A Review of Operational Ensemble Forecasting Efforts in the United States Air Force

Evan L. Kuchera *, Scott A. Rentschler, Glenn A. Creighton and Steven A. Rugg

16th Weather Squadron, 101 Nelson Drive, Offutt Air Force Base, Bellevue, NE 68113, USA; scott.rentschler.1@us.af.mil (S.A.R.); glenn.creighton.1@us.af.mil (G.A.C.); steven.rugg@us.af.mil (S.A.R.)
* Correspondence: evan.kuchera@us.af.mil

Abstract: United States Air Force (USAF) operations are greatly influenced and impacted by environmental conditions. Since 2004, USAF has researched, developed, operationalized, and refined numerical weather prediction ensembles to provide improved environmental information for mission success and safety. This article reviews how and why USAF capabilities evolved in the context of USAF requirements and limitations. The convergence of time-lagged convection-allowing ensembles with inline diagnostics, algorithms to estimate the sub-grid scale uncertainty of critical forecasting variables, and the distillation of large quantities of ensemble information into decision-relevant products has led to the acceptance of probabilistic environmental forecast information and widespread reliance on ensembles in USAF operations worldwide.

Keywords: ensemble prediction systems; ensemble post-processing; ensemble products; forecast uncertainty

1. Introduction

Weather impacts United States Air Force (USAF) operations worldwide, both via threats posed to life and property and via advantages conferred (with superior understanding compared to an adversary) in the sphere of conflict.

The canonical example of the latter is the D-Day invasion [1] in June 1944 during WWII, when in the midst of an overall stormy pattern that led Nazi leaders in Europe to let down their guard, Allied forecasters found a brief window of adequate weather for attack. The Allies both reduced risks to property and safety and improved their likelihood of mission success by exploiting the opportunities gained from superior weather information.

An example of failure occurred during Operation Eagle Claw in April 1980 [2] when mission planners and meteorologists did not anticipate and account for convectively-generated dust storms that could not be seen by the satellites of that era. The zero visibility conditions were the primary factor that led to the mission being aborted, indirectly led to eight fatalities when a helicopter later crashed into a transport aircraft, and arguably cost a US President his re-election bid [3].

An unpublished study of 266 major battles from 1479 BC to 2003 AD [4] found that weather was a factor in the outcome of 36% of them, evidence that the D-Day and Operation Eagle Claw scenarios were not isolated in history. Generating superior environmental information, and also appropriately exploiting it for advantages in conflict, is therefore integral to USAF mission success.

Since the days of WWII, Numerical Weather Prediction (NWP) has provided great advancements in weather forecasting skill, but until recently weather hazard probabilities were usually inferred from deterministic NWP output by the forecaster or decision maker based on their own subjective experiences, per the US National Research Council’s report “Completing the Forecast” [5].

As stated in that report:
“The chaotic character of the atmosphere, coupled with inevitable inadequacies in observations and computer models, results in forecasts that always contain uncertainties. These uncertainties generally increase with forecast lead time and vary with weather situation and location. Uncertainty is thus a fundamental characteristic of weather...and no forecast is complete without a description of its uncertainty.”

Risk management principles addressing situational uncertainty are utilized widely for decision making in USAF. Doctrine [6] teaches that the degree of risk should be assessed by cross-checking the impact of a hazard against the probability of that hazard occurring (Figure 1). Despite the clear “probability” of occurrence language therein, weather information is nonetheless presented to decision makers deterministically, by regulation [7]. This is far from optimal. Thompson [8] discussed the cost-loss ratio as a methodology to use probabilistic information for decision making in weather, adapted to the USAF by Scruggs [9] whose statement in 1967 still echoes today:

“...some operators may refuse to accept weather forecasts worded in probabilistic terms...Unfortunately, this attitude makes the forecaster—not the operator—the decision-maker.”

The culture of deterministic thinking for weather in the USAF is rooted deeply both in the traditions of how weather information has been presented historically, and in the preference of some decision makers to push decision responsibility to the weather forecaster. This culture is contrary to the likely benefits of a probabilistic and risk-based approach to weather for reducing costs and increasing mission effectiveness [10–13]. What is the magnitude of possible benefits that may be unrealized due to these traditions?

As of 2021 the US military overall has $1.7 trillion in weapons system assets [14] and $1 trillion in facilities [15] that have varying degrees of exposure to weather risk. As a thought experiment applying the Lazo et al. [16] estimate of 3.4% weather sensitivity for US economic activity to the combined $2.7 trillion of US military assets, $89 billion could be sensitive to weather.

More tangible examples of weather exposure costs include $5 billion dollars of damage to Tyndall Air Force Base in Florida from 2017’s Hurricane Michael [17], up to $2 billion in damage to 17 F-22 aircraft undergoing maintenance at Tyndall during Michael [18], $46 million in hail damage to aircraft that were not evacuated or sheltered in advance of a hail storm at Laughlin Air Force Base in Texas in 2016 [19], and $10 million in damage to unsheltered and partially sheltered aircraft from a tornado at Offutt Air Force Base in

Figure 1. Doctrinal USAF risk assessment matrix. Adapted from: United States, Department of Defense, Department of the Air Force [6].
The value of societal benefits or losses derived from weather-sensitive mission success (D-Day) or failure (Eagle Claw) are much harder to quantify. A $934 billion dollar annual budget [21] and $2.7 trillion in combined weapons/facilities assets indicate that the societal value of achieving military objectives (using weather information when needed) is perceived to be quite high in the US.

Due to the rapid development of ensemble modeling enabled by computing advancements in the 1990s and early 2000s (see Kalnay [22] for a review) and concrete evidence that real decision making with uncertainty information leads to better economic outcomes [23,24], the USAF has invested in the development and operationalization of ensembles over the past 17 years. The methods and results sections are divided into three distinct periods of these USAF ensemble efforts, documenting the lessons learned in each. The first is the development era (2004–2008), the second is the convection-allowing (sometimes referred to as convection-permitting) modeling (CAM; 2008–2015) era, and the third is the “Rolling” ensemble era (2015-present) where a time-lagged approach prevailed, harkening back to some of the very first methods used to generate uncertainty information [25] with ensembles.

2. Development Era (2004–2008)

2.1. Methods

Following findings that diversity in regional ensemble model analyses and physics improved forecast reliability and dispersion [26–31], ideas showing the potential of pairing finer horizontal resolution and ensembles together [32], and conclusions that there was a solid theoretical basis for applying ensemble forecasts to USAF decision making processes [13,33], an effort was undertaken jointly with the US Navy (USN) to run a two-center co-located fine scale regional ensemble [34]. This was to be composed of members from the National Centers for Atmospheric Research (NCAR)’s Weather Research and Forecasting (WRF) model [35], and the USN’s Coupled Ocean Atmosphere Mesoscale Prediction System [36] but sadly, communication challenges prohibited the sharing of regional model runs.

USAF and NCAR partnered to evaluate many different ways of setting up a regional ensemble to improve forecasts and decision making. Global initial and lateral boundary conditions from the US National Centers for Environmental Prediction (NCEP)’s Global Ensemble Forecast System (GEFS [37]) were selected, while the regional model was WRF. Varied data assimilation techniques were tested, along with ensembles consisting of physics suites where each member had empirical parameters varied within bounds specified by their developers, ensembles where members had different physics suites, members where unresolved energy “backscatter” was introduced, and combinations thereof [38]. Typical ensemble metrics were examined, in addition to subjective evaluation from forecasters and users, to judge performance.

This was more than just a modeling endeavor. For the USAF the effort was end to end, from designing the best possible forecasting ensemble, to training and educating users, and jointly developing with those users the most useful extractions of information and products from the ensemble for optimized decision making [39]. As the USAF has many mission-unique environmental information needs, the ability to design the ensemble with the end-state products and decisions in mind enhanced the likelihood of success. Additional sub-grid scale, tailored end-state products were developed using inputs from the regional ensembles described above, and also from existing operational global ensembles. The initial global ensemble included the NCEP’s GEFS, the Canadian Meteorological Centre (CMC)’s Global Environmental Multiscale (GEM) [40] ensemble, and eventually USN’s Navy Operational Global Atmospheric Prediction System (NOGAPS [41]). Simple algorithms were developed and applied to outputs from each of the ensemble members (all equally weighted) to estimate smaller-scale, high-impact phenomena such as lightning, hail, snowfall, turbulence [42], and cloud ceilings that were relevant to USAF decision
thresholds. USAF primarily evaluated reliability/attribute diagrams [43] and Brier Skill Scores [44] in addition to subjective impressions of performance in a variety of case studies.

2.2. Results

Hacker et al. [38] contains a description of the myriad of regional ensemble designs that were tested. Key results included:

- Ensemble Transform Kalman Filter methods generally performed better than Perturbed Observations
- All methods of accounting for model uncertainty improved forecast skill at least marginally
- Model physics diversity was critical for increasing/obtaining forecast skill in variables in the planetary boundary layer
- Utilizing Stochastic Backscatter provided for the highest degree of skill in variables aloft
- Diverse physics, parameter perturbations, and backscatter employed together led to the most skillful ensemble

Evaluations from USAF confirmed these findings [39] and added another—that the use of GEFS for regional ensemble initial conditions (ICs) was the primary contributor to forecast underdispersion (Figure 2) for the short-term sensible weather variables that USAF was most interested in. It was hypothesized that the initial condition spread in the global ensemble was perhaps optimized for appropriate dispersion at longer lead times and for the larger scales as had been observed in the European Center for Medium-range Weather Forecasting ensemble [45], and that a new methodology for generating initial condition diversity was needed before this capability could become operational. Regardless, as expected, the regional ensemble was found to be able to resolve finer scale features that the global ensemble could not (Figure 3), promising superior forecast utility if the underdispersion problem could be addressed.

![Figure 2](https://example.com/figure2.png)

**Figure 2.** Rank histogram of 48 h ensemble wind speed forecasts from multi-physics WRF-based ensemble systems using GEFS ICs with no data assimilation (NoDA), model parameter perturbations (MP), and Ensemble Transform Kalman Filter (ETKF) data assimilation. DGFS used GEFS ICs but with the same WRF configuration for all members, serving as a baseline for comparison.
Figure 2. Rank histogram of 48 hour ensemble wind speed forecasts from multi-physics WRF-based ensemble systems using GEFS ICs with no data assimilation (NoDA), model parameter perturbations (MP), and Ensemble Transform Kalman Filter (ETKF) data assimilation. DGFS used GEFS ICs but with the same WRF configuration for all members, serving as a baseline for comparison.

Figure 3. (Top left): 42-h global ensemble probability (percent) forecast of snow exceeding 6 inches (15.24 cm) for the 24 h preceding 1200 UTC on 6 November 2008, along with the Mean Sea Level Pressure from the GEFS control member (labelled “Consensus”). (Top Right): 24-h regional ensemble (10-km) probability (percent) forecast of snow exceeding 12 inches (30.48 cm) for the 24 h preceding 1200 UTC on 6 November 2008. (Bottom): Observed snowfall for the 5–7 November 2008 period.

The global ensemble was found to perform reliably and skillfully on the larger scales that it could resolve. The technique of combining ensembles from multiple centers served both to increase membership and diversity (Figure 4), and long-range forecasts of high impact events like a blizzard during Christmas 2009 (Figure 5) were sometimes provided with a level of confidence that would never be possible solely with deterministic models, given their typical error characteristics. This result supported the initial hypothesis of Palmer and Tibaldi [46] that the forecast could indeed be forecast using ensembles. Due to these positive results, plans were made to formally operationalize this blended ensemble into the Global Ensemble Prediction Suite (GEPS).

User training and outreach was the most successful part of the development effort. A webpage was set up with real-time forecast products, and training modules were developed and socialized. Training was most successful when one person in a forecasting unit was motivated to embrace and incorporate the new ideas into his or her practice, and thus served as an information bridge between the USAF ensemble experts and the forecasters in their unit. Numerous requests for product enhancements were fulfilled, with feedback loops established for ongoing refinements to better meet needs. A skeptic of the project set out to prove it was wasted effort by exploring the (presumed lack of) webpage visits, but instead was turned into a proponent when he or she saw its popularity.
User training and outreach was the most successful part of the development effort. A webpage was set up with real-time forecast products, and training modules were developed and socialized. Training was most successful when one person in a forecasting unit was motivated to embrace and incorporate the new ideas into his or her practice, and while the improved information on forecast confidence from the ensemble was certainly a part of this success, the authors believe that the key was developers utilizing “teal” practices [47] to be responsive to and collaborate with users to create mission-tailored ensemble products. Probability products for sensible weather parameters like lightning or
wind gusts exceeding a key threshold were most popular, while statistical products like mean and spread were less so, similar to the findings of Evans et al. [48].

Despite these successes, the ensemble information was generally not making it all the way to the end decision; rather, forecasters were using it to refine their deterministic forecasts and to help them manage their time by using the ensembles to identify areas that most needed their attention. Additionally, probability interpretation issues were revealed. The ensembles generally forecasted probabilities on smaller spatial and temporal scales than the products forecasters were used to issuing and they would therefore misinterpret the ensemble probabilities as being too low. For example, the probability of winds exceeding a certain threshold were output hourly from the ensemble, but warnings for strong winds were typically issued for periods covering a multi-hour event. Any exceedance during the warning period verified it, but the ensemble output covered the probability for just a one-hour output interval, not their warning interval. This unveiled a conundrum—is it better to output at frequent intervals to enable refinement of event timing? Or better to output over longer intervals so that probabilities more closely match observed relative frequencies over typical warning time intervals? This issue was explored more in future efforts. By 2008, even though the capability was a prototype, forecasters were using it to inform their operational decision making (forecaster quotes in Kuchera et al. [39]).

3. CAM Era (2008–2015)

3.1. Methods

Following the hypothesis that GEFS ICs were under-dispersive for short term regional runs, there was a significant amount of investigation in 2009–2010 on how to best generate ICs for the mesoscale ensemble. USAF ultimately changed its configuration to use just three ICs–deterministic runs of GFS, GEM, and the United Kingdom’s Unified Model (UM [49]) which had just recently been installed and run operationally by USAF, dubbed the Global Air-Land Weather Exploitation Model (GALWEM). This IC generation methodology was identical to another being developed and evaluated at this same time for the German Weather Service (DWD [50]). The choices of WRF physics packages for the ensemble members [51,52] were changed periodically but the philosophy of combining them in unique configurations of WRF to maximize dispersion of sensible variables of interest to USAF was unwavering. For simplicity of maintenance and concern about realism, the regional ensemble did not use physics parameter perturbations or backscatter. Ensemble data assimilation, while showing promise in initial tests, had been computationally unstable during tropical cyclones and for high resolution runs and was subsequently scrapped.

These choices made the USAF regional ensemble akin to a “poor man’s ensemble” [53] where rather than systematically designing an ensemble system to account for known uncertainties, the ensemble was now basically a collection of deterministic runs. While perhaps sacrificing some forecasting skill by abandoning advanced ensemble design techniques, ease of sustainment and maintenance in operations demanded a simpler design. The most impactful event thus far in the USAF ensemble journey was the visit to the NOAA Hazardous Weather Testbed (HWT) in 2008 for their annual Spring Forecasting Experiment. Each spring forecasters and researchers gather to jointly evaluate and discuss novel models and techniques for convective forecasting. Two critical advancements for future USAF use were in their initial evaluation stages during this experiment. The first was the use of convection-allowing (i.e., convective parameterization disabled to allow the model to explicitly simulate convective processes) ensembles at 4-km resolution for thunderstorm forecasting [54–57]. The second was the use of “inline” diagnostics calculated at each time step (~30 sec) in the model integration that captured key details about the strength and character of the convection as it evolved [58–60]. Both of these addressed two critical USAF needs—improved forecasting of convection (e.g., Eagle Claw) and decision-tailored information about that convection (e.g., peak gust strength, lightning/hail potential). With assistance from the HWT group, USAF rapidly imported these capabilities.
and began generating prototype 4-km ensemble forecasts in the US and in convectively prone areas overseas like southwest Asia.

Two key challenges to operationalize the 4-km ensembles were high computational cost and writing large output files in a timely manner. The first challenge was overcome by reducing the vertical resolution of the member runs substantially, to between 21 and 27 levels depending on the member [52]. Aligo et al. [61] found that convective precipitation forecast skill was similar between 31 and 62 levels, highlighting this as an area of potential cost savings to explore and exploit. Additionally, the WRF executable was compiled to run 20% faster at the expense of bit-for-bit reproducibility. For generating ensemble members, small variations due to computational shortcuts are actually a desirable trait.

The large file challenge was addressed by utilizing the “quilting” option in WRF (which uses a designated number of processors to perform file writes at output times in the background while the model integration proceeds). In addition, the output variable list was reduced to the bare minimum by deriving input-heavy algorithms (e.g., winter precip type, radar reflectivity, cloud cover) as inline diagnostics, extending the techniques learned at HWT 2008 [62]. This eliminated a standard deterministic post-processor used for each member, as simpler derivations (e.g., wind gust, wind chill) not already derived as inline diagnostics, could be calculated in the ensemble post-processor when probabilistic threshold variables were calculated.

While NWP ensembles are designed to account for uncertainties that exist on resolved scales, uncertainties also exist on unresolved (sub-grid) scales, and these uncertainties/scales often produce sensible weather that greatly impacts USAF operations. To more reliably forecast sensible weather variables that were not explicitly resolved, the Weibull [63,64] distribution was selected to enable output of a tailorable distribution of possible values for each member’s forecast.

Take for example an algorithm to predict snow accumulation. A deterministic algorithm (e.g., 10 cm of snow for every cm of liquid precipitation output from the model) does not account for the uncertainty inherent in predicting snow accumulation with varying ice crystal habits, melting, compaction, etc. [65]. By subjectively specifying the parameters that describe a Weibull probability distribution function (PDF), this uncertainty can be estimated and accounted for to make probabilistic predictions of exceeding specific snow accumulation thresholds (e.g., a 30% chance of seeing more than 11 cm of snow with 1 cm of liquid, a 10% chance of seeing more than 12 cm, etc.) for a single ensemble member. Most probabilistic variables in the USAF ensemble post-processor were calculated this way, as outlined in Creighton et al. [62] and Creighton [66]. Many variables were assigned Weibull parameters for a Gaussian-like distribution, while others were assigned parameters to modify the distribution shape, all subjectively based on developer experience with the variable. This approach is admittedly not rigorous, but follows from Weibull [64]: “It is believed that in such cases (distribution functions of random variables) the only way of progressing is to choose a simple function, test it and stick to it as long as none better has been found.” This was the approach taken to provide a first-order estimate of sub-grid scale uncertainty for key variables. The ultimate result is that each member, rather than contributing a “yes” or “no” vote to the final ensemble probability of exceeding a threshold, instead contributes a probability that is averaged with all other member probabilities to comprise the final ensemble probability. This methodology accounts both for sub-grid scale and (as part of an ensemble) flow-dependent uncertainties and was expected to improve forecast reliability and skill.

Previous work used deterministic threshold values of diagnostic parameters (e.g., CAPE, updraft helicity) to determine if a discrete severe weather event (e.g., tornado, large hail) should be forecast from a single ensemble member [59,67,68]. In reality, the probability of occurrence of a sub-grid scale weather event will not likely jump from 0 to 1 as the environmental conditions pass a diagnostic threshold value. To attempt to address this more realistically, empirically derived algorithms were created not to predict sub-grid scale event occurrence, but instead its magnitude (e.g., maximum tornado wind speed or
number of lightning strikes in a given area over a period of time) using inputs from larger scale diagnostics. Then, by empirically shaping the Weibull PDF for those variables, any threshold probability value could be obtained (e.g., the probability of a tornado exceeding EF1 wind speed, the probability of more than 5 or 10 lightning strikes in a given area over a period of time) for each ensemble member. Details can be found in Creighton et al. [62].

Additionally, forecasting a very rare event like a tornado in a small area (i.e., a 4-km box) at a given time will almost always result in very low probabilities that can be difficult to apply to decision making (as was learned in user feedback during the development era). Thus, probabilities for lightning, hail, supercell, and tornado were artificially upscaled to decision-tailored ranges of occurrence within 10 and 20 nautical miles (NM). Since there was an unknown degree of dependence on the probability of occurrence from grid box to grid box in the larger area, the probability in the larger area for each member was calculated twice. First, all grid points were assumed to be completely independent (i.e., the probability of occurrence is one minus all of the probabilities of “no” occurrence in the radius multiplied together), and then they were assumed to be completely dependent (i.e., the maximum probability in the upscaling radius is the probability for the total area). These two probabilities were then averaged to obtain the final probability of occurrence for that member in the 10 or 20 NM radius. Probability values from each member were then simply averaged to calculate the total ensemble probability in the radius.

Parallel efforts were undertaken during this time to improve dust forecasting due to ongoing hostilities in Iraq and Afghanistan. Source region and saltation/lofting improvements highlighted in Hunt et al. [69] and LeGrand et al. [70] were combined with convection-allowing ensemble runs in an attempt both to highlight areas where convective outflows would generate severe dust storms in Iraq, and where airborne dust would flow in areas of highly variable terrain in Afghanistan.

In this era, frequent requests for new mission-specific probability thresholds highlighted that an on-demand capability was required to provide the potentially limitless combinations of variables, thresholds, and time periods that could be of interest. A prototype capability called iPEP (interactive Point Ensemble Probability, an enhancement of an existing static PEP product) was developed that consisted of a front-end interface that allowed a user to choose their variable, threshold, and time period of interest (solving the output interval dilemma discussed earlier), and a back-end that would extract the relevant variables from all ensemble members and make tailored probabilistic calculations based on the inputs. It also could generate joint probabilities of seeing all of 2 or more thresholds met, or any of 2 or more thresholds met, depending on the request. Details on the methodology including statistical correlation assumptions for multi-variable joint probability calculations can be found in Creighton [66].

3.2. Results

The combination of improved diversity in ICs, convection-allowing resolution, and reliable algorithms tailored to mission impacts resulted in a regional ensemble capability that was well-received by users, and showed useful skill across many variables (examples follow). Together with GEPS, this new Mesoscale Ensemble Prediction Suite (MEPS) was operationalized on 28 March 2012 [52], one of several CAM ensemble systems being operationalized around this time [71,72]. Four case studies are selected to highlight how MEPS provided environmental intelligence guidance for key events during this era.

3.2.1. US Tornado Outbreaks—April 2011 and 2012

In February 2011, a relocatable 4-km mesoscale ensemble covering a portion of the US had been designed using ICs from deterministic runs of global models (as outlined earlier) instead of members of GEFS. This ensemble was run in the southeastern US in advance of the deadliest day for tornadoes in 75 years [73]. The CAM ensemble was able to simulate updraft helicity tracks [58], and in conjunction with the newly developed USAF tornado algorithm [62] was able to produce hourly probabilities of seeing a tornado within a 20 NM
radius (Figure 6). Analysis of this event (Table 1) found that non-zero tornado probabilities were forecast for every tornado that occurred in the model domain within 20 nautical miles of its occurrence on 27 April 2011. Hourly probabilities exceeding 20% were observed over a broad area with 42 h lead time (Figure 7), providing significant notice for decision makers to take protective action.

![Figure 6](image-url)

**Figure 6.** PEP bulletin from the 27 April 0600 UTC run of the 4-km regional ensemble valid for Birmingham, AL. Variables listed on the left hand side, with probabilities (percent) of exceedance for each valid time in rows to the right of the label. Yellow and red coloring designed to draw user attention to higher probabilities. Note the elevated potential for lightning, strong surface winds, large hail, and tornadoes.

| Date     | Tornado Reports | Misses (0% Forecast) | Hit Rate | Average Forecast Probability Per Tornado |
|----------|-----------------|----------------------|----------|-----------------------------------------|
| 22 April | 28              | 1                    | 0.964    | 0.066                                   |
| 23 April | 4               | 0                    | 1        | 0.019                                   |
| 24 April | 11              | 0                    | 1        | 0.024                                   |
| 25 April | 34              | 5                    | 0.853    | 0.068                                   |
| 26 April | 44              | 15                   | 0.659    | 0.074                                   |
| 27 April | 153             | 0                    | 1        | 0.166                                   |
| 28 April | 5               | 0                    | 1        | 0.027                                   |

Table 1. 4-km ensemble tornado probability forecast (within 20 nautical miles for the preceding hour) performance for 22–28 April 2011. Forecasts generated 0600 UTC, valid for the 2100-0300 UTC (39–45 h lead time) period. Preliminary observed tornado reports as compiled by the US National Weather Service.
Table 1. 4-km ensemble tornado probability forecast (within 20 nautical miles for the preceding hour) performance for 22-28 April 2011. Forecasts generated 0600 UTC, valid for the 2100-0300 UTC (39–45 hour lead time) period. Preliminary observed tornado reports as compiled by the US National Weather Service.

| Date     | Tornado Reports | Misses (0% Forecast) | Hit Rate | Average Forecast Probability Per Tornado |
|----------|-----------------|----------------------|----------|------------------------------------------|
| 22 April | 28              | 1                    | 0.964    | 0.066                                    |
| 23 April | 4               | 0                    | 1        | 0.019                                    |
| 24 April | 11              | 0                    | 1        | 0.024                                    |
| 25 April | 34              | 5                    | 0.853    | 0.068                                    |
| 26 April | 44              | 15                   | 0.659    | 0.074                                    |
| 27 April | 153             | 0                    | 1        | 0.166                                    |
| 28 April | 5               | 0                    | 1        | 0.027                                    |

Figure 7. 42-h 4-km mesoscale ensemble probability (percent) of a tornado within 20 nautical miles (37 km) along with maximum hourly 2–5 km above ground level updraft helicity exceeding 50 m²/s² from every ensemble member within the preceding hour valid on 0000 UTC 28 April 2011. Observed tornado reports (center of white triangles) as compiled by the US National Weather Service from 2300-0000 UTC 27 April 2011.

Partly due to the skill demonstrated in the 2011 event, when a subsequent event threatened in April 2012 (Figure 8) after 4-km MEPS was operational, aircraft valued at $640 million were evacuated at significant cost from McConnell Air Force Base [74] in the morning when a high-impact tornado event was forecast later in the day. A tornado struck the base [75] late that evening.
Figure 8. (Top): NOAA Storm Prediction Center outlook for 1200 UTC 14 April 2012 to 1200 UTC 15 April 2012 issued at 0543 UTC on 14 April 2012 along with US National Weather Service preliminary observed tornadoes (red dots) from 1200 UTC 14 April 2012 to 1200 UTC 15 April 2012. Slight risk corresponds to a 2% chance of a tornado within 25 nautical miles during the valid period, moderate a 15% chance, and high a 30% chance. (Bottom): 36 to 60 h forecast of tornado probabilities (percent) within 20 nautical miles from the 0000 UTC 13 April 2012 run of 4-km MEPS (valid 1200 UTC 14 April to 1200 UTC 15 April). 24-h probabilities calculated using the hourly probabilities, assuming complete independence from hour to hour.

3.2.2. Iraq Convective Dust—19 September 2009

On 19 September 2009 around 1500–1600 UTC, convective outflow (Figure 9) with a maximum observed gust of 27 m/s lofted substantial dust (pink dust enhancement from Ashpole and Washington [76]) leading to zero visibility conditions in and around Balad Air Base, Iraq. An aircraft attempting to land crashed killing Spc Michael Shane Cote Jr, and injuring 12 others. USAF had just set up an experimental 4-km ensemble domain over Iraq in August 2009 with GEFS ICs. Though still underdispersed, about half the members indicated convection and outflow winds similar to what was observed (Figure 9), while the other half failed to generate any convection (not shown). Dust was not explicitly modeled in the CAM members at this time. Operational dust forecasting tools were fed by winds
from coarser, non-CAMS and thus did not show any dust potential (not shown). A solely deterministic CAM run in this case may not have captured the convection and outflow winds given the variation in ensemble forecasts.

Figure 9. (Left): Dust enhanced infrared satellite from METEOSAT-9 valid 1500 UTC on 19 September 2009. Balad Air Force Base noted with the yellow star. (Right): Sustained wind speed forecast from member 1 of the 1800 UTC 18 September 2009 4-km regional ensemble valid at 1500 UTC on 19 September 2009. Light gray is 10–15 knots, dark gray is 15–20 knots, light purple is 20–25 knots, blue is 25–30 knots, green is 30–35 knots, and yellow is 35–40 knots.

This case was the catalyst to use the WRF-CHEM [77] model with six dust size bins turned on [70] to evaluate CAM performance for convectively generated dust storms. Results (Figure 10) for dust with a 4-km CAM run were realistic and substantially different from the otherwise equivalent 12-km run with convective parameterization turned on. This advantage stands in contrast to the lack of surface observations in this region, and that convective anvils can block satellite diagnosis of dust. Thus, the ability to model both the initiation of convectively generated dust, and also its evolution long after the convection has ended, has been a valuable modeling capability for USAF since its inception in operations in 2012, and will hopefully ensure another Operation Eagle Claw failure [2] does not happen again.

Figure 10. Comparison of 100 m above ground level dust concentration (ug/m^3) between a 4-km WRF run with convective parameterization turned off (left) and a 12 WRF run with convective parameterization turned on (right). Both models initialized with analyses valid 1800 UTC on 18 September 2008, and forecasts valid 1900 UTC 19 September 2008.
3.2.3. Derecho—29 June 2012

On 29 June 2012 a derecho developed in Indiana and moved at 27 m/s east-southeast to the Atlantic coast causing 4 million power outages and 13 deaths over a 1000-km path [78]. Another 34 deaths were attributed to heat in areas where power outages were prolonged. Warm season derechos are often weakly forced [79] making them a difficult forecast challenge. A review by Furgione [78] noted that this event was “not well forecast in advance” because the primary US operational models at that time were not convection allowing. In the USAF, the 4-km MEPS in operations over CONUS had several members (but only a minority) simulating a derecho with 17 h of lead time (Figure 11). Note that the ensemble also correctly indicated a severe wind threat from another system that developed in the wake of the derecho, and had a “false alarm” derecho to the southeast of the main one that occurred. This case highlights the criticality of both having CAMs to enable the simulation of complex upscale convective growth, and also an ensemble to account for the sometimes broad envelope of possible non-linear atmospheric responses to subtle forcing.

Figure 11. Probability (percent) of wind gusts exceeding 65 knots (33.4 m/s) from the operational 4-km MEPS run initialized at 0000 UTC on 29 June 2012 valid at 2100 UTC the same day. Two observed wind report time/locations noted on the image. Inset radar courtesy of the College of DuPage website valid 2054 UTC.

3.2.4. Afghanistan Dust in Variable Terrain

While CAMs are intended to simulate convective processes more explicitly, the finer resolution is also valuable for simulating complex terrain flows. On 8 August 2011, an area of dust moved southwestward from the northern plains area of Afghanistan into the Amu Darya river basin, disrupting USAF efforts during Operation Enduring Freedom. Weather forecasters supporting those efforts recorded that many ensemble members (now using...
WRF-CHEM to model dust) accurately simulated this flow of dust, as indicated by the 30–50% probability of visibility less than 3 miles forecast (Figure 12). Those forecasters noted that no other capability (e.g., Barnum et al [80]) showed any dust in this area, attributed to the 4-km resolution of the underlying terrain (Figure 12). As in the other cases noted earlier, an ensemble was necessary as not all members simulated the event of interest accurately, but very fine resolution was also necessary to simulate the physical processes in terrain. Forecasters reported that there had been four such operationally meaningful visibility-reducing events that summer that only the 4-km ensemble had been able to anticipate [81].

Figure 12. Forecaster-submitted review of a visibility-reducing dust event in Afghanistan. Top are 18-hour visibility forecasts from a post-processing dust advection model (Barnum et al. 2004) using GFS (left) and 15-km WRF (right) inputs. Bottom left is the 24-hour probability (percent) of visibility less than 3 miles (4.8 km) from the 4-km MEPS with explicit dust modeling. Bottom right is a terrain map of the area of interest in Afghanistan with military locations of interest. All forecasts valid at 1800 UTC on 8 August 2011.

3.2.5. External Studies

During the period 2012–2015, USAF contributed its 4-km MEPS over the US to the HWT Spring Forecasting Experiment, where it was evaluated alongside other ensembles being developed and evaluated that were all shown to have useful skill in predicting severe weather [82]. Of six CAM ensembles evaluated in 2015, MEPS scored in the middle of the group [83], anywhere from 2nd to 5th depending on the metric. 4-km MEPS was also evaluated in the Hydrometeorological Testbed at the Hydrometeorological Prediction Center in early 2012, where value was noted in resolving localized snowfall and cold air
damming details in areas of variable terrain, but a high bias in QPF and snowfall totals was also reported [84]. Adams-Selin [85] examined characteristics of each ensemble member for a strongly forced squall line, noting high sensitivity in the strength and evolution of the squall line for different planetary boundary layer physics schemes, despite the strong forcing. Ryerson [86] found that for an ensemble system designed to closely match the one used by USAF, a nighttime warm bias led to an ensemble-wide lack of near-surface cloud water in many observed fog situations, but that post-processing could largely remedy the issue. Guyer and Jirak [87] also looked at US severe weather cases in the cool season for MEPS and another CAM ensemble [88], noting that both were able to provide important details on timing and intensity of high-impact severe weather events. Clements [89] evaluated GEPS and MEPS performance from April to October 2013, noting that winds and precipitation were more skillfully forecast in the higher resolution ensembles (especially in areas of variable terrain) and that the lightning algorithms were over-forecasting lightning probabilities due to a night-time high bias at the locations tested. Homan [90] found that the GEPS ensemble mean reduced the error on forecast inputs to long-haul flight fuel loading calculations by 10% for forecast hours 12–36 as compared to the GFS, enabling less contingency fuel to be loaded, lowering the aircraft weight and leading to fuel/cost savings.

3.2.6. Internal Studies

An IPEP prototype (Figure 13) was evaluated by a focus group on limited test computing resources. There has been widespread recognition by forecasters (e.g., US Army operations, remotely-piloted aircraft, base asset protection, US Aviation Weather Center) and operational leaders of the benefits of precisely tailoring ensemble information to specific mission decision thresholds. Still, sufficient resources to scale up to a fully operational capability have not yet been invested by USAF.

![Figure 13](https://example.com/figure13.png)

**Figure 13.** Interface for the iPEP capability, with tailored probability (percent) thresholds in the left column, and resultant probabilities in the rows on the right, for the approach of Hurricane Irma in Miami, FL in 2017.

An example of using the Weibull distribution to more reliably and skillfully predict PDFs of sub-grid scale phenomena can be seen for the prediction of 10-m winds and gusts over land [62]. The shift parameter (where the PDF for the variable begins if not at zero) is simply the maximum sustained wind from the model over the output period (i.e., the gust has to be higher than the sustained wind). The shape parameter is 3, which is Gaussian but with some right skewness. The scale parameter is the maximum sustained wind raised to the 0.75 power, which causes a slow decrease in the gust factor as the sustained wind gets larger, approximating the observations found in Davis and Newstein [91]. Brier Skill Scores were higher and forecasts more reliable for wind gusts (25 knot threshold) using the Weibull parameters compared to sustained winds (15 knot threshold) using the uniform ranks method (Figure 14) for a one month period for the 4-km ensemble in 2011.
Figure 14. Reliability/Attribute diagram for the 4-km MEPS probability (percent) of 10-m sustained winds (left) exceeding 15 knots (7.7 m/s) and 10-m wind gusts (right) exceeding 25 knots (12.9 m/s) compared to 10-m wind observations in the US from 20 January 2014 to 19 February 2014.

An example of the probability upscaling of severe weather variables can be seen in Figure 15, where the probability of one or more lightning strikes within 4-km (algorithm patterned after results from McCaul et al. [92]) was upscaled to 10 and 20 nautical miles. This methodology has not been robustly evaluated, but the probability values have seemed reasonable in subjective comparisons to forecaster generated probabilistic forecasts from the US Storm Prediction Center (Figure 15).

Figure 15. (Top Left): Probability (percent) of lightning within 4 km from the 4-km MEPS. (Top Right): Probability (percent) of lightning upscaled to within 10 nautical miles (18.5 km) from the 4-km MEPS. (Bottom Left): Probability (percent) of lightning upscaled to within 20 nautical miles (37 km) from the 4-km MEPS. All probabilities shown from an ensemble initialized 1000 UTC on 11 March 2021, valid from 2100 to 2200 UTC on 11 March. (Bottom Right): US Storm Prediction Center thunderstorm probability (within 12 miles or 19.3 km of a point) issued at 1352 UTC on 11 March 2021, valid from 2000 UTC 11 March to 0000 UTC 12 March.
3.2.7. User Feedback

During this period, users provided significant feedback on how ensembles were benefiting their decision making [51]. For instance, one unit reported that grounded aircraft during a dust storm were costing half a million dollars per day, but that the mesoscale ensemble suggested a narrow calmer period that allowed them to be successfully evacuated. Another unit reported a 16% improvement in warning issuance and a 24% reduction in false alarms after using 4-km MEPS during their convective season as compared to the previous season. Efforts made during the initial development phase to engage and involve users in design and testing led to greater acceptance of the capabilities, and a willingness to share results (both good and bad). The CAMs proved to be useful for convective forecasting, and for USAF forecasters working in areas without radar, the CAM output was often used as a surrogate for unobservable storm characteristics even as the convective event was happening [93].

4. Rolling Ensemble (2015–Present)

4.1. Methods

During this era, more effort was put into systematically gathering and evaluating user perceptions, with formal surveys given along with explorations of product usage by examining web server data. Surveys were given via a link on the operational products webpage and users self-selected to respond. Questions covered what missions were supported and how those missions wished to receive forecast information (deterministically, probabilistically, or a mixture), and what modeling suites, products, and variables were most useful.

A novel idea was initially proposed in May 2013, and implemented on 20 July 2015, called a “Rolling” ensemble. This methodology leverages time lagging by running just a single member every 2 h, but including runs from the last 30 h to create ensemble products. This is attractive operationally for a number of reasons. Foremost is the ease of scheduling on the supercomputer. Each 2-hourly domain gets a fixed number of processors based on how long it takes to integrate the full model forecast in 2 h. Another benefit is the ability to weather an occasional system outage without a significant impact to the full ensemble (i.e., only losing one or two runs of 16). Trending information is also useful, as the results of more recent runs can be contrasted with the older runs in the ensemble. Finally, the ensemble always contains current information given the release of new runs every two hours, although the global initial and lateral boundary conditions are only updated every six hours (note—the capability initially had 3DVAR data assimilation which made each run truly current but this was stopped due to resource limitations). This methodology is operational today for USAF in addition to a similar methodology for UKMO [94], with two large domains covering most of the world at 20-km resolution run to 132 h, and fixed/relocatable 4-km domains run to 72 h.

To respond to user requests for improved short term convective forecasting to support training, a time-lagged rapid refresh ensemble using the High Resolution Rapid Refresh [95] system for ICs was evaluated over USAF training ranges [96] during the 2017 summer. These ranges are in areas of variable terrain, and increased accuracy in forecasts of timing/location of convection and attendant threats would enable the USAF to perform more training safely. Three WRF runs with varied physics at 3 and 1-km were initialized from the HRRR 1-h, 2-h, and 3-h old forecasts each hour (details in Kuchera et al. [96]) and then combined with equivalent runs from the previous 4 h to create a 12-member time-lagged “Rolling” rapid refresh ensemble forecast to 12 h.

During this period, USAF operations implemented hybrid ETKF-4DVAR data assimilation (DA) [97] into its deterministic global GALWEM runs. The initial condition perturbations used for DA provided ICs for a 40-km resolution, 70-level global ensemble (GALWEM-GE) out to 16 days that was implemented on 4 November 2020. Design of this ensemble was patterned off the NAEFS [98] ensemble (20 members plus a control run...
twice daily to 384 h with half-degree output) to enable sharing and blending the ensembles together for improved operational products.

4.2. Results

Surveys and webpage product analyses were performed in 2016 [99] and 2020. 100 surveys were completed from August to October 2020, with users reporting that 89% of supported missions wanted a combination of “yes/no” and “confidence” information, that the 4-km MEPS was a more critical forecasting tool than the 20-km MEPS or GEPS (Figure 16), and that probability threshold information was more important than the individual ensemble members or mean/spread products (Figure 16). Additionally, lightning was cited as critical by 74% of survey respondents, with no other variable reaching 50% critical (Figure 16). For the first 6 months of 2020, 6020 unique users who authenticated into the USAF website used an ensemble product, with 87.8 million overall product downloads. 46.9 million of these downloads were for a 4-km MEPS product, and 95% of users viewed a PEP product, the most popular single product in the suites.

Figure 16. Ratings distribution for 100 survey responses received from August to October 2020 to the question “Please rate each of the following on how important it is to your efforts” where the “following” is each suite, product, and variable listed.

The initial evaluation of the Rolling ensemble came during a military exercise with South Korea [100]. The squadron providing weather support used both the test Rolling ensemble output and the operational output for the exercise, comparing and contrasting utility. The squadron commander reported that “my biggest take-away from the experiment was the ability to see trends (both geographic coverage and probabilities) over time at a specific valid time” and that he “hoped the lagged ensemble gains traction” due to his perception of its benefits [101]. To facilitate examining these trends, a web interface was developed with a double-slider bar (Figure 17) where one loop was the forecast hour, and the second loop the member of the ensemble or the ensemble products themselves...
(which were updated every two hours by incorporating the newest member, and removing the oldest). By looping through changes in the lagged forecasts on a fixed forecast hour, trends could be discerned quickly. Additionally the PEP bulletin was modified to denote where probabilities had either increased or decreased by more than 15% in the previous 12 h (Figure 18).

Figure 17. An example of the USAF “Stamp Chart” where instead of each member being laid out like postage stamps, a double-slider bar is used where the top looper displays all ensemble members valid at the selected time. In this example a “Rolling” ensemble member from a 6-h old cycle is displayed valid 18 h into the future, relative to the most current run.

Evaluation of winds over a test domain in CONUS in spring 2015 comparing the operational to the Rolling methodology indicated similar forecasting skill, with a very small improvement for the Rolling (Figure 19). Interestingly, precipitation skill increased in the earlier forecast hours but then decreased at later hours (Figure 20), perhaps indicating that the Rolling methodology was mitigating spin-up issues [102] with older runs included in the earlier forecast hours, but then reducing skill by including forecasts where error growth was dominating in later hours. Hepper [103] evaluated snowfall forecasts from three different members of the Rolling ensemble both with and without 3DVAR data assimilation, and found that it made little difference in the forecasts, with the model ICs and physics variations making a much more substantial impact. A weather squadron responsible for forecasting in the southeastern US reviewed 29 cases during the 2015 severe weather season and found both the operational and Rolling methodologies were similar and equally skillful [104]. Burns [105] found that for the 20-km MEPS, the new Rolling methodology was generally better than the previous methodology for ceiling, lightning, precipitation, and wind forecasts. All in all, the decision to implement Rolling came more from the pragmatic reasons cited earlier than from the minor forecast improvements seen in evaluations. Subsequent to implementation, Melick et al. [106] studied a tornado outbreak
in the southeast US and found MEPS updraft helicity forecasts to be skillful using the fractions skill score (FSS) [107] for neighborhood areas with 40-square-km and larger.

Figure 18. PEP bulletin from the 14 February 2021 1400 UTC run of the “Rolling” MEPS valid for Minneapolis, MN. Variables listed on the left hand side, with probabilities of exceedance for each valid time in rows to the right of the label. Yellow and red coloring designed to draw user attention to higher probabilities. Maroon (blue) box outlines indicate the rolling ensemble has increased (decreased) in probability by at least 15% in the last 12 h of model runs. Note the elevated risk of snow, somewhat strong surface winds, and reduced visibility in the MEPS forecasts.

Figure 19. Brier skill score for 10-m wind gusts exceeding 25 knots (12.9 m/s) for the period 1 April 2015 to 28 May 2015 for the operational (at that time) 4-km MEPS compared to 10-m wind gust observations in the US (CONUSD; in red) versus the “Rolling” ensemble (CONUSDIAG; in blue). Error bars represent the 95% confidence interval for the skill score.
For the rapid refresh experiment [96], 22 forecaster surveys were completed. 91% rated the rapid refresh forecasts at least somewhat better than their existing tools (which included the operational 4-km MEPS), and 55% rated them significantly better. When comparing their training results to the previous convective season, one group of forecasters reported a 36% decrease in cancellations, and another group of forecasters reported an average of 2 more flying hours per day. This could be attributed just to changes in the weather between the two years, but the impression of the forecasters was that the rapid refresh information did contribute significantly to the improvement. While the rapid refresh component of these findings is still under investigation and development, 1-km MEPS domains were implemented over the Korean Peninsula, parts of Japan and Alaska, and in the Cape Canaveral area in March 2021 using the Rolling ensemble design to improve forecasting in areas of variable terrain, but with forecasts only to 30 h instead of 72 (Figure 21).

A one-month GALWEM-GE evaluation was performed in November 2017, comparing GEPS alone to GEPS with GALWEM-GE included in its ensemble [99]. Slight improvements were noted for all variables evaluated, with an average 4.4% improvement in the continuous ranked probability scores (CRPS) [108,109]. Another evaluation occurred in the summer of 2020 (Figures 22 and 23) that repeated the same consistency in slight skill improvement for CRPS scores, informing the decision to implement later that year.
Figure 21. Sample forecast of 1-km MEPS wind gust probabilities (percent) exceeding 35 knots over the Korean Peninsula.

Figure 22. Continuous Ranked Probability Scores for GEPS (blue) and GEPS with GALWEM-GE included (orange) for 250 hPa wind speed as compared to the corresponding UM analysis for the period 28 July 2020 through 28 August 2020. Error bars represent the 95% confidence interval for the CRPS values.
Figure 22. Continuous Ranked Probability Scores for GEPS (blue) and GEPS with GALWEM-GE included (orange) for 250 hPa wind speed as compared to the corresponding UM analysis for the period 28 July 2020 through 28 August 2020. Error bars represent the 95% confidence interval for the CRPS values.

Figure 23. Continuous Ranked Probability Scores for GEPS (blue) and GEPS with GALWEM-GE included (orange) for 700 hPa relative humidity as compared to the corresponding UM analysis for the period 28 July 2020 through 28 August 2020. Error bars represent the 95% confidence interval for the CRPS values.

5. Conclusions

Nearing the 10-year mark of operational ensemble modeling at USAF, a number of lessons and themes have emerged. The convergence of CAMs with inline diagnostics, algorithms to estimate the sub-grid scale uncertainty of critical forecasting variables, and the distillation of large quantities of ensemble-generated information into decision-relevant products has led to widespread reliance on ensembles (particularly 4-km MEPS) in USAF operations worldwide. Useful ensemble products tailored to the phenomena that drive decisions (e.g., probability of lightning as opposed to mean/spread of an instability diagnostic) that have to be made have helped evolve the USAF culture towards acceptance of probabilistic environmental forecast information. Rationales for critical choices that were made in three major aspects of development will be discussed in the following sections, along with a few comments on future efforts.

5.1. Ensemble Design

Du et al. [110] provides a comprehensive review of the diversity of methodologies for ensemble design. Focusing on operational CAM ensembles, two basic and contrasting approaches are currently in use. The first approach utilizes a single model and data assimilation system, with attempts to account for all relevant uncertainties therein. Walters et al. [111] describes some benefits of this approach which is employed for the UKMO, DWD [50], and MetCoOp (Finland/Norway/Sweden [112]) operational CAM ensembles:

“By studying the same model formulation across a range of timescales and system applications, one can learn about the rate of growth and nature of both model errors and desirable behaviours. Also, by constraining configurations to perform adequately across a wide variety of systems, scientists can be more confident that
model developments seen to improve performance metrics in any one system are doing so by modelling a truer representation of the real atmosphere.”

This is contrasted with systems using multiple models for ICs and to comprise the ensemble model runs themselves in the operational HREF [60] CAM ensemble and in the MEPS described herein for USAF. Du et al [110] cites increased spread, bias cancellation, and reduction of systematic errors that are often present in single model systems as reasons to choose this approach. Drawbacks include unequal skill amongst members, bias clustering by model, and the increased cost of maintaining and ensuring compatibility between multiple models. Still, the multi-model approach has been pragmatic and easily adaptable to the fast-changing operational situations USAF is faced with, and as such has been the method of choice. From Hansen [113]:

“Even if a given multimodel ensemble is unable to bound truth, if each ensemble member is consistent with its model attractor, the ensemble’s distribution can provide information about the sensitivity of regions of state space to the different models making up the ensemble.”

This imperfect highlighting of sensitivity has proven valuable to USAF forecasters, as systematic biases in single model systems may not identify areas of uncertainty. When forecasters are alerted to areas with uncertainties, they can prioritize their limited time to more deeply analyze them, and ignore areas of greater certainty with confidence. Finally, USAF experience has not borne out the oft proclaimed difficulty in maintaining a multi-model system, which is likely due to the robustness and versatility of the WRF modeling framework and components. Since USAF does not develop its own NWP models, many of the meaningful benefits cited by Walters et al. [111] are not applicable.

5.2. Scientific Rigor in an Operational Environment

Another aspect of the forecasting mission in the USAF is the question of scientific rigor when developing new capabilities and techniques to meet ever-changing operational requirements. Ideally, one would want to thoroughly evaluate new methods but these evaluations take time and resources while a mission need is going unmet. Where is the point of diminishing returns for operational benefits with further rigor, and how should that point be determined? This point of view is especially valid when the lack of observational truth data in many areas of the world increases the cost and reduces the benefit of rigor. USAF rigor has varied from significant development and testing over a period of years for explicit dust modeling, to creating and empirically tuning a dust lofting potential index over the course of just an afternoon. A key lesson has been that by using the knowledge and experience of scientists, along with the perspectives of users, it is worth considering if a simple solution is available and likely effective that could potentially meet a need. The employment of rigor is a cost-benefit analysis in an operational environment. The costs of more complexity or more evaluation have to be weighed against the risk of not providing something to benefit a potentially pressing (and possibly transient) operational need.

5.3. Post-Processing

One area where other organizations [114] have focused but USAF has not is statistical correction of ensemble output. One reason has been that many of the priority variables USAF forecasts do not have enough corresponding quality observations, either due to the difficulty of sensing the variable or the lack of available sensing in USAF operational areas. Another reason is that bias correction is not addressing the underlying cause of model bias which almost certainly reduces the overall skill of the model simulation. Investing development resources into addressing the root causes of model bias may be a more cost-effective way to address forecast quality in the long term. There is also concern about bias correcting the inputs to the algorithms used to tailor information to decision making. Algorithms need raw, physically consistent data as inputs before any bias corrections can be made. For instance, a bias correction to low level temperatures could create convective
instability in an area where the physical model produced convection and rain-cooled, stable low-level temperatures (Figure 24) when those corrected low-level temperatures are used in the calculation of instability diagnostics. To avoid this possibility corrections should be made only on the variable of interest, not on the inputs used to calculate the variable of interest. Finally, there is risk that perceived improvements from bias correction are illusory because of poorly chosen metrics. Many variables, like 500 hPa height, are useful for verification of overall model performance, but may not be relevant to decision-making. To evaluate the true usefulness of a bias correction, it is important to design metrics appropriately so that statistical improvements do not “teach to the test” rather than demonstrate real value for improved decision-making.

USAF is interested in furthering its post-processing efforts by enabling more decision-specific product tailoring (e.g., iPEP). One use case is the optimization of a flight path by finding the path in space and/or time that has the lowest risk. As seen in Figures 25 and 26, to calculate this accurately each ensemble member is required or the probability estimates could be significantly in error. Bandwidth limitations with large CAM ensemble datasets also pose a challenge, which some have proposed to solve by creating statistical summaries of the full ensemble (see Vannitsem et al. [114]). For USAF use-cases (e.g., flight paths, line of sight calculations, joint probabilities), those methodologies could be insufficient. Instead, computing solutions that enable tailored “what is the risk for my situation?” data extractions based on user inputs from full ensemble member datasets that retain their space/time correlations would appear to be a more optimal approach to the bandwidth issue.
Figure 25. Five hypothetical ensemble convection forecasts, with two hypothetical flight paths. Despite there being a 20% chance of convection at a single point in time over a large area, one flight path has a 100% chance of being impacted by convection due to the ensemble certainty in existence and orientation of convection.

Figure 26. Five hypothetical ensemble convection forecasts, with two hypothetical flight paths. Each has a 20% chance of being impacted due to uncertainty in convection happening rather than uncertainty in location.

5.4. Future Efforts

USAF intends to continue to leverage its Rolling ensemble paradigm in operations, and extend it to rapid refresh (hourly or greater) capabilities. The multi-model approach for ensemble design remains preferred and USAF will be studying and evaluating existing
and emerging models both for providing initial conditions, and for comprising the ensemble membership, with no specific models targeted save continuing to use WRF for the foreseeable future. USAF also is evaluating ways to participate in collaborative projects for ensemble post processing using IMPROVER [115] and verification using the Model Evaluation Toolkit (MET) and METplus [116,117]. Finally, USAF is interested in exploring the potential of cloud computing resources to enable situationally-dependent model/ensemble design to best address a specific forecast challenge and its importance [118] through finer/coarser resolution, more/less members, etc., as appropriate.

Author Contributions: Conceptualization, E.L.K. and S.A.R. (Steven A. Rugg); Methodology, E.L.K., S.A.R. (Scott A. Rentschler), G.A.C. and S.A.R. (Steven A. Rugg); Software, S.A.R. (Scott A. Rentschler), G.A.C. and E.L.K.; Validation, E.L.K.; Formal Analysis, E.L.K.; Investigation, E.L.K.; Resources, S.A.R. (Steven A. Rugg); Data Curation, S.A.R. (Scott A. Rentschler) and G.A.C.; Writing—Original Draft Preparation, E.L.K.; Writing—Review & Editing, E.L.K., S.A.R. (Scott A. Rentschler), G.A.C. and S.A.R. (Steven A. Rugg); Visualization, E.L.K., S.A.R. (Scott A. Rentschler) and G.A.C.; Supervision, S.A.R. (Steven A. Rugg) and E.L.K.; Project Administration, E.L.K.; Funding Acquisition, S.A.R. (Steven A. Rugg). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: In addition to the collaborators whose work was cited herein, the authors wish to acknowledge the many and substantial contributions of the following individuals in the areas of modeling, software/product development, and verification over the course of the USAF ensembles effort: Stephen Augustyn, Gordon Brooks, Robert Craig, Jeff Hamilton, James “JimBob” Hughes, Jay Martinelli, Mickey Mitani, Julie Schramm, Willy Sedlacek, Mike Sestak, Matt Sittel, Jerry Wegiel, Braedi Wickard, and Adam Wilson. Also, the authors wish to specifically thank Jay Martinelli, Chris Melick, and Willy Sedlacek, and three anonymous reviewers for their thorough reviews and suggestions which greatly improved the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Cox, J.D. Storm Watchers: The Turbulent History of Weather Prediction from Franklin’s Kite to El Niño; John Wiley: Hoboken, NJ, USA, 2002.
2. Benson, J.T. Weather and the Wreckage at Desert-One. Air Space Power J. 2007. Available online: https://www.airuniversity.af.edu/Portals/10/ASPJ/journals/Chronicles/benson.pdf (accessed on 21 February 2007).
3. Kamarck, E. The Iranian Hostage Crisis and Its Effect on American Politics. Brookings. 4 November 2019. Available online: https://www.brookings.edu/blog/order-from-chaos/2019/11/04/the-iranian-hostage-crisis-and-its-effect-on-american-politics/ (accessed on 12 March 2021).
4. Nobis, T. Conclusions of Weather in Battle History Survey; Internal Document; Air Force Research Laboratory: Rome, NY, USA, 2010.
5. National Research Council. Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts; The National Academies Press: Washington, DC, USA, 2006. [CrossRef]
6. United States Department of Defense; Department of the Air Force. Air Force Instruction 90-802: Risk Management. Available online: https://static.e-publishing.af.mil/production/1/af_se/publication/af90-802/af90-802.pdf (accessed on 1 April 2019).
7. United States Department of Defense; Department of the Air Force. Air Force Manual 15-129: Air and Space Weather Operations. Available online: https://static.e-publishing.af.mil/production/1/af_a3/publication/afman15-129/afman15-129.pdf (accessed on 9 July 2020).
8. Thompson, J.C. On the Operational Deficiencies in Categorical Weather Forecasts. Bull. Am. Meteorol. Soc. 1952, 33, 223–226. [CrossRef]
9. Scruggs, F.P. Decision Theory and Weather Forecasts: A Union with Promise. Air University Review. July–August 1967. Available online: https://web.archive.org/web/20170126062712/http://www.airpower.maxwell.af.mil/airchronicles/areview/1967/jul-aug/scruggs.html (accessed on 22 March 2021).
10. Murphy, A.H.; Katz, R.W.; Winkler, R.L.; Hsu, W.-R. Repetitive Decision Making and the Value of Forecasts in the Cost?Loss Ratio Situation: A Dynamic Model. Mon. Weather Rev. 1985, 113, 801–813. [CrossRef]
11. Pielke, R.A. Who decides? Forecasts and responsibilities in the 1997 Red River flood. *Appl. Behav. Sci. Rev.* **1999**, *7*, 83–101. [CrossRef]

12. Zhu, Y.; Toth, Z.; Wobus, R.; Richardson, D.; Mylne, K. The Economic Value Of Ensemble-Based Weather Forecasts. *Bull. Am. Meteorol. Soc.* **2002**, *83*, 73–83. [CrossRef]

13. Eckel, F.A.; Gunningham, J.G.; Hette, D.E. Weather and the Calculated Risk: Exploiting Forecast Uncertainty for Operational Risk Management. *Air Space Power J.* **2008**, *22*, 71–82.

14. United States Government Accountability Office. Weapon Systems Annual Assessment: Limited Use of Knowledge-Based Practices Continues to Undercut DOD’s Investments. Available online: https://www.gao.gov/products/gao-19-336sp (accessed on 7 May 2019).

15. United States Government Accountability Office. Defense Real Property: DOD Needs to Take Additional Actions to Improve Management of Its Inventory Data. Available online: https://www.gao.gov/products/gao-19-73 (accessed on 13 November 2018).

16. Lazo, J.K.; Lawson, M.; Larsen, P.H.; Waldman, D.M. U.S. Economic Sensitivity to Weather Variability. *Bull. Am. Meteorol. Soc.* **2011**, *92*, 709–720. [CrossRef]

17. Shapiro, A. Tyndall Air Force Base Still Faces Challenges In Recovering From Hurricane Michael. NPR Organization. Available online: https://www.npr.org/2019/05/31/728794872/tyndall-air-force-base-still-faces-challenges-in-recovering-from-hurricane-michael (accessed on 12 March 2021).

18. Mizokami, K. Hurricane Michael Mangled at Least 17 F-22 Raptors That Failed to Flee Their Base. *Popular Mechanics*. 15 October 2018. Available online: https://www.popularmechanics.com/military/aviation/a23792532/f-22s-damaged-destroyed-hurricane-michael/ (accessed on 12 March 2021).

19. 309th AMARG the First FAA Military Repair Station in AFSC. *Standard-Examiner*. Available online: https://www.standard.net/hilltip/news/309th-amarg-the-first-faa-military-repair-station-in-asfc/article_8e2bbf9b-de36-5a90-a57b-746c9d076b88.html (accessed on 12 March 2021).

20. Liewer, S. Tornado Caused Almost $20 Million in Damage at Offutt Air Force Base. Available online: https://omaha.com/news/military/tornado-caused-almost-20-million-in-damage-at-offutt-air-force-base/article_d05a175-0658-5595-b585-9f0f878e4b6.html (accessed on 12 March 2021).

21. Amadeo, K. Why Military Spending Is More Than You Think It Is. Available online: https://www.thebalance.com/u-s-military-budget-components-challenges-growth-3306320 (accessed on 12 March 2021).

22. Kalnay, E. Historical perspective: Earlier ensembles and forecasting forecast skill. *Q. J. R. Meteorol. Soc.* **2019**, 145, 25–34. [CrossRef]

23. Joslyn, S.L.; Leclerc, J.E. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *J. Exp. Psychol. Appl.* **2012**, *18*, 126–140. [CrossRef] [PubMed]

24. Marimo, P.; Kaplan, T.R.; Mylne, K.; Sharpe, M. Communication of Uncertainty in Temperature Forecasts. *Weather Forecast* **2015**, *30*, 5–22. [CrossRef]

25. Hoffman, R.N.; Kalnay, E. Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus A Dyn. Meteorol. Oceanogr.* **1983**, *35*, 100–118. [CrossRef]

26. Du, J.; Mullen, S.L.; Sanders, F. Short-Range Ensemble Forecasting of Quantitative Precipitation. *Mon. Weather Rev.* **1997**, *125*, 2427–2459. [CrossRef]

27. Stensrud, D.J.; Brooks, H.E.; Du, J.; Tracton, M.S.; Rogers, E. Using Ensembles for Short-Range Forecasting. *Mon. Weather Rev.* **1999**, *127*, 433–446. [CrossRef]

28. Wandishin, M.S.; Mullen, S.L.; Stensrud, D.J.; Brooks, H.E. Evaluation of a Short-Range Multimodel Ensemble System. *Mon. Weather Rev.* **2001**, *129*, 729–747. [CrossRef]

29. Mylne, K.R.; Evans, R.E.; Clark, R.T. Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Q. J. R. Meteorol. Soc.* **2002**, *128*, 361–384. [CrossRef]

30. Du, J.; McQueen, J.; Geoff DiMego, G.; Black, T.; Juang, H.; Rogers, E.; Ferrier, B.; Zhou, B.; Toth, Z. The NOAA/NWS/NCEP Short Range Ensemble Forecast (SREF) system: Evaluation of an initial condition vs multiple model physics ensemble approach. In *Proceedings of the 16th Conference on Numerical Weather Prediction*, Seattle, WA, USA, 11–15 January 2004; CD-ROM, 21.3; American Meteor Society: Geneseo, NY, USA.

31. Eckel, F.A.; Mass, C.F. Aspects of Effective Mesoscale, Short-Range Ensemble Forecasting. *Weather Forecast* **2005**, *20*, 328–350. [CrossRef]

32. Roeber, P.J.; Schultz, D.M.; Colle, B.A.; Stensrud, D.J. Toward Improved Prediction: High-Resolution and Ensemble Modeling Systems in Operations. *Weather Forecast* **2004**, *19*, 936–949. [CrossRef]

33. Cunningham, J.G. Applying Ensemble Prediction Systems to Department of Defense Operations. Master’s Thesis, Naval Postgraduate School, Monterey, CA, USA, March 2006. Available online: https://apps.dtic.mil/sti/pdfs/ADA445411.pdf (accessed on 22 March 2021).

34. Nobis, T.E.; Kuchera, E.L.; Rentschler, S.A.; Rugg, S.A.; Cunningham, J.G.; Synder, C.; Hacker, J.P. Towards the Development of an Operational Mesoscale Ensemble System for the DoD Using the WRF-ARW Model. In *Proceedings of the 2008 DoD HPCMP Users Group Conference*, Seattle, WA, USA, 14–17 July 2008; pp. 288–292.
35. Skamarock, W.C.; Skamarock, W.C.; Klemp, J.B.; Dudhia, J.; Gill, D.O.; Barker, D.M.; Wang, W.; Powers, J.G. A description of the Advanced Research WRF Version 2. NCAR Tech. Notes 2005, 88. [CrossRef]

36. Hodur, R.M. The Naval Research Laboratory’s Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). Mon. Weather Rev. 1997, 125, 1414–1430. [CrossRef]

37. Wei, M.; Toth, Z.; Wobus, R.; Zhu, Y. Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. Tellus A Dyn. Meteorol. Oceanogr. 2008, 60, 62–79. [CrossRef]

38. Hacker, J.P.; Ha, S.-Y.; Snyder, C.; Berner, J.; Eckel, F.A.; Kuchera, E.; Pacornich, M.; Rugg, S.; Schamm, J.; Wang, X. The U.S. Air Force Weather Agency’s mesoscale ensemble: Scientific description and performance results. Tellus A Dyn. Meteorol. Oceanogr. 2011, 63, 625–641. [CrossRef]

39. Kuchera, E.L.; Nobis, I.; Rugg, S.; Rentschler, S.; Cunningham, J.; Hughesand, H.; Sittel, M. AFWA’s Joint Ensemble System Experiment (JEFS) Experiment. In Proceedings of the 19th Conference on Numerical Weather Prediction, Omaha, NE, USA, 1–9 June 2009; Available online: https://ams.confex.com/ams/pdfpapers/152656.pdf (accessed on 22 March 2021).

40. Peng, M.S.; Ridout, J.A.; Hogan, T.F. Recent Modifications of the Emanuel Convective Scheme in the Navy Operational Global Atmospheric Prediction System. Mon. Weather Rev. 2004, 132, 1254–1268. [CrossRef]

41. Côté, J.; Gravel, S.; Méthot, A.; Patoine, A.; Roch, M.; Staniforth, A. The Operational CMC–MRB Global Environmental Multiscale (GEM) Model. Part I: Design Considerations and Formulation. Mon. Weather Rev. 1998, 126, 1373–1395. [CrossRef]

42. McCormick, J.R. Near Surface Forecast Challenges at the Air Force Weather Agency. In Proceedings of the 15th Conference on Aviation, Range, and Aerospace Meteorology, Los Angeles, CA, USA, 1 August 2011; Available online: https://ams.confex.com/ams/14Meso15ARAM/webprogram/Manuscript/Paper190769/McCormickPreprint.pdf (accessed on 22 March 2021).

43. Hsu, W.-R.; Murphy, A.H. The attributes diagram A geometrical framework for assessing the quality of probability forecasts. Int. J. Forecast. 1986, 2, 285–293. [CrossRef]

44. Wilks, D.S. Statistical Methods in the Atmospheric Sciences, 2nd ed.; Academic Press: Cambridge, MA, USA, 2006.

45. Arribas, A.; Robertson, K.B.; Mylne, K.R. Test of a Poor Man’s Ensemble Prediction System for Short-Range Probability Forecasting. Mon. Weather Rev. 2005, 133, 1825–1839. [CrossRef]

46. Palmer, T.N.; Tibaldi, S. On the Prediction of Forecast Skill. Mon. Weather Rev. 1988, 116, 2453–2480. [CrossRef]

47. Laloux, F. Reinventing Organizations: A Guide to Creating Organizations Inspired by the Next Stage of Human Consciousness; Nelson Parker: Millis, MA, USA, 2014.

48. Evans, C.; Van Dyke, D.F.; Lericos, T. How Do Forecasters Utilize Output from a Convection-Permitting Ensemble Forecast System? Case Study of a High-Impact Precipitation Event. Weather Forecasting 2014, 29, 466–486. [CrossRef]

49. Brown, A.; Milton, S.; Cullen, M.; Golding, B.; Mitchell, J.F.B.; Shelly, A. Unified Modeling and Prediction of Weather and Climate: A 25-Year Journey. Bull. Am. Meteorol. Soc. 2012, 93, 1865–1877. [CrossRef]

50. Deutscher Wetterdienst. Ensemble Prediction. Available online: https://www.dwd.de/EN/research/weatherforecasting/num_modelling/04_ensemble_methods/ensemble_prediction/ensemble_prediction_en.html;jsessionid=0E5AD903B9A7A241626C1D35E020FAC9.live31083?nn=484822#COSMO-D2-EPS (accessed on 12 March 2021).

51. Rentschler, S.A. Air Force Weather Ensembles. In Proceedings of the 5th NCEP Ensemble Workshop, College Park, MD, USA, 10 May 2011.

52. Kuchera, E.L. Air Force Weather Ensembles. In Proceedings of the 15th WRF Users Workshop, Boulder, CO, USA, 23–27 June 2014; Available online: https://www2.mmm.ucar.edu/wrf/users/workshops/WS2014/ppts/2.3.pdf (accessed on 22 March 2021).

53. Ebert, E.E. Ability of a Poor Man’s Ensemble to Predict the Probability and Distribution of Precipitation. Mon. Weather Rev. 2001, 129, 2461–2480. [CrossRef]

54. Kain, J.S.; Weiss, S.J.; Bright, D.R.; Baldwin, M.E.; Levit, J.J.; Carbin, G.W.; Schwartz, C.S.; Weissman, M.L.; Droegemeier, K.K.; Weber, D.B.; et al. Some Practical Considerations Regarding Horizontal Resolution in the First Generation of Operational Convection-Allowing NWP. Weather Forecasting 2003, 23, 931–952. [CrossRef]

55. Clark, A.J.; Gallus, W.A.; Xue, M.; Kong, F. A Comparison of Precipitation Forecast Skill between Small Convection-Allowing and Large Convection-Parameterizing Ensembles. Weather Forecasting 2009, 24, 1121–1140. [CrossRef]

56. Schwartz, C.S.; Kain, J.S.; Weiss, S.J.; Xue, M.; Bright, D.R.; Kong, F.; Thomas, K.W.; Levit, J.J.; Coniglio, M.C. Next-Day Convection-Allowing WRF Model Guidance: A Second Look at 2-km versus 4-km Grid Spacing. Mon. Weather Rev. 2009, 137, 3351–3372. [CrossRef]

57. Clark, A.J.; Gallus, W.A.; Xue, M.; Kong, F. Convection-Allowing and Convection-Parameterizing Ensemble Forecasts of a Mesoscale Convective Vortex and Associated Severe Weather Environment. Weather Forecasting 2010, 25, 1052–1081. [CrossRef]

58. Kain, J.S.; Dembek, S.R.; Weiss, S.J.; Case, J.L.; Levit, J.J.; Sobash, R.A. Extracting Unique Information from High-Resolution Forecast Models: Monitoring Selected Fields and Phenomena Every Time Step. Weather Forecasting 2010, 25, 1536–1542. [CrossRef]

59. Sobash, R.A.; Kain, J.S.; Bright, D.R.; Dean, A.R.; Coniglio, M.C.; Weiss, S.J. Probabilistic Forecast Guidance for Severe Thunderstorms Based on the Identification of Extreme Phenomena in Convection-Allowing Model Forecasts. Weather Forecasting 2011, 26, 714–728. [CrossRef]

60. Roberts, B.; Jirak, I.L.; Clark, A.J.; Weiss, S.J.; Kain, J.S. PostProcessing and Visualization Techniques for Convection-Allowing Ensembles. Bull. Am. Meteorol. Soc. 2019, 100, 1245–1258. [CrossRef]
61. Aligo, E.A.; Gallus, W.A.; Segal, M. On the Impact of WRF Model Vertical Grid Resolution on Midwest Summer Rainfall Forecasts. *Weather Forecast* 2009, 24, 575–594. [CrossRef]

62. Creighton, G.A.; Creighton, G.; Kuchera, E.; Adams-Selin, R.; McCormick, J.; Rentschler, S.; Wickard, B. AFWA Diagnostics in WRF. Available online: https://www2.mmm.ucar.edu/wrf/users/docs/AFWA_Diagnostics_in_WRF.pdf (accessed on 22 March 2021).

63. Fréchet, M. Sur la loi de probabilité de l’écart maximum. *Ann. Soc. Pol. Math.* 1927, 6, 93–116.

64. Weibull, W. A statistical distribution function of wide applicability. *J. Appl. Mech.* 1951, 18, 293–297. [CrossRef]

65. Roebber, P.J.; Bruening, S.L.; Schultz, D.M.; Cortinas, J.V. Improving Snowfall Forecasting by Diagnosing Snow Density. *Weather Forecast* 2003, 18, 264–287. [CrossRef]

66. Creighton, G.A. Redesign of the AFWA-EPS Ensemble Post Processor; Internal Document; Air Force Weather Agency, Offutt AFB: Bellevue, NE, USA, 15 January 2015.

67. Bright, D.R.; Wandishin, M.S.; Jewell, R.E.; Weiss, S.J. A physically based parameter for lightning prediction and its calibration in ensemble forecasts. In Proceedings of the Conference on Meteorological Applications of Lightning Data, San Diego, CA, USA, 9–13 January 2005.

68. Gallo, B.T.; Clark, A.J.; Dembek, S.R. Forecasting Tornadoes Using Convection-Permitting Ensembles. *Weather Forecast* 2016, 31, 273–295. [CrossRef]

69. Hunt, E.D.; Adams-Selin, R.; Sartan, J.; Creighton, G.; Kuchera, E.; Keane, J.; Jones, S. The Spring 2014 Mesoscale Ensemble Prediction System ‘Dust Offensive’. In Proceedings of the AGU Fall Meeting Abstracts, San Francisco, CA, USA, 15–19 December 2014; Volume 41, p. A41F–3130.

70. Legrand, S.L.; Polashenski, S.; Letcher, T.W.; Creighton, G.A.; Peckham, S.E.; Cetola, J.D. The AFWA dust emission scheme for the GOCART aerosol model in WRF-Chem v3.8.1. *Geosci. Model Dev.* 2019, 12, 131–166. [CrossRef]

71. Schumann, T. COSMO-DE EPS—A New Way Predicting Severe Convection. 2013. Available online: https://www.ecmwf.int/node/13851 (accessed on 12 March 2021).

72. Hagelin, S.; Son, J.; Swinbank, R.; McCabe, A.; Roberts, N.; Tennant, W. The Met Office convective-scale ensemble, MOGREPS-UK. *Q. J. R. Meteorol. Soc.* 2017, 143, 2846–2861. [CrossRef]

73. National Centers for Environmental Information (NCEI). On This Day: 2011 Tornado Super Outbreak. 25 April 2017. Available online: http://www.ncei.noaa.gov/news/2011-tornado-super-outbreak (accessed on 22 March 2021).

74. Witt, C. McConnell Takes Tornado Precautions. McConnell Air Force Base. 14 April 2012. Available online: https://www.mcconnell.af.mil/News/Article/224421/mcconnell-takes-tornado-precautions/ (accessed on 12 March 2021).

75. Wenzl, R.; Plumlee, R.; The Wichita Eagle. WichitaTornado Brings Destruction, No Deaths. 15 April 2012. Available online: https://www.kansas.com/news/article1090380.html (accessed on 12 March 2021).

76. Ashpole, I.; Washington, R. An automated dust detection using SEVIRI: A multiyear climatology of summertime dustiness in the central and western Sahara. *J. Geophys. Res. Space Phys.* 2012, 117. [CrossRef]

77. Peckham, S.E. WRF-Chem Version 3.3 User’s Guide. Available online: https://repository.library.noaa.gov/view/noaa/11119 (accessed on 12 March 2021).

78. United States Department of Commerce; National Oceanographic and Atmospheric Administration. The Historic Derecho of 29 June 2012. By Laura K. Furgione, January 2013. Available online: https://www.weather.gov/media/publications/assessments/derecho12.pdf (accessed on 22 March 2021).

79. Coniglio, M.C.; Stensrud, D.J. Interpreting the climatology of derechos. *Weather Forecast* 2004, 19, 595–605. [CrossRef]

80. Barnum, B.; Winstead, N.; Wesely, J.; Hakola, A.; Colarco, P.; Toon, O.; Ginoux, P.; Brooks, G.; Hasselbarth, L.; Toth, B. Forecasting dust storms using the CARMA-dust model and MM5 weather data. *Environ. Model. Softw.* 2004, 19, 129–140. [CrossRef]

81. Burton, K. AFWA Dust Model Comparison; Internal Document; 19th Expeditionary Weather Squadron: Bagram Air Base, Afghanistan, 13 August 2011.

82. Melick, C.J.; Jirak, I.L.; Dean, A.R.; Correia, J., Jr.; Weiss, S.J. Real Time Objective Verification of Convective Forecasts: 2012 HWT Experiment_summary.pdf (accessed on 22 March 2021).

83. Adams-Selin, R. Use of the AFWA-AWC testbed mesoscale ensemble to determine sensitivity of a convective high wind event simulation to boundary layer parameterizations. In Proceedings of the 12th WRF Users’ Workshop, Boulder, CO, USA, 20–24 June 2011; p. 11.

84. Rotsen, W.R. Toward Improving Short-Range Fog Prediction in Data-Denied Areas Using the Air Force Weather Agency Mesoscale Ensemble; Naval Postgraduate School: Monterey, CA, USA, 1 September 2012; Available online: https://apps.dtic.mil/sti/citations/ADA567345 (accessed on 22 March 2021).
113. Hansen, J.A. Accounting for Model Error in Ensemble-Based State Estimation and Forecasting. *Mon. Weather Rev.* **2002**, *130*, 2373–2391. [CrossRef]

114. Vannitsem, S.; Bremnes, J.B.; Demaeyer, J.; Evans, G.R.; Flowerdew, J.; Hemri, S.; Lerch, S.; Roberts, N.; Theis, S.; Atencia, A.; et al. Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bull. Am. Meteorol. Soc.* **2021**, *102*, E681–E699. [CrossRef]

115. Flowerdew, J. Initial Verification of IMPROVER: The New Met Office Post-Processing System. In Proceedings of the 25th Conference on Probability and Statistics, Austin, TX, USA, 8 January 2018; Available online: https://ams.confex.com/ams/98Annual/webprogram/Paper325854.html (accessed on 22 March 2021).

116. Jensen, T. The Use of the METplus Verification Capability in Both Operations and Research Organizations. In Proceedings of the 99th AMS Annual Meeting, Phoenix, AZ, USA, 9 January 2019; Available online: https://ams.confex.com/ams/2019Annual/webprogram/Paper353523.html (accessed on 22 March 2021).

117. Brown, B.; Jensen, T.; Gotway, J.H.; Bullock, R.; Gilleland, E.; Fowler, T.; Newman, K.; Adriaansen, D.; Blank, L.; Burek, T.; et al. The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bull. Am. Meteorol. Soc.* **2020**, *1*, 1–68. [CrossRef]

118. West, T. 16th Weather Squadron Advancements in Providing Actionable Environmental Intelligence for Unique Air Force and Army Mission Requirements. In Proceedings of the 101st Annual AMS Meeting, Online. 10–15 January 2021.