1.3 EVALUATION OF CONVECTIVE WEATHER FORECAST PERFORMANCE BY OPERATIONALLY RELEVANT METEOROLOGICAL CHARACTERISTICS

Colleen Reiche *, Michael Robinson, Victor Klimenko, Scott Percic, Mike Kay
AvMet Applications, Inc., Reston, Virginia

Rafal Kicinger
Metron Aviation, Dulles, VA

1. INTRODUCTION

Adverse weather remains the most disruptive constraint in the National Airspace System (NAS), contributing to the vast majority of impacts on air traffic operations that occur annually. These disruptions, and thus the need to effectively predict the occurrence of adverse weather has motivated the development of multiple weather forecast systems and products which, due to the highly nonlinear nature of the atmosphere, have associated forecast errors and uncertainty. To date, operational use of these forecasts in making risk-managed, air traffic impact mitigation decisions has been hampered by the lack of explicit, objective awareness of the weather forecast errors. Moreover, the availability of multiple deterministic and probabilistic forecast products, without accompanying, objective guidance for how to collectively use them to increase user confidence in the forecast through direct accountability of individual and relative forecast errors limits the use of multiple data sources to operational decision-making. These shortfalls erode abilities to define, coordinate, and execute effective Air Traffic Management (ATM) strategies, resulting in inconsistent solutions which do not take full advantage of the opportunities to mitigate impacts.

With an initial focus on convective weather, forecast performance for several pertinent convective weather forecast systems, including both probabilistic (SREF, LAMP) and deterministic (HRRR), in three relevant ATM planning time horizons (tactical: 1-3 hour, strategic: 4-8 hour, long-strategic: 9-12 hour) are quantified and evaluated. This was accomplished by first identifying and defining specific characteristics associated with probabilistic and deterministic forecasts that are meaningful from both a meteorological and ATM perspective. This performance was evaluated for a set of weather events from the 2012 convective weather season (April – September), classified by target convective characteristics - weather pattern, time of day, and region. Performance variability for each forecast product was assessed for all lead times within each planning window through comparison of results across each subset of classified weather events with common combinations of the target convective characteristics.

* *Corresponding Author Address:* Colleen Reiche, AvMet Applications, 1800 Alexander Bell Dr., Suite 130, Reston, VA 20191; e-mail: reiche@avmet.com

This study investigates the technical feasibility of combining pertinent convective weather forecast performance for the three target forecast products into a consolidated, more insightful, expression of forecast performance by planning period through development of an initial aggregate forecast performance summary "scorecard". The utility of this fully-populated performance "scorecard" for several realistic operational convective weather impact scenarios, along with its role in supporting development of a common, combined, convective weather prediction, will also be described.

2. MOTIVATION FOR AGGREGATE PERFORMANCE SUMMARY

Key challenges in today's ATM environment relative to convective weather impact management include:

1. Lack of awareness of the historical performance of operational forecast products and associated level of uncertainty expected, given convective weather events of specific organization type, location, and relative time of occurrence;

2. Inability to assess the relative strengths and weaknesses of the multiple forecast products available to decision-makers, as they may vary by specific convective weather impact scenarios and the planning lead-times;

3. Inability to assimilate and combine multiple forecast inputs, accounting for error tendencies, and produce an operationally-meaningful, aggregated forecast product that leverages performance strengths from the individual, contributing predictions.

Specifically, today's traffic managers have access to several forecast products (Figure 1), each of which require excess effort to access, interpret, consider, and re-assess. Moreover, both traffic managers and supporting meteorologists are required to mentally recall (if possible) performance tendencies (e.g., forecast too late, too strong, too large of line?) given the skill of past predictions relative to a specific event expected during the air traffic operation being managed. These subjective (and thus often inconsistent) considerations must be performed using forecast products that typically have no explicit information about forecast uncertainty that may help guide this practice. The mental model for these considerations becomes over-taxed when decision-makers then try to simultaneously consider these performance "adjustments" for multiple forecast products – and subsequently try to separate out "good" guidance from "poor" guidance before reaching an impact management decision; All to be repeated weather events and NAS constraints evolve during an operational day.

Given the complexities in manually tracking and completing these forecast-to-impact assessment tasks, the most useful forecast information can (and do) go unused, while erroneous predictions can (and are) weighted too heavily. This results in inefficient weather impact mitigation plans and undesirable traffic management outcomes, often in the form of increased air traffic delays and operating costs and decreased customer (airlines, passengers) satisfaction.

This technical feasibility effort seeks to assess the viability of combining derived, operationally-relevant convective weather forecast performance results (which are approximated subjectively in the current ATM operation) from multiple, "stovepiped" products into aggregate forecast performance summary or "scorecard". With this performance summary, operational users would then be able to explicitly consider the forecast performance tendencies – for a specific type of convective weather event – across multiple forecast products, for improved performance-awareness and forecast uncertainty guidance.

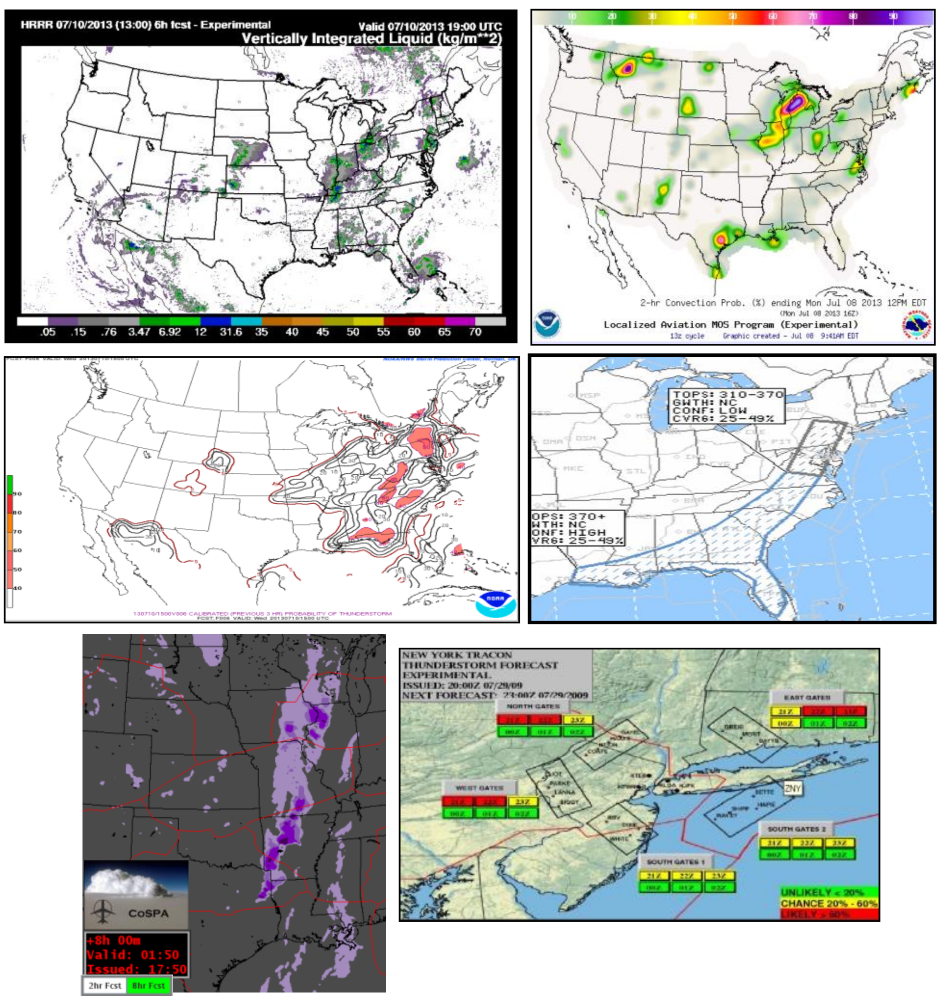

Figure 1. A sample of individual and unique convective weather forecast products available to operational decision-makers supporting air traffic management.

3. PREVIOUS FORECAST PERFORMANCE ASSESSMENTS

Progress in technology and science have greatly improved the ability to forecast weather, and meteorologists have made strides in reducing forecast uncertainty through enhanced observation sensing, numerical modeling, and data assimilation. However, the highly non-linear nature of the atmospheric system ensures that forecast errors and associated uncertainty will always exist. Conveying information on these errors and uncertainty to users who are making risk-based decisions with varying tolerance for cost/action allows them to weigh this weather forecast uncertainty information in their decision process, which effectively combines the expertise of both the forecaster and decision maker. Forecast errors not only impact these decisions and outcomes, but also the confidence of the decision maker in the forecast information itself. For these reasons, the American Meteorological Society (AMS) requested the National Research Council (NRC) to investigate the societal needs and potential benefits to using forecast uncertainty (NRC 2006). This NRC report, as well as one conducted by the AMS itself (AMS 2008), determined that both forecasters and decision-making users can benefit from the combined usage of probabilistic and deterministic forecasting to convey uncertainty information, and recommend improvements and research on effectively communicating this forecast uncertainty information to users. These reports led the AMS to develop an implementation plan for communicating forecast uncertainty to users as an important component of overall forecast information (Hirshberg et al. 2011).

Several previous studies have examined similar forecast errors to those examined in this study, such as spatial errors associated with predictions of precipitation (Micheas et al. 2007), and their associated variability by relevant characteristics such as time of day, region, or weather type/morphology for predictions of precipitation or rainfall (Grams et al. 2005, Davis et al. 2006) and for the National Digital Forecast Database Thunderstorm Probability field (Lack et al. 2012b). Some examination of ATM-related convective weather forecast errors has been conducted with stratifications by some of these characteristics, such as planning window and geographic region (Lack et al. 2011, Lack et al. 2012a), but these have only minimally examined the error variability. The expression of errors in this study, combined with use of “intelligent” stratifications, potentially makes this a valuable addition to these previous studies.

Eliciting relationships between convective weather forecast errors and complex ATM decisions is non-trivial due to the complexity of both. Investigation into use of a Bayesian decision network to predict ATM decisions based on an analysis of discretized meteorological input variables from historical ATM weather impact mitigation events has been conducted toward this goal (Pepper et al. 2003). Results from this investigation suggest that examining specific elements of strategic ATM decisions, such as the conveyance of weather uncertainty, may be a more effective method for extracting relevant information on ATM responses than examining the comprehensive and complex ATM process as a whole. Building on these results, this research will provide a targeted and comprehensive examination of forecast errors, each of which can then be related to specific ATM decisions and responses to explore straightforward methods for its conveyance to ATM decision makers. An earlier approach for translating strategic probabilistic forecast information into statements of air traffic constraints utilized the concept of a decision tree and converted the probabilistic forecast into discrete categories to characterize potential outcomes and identify logical ATM responses (Davidson et al. 2004). This work will build on this approach by improving the characterization of potential weather outcomes through improved depiction of forecast errors which can later by mapped to specific ATM decisions and decision points.

Translation of forecast error and uncertainty information to air traffic decision-making is not unique to en route ATM responses, but is also necessary for terminal applications. Several studies have been conducted in the use of probabilistic forecasts of stratus cloud layer clearing for terminal-based traffic management decisions in San Francisco (Evans et al. 2006, Cook & Wood 2009). Automation of translating stratus clearing probabilities into likely terminal management responses was examined in one of these studies through use of Monte Carlo simulations (Cook & Wood 2009). These simulations generated possible Ground Delay Program (GDP) responses and among these, highlighted the most cost-efficient GDP decision. The methods described in this work builds on these prior studies, and could be leveraged for terminal ATM decisions or extended to other types of forecasts beyond convective weather.

A need also exists for forecast error and uncertainty translation to improve the efficiency of risk-based decision-making methodologies to applications outside of aviation: Lightning probabilistic forecasts of lightning are being used to aid decision-making for electrical systems protection in Brazil (Leite et al. 2007) and fire prevention (Gibson et al. 2008, Pence & Zimmerman 2011). High seas probabilities are being used to predict pirate activity off Africa for military applications (Hirschberg et al. 2011, Petry et al. 2010). Hurricane location probabilities are being used for evacuation decisions (Regnier & Harr 2006). Finally, weather probabilities have been applied to space shuttle operations (Brody et al. 1997). The forecast error variability quantified here could be also be applied to these other application spaces. In turn, alternative decision-making considerations developed for these other operational domains, in context with potential error-awareness advancements from research presented in this report, can be evaluated for potential opportunities for enhanced air traffic management applications.

4. EXPERIMENT DESIGN

4.1 *Target Precipitation Products*

As convective weather is the target of this study, precipitation-based products, both observed and forecast, are analyzed. Because multiple types of convective weather forecasts, including "radar-like" deterministic and probabilistic, provide different types of valuable predictive information and are relevant and useful to aviation, specific forecast systems of these two types are evaluated in this study based on their current and potential future prevalence as forecast decision aids for aviation planners and air traffic decision-makers.

The target deterministic model for this study is the High Resolution Rapid Refresh (HRRR) model, generated by the Earth Systems Research Laboratory (ESRL). The HRRR is a high resolution (3 km) experimental forecast model capable of explicitly depicting convection which produces hourly fresh model realizations. The model receives its lateral boundary conditions from the Rapid Refresh (RAP) 13-km resolution model, inside which it is nested, and assimilates radar and satellite observations. This project focuses on the deterministic HRRR VIL forecast, given the pertinence of this product to aviation applications, as noted above. An overview of the HRRR forecast, also outlining forecasts improvements released during the 2012 season which is the focus in the current study, can be found in Alexander et al. (2012) and Weygandt et al. (2012).

The experimental Localized Aviation MOS Program (LAMP) convective forecast, developed and generated by the National Weather Service (NWS) Meteorological Development Laboratory (MDL), is one of the two probabilistic forecasts targeted in this study. It predicts convection over a 2-hour time valid time window based on Model Output Statistics (MOS) from the Global Forecast System (GFS) and the North American Mesoscale (NAM) numerical weather prediction models. A convective event is defined as the occurrence of a radar reflectivity value at or above 40 dBZ and/or one or more cloud to ground (CTG) lightning strikes within a 2.5-km grid box. This product is issued hourly with each forecast valid over a 2-hour window out to 24 hours. Details on the methodology used to generate these probability forecasts can be found in Charba et.al (2011) and the general statistical approach for LAMP is described in Ghiradelli and Glahn (2010).

The second target probabilistic forecast for this study is the Short Range Ensemble Forecast (SREF), a multi-model, multi-physics ensemble comprised of 21 members produced by the National Centers for Environmental Prediction (NCEP). The calibrated SREF Thunderstorm Probability Forecast is produced by the Storm Prediction Center (SPC) through post-processing of both the SREF forecast and the 3-hour time-lagged North American Model (NAM) ensemble forecast. Forecasts are issued every six (6) hours (at 0300, 0900, 1500, and 2100 UTC) for lead times every three (3) hours out to a maximum lead time of 87 hours using a 40 km grid resolution. A thunderstorm event for this product is defined as having at least one lightning strike within a grid box. The calibration technique for these probabilities is described in Bright et al. (2005) and details on recent technique refinements and verification results are described in Bright and Grams (2009).

The Multiple-Radar Multiple-Sensor (MRMS) VIL product, a 2D product derived from the 3D radar observations using a method described in Lakshmanan et al. (2006), will serve as the observation product in this study, against which all forecasts will be evaluated to compute the errors. The MRMS system, developed and produced by the National Severe Storms Laboratory (NSSL), leverages the overlapping WSR-88D NEXRAD radar coverage to generate a seamless, rapidly updating, high resolution 3D depiction of the radar data and objectively blends this radar information with other surface, upper air, and satellite observations. The update cycle for the MRMS VIL product is every five (5) minutes and its spatial resolution is 1 km.

4.2 *Convective Weather Characteristics*

In order to assess variations in forecast performance across target convective characteristics, historical convective weather events from the 2012 convective season (April – September) have been identified and classified based on three target characteristics which are both meteorologically meaningful and applicable to ATM planning – weather pattern, region, and time of day.

Weather patterns, in the context of this analysis, refer to the dominant mode of storm organization associated with convective weather events. To capture both a range in meteorological conditions (and associated atmospheric, thermodynamic, and kinematic variety found with different modes of convective weather organization) and a range in air traffic impact considerations and responses, three target weather patterns were chosen – air mass, organized storms with gaps, and line storms.

To understand variations in forecast error due to the variations in frequency and timing of the target weather pattern occurrence by location, the three types of weather patterns were classified according to their location in three (3) regions – Northeast, Southeast, and West (Figure 2). The target regions experience all three target weather patterns, but often have different meteorological forcing mechanisms which accounts for the seasonal differences in their timing in each region. Identification of three regions captures this spatial meteorological variability while ensuring a sufficient sample size among the events in each category (region-pattern pair) for the project needs, which may not have been possible with more target regions.

These regions were also chosen based on their unique ATM operational considerations which will also impact how a more robust characterization of forecast performance would be applied in each region. The Northeast region of the NAS includes resources and operations which are complex and densely-packed, and experiences high traffic volume, which can quickly become constrained and is likely to have cascading impacts affecting other airspace regions. Because of this, the Northeast may be most sensitive to small variations in forecast performance and thus may require a more strict risk management approach to decision-making given the forecast uncertainties. The Southeast region also experiences high traffic volume, but has a less complex network of resources enabling it to be less constrained than the Northeast. This combination may allow for more tolerance of forecast performance variability in this region. The West region, as a whole, experiences the lowest volume of NAS traffic and possesses the least constrained airspace resources. As a result, weather impact management accounting for forecast errors may be served by an alternative model for risk-aversion and forecast uncertainty tolerances.

Figure 2. Three target regions defined for this study.

The evolution and formation of weather patterns, along with air traffic volume and ATM constraints, can vary by the time of day. To capture potential variability in forecast errors by time of day, two unique time of day periods are used to classify convective weather events – overnight/morning (0300-1500 Z) and late morning/afternoon (1500-0300 Z). Daytime heating during the late morning/afternoon hours facilitates convective growth and storms of all patterns can often reach their peak intensity during this period. Storms can also form due to large scale forcing and persist through the overnight and morning hours. Air traffic demand peaks in the late afternoon and evening, making this period of the day potentially sensitive to convective weather constraints and associated forecast errors. However, during the early morning, many large airports experience high volume "pushes" indicating that weather constraints, potentially compounded by inefficient traffic management decisions made without better accountability of forecast error tendencies, could have cascading effects into the rest of the day depending on the region.

Because the applicability and utility of forecast performance information to ATM planning varies by planning period, three planning lead time windows have been defined here – tactical (1-3 hours), strategic (4-8 hours), and long-strategic (9-12 hours). These planning periods have been defined based on common decisions periods for ATM planning. Forecast performance variability was evaluated within each of these planning periods across the target convective characteristics. Stratifying in this manner facilitates assessing forecast performance in the context of specific operational needs in each planning period and will thus facilitate intelligent aggregation.

5 FORECAST PERFORMANCE EVALUATION METHODOLOGY

The availability of both probabilistic and deterministic forecasts is important for ATM applications and decision-making as they both provide unique and complimentary forecast information that together provides a robust characterization of the predicted convective mode, intensity, scale, and likelihood. While information on these types of characteristics could be gleaned from converting a probabilistic forecast into a dichotomous convection prediction, a deterministic forecast is more readily interpreted by ATM decision makers largely because of the extent to which these forecasts have been used to support ATM operations historically. Therefore, full characterization of forecast performance requires consideration of characteristics of each type.

5.1 *Evaluating Deterministic Forecast Performance*

Expectations and predictions of significant convective weather coverage are critical input for ATM decision-making. Traffic managers routinely make decisions about available fix, route, sector, and en route center capacity degradation (and thus, the need for airspace management responses should demand-capacity imbalances result) based upon the expected coverage and extent of storms in and across key airspace regions. As a result, the ability to effectively monitor, and if needed, act on weather constraints that result in airspace congestion or controller workload conditions is sensitive to how well the coverage ("weather impact region") of convective weather elements and systems can be predicted. Lack of knowledge of coverage forecast reliability will likely result in under or over-utilization of a given NAS resource.

Related to this of course is storm intensity and the ability of traffic managers to anticipate that storms affecting key airspace regions will be strong enough to require pilot avoidance and deviations and will limit capacity available to serve pending traffic demand. User confidence in anticipated storm intensity, above or below key operational thresholds (e.g., Level 3 convection) enables more proactive and aggressive impact mitigation actions that effectively reduce delay, controller workload, costs, and safety concerns. Traffic management decisions derived from poorly performing storm coverage or intensity forecasts can often result in increased impacts and operational disruptions.

Storm coverage and intensity characteristics were assessed for both MRMS observations and 1-12 hour lead time HRRR forecasts of classified April – September 2012 convective weather events within appropriately sized hexagonal cells across the analysis domain (CONUS). Storm coverage is defined here as the percentage of the HRRR or MRMS grid cells within a hex cell that have VIL values at/above 3.5 kg m^{-2}, equivalent to Video Integrated Processor (VIP) Level 3. VIP Level 3 is recognized as the intensity threshold at which convection may first be considered hazardous – thus requiring pilot avoidance (Robinson et al. 2002). Forecast performance at predicting storm coverage is quantified as the difference between the HRRR and MRMS coverage within corresponding hex cells.

To determine appropriate hex cell sizes for each target planning period, a sensitivity study was conducted to assess the distribution of these coverage differences for a range of operationally meaningful hex cell sizes for forecasts in each target planning period. The most appropriate hex cell size in each planning period was determined as that which was large enough to provide a smooth, monotonic distribution in coverage differences while being small enough to produce a sufficient range in coverage differences. The final hex cell sizes in each planning period are:

- Tactical (1-3 hour) – **40 km**
- Strategic (4-8 hour) – **70 km**
- Long-Strategic (9-12 hour) – **70 km**

Understanding the errors associated with intensity are important to ATM decision-making as intensity is an important factor in the likelihood of deviations. Despite the assertion in standard operating procedures and pilot handbooks that any VIP greater than Level 3 should be considered hazardous convection and thus avoided, aircraft are often observed to penetrate storms with intensity greater than this threshold. This behavior motivates the evaluation of storm intensity at the larger end of the VIL distribution, which may be more crucial in determining the likelihood of aircraft deviation than the "average" or background VIL values. For this reason, the 95th percentile was chosen to characterize these most intense VIL values from distributions of non-zero VIL within hex cells in both the HRRR forecast and MRMS observation grids. Forecast performance at predicting storm intensity is quantified as the difference between the 95th percentile VIL value in the forecast and the 95th percentile VIL value in the MRMS observation grid.

5.2 *Evaluating Probabilistic Forecast Performance*

Unlike deterministic forecasts, which provide radar-like depictions of future weather including specific attributes like intensity and size, probabilistic forecasts predict the likelihood that any type of convection will occur in a given grid box. Due to this fundamental difference, these forecasts do not have the same types of characteristics (intensity, coverage) as a deterministic forecast and thus a unique set of performance metrics are required. Within most operational domains of interest (e.g., air traffic management), the utility of weather forecast probabilities can be optimized through the development and accompaniment of functional "rule sets" that translate (even low) weather probabilities into response thresholds, accounting for domain-specific forecast needs, risks, and decision costs. Moreover, a formal method for devising probabilistic rule sets can be optimized by targeting key forecast performance characteristics that, when translated, provide specialized decision support for specific domain needs. Two such performance characteristics pertinent to convective weather probabilistic forecasts, which will provide the basis for probabilistic forecast performance quantification, are reliability and skill (Wilks 2005). The errors associated with these characteristics will be quantified and assessed for both target probabilistic forecasts – the SREF thunderstorm probability and the LAMP convection probability.

The reliability of a probabilistic forecast refers to how closely the forecast of an event (in this case, convection) corresponds to the actual event occurrence frequency and is a valuable property of any probabilistic forecast (AMS 2008). Similar to bias for a deterministic forecast, which describes the magnitude of over or under-forecasting storm intensity, understanding the reliability of a probabilistic forecast will impact what probability values are actionable in ATM. A perfectly reliable forecast is one where the forecast probabilities verify with the exact same frequency as climatology. For instance, a 40% probability would verify (with convection occurring) 40% of the time.

To facilitate interpretation of the probabilistic forecast reliability, the weighted mean difference between a forecast and the actual observed frequency of convection was computed for all classified 2012 convective weather events. The difference between forecast probability values, binned in 5% increments, and the actual frequency of convection occurrence across all bins was first calculated. These differences are then averaged across all bins and weighted by the total number of predictions in that bin giving the most weight to the most frequently issued probability values.

$$\overline{X_D} = \sum(P_{fcst} - P_{obs}) \tag{1}$$

This value reflects the average amount of over or under-forecasting (bias) that was observed with the probabilistic forecast for a given combination of convective weather characteristics.

The forecast skill is commonly quantified using a Relative Operating Characteristic (ROC) diagram (Mason 1982), which plots the probability of detection (POD) against the false alarm rate (FAR) for a given probabilistic forecast over a historical evaluation period using incrementally increasing probability thresholds used to define the prediction of an event., defined here as a forecast probability exceeding the incrementally varying probability threshold. The POD, or true positive rate, is defined as the ratio of hits to the total number of observations where the event occurred, as

$$\text{POD} = \text{YY} / (\text{YY} + \text{NY}) \tag{2}$$

where *YY* indicates the number of true positive forecasts and *NY* indicates the number of false negative forecasts for a given probability threshold. The *FAR*, or false positive rate, is defined as the ratio of false alarm forecasts to the total number of observed non-events, as

$$\text{FAR} = \text{YN} / (\text{YN} + \text{NN}) \qquad (3)$$

where *YN* indicates the number of false positive forecasts and *NN* indicates the number of true positive forecasts. For each threshold probability from 0-100% in 1% increments, a unique POD and FAR are computed based on that threshold and plotted on the ROC diagram (Figure 3).

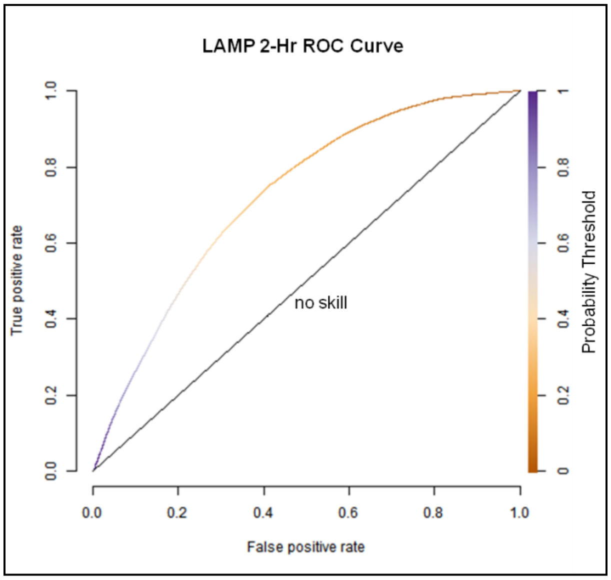

Figure 3. Sample ROC diagram for LAMP 2-hr forecasts (curved line) and no skill line (dark purple line).

The overall skill of the ROC curve, across all probability thresholds, can be evaluated by computing the Area Under the Curve (AUC) which quantifies the likelihood that a given forecast can accurately produce higher probabilities for events and lower probabilities for non-events. In other words, it describes the probability that a randomly selected forecast probability value chosen from the pool of those that correctly predicted the occurrence of convection (an event) will be greater than a randomly selected forecast probability value chosen from the pool of those that correctly predicted the non-occurrence of convection (non-events). The AUC for a forecast with perfect skill is 1.0, indicating a perfect 100% chance that the frequency of convective events varies with the forecast probability and of issuing greater probability values for events than non-events. On the other hand, the AUC for a forecast with no skill is 0.5 (the area under the diagonal "no skill" line on the ROC diagram), indicating that for a randomly selected pair of probabilities (one from the pool of events and one from non-events), the probability value associated with convection (event) is equally as likely (50%) to be larger than the one associated with the lack of convection as it is to be smaller, reflecting a complete lack of skill. These AUC values were computed for all classified 2012 convective weather events for all LAMP and SREF forecasts in each planning period.

6. IDENTIFYING AGGREGATE FORECAST PERFORMANCE

To collect, organize, and visualize the overall and aggregate forecast performance of each considered, individual convective weather forecast by key operational weather event characteristics, a summary table is constructed to consolidate historical forecast errors (Table 1). In this summary table, the forecast performance category is determined by placing a single summarized metric value specific to that forecast characteristic into predefined ranges which will be defined in subsequent sections. Each entry is determined independently to guide targeted interpretation and potential forecast performance adjustments.

An indication of over or under-forecasting for at least one forecast characteristic for each forecast model (reliability, intensity errors, and coverage errors) is also included in performance summary table entries as a positive or negative sign, reflecting whether the forecasted values were typically greater ("+") or less than ("-") the observed. Where both performance values are available (skill category and bias) for a forecast characteristic, these categories and "+/-" values would be used in tandem to guide interpretation and potential forecast performance adjustments of each forecast product. "Good" ("poor") error entries would require minimal (significant) uncertainty accountability and/or potential performance-adjustments in the direction dictated by the "+/-".

Table 1. Aggregate forecast performance summary table.

			Tactical (1-3 hr)							Strategic (4-8 hr)							Long-Strategic (9-12 hr)						
			LAMP		SREF		HRRR			LAMP		SREF		HRRR			LAMP		SREF		HRRR		
			Reliability	Skill	Reliability	Skill	Intensity		Coverage	Reliability	Skill	Reliability	Skill	Intensity		Coverage	Reliability	Skill	Reliability	Skill	Intensity		Coverage
							VIP 3	VIP 5						VIP 3	VIP 5						VIP 3	VIP 5	
Northeast	Line	15-03Z																					
		03-15Z																					
	Org with Gaps	15-03Z																					
		03-15Z																					
	Air Mass	15-03Z																					
		03-15Z																					
Southeast	Line	15-03Z																					
		03-15Z																					
	Org with Gaps	15-03Z																					
		03-15Z																					
	Air Mass	15-03Z																					
		03-15Z																					
West	Line	15-03Z																					
		03-15Z																					
	Org with Gaps	15-03Z																					
		03-15Z																					
	Air Mass	15-03Z																					
		03-15Z																					

6.1 *Probabilistic Forecast Performance Summary Representation – SREF, LAMP*

To aggregate SREF and LAMP forecast skill in each target planning period, the Area Under the (ROC) Curve (AUC) values computed for each event type were placed into three performance categories (Table 2). To compensate for the upper limit of possible skill (AUC) values of 1.0 being not realistically attainable, as it would indicate that a forecast can always generate higher (lower) probabilities when convection does (does not) occur with perfect accuracy, the ranges for the top two performance categories (good, moderate) are wider than for the lowest (poor). The minimum possible skill (AUC) value of 0.5 indicates that the likelihood of a forecast producing higher (lower) probabilities when convection does (does not) occur is only 50%, or equivalent to a coin flip, meaning it has poor skill and little potential up-side from error-adjustment of probability values (reliability). As with most forecast performance category definitions in this technical feasibility study, each of these thresholds can be considered initial settings, which can be modified, parameterized, and even made to be dynamic, with more research and additional operational considerations.

Table 2. Skill metric ranges for each probabilistic forecast performance category.

	Skill (AUC)
Good	0.8-1.0
Moderate	0.6-0.8
Poor	0.5-0.6

The weighted mean difference values computed for all classified 2012 convective weather events in the three target planning periods were also placed into three performance categories containing a bias indicator denoted as a "+/-" that was not included in the skill entries (Table 3). The ranges were selected based on the original binning of the probability values for analysis in 5% increments and the corresponding "+/-" value is based on the sign of the weighted mean difference value. For example, the average weighted mean difference statistic for strategic LAMP predictions of Northeast line storms occurring from 15-03Z is +0.17, meaning that a given forecast probability value, on average, is 17% above the actual frequency of convection occurrence. This indicates that these strategic LAMP predictions over-forecast the expected frequency of convection occurring for this event type by at least 10%, so this aggregate performance summary entry includes a red "+".

Table 3. Probabilistic forecast reliability performance category ranges for weighted difference values.

	Over-Forecast "+"	Under-Forecast "-"
Good	0 to 5%	-5% to 0%
Moderate	5% to 10%	-10% to -5%
Poor	> 10%	< -10%

6.2 *Deterministic Forecast Performance Summary Representation – HRRR*

The full range of potential storm coverage differences (forecast-observed) are placed into categories representing good, moderate, and poor skill for both over-forecast ("+") and under-forecast ("-") bias indicators (Table 4). The coverage difference ranges in each category were determined through assessment of storm coverage differences among synthetic storms of both air mass and line storms to represent operational tolerance to these coverage differences.

Table 4. Coverage difference ranges for each performance category.

	Over-Forecast "+"	Under-Forecast "-"
Good	0 to 10%	-10% to 0%
Moderate	10% to 30%	-30% to -10%
Poor	> 30%	< -30%

To determine the aggregate forecast performance summary coverage entry for a given combination of convective characteristics and planning period, the percentage of errors falling in each of the six coverage difference ranges (represented in each performance category) is first computed. For tactical predictions of line storms in the West occurring from 15-03Z, coverage differences were observed in all six categories (Table 5). The coverage entry category is determined to be that which contains the greatest percentage of all coverage differences. Thus, the coverage entry in the aggregate forecast performance summary for West line storms occurring from 15-03Z would be yellow "+", indicating that the majority of coverage differences (40% of all) associated with predictions of this convective weather scenario reflect moderate over-forecasting (10-30% coverage).

Table 5. Percentage of all coverage differences in each category for predictions of West line storms occurring from 15-03Z.

	Over-Forecast "+"	Under-Forecast "-"
Good	1%	1%
Moderate	40%	29%
Poor	15%	14%

In order to better represent storm intensity differences (forecast – observed) for each combination of convective characteristics in an operational context in the aggregate forecast performance summary, these values are expressed categorically through two critical Video Integrated Processor (VIP) levels: Level 3 and Level 5. VIP Level 3 corresponds to the minimum storm intensity that should be avoided according to the pilot's handbook but is sometimes still traversed by air traffic, and VIP Level 5 corresponds to the storm intensity that the majority of commercial air traffic will more certainly avoid. Specifically, these intensity aggregate forecast performance summary entries categorically (good, moderate, poor) reflect the average difference in VIP levels between HRRR forecast and observation across all hex cells in a given event category. As with the previously described coverage differences, intensity differences in the aggregate forecast performance summary also include an indication of over ("+") or under ("-") forecasting of VIP level by the HRRR.

To generate these entries, VIL distributions from within hex cells were first converted to VIP levels using the commonly accepted ranges (Table 3 9). When the HRRR forecasted VIP Level 3 or VIP Level 5 in a hex cell for a given event category, the percentage of hex cells that were observed at each VIP Level (0-6) was computed to characterize the distribution of observed precipitation intensities when these forecasts were issued.

Table 6. Conversion table from VIL(kg m^{-2}) to VIP Level (from Troxel and Engholm 1990).

VIP level	1	2	3	4	5	6
VIL threshold (kg/m^2)	0.14	0.7	3.5	6.9	12.0	32.0

In order to consolidate these distributions of observed intensity into one forecast aggregate performance summary entry for VIP Level 3 and VIP Level 5, the average frequency-weighted intensity difference is computed as the difference between forecasted and observed intensity levels. For all possible forecast-observed VIP Level pairs, the intensity difference in VIP levels is multiplied by the frequency with which that pair occurred, as

$$Weighted\ Difference = \sum Freq * (VIP_{pred} - VIP_{obs}) \quad (4)$$

For some event categories and planning periods, the HRRR forecasted either VIP Level 3 or Level 5 too infrequently for the 2012 convective season events analyzed in this study to generate a distribution of observed VIP Levels. To ensure a sufficient sample size from which to generate aggregate performance summary entries, weighted intensity difference values were only computed if at least ten (10) hex cells contained VIP Level 3 or VIP Level 5. Entries for event categories not meeting this criteria are shown as empty in the aggregate forecast performance summary.

These weighted VIP intensity differences are subsequently placed into performance categories for each combination of convective characteristics and planning period, including an expression of forecast bias, based on operational tolerance to various error ranges (Table 7). Any weighted differences that are less than one VIP level are considered "good" and operationally tolerable as they require no performance-adjustment, differences of one VIP level are considered "moderate" and marginally-tolerable operationally, and differences of 2 VIP levels or more are considered "poor" and would require significant performance-adjustment. Because the range of intensity differences in the good category are less than 1 VIP level for either over or under-forecasting ("+/-"), any aggregate forecast performance summary entry in this category would not require any VIP based performance-adjustment. In these situations, the "+/-" bias indication is retained to provide enhanced forecast performance awareness of the underlying VIL distributions which may be leveraged when combining with other forecast products in generating a common, combined convective weather forecast region.

Table 7. Weighted VIP difference ranges for each aggregate forecast performance summary category.

	Over- Forecast "+"	Under-Forecast "-"
Good	0 to 1	-1 to 0
Moderate	1 to 2	-2 to -1
Poor	> 2	< -2

6.3 *Aggregate Forecast Performance Summary*

The completed aggregate forecast performance summary table for convective weather provides guidance on uncertainty-awareness and potential error-adjustments based on the target error types and included forecast products (Table 8). This information, derived from detailed and focused forecast performance analyses and characterizations, could in turn support the development of a common, combined convective weather prediction for ATM decision-makers. It also provides information on (a) the varied performance of an individual forecast product given different types of operationally-relevant weather events (i.e., by reading the summary table down the columns for an individual forecast) and (b) the relative performance of each forecast model, the latter of which can guide the relative weighting of each model in a common, combined forecast "polygon" (i.e., by reading the summary table across the rows to account for all forecasts for a given weather event type). For example, long-strategic (9-12 hour) LAMP forecasts of Southeast line storms occurring from 15-03Z have both poor reliability and skill while the HRRR is good at predicting storm intensity and has moderate skill in storm coverage prediction. This suggests that the LAMP should be deemphasized and the HRRR emphasized in generating the combined convective prediction for this event scenario.

Table 8. Aggregate forecast performance summary table populated for target forecast products.

			Tactical (1-3 hr)							Strategic (4-8 hr)							Long-Strategic (9-12 hr)						
			LAMP		SREF		HRRR			LAMP		SREF		HRRR			LAMP		SREF		HRRR		
							Intensity							Intensity							Intensity		
			Reliability	Skill	Reliability	Skill	VIP 3	VIP 5	Coverage	Reliability	Skill	Reliability	Skill	VIP 3	VIP 5	Coverage	Reliability	Skill	Reliability	Skill	VIP 3	VIP 5	Coverage
Northeast	Line	15-03Z	+		-		-	+	+	+		-		-	+	+	+		-		-	+	+
		03-15Z	+		-		-	+	+	+		-		-	+	+	+		-		-		+
	Org with Gaps	15-03Z	+		-		-	+	+	+		-		-	+	+	+		-		-	+	+
		03-15Z	+		-		-	+	+	+		-		-	+	+	+		-		-	+	+
	Air Mass	15-03Z	+		-		-	+	+	+		-		-		+	+		-		-	+	-
		03-15Z	+		-				+	+		-				-	+		-				-
Southeast	Line	15-03Z	+		-		-	+	+	+		-		-	+	+	+		-		-	+	+
		03-15Z	+		-				+	+		-			+	+	+		-				+
	Org with Gaps	15-03Z	+		-		-	+	+	+		-				+	+		-				+
		03-15Z	+		-		-	+	+	+		-		-		+	+		-		-		+
	Air Mass	15-03Z	+		-		-		+	+		-		-		+	+		-		-		+
		03-15Z	+		-				+	+		-				+	+		-				+
West	Line	15-03Z	+		+		-	+	+	+		-		-	+	-	+		-		-	+	-
		03-15Z	+		-		-	+	-	+		-		-	+	-	+		-		-		-
	Org with Gaps	15-03Z	+		-		-	+	+	+		-		-	+	-	+		-		-		-
		03-15Z	+		-		-	+	-	+		-		-	+	-	+		-		-	+	-
	Air Mass	15-03Z	+		-		-	+	+	+		-		-	+	+	+		-		-		+
		03-15Z	+		-		-	+	+	+		+		-	+	+	+		+		-	+	-

Because each forecast performance aggregate summary entry was evaluated independently, a given forecast model may have different performance categories for multiple forecast characteristics. For example, the LAMP has good skill at tactical predictions of West line storms occurring from 15-03Z while having poor reliability, greatly over-forecasting probability values (Table 8). The gross over-forecasting indicates that the observed frequency of convection is at least 10% lower than indicated by the forecast probability values, so a probability forecast value of 40% for this type of event would have an actual frequency of convection below 30%. Despite the fact that the forecast probability values have limited accuracy in representing the expected convection likelihood, the forecast does have good skill for this type of event which indicates that it can effectively discriminate between events and non-events, predicting greater (lower) probability values where convection does (does not) occur.

The aggregate forecast performance summary helps to improve understanding of how forecast performance may vary by lead time, which in turn may facilitate improved operational interpretation of these forecasts from the longer (advanced planning for significant weather-induced congestion events) to the shorter (tactical management of route, fix, airport impacts) lead times. A summary table interpretation example of forecast performance considerations by planning lead time is as follows:

- **For long-strategic and strategic predictions of organized storms with gaps in the West occurring from 15-03Z** (see Table 8), both the LAMP and SREF have moderate skill while the HRRR moderately under-forecasts the spatial coverage of the storms, depicting them as 10-30% smaller than observed. The SREF has good reliability for this event scenario, only slightly under-forecasting the likelihood of convection (< 5%) while the LAMP moderately over-predicts the convection likelihood (5-10%). Therefore, in considering the performance categories of all three forecast products, the HRRR and SREF would likely receive more emphasis than the LAMP in generating a common convective weather prediction region and require little performance-adjustment .

- **As forecast lead times progress into the tactical planning period (1-3 hr) for these predictions of West organized storms with gaps occurring from 15-03Z**, the forecast performance summary entries for several of the forecast characteristics change, indicating that they must be interpreted and potentially combined differently closer to the time of convective constraints(see Table 8). Unlike in the strategic and long-strategic planning periods, tactical LAMP forecasts show good skill in predicting these events, indicating that a LAMP forecast should be emphasized much more strongly in this planning period than in the longer planning periods where it had poor skill. However, due to its poor reliability, the forecast probabilities would need to be adjusted downward at least 10% to compensate for its over-prediction and thus more accurately reflect the expected convection likelihood. On the other hand, the HRRR shows a reduction in skill at predicting VIP Level 5 intensity in the tactical planning period, moderately over-forecasting these intensities (1 VIP level) where there was only slight over-forecasting in the strategic planning period. The HRRR coverage errors also change signs from the strategic to tactical planning periods, so storms depicted in tactical predictions are now 10-30% larger than observed.

In event scenarios like this, the concise guidance available in the aggregate performance summary facilitates improved performance awareness of each forecast model individually as well as enables understanding of relative emphasis across the various forecast models.

The aggregate performance summary also facilitates enhanced performance-awareness and interpretation of each forecast model for a given lead time forecast when different patterns of convection are predicted to occur in separate regions. A summary table interpretation example of forecast performance considerations given different this type of scenario is as follows:

- A common convective weather scenario in the spring and early summer is a **line of convection in the Northeast and air mass storms in the Southeast occurring between 15-03Z**. According to the forecast performance summary table (Table 8), the performance of each individual model product would be different for both weather patterns.

- For a **strategic** forecast of this type of convective weather scenario, the HRRR has similar performance for both the **Northeast line and the Southeast air mass storms** as both would require a reduction in spatial extent of storms to compensate for moderate coverage over-forecasting (10-30%).

- There is insufficient data to guide improved performance awareness of VIP Level 5 storm intensity predictions for the **Southeast air mass storms**, but any high intensities in the **Northeast line** would need to be reduced to compensate for the moderate over-forecasting of these intensities (reduce by 1 VIP level).

- Both probabilistic forecasts have different performance for the two types of weather events. While the SREF has moderate skill for both the **Northeast line and Southeast air mass storms**, it moderately (5-10%) under-predicts the likelihood of convection associated with Southeast air mass storms and only slightly (< 5%) under-predicts the convection likelihood for Northeast line storms.
    - This means that when interpreting a forecast for this broad (overall) convective weather scenario, the actual likelihood of convection is slightly greater than the probability values predicted in the Northeast and moderately greater than those predicted in the Southeast.

• The LAMP greatly over-predicts (> 10%) the likelihood of convection for both **Northeast line storms and Southeast air mass storms**, so the actual convection likelihood is far lower than the predicted probability values. However, the LAMP also has poor skill in predicting Southeast air mass storms, meaning it has limited ability to issue greater (lower) probability values where there is (is not) actual convection, while having moderate skill and thus potential utility for the Northeast line storm prediction. This suggests that the LAMP should be deemphasized in the Southeast relative to the other two forecast models but could provide useful information when combined with the HRRR prediction in the Northeast for this convective weather scenario.

There is also operational utility of the aggregate forecast summary table when all available forecast products have limited capabilities to accurately predict event characteristics for a given weather impact scenario. When this occurs, the overall information highlighting prediction shortfalls from multiple forecast products may help to manage risk and drive cautious and iterative impact planning which is more appropriate than aggressive tactics given awareness of the elevated weather uncertainty. A summary table interpretation example of forecast performance considerations given different this type of high uncertainty scenario is as follows:

• Probabilistic predictions of **Southeast line storms occurring from 03-15Z for tactical planning periods** are poor for both reliability (over-forecasting by more than 10%) and skill, suggesting that the LAMP and the SREF forecasts should be de-emphasized in this planning period (Table 3 13).

• There was an insufficient amount of historical data to guide the assessment of HRRR intensity performance, at either VIP Level 3 or 5, so no guidance on potential intensity performance-adjustment is available.

• The HRRR moderately over-predicts (10-30%) the spatial coverage of these types of storms.

• The poor performance of both LAMP and SREF along with the marginal HRRR performance (what is available) suggests forecast information for tactical impact planning can be highly uncertain and decision-makers should account for this when considering impact mitigation solutions.

Development of this preliminary aggregate forecast performance summary, along with its potential utility in realistic operational convective weather constraint scenarios discussed in this section, demonstrate that it is technically feasible to summarize and combine pertinent convective weather forecast performance metrics into a consolidated expression of forecast performance to enhance operational performance-awareness. While this preliminary aggregate performance summary includes the specific convective weather characteristics previously identified in this study to demonstrate technical feasibility, these parameters could be expanded and extended in future analysis to include more forecast characteristics, additional convective weather characteristics, additional weather forecast products, and predictions of other weather phenomena, such as surface winds, ceiling and visibility, winter-weather (or specific types), etc.

7. SUMMARY

This study investigated the technical feasibility of combining forecast performance relative to previously identified types of pertinent convective weather forecast characteristics into a consolidated, more insightful, expression of forecast performance. This has been demonstrated by combining and preliminarily thresholding performance characteristics from three forecast products targeting operationally-relevant ATM planning periods to form an initial aggregate forecast performance summary "scorecard" for improved forecast uncertainty-awareness. This "scorecard" contains a single categorical expression, or entry, denoting forecast performance associated with each forecast characteristic for all convective weather scenarios (weather pattern, time of day, region) and operational planning periods (tactical, strategic, long-strategic). For each forecast model product included in this study (HRRR, SREF, LAMP), aggregate forecast performance summary entries for one forecast characteristic include a combined expression of categorical performance (good, moderate, poor) and forecast bias (over or under-forecasting). This guidance enhances performance-awareness and operational interpretation of these forecasts.

The utility of fully-populated aggregate forecast performance summary table was described for several realistic operational convective weather impact scenarios. This guidance could support enhanced operational interpretation of target forecast products across planning periods, from consideration of capacity reduction at long lead times to potential blockage of air routes at short lead times as well as across convective weather scenarios, such as for a forecast predicting various weather types in different regions. There is also operational utility of the aggregate forecast summary table when all available forecast products have limited capabilities to accurately predict event characteristics for a given weather impact scenario. When this occurs, the overall information highlighting prediction shortfalls from multiple forecast products may help to manage risk and drive cautious and iterative impact planning which is more appropriate than aggressive tactics given awareness of the elevated weather uncertainty.

It was also demonstrated that the aggregate forecast performance information may support development of a common, combined, performance-adjusted convective weather prediction. As a result, this technical feasibility study is an explicit substantiation for the concept of an aviation weather "Single Authoritative Source" (SAS) – where the SAS receives multiple weather data inputs and then outputs an appropriate, unified weather solution to ATM users seeking to address specific weather constraint conditions.

The improved forecast performance awareness provided by this operations-relevant performance analysis, the aggregate performance summary table, and the concept for a consolidated forecast output based on this information, is useful to both:

1. ATM decision-makers considering explicit weather impact management strategies and actions and;

2. Aviation meteorologists supporting the air traffic mission and counseling traffic managers about the validity of operational forecast products given specific weather impact events.

Moreover, opportunities to "performance-adjust" both probabilistic and deterministic predictions of operationally-relevant convective weather characteristics may support more appropriate weather forecast usage by automated DSTs seeking out critical thresholds for impact "alarming" which may trigger specific ATM alerts or impact mitigation recommendations.

8. REFERENCES

Alexander, C., S. Weygandt, S. Benjamin, D. Dowell, T. Smirnova, M. Hu, J. Brown, P. Hofmann, E. James, H. Lin, 2012. The 2012 High-Resolution Rapid Refresh (HRRR): WRF enhancements and challenges. 13th WRF Users Workshop. Boulder, CO.

AMS, 2008: Enhancing weather information with probability forecasts: An information statement of the American Meteorological Society. Bull. Amer. Meteor. Soc., 89, 1049-1053.

Bright, D.R., M. S. Wandishin, R. E. Jewell, S. J. Weiss, 2005. A physically based parameter for lightning prediction and its calibration in ensemble forecasts. Conf. on Meteor. Appl. of Lightning Data, San Diego, CA, Amer. Meteor. Soc., 4.3.

______, and J.S. Grams, 2009. Short Range Ensemble Forecast (SREF) Calibrated Thunderstorm Probability Forecasts: 2007-2008 Verification and Recent Enhancements. Preprints, 4th Conf. Meteorological Applications of Lightning Data, Phoenix, AZ, Amer. Meteor. Soc., 6.3.

Brody, F., R. Lafosse, D. Bellue, and T. Oram, 1997: Operations of the National Weather Service Spaceflight Meteorology Group. Wea. Forecasting, 12, 526–544.

Charba, J. P., F. G. Samplatsky, and P. E. Shafer, 2011. Experimental LAMP 2-H convection guidance forecasts on a 20-km grid. 24th Conf. on Weather and Forecasting, Seattle, WA, Amer. Meteor. Soc, J19.3.

Cook, L. and B. Wood, 2009: A model for determining ground delay program parameters using a probabilistic forecast of stratus clearing. 8th USA/Europe Air Traffic Management Research and Development Seminar, Napa, CA

Davidson, G., J. Krozel, S. M. Green, C. K. Mueller, 2004: Strategic flow management concept of operations. AIAA Aircraft Technology, Integration, and Operations Conf., Chicago, IL., American Institute of Aeronautics and Astronautics.

Davis, C., B. Brown, and R. Bullock, 2006. Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. Mon. Wea. Review, 134, 1785-1795.

Evans, J., M. E. Weber, and W. Moser, 2006: Integrating advanced weather forecast technologies into air traffic management. Lincoln Laboratory Journal, 16, 81-96.

Fahey, T., M. Phaneuf, W. Leber, M. Huberdeau, D. Morin, and D. Sims, 2006: Forecasts of Convection for Air Traffic Management Strategic Decision Making: Comparison of User Needs and Available Products. 12th Conference on Aviation, Range, and Aerospace Meteorology, Amer. Meteor. Soc., Atlanta, GA.

Ghiradelli, J.E., and B. Glahn, 2010. The Meteorological Development Laboratory's aviation weather prediction system. Wea. and Forecasting, 25, 1027-1051.

Gibson, C. and S. Sharples, 2010: A new tool for fire management decision support. The '88 Fires: Yellowstone and Beyond, Conference Proceedings. Tall Timbers Miscellaneous Publication, 16, Tall Timbers Research Station, Tallahassee, Florida.

Grams, J.S., W.A. Gallus Jr., S.E. Koch, L.S. Wharton, A. Loughe, and E.E. Ebert, 2006. The use of a modified Ebert-McBride technique to evaluate mesoscale model QPF as a function of convective system morphology during IHOP 2002. Wea. and Forecasting, 21, 288-306.

Greene, D. R. and R. A. Clarke, 1972. Vertically integrated water – A new analysis tool. Mon. Wea. Review, 100, 548-552.

Hirschberg, P., E. Abrams, A. Bleistein, W. Bua, L. Delle Monache, T. W. Dulong, J. E. Gaynor, B. Glahn, T. M. Hamill, J. A. Hansen, D. C. Hilderbrand, R. N. Hoffman, B. Hearn Morrow, B. Philips, J. Sokich, and N. Stuart, 2011: A weather and climate enterprise strategic implementation plan for generating and communicating forecast uncertainty information. Bull. Amer. Meteor. Soc., 92, 1651-1666.

Jolliffe, I.T. and Stephenson, D.B., eds, 2003, Forecast verification: A practitioner's guide in atmospheric science. John Wiley & Sons: Chichester, England.

Kutner, Michael H., Christopher J. Nachtsheim, and John Neter, 2004. Applied Linear Regression Models (Fourth Edition). Boston: McGraw-Hill. Print.

Lack, S.A., G.J. Layne, S. Madine, and J.L. Mahoney, 2011: Quality Assessment of CoSPA, NOAA Technical Memorandum OAR GSD-38.

______, M.P. Kay, G.J. Layne, M.A. Petty, and J. L. Mahoney, 2012a. 2011 quality assessment of CoSPA. Prepared by Quality Assessment Product Development Team, NOAA ESRL, 62pp.

______, A.F. Loughe, M.P. Kay, M.A. Petty, and J.L. Mahoney, 2012b. The NDFD thunderstorm probability field: An assessment of forecast lead time and spatial displacement. Prepared by Quality Assessment Product Development Team, NOAA ESRL, 31pp.

Lakshmanan, V., T. Smith, K. Hondl, G. Stumpf, A. Witt, 2006: A real-time, three dimensional, rapidly updating, heterogeneous radar merger technique for reflectivity, velocity and derived products. Wea. Forecasting, 21, 802-823.

Leite, E., C. Assis Beneti, O. C. Rotunno Filho, 2007: Probabilistic forecast of lightning events applied to structures and electrical systems protection. 9th International Symposium on Lightning Protection, Foz do Iguacu, Brazil.

Mason, I., 1982: A model for assessment of weather forecasts. Aust. Meteor. Mag., 30, 291–303.

Micheas, A.C., N.I. Fox, S.A. Lack, C.K. Wikle, 2007. Cell identification and verification of QPF ensembles using shape analysis techniques. J. of Hydrology, 343, 105-116.

NRC, 2006: Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts. National Academies Press, 124pp.

Operational Readiness and Impact Team, 2011. 2010 CoSPA – LCH operational evaluation: Consolidated Summary Report for 2010 assessment activities. FAA ATO Operations Planning, Atlantic City, NJ.

Pence, M. and T. Zimmerman, 2011: The wildland fire decision support system: Integrating science, technology, and fire management. Fire Management Today, 71, 18-22

Pepper, J. W., K. R. Mills, L. A. Wojcik, 2003: Predictability and uncertainty in air traffic flow management. 5th USA/Europe Air Traffic Management R&D Seminar, MITRE Corporation, 10 pp.

Petry, F., T. Tsui, J. Cook, L. Smedstad, and J. Dykes, 2010: Smart climatology system. Naval Research Laboratory Report NRL/MR/7540-10-9282, Monterey, CA.

Regnier, E., P. Harr., 2006: A Dynamic Decision Model Applied to Hurricane Landfall, Weather and Forecasting, 21, 764-780.

Robinson, M., J. E. Evans, B. A. Crowe, 2002. En route weather depiction benefits of the NEXRAD vertically integrated liquid water product utilized by corridor integrated weather system. 10th Conf. on Aviation, Range, and Aerospace Meteor., Portland, OR, Amer. Meteor. Soc., 5.2.

______, M., J. Venuti, and S. McGettigan, 2011: Use of a High-Resolution Deterministic Weather Forecast (CoSPA) for Strategic Air Traffic Management Decision Support, 2nd Special Symposium on Weather – Air Traffic Management Integration, AMS, Seattle, WA.

Sheth, K., B. Sridhar, D. Mulfinger, 2006: Impact of probabilistic weather forecasts on flight routing decisions. 6th AIAA Aviation, Technology, Integration, and Operations Conference, American Institute of Aeronautics and Astronautics, Wichita, KS.

Stanski, H.R., L.J. Wilson, and W.R. Burrows, 1989: Survey of common verification methods in meteorology. World Weather Watch Tech. Rept. No.8, WMO/TD No.358, WMO, Geneva, 114 pp.

Schwedler, B., R. Solomon, D. Bright, J. Lichty, J. Levit, S. Lack, and T. Lloyd, 2013: Application and Assessment of the Aviation Weather Center's Aviation Winter Weather Dashboard. 16th Conf. on Aviation, Range, and Aerospace Meteor., Austin, TX. Amer. Meteor. Soc., J2.1.

Troxel, S.D., and C.D. Engholm, 1990. Vertical reflectivity profiles: Averaged storm structures and applications to fan beam radar weather detection in the U.S. 16th Conference on Severe Local Storms, Kananaskis Park, Alta, Canada, Amer. Meteor. Soc.

Wang, Y. and Sridhar, B., 2010, Convective weather forecast accuracy analysis at center and sector levels. 29th Digital Avionics Systems Conference (DASC), October 3-7, 2010. Salt Lake City, UT.

Weygandt, S., C. Alexander, S. Benjamin, 2012. Preliminary evaluation of HRRR 2012 warm season. Support of FAA AWRP MDE.

Wilks, D.S., 1995. Statistical Methods in the Atmospheric Sciences: An Introduction. Academic Press, 467 pp.