual tornadoes are placed in a clustered manner, in which an initial seed tornado is placed randomly, weighted by continuous probabilities from the Day 1 tornado coverage forecast. The system then utilizes historical data to determine the probability of another tornado occurring downstream, which serves as a weight for randomly determining if another tornado occurs in the cluster. Once the cluster is terminated (i.e., it was determined that another tornado does not occur downstream), the process repeats until all tornadoes in the simulation are placed. For direction and distance, these simulated tornadoes are combined with storm motion fields from the HRRR to produce quasi-realistic tornado paths for the background environment.

Each realization is overlaid on 1-km gridded societal data (e.g., population, schools) from the US Census, such that the potential impact to society can be quantified. Additionally, a machine-learning workflow (trained on historical tornadoes from 1999 to 2021) is used to predict a number of injuries and fatalities associated with each tornado.

With impacts quantified across each realization, distributions of tornado counts and their respective potential impacts can be visualized. Additionally, these impacts are compared to historical data to construct recurrence rate information. Thus, this system can be used to convert SPC's operational tornado forecasts into quantifiable impact data that can be communicated to partners in emergency management, etc. for improved preparedness.



	1+	3+	5+	10+	20+
EF3+	100%	98%	90%	56%	11%
EF2+	100%	100%	100%	94%	60%
EF1+	100%	100%	100%	100%	97%
All	100%	100%	100%	100%	100%

Figure 3 An example of tornado counts and impacts estimated from the 31 March 2023 1630 UTC outlook. For the box and whiskers plots, the whiskers represent the 5th and 95th percentiles, or the reasonable best case and worst case scenarios, respectively. The black box indicates the 25th-75th percentile range, while the vertical white line represents the median scenario.

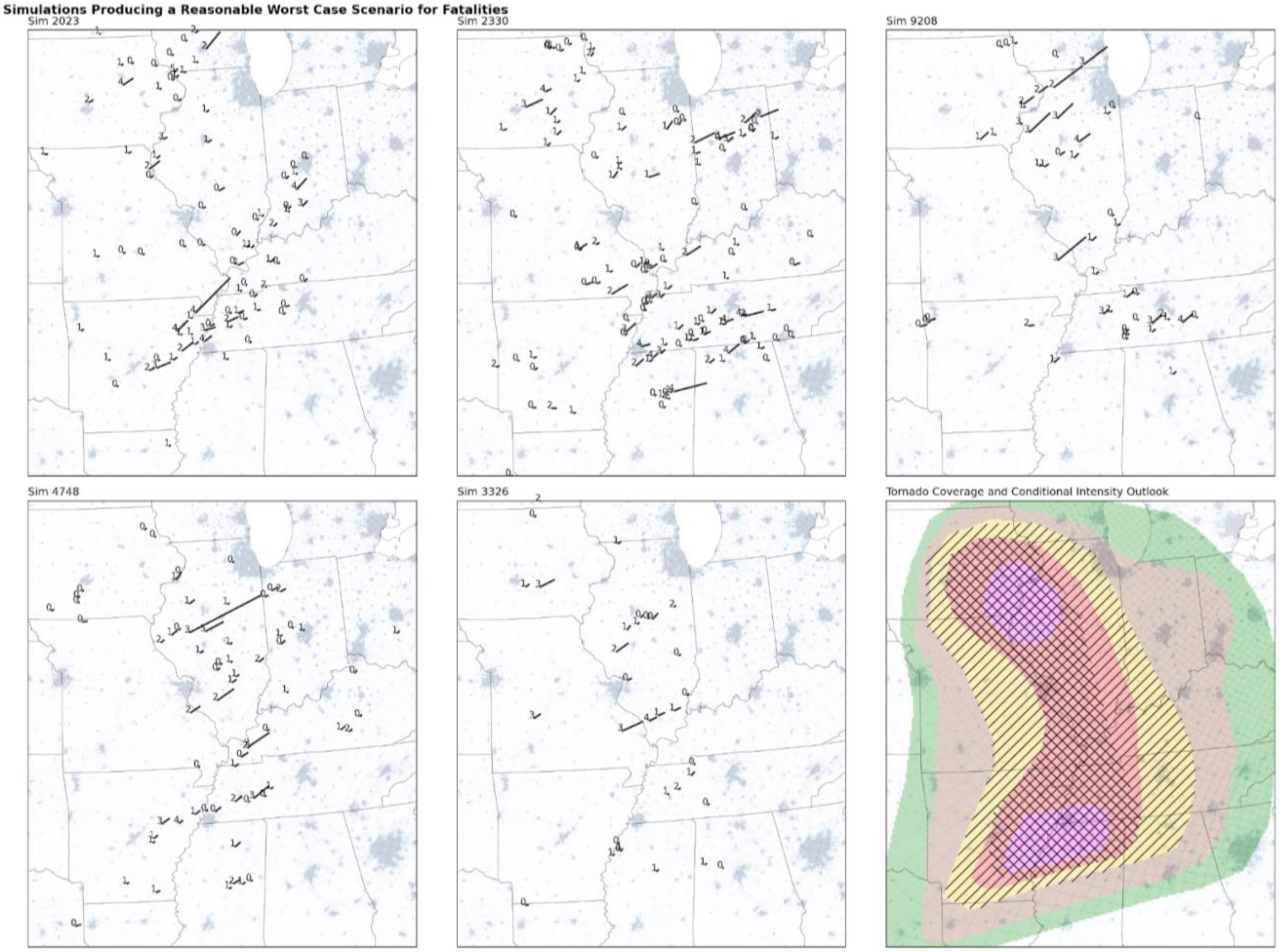

Simulations Producing a Reasonable Worst Case Scenario for Fatalities
Sim 2023
Sim 2330
Sim 9208
Sim 4748
Sim 3326
Tornado Coverage and Conditional Intensity Outlook

Figure 4 Five simulations that produced the 95th percentile (i.e., the reasonable worst case scenario) for fatalities, driven by the 31 March 2023 1630Z outlook (lower right plot). In the simulation plots, lines denote simulated tornado paths, while numbers annotated near the tornado start point indicate the rating.

3. SFE 2025 Core Interests and Daily Activities

2025 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 option 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 panel discussion from 1:15-2pm. Tables 21-23 provide a daily schedule for Monday, Tuesday-Thursday, and Friday, respectively. Further details are provided in subsequent sections.

Table 21 Schedule for Monday.

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* (Tom Galarneau)
11:15 AM – 12:30 PM	**Group Forecasting Activity** (Coverage and Conditional Intensity Outlooks) *In-Person R2O* (Day 1); *Virtual 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** (Mesoscale Discussions and Training) *In-Person R2O* (Meso-beta MD); *Virtual 1* (WoFS); *Virtual 2* (WoFS)
3:15 PM – 4:00 PM	**Individual Forecasting Activity Continued** (Day 1 Updates and MDs) *In-Person R2O* (Day 1 Update); *Virtual 1* (WoFS); *Virtual 2* (WoFS)

Table 22 Schedule for Tuesday - Thursday.

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* (Tom Galarneau)
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* (Tom Galarneau)
11:15 AM – 12:30 PM	**Group Forecasting Activity** (Coverage and Conditional Intensity Outlooks) *In-Person R2O* (Day 1); *Virtual 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** (Mesoscale Discussions) *In-Person R2O* (Meso-beta MD); *Virtual 1* (WoFS); *Virtual 2* (WoFS)
3:15 PM – 4:00 PM	**Individual Forecasting Activity** (Mesoscale Discussions and Training) *In-Person R2O* (Meso-beta MD); *Virtual 1* (WoFS); *Virtual 2* (WoFS)

Table 23 Schedule for Friday.

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* (Tom Galarneau)
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 00Z Total Severe

Three different sets of extended range total severe probabilities for Day 3-7 lead times are subjectively rated. These methods include: (1) GEFS Reforecast MLP, (2) GEFS Operational MLP, and (3) GEFS NN.

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?

C2. Medium Range CAM Severe ML Guidance

A machine-learning algorithm using predictors from the extended-range, 3-km grid-spacing NCAR-MPAS ensemble is used to derive total severe probabilities for Day 3-5 lead times. These will be compared to probabilities derived from UH using neighborhood maximum ensemble probabilities (NMEPs), and GEFS NN.

Primary Science Question(s): Does CAM-based, machine-learning guidance provide value relative to ML guidance derived from coarser global ensembles like the GEFS, and more simple NMEP-based methods?

C3. Tornado Conditional Intensity

A random-forecast model, torCI, that uses individual environmental parameters and climatology to predict the probability of a significant (EF2+) tornado, given that a tornado occurs. It is trained on only tornado environments 2007-2024, no nontornadic environments, so that it predicts conditional intensity directly. Features in the current version are 0-1-km SRH, 100-hPa MLCAPE, MLCIN, MLLCL, u and v components of 0-6-km bulk shear, u and v components of 300-hPa wind, u and v components of 10-m wind, and a gridded, smoothed climatology of conditional intensity in rolling three-month windows. Daily outlooks are created using the 09 UTC RAP. At the end of each week, the torCI conditional intensity forecasts will be compared to 09 UTC RAP forecasts of the significant tornado parameter (STP), which forms the current foundation for the conditional intensity distributions at SPC.

Primary Science Question(s): Does a ML method provide benefit for conditional intensity forecasting of tornadoes over STP?

CAM (E)nsembles

E1 & E2. CLUE: Day 1 REFS vs. HREF

This evaluation will feature an in-depth examination of severe storm-attribute and environmental fields from 0000- and 1200-UTC initialized versions of REFS, SPC REFS, and HREF 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. A greater number of fields will be available for this comparison relative to other comparisons, allowing for participants to examine more facets of the guidance and identify potential contributions to severe convective hazard forecast success or failure.

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?

E3. CLUE: Day 2 REFS vs. HREF

This evaluation is similar to the Day 1 REFS vs. HREF, but for Day 2 lead times and limited to 1200 UTC initializations. In addition, the 0000 UTC initialized NCAR MPAS ensemble will be included in the Day 2 comparisons.

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? How does an MPAS-based single-physics ensemble compare to REFS and HREF in terms of spread and skill?

(D)eterministic CAMs

D1. CLUE: 0000 UTC Day 1 Deterministic Flagships

This activity will focus on rating the primary deterministic CAMs provided by several SFE collaborators – NSSL (*NSSL-MPAS-RT*), EMC (*RRFS*), NASA (*NASA-FV3*), and GSL (*GSL-MPAS-RRFS*) – 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., WRF-ARW-based HRRRv4)?

D2. CLUE: 1200 UTC Day 2 Deterministic Flagships

Five deterministic, 1200-UTC initialized, CAM configurations are subjectively evaluated for Day 2 lead times. These configurations include: (1) RRFS, (2) NSSL-MPAS-RT, (3) NASA-FV3, (4) GSL-MPAS-RRFS, and (5) HRRR.

Primary Science Question(s): What strategies for CAM configurations perform the best at Day 2 lead times, and what are their forecast characteristics at Day 2 lead times for severe weather forecasting applications?

D3. CLUE: 0000 UTC Day 3 Deterministic Flagships

Five deterministic, 0000-UTC initialized, CAM configurations are subjectively evaluated for Day 3 lead times. These configurations include: (1) RRFS, (2) NSSL-MPAS-RT, (3) NASA-FV3, (4) GFDL-FV3, and (5) NCAR MPAS01.

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

D4. CLUE: 0000 UTC RRFS vs. Operational CAMs

The RRFS will be compared to operational CAMs that EMC plans to retire once RRFS is implemented, which includes the NAM Nest, HRW ARW, HRW NSSL, and HRW FV3. This activity will feature a “deeper dive” into storm attribute and environmental fields and serve to unearth ways in which the currently operational CAMs differ from the candidate to replace them – RRFS. Specifically, whether the RRFS improves upon or degrades forecast of the operational CAMs for fields relevant to severe weather forecasting will be examined. A greater number of fields will be available for this comparison relative to other comparisons, allowing for participants to examine more facets of the guidance and identify potential contributions to severe convective hazard forecast success or failure.

Primary Science Question(s): How do 0000-UTC initialized forecasts of the RRFS compare to those of the operational CAMs? Are there systematic shortcomings or advantages of the RRFS?

D5. CLUE: 1200 UTC RRFS vs. Operational CAMs

This evaluation is the same as D4, except for 1200 UTC initialized models are examined.

Primary Science Question(s): How do 1200 UTC initialized forecasts of the RRFS compare to those of the operational CAMs? Are there systematic shortcomings or advantages of the RRFS?

D6. CLUE: RRFS vs. HRRR DA

The HRRR and RRFS are examined in the first 12 hours of the forecast period for 2100 and 0000 UTC initializations to evaluate the impact of their data assimilation.

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

D7. CLUE: 0000 UTC MPAS Resolution Sensitivity

The 3-km grid-spacing GSL-MPAS-RRFS and 3.5-km grid-spacing GSL-MPAS3.5, which are identically configured except for their grid-spacing, will be compared. MPAS is believed to have higher effective resolution than WRF because of its unstructured grid and numerics, thus, this resolution sensitivity test will examine whether 3.5-km grid-spacing could meet performance requirements while saving computational time relative to 3-km runs.

Primary Science Question(s): Does GSL-MPAS3.5 have similar performance characteristics to GSL-MPAS-RRFS? Could GSL-MPAS3.5 meet performance requirements for RRFSv2 while saving computational time relative to 3-km runs?

D8. CLUE: 0000 UTC Day 2 Global vs. Regional CAM

A uniform 3-km grid-spacing *global* MPAS configuration will be compared to an identical configuration that uses a 13-km grid-spacing global mesh with refinement to 3-km grid-spacing over the CONUS for the Day 2 period.

Primary Science Question(s): Do any differences show up by Day 2 over the CONUS between these configurations?

(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.

Primary Science Question(s): What are the optimal methods for producing quality mesoscale analyses for convective forecasting applications?

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-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-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 NWP Emulators: GFS ICs

GraphCast, Pangu-Weather, and Aurora AI-driven global weather predictions 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 NWP Emulators initialized from the GFS compare to traditional NWP forecasts? Is there value in the NWP emulators for extended range severe weather forecasting applications? Does GraphCast tuned with GDAS improve forecasts initialized with GFS?

AI2. Global NWP Emulators: EC ICs

This evaluation is the same as the previous one, except for global NWP emulators starting with EC initial conditions, including the EC AIFS.

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

AI3. Global NWP Emulators: IC Comp

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 Ensembles NWP Emulators

WeatherNext GenCast ensemble forecasts 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 does a global ensemble based on AI NWP emulation compare to traditional NWP ensembles in terms of the mean pattern and overall spread?

AI5. WoFS vs. WoFS-Cast

This evaluation examines the quality of deterministic and probabilistic reflectivity forecasts from 0000 UTC initializations of WoFS and an AI-NWP system called WoFSCast. WoFSCast uses a machine learning algorithm trained on 3 years of WoFS forecasts to generate the same products as WoFS more quickly and at a fraction of the computational cost.

Primary Science Question(s): How do forecasts derived from a machine learning algorithm designed to emulate WoFS compare to WoFS itself?

(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.

O2. Day 1 Outlook Update (w/ WoFS)

The Day 1 outlooks for tornado, wind, and hail are compared to the Day 1 outlook updates, which are use of WoFS by an operational forecaster and a consensus of non-forecaster participants.

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?

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

The SPC Impacts System is run on the Day 1 tornado outlooks with conditional intensity information to estimate the number of tornadoes by EF scale and the potential societal impacts.

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?

b. Forecast Products and Activities

There will be two periods of experimental forecast activities during SFE 2025. 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 2019-2024. 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. Forecasters can also delineate "hatched" areas, which represent 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 current coverage forecasts include intensity information that is not explicitly communicated to users, so coverage forecasts and intensity forecasts could be better labeled/communicated. These results have been used to identify four conditional intensity groups (CIG) that can be forecast via examination of the atmospheric environment: no CIG, 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, all In-Person participants will create their own Mesoscale Discussion (MD) Product using WoFS and other available observations and CAM guidance within the SFE Drawing Tool. Then, during the 3:15-4pm CDT time period, each In-Person participant will use WoFS and other available guidance to update the Day 1 individual hazard coverage and conditional intensity forecasts for the period 2100 – 1200 UTC.

During the 2:15-4pm CDT time period virtual participants will split into two groups for an activity using the newly designed, experimental WoFS viewer. Both groups will complete the same activity but will do so separately to keep the group size manageable. During the first day each week, all virtual participants will complete a short training on WoFS and the new WoFS Viewer from 2:15-3pm CDT. Then, from 3-3:45pm CDT, participants will create their own MD using the new WoFS Viewer and drawing tools in Google Slides. From 3:45-4:00pm CDT, virtual participants will complete a short survey on their experiences with the new WoFS Viewer. After the first day, the virtual participants will create their own MD from 2:15-3pm CDT, share their MDs with their group in a weather discussion from 3-3:15pm CDT, create a second MD from 3:15-3:45pm CDT, and then complete a survey on the new WoFS Viewer from 3:45-4:00pm CDT.

These WoF activities are the ninth year WoFS has been tested in the SFE to explore the potential utility of WoF 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 is composed of three overlapping programs (Fig. B1). The Experimental Forecast Program (EFP) is 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. The EFP embodies the collaborative experiments and activities previously undertaken by the annual SPC/NSSL Spring Experiments. For more information see https://hwt.nssl.noaa.gov/efp/.

The Experimental Warning Program (EWP) is 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. The EWP embodies the collaborative warning-scale experiments and technology activities previously undertaken by the OUN and NSSL. For more information about the EWP see https://hwt.nssl.noaa.gov/ewp/. A key NWS strategic goal is to extend warning lead times through the "Warn-on-Forecast" concept (Stensrud et al. 2009), which involves using

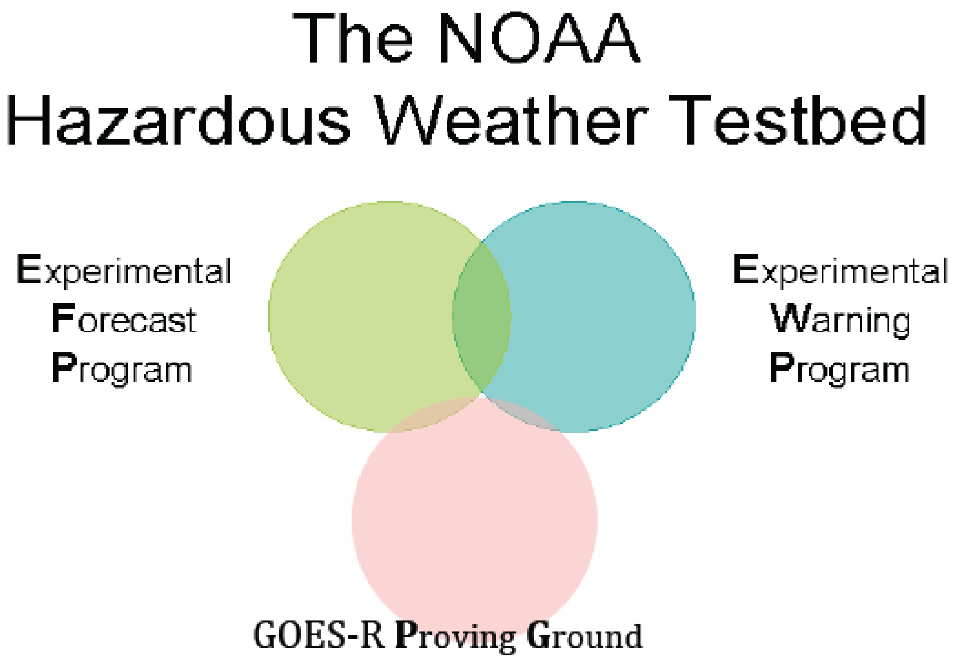

The NOAA
Hazardous Weather Testbed
Experimental Forecast Program
Experimental Warning Program
GOES-R Proving Ground

Figure A1: The umbrella of the NOAA Hazardous Weather Testbed (HWT) encompasses two program areas: The Experimental Forecast Program (EFP), the Experimental Warning Program (EWP), and the GOES-R Proving Ground (GOES-R).

frequently updated short-range forecasts (≤ 1h lead time) from convection-resolving ensembles. This provides a natural overlap between the EFP and EWP activities.

The GOES-R Proving Ground (established in 2009) exists to provide demonstration of new and innovative products as well as the capabilities available on the next generation GOES-16 satellite. The PG interacts closely with both product developers and NWS forecasters. More information about GOES-R Proving Ground is found at http://cimss.ssec.wisc.edu/goes_r/proving-ground.html.

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)
- 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 2025 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)	

Appendix C: References

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*, (Accepted).

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

Han, J., and C. S. Bretherton, 2019: TKE-based moist Eddy-Diffusivity Mass-Flux (EDMF) parameterization for vertical turbulent mixing, *Wea. Forecasting*, **34**, 869-886.

Harris, L. M., S. L. Rees, M. J. Morin, L. Zhou, and W. F. Stern, 2019: Explicit prediction of continental convection in a skillful variable-resolution global model. *Journal of Advances in Modeling Earth Systems*, **11(6)**, DOI:10.1029/2018MS001542.

Harris, L. M., and Coauthors, 2020: GFDL SHiELD: A Unified System for Weather-to-Seasonal Prediction. *Journal of Advances in Modeling Earth Systems*, **12(10)**, DOI:10.1029/2020MS002223.

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

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.

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.

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

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. *Natur*e, **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.

Sobash, R. A., and D. A. Ahijevych, 2024: Evaluating Machine Learning–Based Probabilistic Convective Hazard Forecasts Using The HRRR: Quantifying Hazard Predictability and Sensitivity to Training Choices. *Wea. Forecasting*, **39**, 1399–1415, https://doi.org/10.1175/WAF-D-23-0221.1.

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.

Zhou, L., S.-J. Lin, J.-H. Chen, L. Harris, X. Chen, and S. L Rees, 2019: Toward Convective-Scale Prediction within the Next Generation Global Prediction System. *Bull. Amer. Meteor. Soc*. DOI:10.1175/BAMS-D-17-0246.1.

Zhou, L., Harris, L., Chen, J.-H., Gao, K., Guo, H., Xiang, B., et al. (2022). Improving global weather prediction in GFDL SHiELD through an upgraded GFDL cloud microphysics scheme. Journal of Advances in Modeling Earth Systems, 14, e2021MS002971.