Tenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2013)

An Operational Evaluation of the Ground Delay Program Parameters Selection Model (GPSM)

Assessment, Benefits, and Lessons Learned Evaluating an Air Traffic Management-Weather Integrated Decision Support Tool

Lara Shisler, Christopher Provan
Mosaic ATM, Inc.
Leesburg, VA, USA
lshisler@mosaicatm.com, cprovan@mosaicatm.com

David A. Clark
MIT Lincoln Laboratory
Lexington, MA, USA
davec@ll.mit.edu

William N. Chan, Shon Grabbe
NASA Ames Research Center
Mountain View, CA, USA
william.n.chan@nasa.gov, shon.grabbe@nasa.gov

Kenneth Venzke, Christine Riley
National Weather Service
Monterey, CA, USA
kenneth.venzke@noaa.gov, christine.riley@noaa.gov

Dan Gilani, Ed Corcoran
Federal Aviation Administration
Washington, D.C., USA
dan.gilani@faa.gov, ed.corcoran@faa.gov

Abstract—**The results of an operational evaluation of an Air Traffic Management (ATM)-Weather integrated tool, the Ground Delay Program (GDP) Parameters Selection Model (GPSM), are presented. A shadow evaluation was conducted in 2011, followed by an operational evaluation in 2012. The execution of these evaluations required collaboration and joint support across various agencies and organizations, including the National Aeronautics and Space Administration (NASA), the Federal Aviation Administration (FAA), the National Weather Service (NWS), Mosaic ATM, and MIT Lincoln Laboratory, along with the participation of the National Airspace System (NAS) user community. The shadow evaluation in 2011 showed that ground delays issued during the initial GDP could have been reduced by 20% if GPSM's recommendations had been used operationally. These promising results led to an operational evaluation the following year. Despite challenges related to unexpected weather patterns, weather sensor outages, and slow user acceptance, analytical results show that GPSM provided benefits when used in operational decision making. On days where GPSM recommendations were closely followed, ground delays were on average 20% lower relative to days where recommendations were not followed, consistent with expectations set in 2011. The gap between planned and observed arrival rates fell by 29% relative to the preceding three years.**

Keywords-field evaluation; SFO Marine Stratus Forecast System; ground delay program; probabilistic forecast; weather integration; automated decision support tool

I. Introduction

A major challenge for air traffic managers in the San Francisco area is the summertime low altitude cloud layer that develops overnight in the San Francisco Bay. This layer, called marine stratus, has a tremendous impact on San Francisco Airport (SFO) arrivals since it precludes simultaneous arrival operations on closely spaced parallel runways, which reduces the arrival capacity from 60 to 30 flights per hour. Frequently, the stratus layer is anticipated to burn off after the first bank of scheduled arrivals, and a strategic Ground Delay Program (GDP) must be implemented. A GDP sets target arrival rates that keep traffic at or below capacity by assigning ground delays to flights at their departure airports in order to defer excess demand to later time periods with available capacity. If a GDP is issued with rates that are too conservative, unnecessary delay is absorbed on the ground at the departure airports, and arrival capacity is wasted. On the other hand, if the GDP rate rises above 30 flights per hour before the stratus burn-off time, then airborne holding and diversions may be necessary. Air traffic controllers prefer to avoid the high cost and risk of holding and diversions, and thus historic practices tended to result in overly conservative programs.

The frequency of marine stratus occurrence at SFO and its tremendous impact on aviation operations motivated the Federal Aviation Administration (FAA) Aviation Weather Research Program (AWRP) to sponsor development of the prototype SFO Marine Stratus Forecast System (MSFS) [1], an automated forecast product designed specifically to predict the time of stratus clearing in the SFO approach zone. Technical

This research has been funded by NASA Ames Research Center, Federal Aviation Administration System Operations, and the National Weather Service.

development of the prototype was led by MIT Lincoln Laboratory in collaboration with the National Weather Service (NWS), San Jose State University (SJSU), and the University of Quebec at Montreal (UQAM). The topology and meteorology of the San Francisco Bay area allowed for the development of a model that would predict stratus clearing with reasonable accuracy. National Aeronautics and Space Administration (NASA) Ames Research Center first recognized the opportunity presented by the MSFS to investigate the integration of a probabilistic weather forecast with air traffic management (ATM) decision making. Analysis showed that despite the deployment of the MSFS in 2004, and the improved accuracy in forecast stratus clearing times provided by the system, there was no measurable improvement in the GDP planning process or the efficiency of the resulting GDPs. The conclusion was that the probabilistic nature of the forecast product was difficult to interpret for traffic managers and that in order to improve GDP efficiency, a model would need to be developed to translate the probabilistic forecast into traffic flow management (TFM) decisions [2].

NASA-funded research conducted by Mosaic ATM toward this goal led to the development of the GDP Parameters Selection Model (GPSM). GPSM integrates the forecast of stratus clearing from the MSFS into the current process of modeling and issuing GDPs at SFO. Utilizing historical forecast performance to build a probabilistic error distribution, the model selects GDP parameters that best balance the objectives of minimizing delay and managing risk. GPSM represents one of the first fully developed tools to achieve the major Next Generation Air Transportation System (NextGen) goal of integrating probabilistic weather forecasts into TFM decision making [3].

The GPSM model and preliminary benefits assessment were described in a paper at the 2009 ATM Seminar [4]. This led to interest by the FAA, who then sponsored the development of a prototype of the GPSM that could be tested by the operational community. GPSM was designed to fit within today's environment with no required changes to the tools used to issue GDPs and only minor procedural changes. This allowed for an operational evaluation of GPSM prior to any NextGen changes to the United States National Airspace System (NAS) infrastructure.

In this paper we report on the conduct of the GPSM operational evaluation over 2011 and 2012. We first provide the NextGen vision for ATM-Weather integration in the NAS. This is followed by a brief overview of the design and implementation of GPSM. We then describe the design of the operational evaluation, followed by the results and benefits of that evaluation. We close with a discussion of lessons learned by this evaluation and conclusions.

II. ATM-Weather Integration in the NAS

Aviation operations are significantly impacted by weather. Weather delays account for 70 percent[1] of the $41 billion annual cost of air traffic delays within the NAS [5]. Approximately two thirds ($19 billion) of weather delays are considered to be avoidable, i.e., unnecessary if weather was forecast with 100% accuracy [6].

The NextGen Concept of Operations [3] defines eight new capabilities to be developed, one of which focuses on assimilating weather into decision making. Because of the profound impact adverse weather has on transportation, there is a major focus on developing new aviation weather information capabilities that will help stakeholders at all levels make better weather-related TFM decisions [7]. Those capabilities will be developed based on three major tenets:

- A common weather picture for all air transportation decision makers and aviation system users;
- Weather directly integrated into sophisticated decision support capabilities to assist decision makers;
- Use of internet-like information dissemination capabilities to realize flexible and cost-efficient access to all necessary weather information.

NextGen decision support tools (DSTs) will directly incorporate probabilistic weather data and aid in the human interpretation of probabilistic weather. This will allow decision makers to determine the best response to mitigate the potential operational impact of weather on both a tactical and strategic time horizon while minimizing delays and restrictions. Using automation to better manage uncertainties associated with weather minimizes capacity limitations and reduces the likelihood of overly conservative actions while maintaining acceptable levels of risk across the NAS.

The NextGen Joint Planning and Development Office (JPDO) sponsored the development of a plan for the integration of ATM and Weather in 2010 [8]. As noted in this plan, the Weather–ATM Integration Working Group (WAIWG) of the NAS Operations Subcommittee of the FAA's Research, Engineering and Development Advisory Committee (REDAC) conducted a 12-month study to examine the potential benefits of integrating weather and ATM. The report of this committee made several recommendations regarding the potential for weather integration to help reduce delays by improving the quality and method of use of weather information and integrating weather support in the NAS.

This plan includes a conceptual flow of weather integration, shown in Fig. 1, which has been accepted by the weather and ATM communities. It serves as an overview of the envisioned NextGen weather concept for enhancing ATM decision making

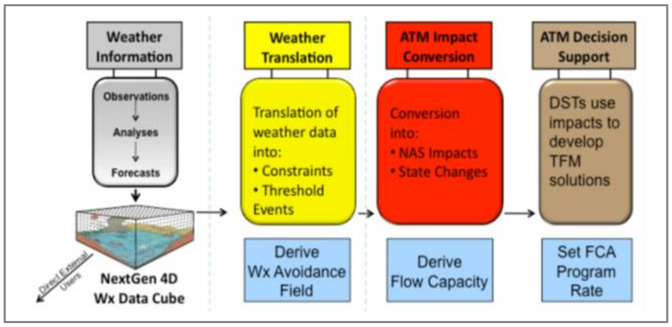

Figure 1. Conceptual flow of weather integration.

[1] The FAA Operations Network (OPSNET)

in the face of adverse weather.

Five levels of weather integration were also defined, each moving closer towards the conceptual vision depicted in Fig. 1.

- Level 0: Stand-Alone Displays – Weather data are displayed on dedicated interfaces separate from any ATM data
- Level 1: On-the-Glass Weather Integration – Weather overlays are added to ATM tools. Examples include the Corridor Integrated Weather System (CIWS) added to the Traffic Situational Display (TSD) and the Weather and Radar Processor (WARP) displays on controllers' Display System Replacement (DSR).
- Level 2: Translated Weather Integration – Automation translates weather data into a constraint, such as a Weather Avoidance Field (WAF). Other examples include the wind shear function of the Integrated Terminal Weather System (ITWS) and the Route Availability Planning Tool (RAPT).
- Level 3: Impact Integration – User-in-the-Loop Tools – Built upon Level 2 technologies, they ingest NAS traffic and other data to determine impact. The Integrated Departure Route Planning (IDRP) is an example of this level of integration.
- Level 4: Machine-to-Machine (M2M) Integration – Constraints (Level 2) and impacts (Level 3) are used by DSTs through M2M integration. Tools provide automated recommendations for ATM decisions without the need of human interpretation or translation.

There are currently no operational Level 4 integration tools in use. The ATM-Weather Integration Plan identified three maturing capabilities that are not yet in operational use but are the most mature new concepts in ATM-Weather integration. These included Integrated Departure Route Planning (IDRP), Collaborative Trajectory Options Program (CTOP), and GPSM. Of these three capabilities, only GPSM meets the criteria for a Level 4 DST by moving beyond impact assessment and providing actual automated recommendations for Traffic Management Initiatives (TMIs). Thus, during the summer of 2012, GPSM became the first Level 4 ATM-Weather integration DST to begin operational trials.

III. GPSM Overview

GPSM is built around a core optimization model that evaluates large sets of GDP parameters and selects the parameters that minimize a weighted sum of the expected excess ground delay and airborne holding subject to certain risk mitigation constraints. The parameters that are optimized as part of the model include start time, end time, airport arrival rate (AAR), and, optionally, geographic scope. For each candidate set of GDP parameters, metrics such as excess ground delay (which occurs when GDPs are too conservative), airborne holding (which occurs when GDPs are too aggressive), and a variety of additional risk metrics are calculated for each possible stratus clearing time based on an error distribution built around the MSFS clearing time forecast. The probabilities of each clearing time are used to calculate an expected value for each metric under each GDP scenario. Then the metrics are combined into an objective function that calculates a cost for a particular GDP scenario, and the GDP scenario with the lowest cost is selected as the recommended program to implement. References [2,4,9] provide a more detailed description of the model.

The implementation of GPSM for SFO is actually a combination of two underlying component models: a weather translation model that translates the forecast of stratus clearing into a probabilistic estimate of capacity, and the optimization model that uses that probabilistic capacity estimate and traffic data to determine the optimal GDP parameters. The separation of these underlying models, shown in Fig. 2, is an important element of the design.

The weather translation component in the SFO implementation uses a historical database of MSFS forecast performance since 1997 (a 15-year archive) to build an error distribution in real-time around any newly generated forecast. This error distribution is dependent on forecast run time and the automated MSFS confidence rating assigned to that forecast. A single outcome from the probabilistic clearing time distribution is translated into an arrival capacity scenario by assuming an arrival rate of 30 flights per hour prior to stratus clearing and 60 flights per hour thereafter.

This weather translation component is specific to the SFO implementation of GPSM. However, because it is independent from the core GPSM optimization model, it can be replaced when operating at another airport by a Weather Translation Model (WTM) that uses the appropriate weather forecast products that best capture the weather factors influencing capacity at that airport and that can translate these weather forecasts into probabilistic estimates of airport arrival capacities.

The GDP parameters that are selected by the optimization component model are displayed in a GPSM table integrated into the web interface for the MSFS forecast tool. All FAA and collaborative users can access this web page. The GPSM table also provides two alternative sets of parameters to show users the impact of issuing a more aggressive or conservative GDP than recommended. A variety of delay and risk metrics are displayed for the recommended and alternative parameters as well as for any SFO GDP currently in place.

An area of further research is the application of GPSM to other airports using available or newly designed airport-specific WTMs. SFO provided a unique opportunity for testing the concept of integrating a probabilistic weather product with TFM decision making due to the fact that there was already an operational automated forecast product available (MSFS) with an established forecast error profile, and that the translation to capacity predictions is straightforward due to the dependence of SFO capacity primarily on the single forecast dimension of stratus clearing time. WTMs for other airports will likely require more complex models that consider a range of different weather factors, such as wind speed, wind direction, ceilings,

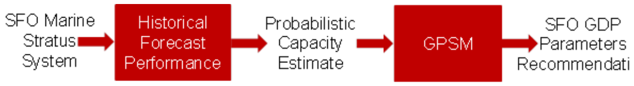
SFO Marine Stratus System
Historical Forecast Performance
Probabilistic Capacity Estimate
GPSM
SFO GDP Parameters Recommendati

Figure 2. GPSM software architecture as implemented at SFO.

visibility, and convection. Probabilistic predictions will be made more difficult by the interdependence of these various forecast dimensions.

IV. Conduct of the Operational Evaluation

The GPSM evaluation was conducted during 2011 and 2012. The MSFS forecasts are available only during the summer stratus season lasting from May 15th to October 15th of each year. The operational evaluation was split into two sections: a shadow evaluation during the 2011 stratus season followed by a full operational evaluation in 2012. The following subsections describe the procedures for each.

A. 2011 Shadow Evaluation

During the 2011 shadow evaluation, GPSM was not used as part of the operational decision making process for issuing SFO GDPs. Instead, a shadow position was designated at the Air Traffic Control System Command Center (ATCSCC) for the personnel conducting the GPSM evaluation. This position neighbored the position responsible for monitoring and issuing west coast airport TMIs, which allowed the staff at the shadow position to monitor the same data sources and to listen to any planning conference calls or discussions. The operational west coast position was not given access to the GPSM recommendations, and GDPs were planned and issued under the existing procedures. The shadow position had all of the same ATM tools at their disposal and additionally had access to GPSM. The recommendations provided by GPSM were monitored from the initial GDP planning time frame until the stratus cleared and the GDP was cancelled. Observations made by the shadow position were captured in the National Traffic Management Log (NTML).

The shadow position personnel were tasked to monitor GPSM to determine answers to the following questions:

- Are GPSM's recommendations for the initial GDP sound?
- How stable are GPSM's recommendations over time as traffic and weather forecasts evolve?
- Does the improved forecast in the 15Z hour, when visible satellite imagery is available, result in GPSM recommendations that are better and that can guide revisions?
- Are there user interface and human factors considerations that should be addressed before operational use?
- Is it clear to the traffic managers on which days GPSM's use is appropriate (i.e., when ceilings are due to "typical" status, for which the forecast system is designed)?

The GPSM technical team spent time throughout the shadow evaluation onsite at the ATCSCC, observing the GDP planning process and interacting with the personnel staffed at the shadow position. The in-person support and gathering of feedback on the tool and procedures was integral to the evaluation process. The NTML comments were also collected and reviewed throughout the season.

A quantitative summary called the GPSM Daily Report was generated on a daily basis to analyze each GDP event during the GPSM evaluation. The generation of the Daily Report is fully automated within the GPSM software. The report is automatically distributed the morning after any stratus-induced GDP, allowing a review of the program while memories are still fresh. This report contains summaries of the previous day's weather forecasts, the actual GDP events implemented, the GPSM-recommended events, and the parameters for the "ideal" GDP – the GDP that would have been issued had there been perfect foresight of stratus clearing. For each set of GDP parameters (actual, GPSM-recommended, and ideal), key metrics are included such as the resulting ground delay, unrecoverable delay, unnecessary delay, and airborne holding. These metrics provided via the Daily Report for each individual program were aggregated at various points in time throughout the season and briefed to the operational community. This process allowed for the technical team to quickly respond to suggested changes to GPSM and to update the software at various times throughout the season. The end result was a much better prototype, one much more suited for actual operational use.

The shadow evaluation revealed some key issues regarding the technology and procedures. On the technology side, minor enhancements were made to the user interface, and the logic for many of the metrics was improved. Better modeling was added for the metrics associated with the actual program issued, which allowed for more meaningful comparison with simulated metrics for GPSM-recommended parameters. A key upgrade added the ability for GPSM recommendations to include a period prior to the end of the program with an arrival rate of the maximum 60 flights per hour. Flight demand immediately following the end of a GDP can sometimes be excessive, exceeding even the maximum capacity. Allowing for a program to be extended at a rate of 60 flights per hour smooths out any excess demand and provides a better transition out of the GDP.

One of the most important outcomes of the shadow evaluation was related to the procedures for communication between meteorologists and traffic managers. As mentioned previously, the MSFS is designed to forecast stratus clearing during typical summer weather patterns, wherein the daily cloud dissipation mechanism is dominated by local physical processes, as is common during the warm season. When larger scale transient weather systems impact the region, as is more common in winter months, there can be a significant degradation of forecast accuracy. The system is designed to automatically recognize some of the conditions under which degraded performance might be expected and issue an indication of lower forecast confidence. However, there are additional days on which the human forecaster can recognize other unusual conditions that may impact the automated forecast quality. Forecast performance in turn impacts the quality of GPSM recommendations at SFO, which rely on the MSFS forecasts to generate probabilistic capacity scenarios. Since the MSFS is fully automated and will generate forecasts of clearing as long as the system's sensors detect that low ceilings are in place in the Bay Area, this provides a challenge for traffic managers who need to recognize the days for which the GPSM recommendation is not intended. Though it may be straightforward for a meteorologist to determine the applicability of the MSFS forecast to GPSM on a given low ceiling day, this is far from obvious for traffic managers. During the shadow evaluation, the staff at the shadow position frequently needed to call the Oakland Center Weather Service Unit (CWSU) to better understand the weather situation on a given day in order to determine whether or not the evaluation of GPSM on that day was appropriate. Though those conversations were helpful, they were time consuming and even sometimes confusing. The meteorologists providing the synopsis of the current conditions and their judgment regarding the forecast often used meteorological terminology, and it was not always clear to the specialists at the GPSM shadow position what the implication was for the use of GPSM in the planning process. At the end of the 2011 stratus season, it was clear that a better procedure was needed in order to more clearly communicate to the traffic management community whether or not GPSM should be used. An improved procedure was designed and implemented for the operational evaluation the following year.

B. *2012 Operational Evaluation*

At the end of the shadow evaluation, the FAA made the decision to proceed with an operational evaluation based on strong quantitative estimates of benefit and on the qualitative feedback from traffic managers and the user community. The period between the end of the shadow evaluation and May 15th, 2012, was spent in preparation for the operational evaluation. This preparation was a collaborative effort between the FAA, NASA, the GPSM technical team, and a variety of other participants. The NWS in particular was fully engaged in the preparation for and execution of the 2012 operational evaluation of GPSM.

One key task was the development of the procedures for communicating whether or not GPSM's use was appropriate on a given day. The work conducted to develop these procedures was an excellent example of inter-agency collaboration between the FAA and the NWS. NWS staff at both Monterey and at the CWSU and the MSFS technical team at MIT Lincoln Laboratory worked together to develop a concept of classifying a low ceiling day as one of the following three categories:

- GPSM: High Confidence – Typical stratus day for which the MSFS was designed. The forecast system and GPSM can be used for guidance with high confidence.
- GPSM: Low Confidence – Typical stratus day, but with some other weather influences contributing to lower confidence in the forecast system. GPSM can still be used for guidance, but with caution.
- Not GPSM – Low ceilings are caused primarily by weather factors other than typical marine stratus, or no ceilings are present. GPSM use is not appropriate.

The meteorologists worked together to define the key weather factors and their threshold parameters that would classify a day into each of these categories. This process facilitated a common understanding of the automation between the developers of the forecast system (MIT Lincoln Lab) and the meteorologist using that system, and it also resulted in clear procedures for the CWSU for doing classification such that individual human judgment or subjective interpretation could be minimized. By making the classification specific to GPSM, it largely removed meteorological interpretation from the role of the traffic manager. A joint training session was conducted between the FAA and NWS on GPSM, and the NWS conducted their own internal training sessions for all of their meteorologists on the classification procedures developed.

Training was also conducted for FAA traffic managers. Twelve training sessions were conducted at the ATCSCC and were open not only to ATCSCC traffic managers but to NAS users as well. The inclusion of these two groups in the same training classes led to lively discussions and provided a viewpoint of GPSM's anticipated impact and objectives of the operational evaluation from both the FAA and user perspectives. Additional training sessions were held at Oakland Center for west coast FAA personnel.

While the software and users were all prepared for the start of the operational evaluation on May 15th, unusual weather patterns contributed to a slow start to the 2012 stratus season, as many low ceiling days were classified as not appropriate for GPSM. By July 15th, there had only been two GPSM days versus an average of nearly 17 typical stratus days up to that point over the previous 5 seasons. The small number of opportunities to use GPSM slowed user acceptance of the tool and made progress at the CWSU towards adjusting to the new procedures surrounding the classification of days difficult. The initial impression of many users was that GPSM could not be relied on to help plan GDP parameters at SFO.

The pattern quickly changed in mid-July as the weather became more typical and the NWS refined their procedures for designating GPSM days. The second half of the month added an additional eight GPSM days, and after August and September, the total count of GPSM days was 30. Even with this increase in activity, the total number of GPSM days in 2012 was well below the number of typical stratus days in previous years, as shown in Fig. 3.

Extraordinary challenges with sensor equipment used by the MSFS also impacted the evaluation. During the stratus season, construction began at SFO in close proximity to a suite of sensor equipment used by the MSFS. Power was cut to the equipment without advance notice on August 31st and was not

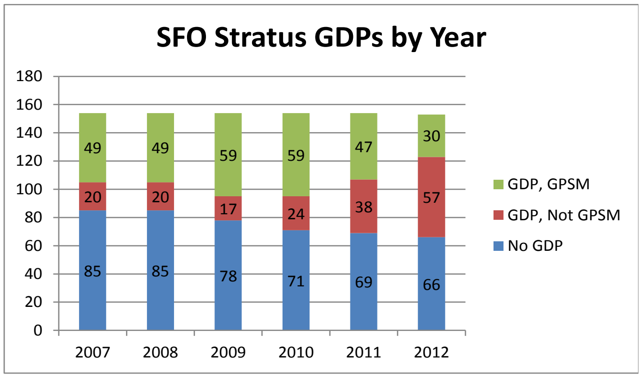

Figure 3. The need for GDPs has been increasing year over year as traffic has grown, but 2011 and 2012 had much higher percentages of days with GDPs that were not caused by typical stratus.

restored for the remaining 6 weeks of the stratus season. The impact was that atmospheric measurements provided by various key sensors were not available to the underlying forecast models in the MSFS. The system was still able to provide overall consensus forecasts issued with lower confidence, but they were based on the less-than-ideal number of measurements, and often only two of the four underlying models were available to create the consensus clearing time. It is unknown exactly how this affected the quality of the forecasts, but the forecast bias shifted from an average of -9 minutes during 2009-2011 (meaning that clearing times were on average 9 minutes later than forecast) with a median absolute error of 38 minutes to +20 minutes with a median absolute error of 40 minutes in 2012, an approximate 30 minute shift in the forecast bias.

Despite these challenges, the data and analyses illustrated that GPSM had an impact in terms of reducing expected ground delay and increasing arrival capacity utilization. The following section contains a quantitative assessment of GPSM's benefits during both the 2011 shadow evaluation and 2012 operational evaluation.

V. BENEFITS ASSESSMENT

The two stages of the operational evaluation of GPSM had to be evaluated differently because of the changes in procedures. The methodology and results for each stage are discussed in the following subsections.

Both evaluations use the same set of metrics to compare GDP efficiency. *Issued ground delay* is the delay assigned to flights affected by the GDP. GDPs are often cancelled before the planned GDP end time. In this case, *absorbed ground delay* measures the amount that flights are actually delayed. Both issued and absorbed ground delay are costly to airlines as large amounts of issued delay create significant planning problems and customer dissatisfaction even if a GDP is cancelled before all of the delay has been absorbed. Absorbed ground delay is computed assuming that flights require 45 minutes in order to respond to GDP cancellation if their controlled time of departure was more than 45 minutes after the cancellation time.

Estimated *airborne holding* is also considered in the evaluation. Because actual airborne holding could not be measured directly from the available data, estimated airborne holding is computed by assuming that, without any airborne delay, flights would arrive at SFO exactly at their issued arrival time (or their estimated arrival time for flights not delayed by the GDP). Airborne delay required to achieve the actual arrival rate implied by the observed stratus clearing time is then computed.

Both evaluations also use the concept of an *ideal GDP* to baseline delay comparisons. This is the GDP that would have been issued on a given day to minimize ground delay without causing any airborne holding if the actual stratus clearing time were known in advance. *Unnecessary delay* is the delay issued or absorbed in a GDP that exceeds the delay that would have been issued under the ideal GDP. The core objective of GPSM is to reduce unnecessary delay. Any delay reduction beyond the unnecessary delay would imply significant increases in airborne holding and potential risk.

A. *2011 Shadow Evaluation*

Decision makers were not able to use GPSM when determining GDP parameters during the shadow evaluation stage in 2011. The benefits assessment therefore focuses on the differences between the efficiency of the GDPs implemented and the GPSM-recommended GDPs. As illustrated in Fig. 3, there were 85 days with GDPs in 2011, of which 47 were determined to be stratus days on which GPSM could have been used.

Fig. 4 compares the total delay issued by traffic managers across the 47 GPSM stratus days in 2011 against the delay that would have been issued under the GPSM recommendations. The delay issued in the initial GDPs would have decreased by approximately 49,000 minutes, or 20%, while the combined delay issued and estimated airborne holding when accounting for the initial GDP and any revisions would have decreased by approximately 42,000 minutes, or 17%. The associated reductions in unnecessary delay issued were 57% for the initial GDP parameters and 48% when accounting for revisions. These delay reductions came with a negligible increase in overall risk – the simulated airborne holding per GDP under the GPSM recommendations was in fact 9% lower than in the actual programs.

A comparison of GDP parameters shows that GPSM achieved these benefits in two ways. First, the initial GPSM recommendations tended to be more aggressive than the actual GDPs. Initial GDPs were on average 40 minutes longer than the GPSM-recommended programs. Second, GPSM-recommended GDPs were more likely than actual GDPs to be revised in response to the improved forecast accuracy at 15Z. This often included recommending revisions with more aggressive GDP parameters when the new forecast warranted such an action. The actual GDPs were revised 8 times across the 47 days included in the analysis with only 1 program revised to an earlier end time whereas GPSM recommended a 15Z revision on 32 of the 47 days, with half of those revisions including an earlier end time.

Absorbed delay under the GPSM-recommended programs would have been reduced by approximately 16,000 minutes. This represents a reduction of 8% in total delay absorbed and a reduction of 37% in unnecessary delay absorbed. In order to

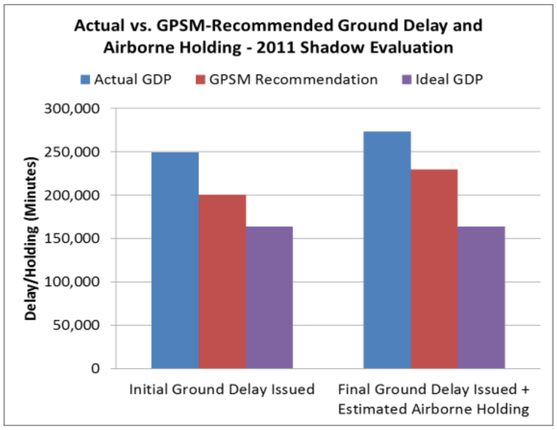

Figure 4. GPSM recommendations called for 20% lower delay in the initial program and 17% less delay overall relative to the actual GDPs.

ensure that these gains were repeatable, GPSM recommendations were computed for past years going back to 2006 using historical weather forecasts and traffic data. Fig. 5 shows that the 2011 absorbed delay benefits were in line with what would have been observed in previous years.

The results of the benefits analysis on data from the 2011 shadow evaluation and previous years provided enough confidence in the GPSM concept and implementation to gain approval for a full operational evaluation in 2012.

B. 2012 Operational Evaluation

During the operational evaluation, all participants in the collaborative GDP planning process were able to view the GPSM recommendations and statistics before and during the SFO planning conference calls. GDP planners were encouraged to base program parameters on these recommendations. As a result, the benefits assessment for the operational evaluation focuses on how the use of GPSM impacted the efficiency of the GDPs issued by ATCSCC specialists compared to previous years.

As shown in Fig. 3, atypical weather patterns during the 2012 stratus season resulted in only 30 days with weather patterns conducive to GPSM use. Those 30 days were further subdivided based on how GPSM was used in the GDP planning process. On 11 of these days, the initial programs issued by ATCSCC specialists aligned closely with the parameters recommended by GPSM. On the remaining 19 days, the initial programs differed substantially from the GPSM recommendations. These two sets of days were kept separate in the benefits analysis in order to better quantify the impact of using GPSM.

While GPSM regularly recommended revisions based on the higher quality forecasts generated in the 15Z hour, observations at the ATCSCC indicated that the GPSM recommendations were not given significant weight during the revision planning process on any of those 30 days. Instead, revisions were typically issued only reactively when it became clear that there was a high likelihood of the stratus persisting well past the initial forecast clearing time. As was the case during the 2011 shadow evaluation, this strategy for revisions reduced the overall efficiency of the implemented programs.

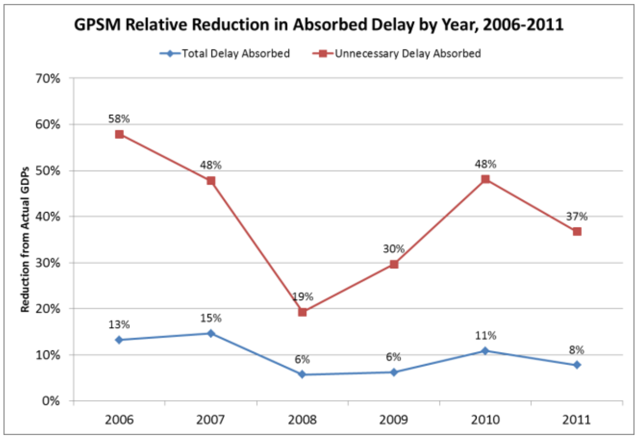

Figure 5. GPSM reductions in total and unnecessary delay absorbed during the 2011 shadow evaluation were in line with data on previous years.

Fig. 6 shows a year-to-year comparison of average initial delay issued per GDP for the actual and ideal programs. The comparison of 2012 operational evaluation results to previous years was made difficult by two changes in the operational environment. First, as mentioned previously, while actual stratus clearing occurred an average of 9 minutes later than forecast during 2009-2011, clearing times on the 30 days included in the 2012 data were an average of 20 minutes earlier than forecast. This artificially pushed up the unnecessary delays during 2012 relative to previous years. Additionally, traffic levels during the morning arrival rush at SFO were 20% higher than in 2011, which led to additional increases in delay.

As a result, the absolute delay comparison in Fig. 6 shows that the average delay issued in the initial GDP was higher for both sets of days in 2012 than for any previous year. However, a comparison between the two subsets of 2012 days shows that initial delay per GDP was 1,630 minutes lower on days the initial GPSM recommendation was followed than on days that it was not. This was a reduction of 20% in initial delay. The reduction was achieved even though the ideal delay was comparable on the two sets of days, and it illustrated a clear benefit of following the GPSM recommendations. However, the unnecessary delay on days when GPSM was followed was still higher than in any previous year studied.

In order to remove the effect of increased traffic, a metric called *missed planned slots* was developed. This metric measures the cumulative difference in slots between the AARs used for the GDP parameters and the actual AARs implied by the observed stratus clearing time. No flight delays are computed, and thus the metric is completely independent from traffic levels. Instead, it is a measure of arrival capacity utilization. A lower value for missed planned slots indicates more efficient GDPs. The blue bars in Fig. 7 compare missed planned slots for all years since 2007 based on the initial GDP parameters. There was a notable decrease between 2007 and 2009 in this metric that coincided with explicit efforts to reduce GDP delays at SFO in reaction to early GPSM research results. The missed planned slots stabilized between 2009 and 2011, and the missed planned slots on days in 2012 where GPSM was not followed for the initial program fell in line with those years. However, there was a clear decrease of 29% on the days on

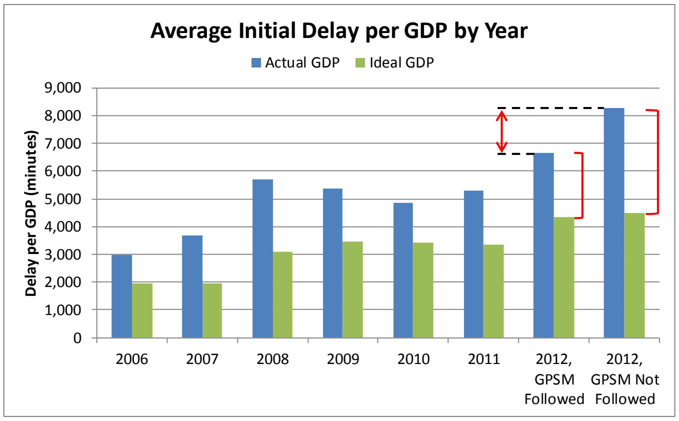

Figure 6. Delay was lower in 2012 on days when GPSM was followed, but absolute 2012 delays were higher than all previous years studied.

which the GPSM-recommended parameters were used relative to the 2009-2011 average. As in the 2011 shadow evaluation, these benefits were achieved without substantially increasing risk – airborne holding per GDP was lower in 2012 than in any year since 2008 and was slightly lower on days for which the initial GPSM recommendation was followed.

To adjust for the impact of the 29-minute change in forecast errors relative to recent years, the metrics for all 2012 GDPs were recomputed with actual clearing times shifted 29 minutes later. This simulates what would have happened in 2012 if the forecasts had been unchanged but forecast errors had matched previous years. This adjustment primarily affects two metrics. Missed planned slots are reduced because the adjusted actual AARs are now lower due to later stratus clearing time. The ideal delay also goes up due to the later stratus clearing time, which results in decreased unnecessary delay.

The red bars in Fig. 7 show the adjusted missed planned slots for both sets of 2012 GDPs. The values are much lower for all 2012 days after the adjustment. There is still a significant difference in adjusted missed planned slots between 2012 days on which the GPSM recommendations were and were not used. However, even on the days that GPSM was not followed, the missed planned slots are now much lower than any previous year. This indicates that the 2012 GDPs would have been more efficient than in previous years across all days had the forecast errors been comparable to previous years. The reduction in adjusted missed planned slots for days that GPSM was not followed also implies that while traffic managers did not follow the recommended parameters precisely on these days, the resulting GDPs were in fact more efficient than in previous years. This suggests that GPSM still had a positive impact on the decision making process on these days.

Fig. 8 looks at the effect of this adjustment on unnecessary delay. Unlike missed planned slots, unnecessary delay does increase as traffic levels increase. The absolute unnecessary delay (blue bars) was nearly 1,000 minutes greater per GDP in 2012 than in any previous year. But unnecessary delay is also impacted by forecast errors. When forecast errors are adjusted to match the 2009-2011 average, unnecessary delay on days when the GSPM recommendation was followed drops to its lowest level since 2007. This decrease happens despite the 20% rise in traffic levels between 2011 and 2012 and a cumulative 30% rise in traffic levels since 2007. On 2012 days that the GPSM recommendation was not strictly followed, the adjusted unnecessary delay is approximately 500 minutes higher per GDP than on days that GPSM was followed. However, unnecessary delay on those days is still comparable to 2008-2011 levels even with the increase in traffic.

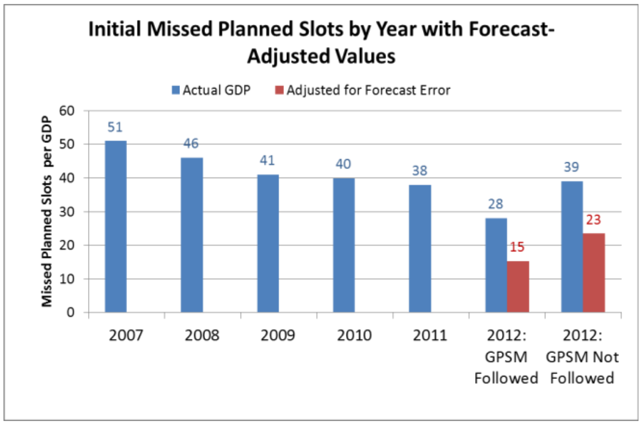

Figure 7. Missed planned slots were noticeably lower in 2012 when GPSM recommendations were followed. Adjusting for changes in forecast error, the missed planned plots for all 2012 days are much lower than previous years.

VI. Lessons Learned

GPSM is one of the first NextGen TFM DSTs to fully integrate probabilistic weather forecasts with automated decision making. Thus an additional benefit of the GPSM field evaluation was lessons learned that can be applied to the development and field evaluation of future NextGen DSTs.

One such lesson was the value of conducting a shadow evaluation ahead of the full operational evaluation. NextGen DSTs represent a paradigm shift in how automation is integrated into the decision making process. Up to this point, most TFM DSTs have focused on providing users with information (scheduled traffic at key points in the system, weather forecasts, etc.) or on presenting this information in a user-friendly way. NextGen tools take the extra step of aggregating necessary data, modeling the effects of uncertainty, and recommending actions to traffic managers. The GPSM shadow evaluation allowed users to begin to understand and build confidence in this type of functionality. Additionally, it allowed the GPSM technical team to test the tool using operational users and improve the methodology and user interface of the tool for the operational evaluation.

The impact of uncertainty and how it is communicated to end users is of particular importance. Pre-NextGen DSTs have left the task of interpreting and addressing uncertainty primarily up to traffic managers. The MSFS forecast tool presented users with measures of uncertainty in the clearing time forecast, but these were not presented in a way that could easily be translated into traffic management impacts or actions. NextGen DSTs must factor uncertainty into their TFM recommendations and decisions. As a result, these decisions will in some individual instances result in suboptimal outcomes. Users must be trained to understand this, and it must be emphasized that judgment on a tool's capabilities cannot be

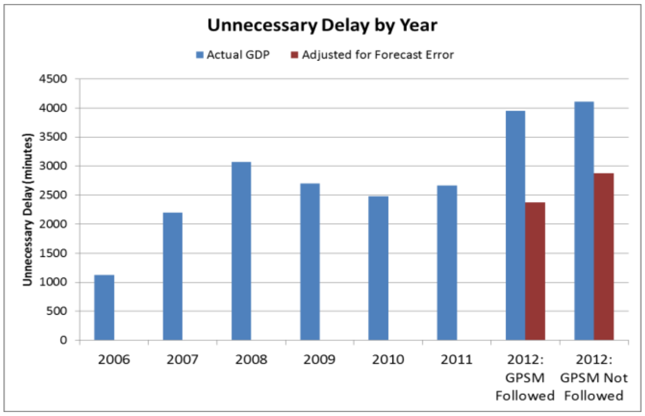

Figure 8. Absolute unnecessary delay was higher in 2012, but when adjusted for forecast errors, it was below 2008-2011 levels despite increased traffic.

made based on one or two outcomes, but should be made based on the aggregate results from a large sample of decisions. If the DST is effective, suboptimal outcomes will be less frequent and less severe than under current operations.

The role of the human weather forecaster continued to be vital in the GDP planning process. NextGen weather TFM DSTs are based on forecasts that have limitations, and the knowledge and experience of meteorologists is required to understand these limitations and, along with TFM experts, determine their impact on the quality of the recommendations and decisions provided. Equally important is the clear definition of how responsibilities are divided between forecasters and traffic managers. In the case of GPSM, this meant that meteorologists at Oakland CWSU with expert knowledge of stratus patterns at SFO were responsible for determining which days were suitable for GPSM, providing a broader weather context for traffic managers, and monitoring the evolving weather scenario throughout the day.

Collaboration with the larger community presents its own set of challenges. Every effort was made to educate all potential GPSM users prior to each phase of the evaluation and to solicit feedback on system and procedure design. Users have a variety of requirements and objectives with the use of a tool, and therefore it is important to seek input from all participants. The feedback received in advance of the trials led to numerous improvements to the software and user interface.

Lastly, the different objectives of collaborative users resulted in different assessments of the results of the field evaluation. Airlines have generally seen GPSM as an important first step towards integrating advanced automation into TFM decisions, resulting in lower levels of delay and greater predictability. For traffic managers, delay reduction is a secondary objective to maintaining system safety. The response from traffic managers ranged from general acceptance with an understanding that the tool would continue to be updated and improved to dislike by many ATCSCC specialists due to the reduced flexibility in determining GDP parameters. While a reduction in assigned ground delay may result in measurable benefits to the user community, these delay reductions are not necessarily evident on a daily basis to the ATCSCC specialist issuing the program, and, as we discovered, this sometimes negatively impacted their perception of GPSM's usefulness.

VII. CONCLUSION

Though it is unclear whether or not GPSM will transition from a prototype to a future operational tool in the Traffic Flow Management System (TFMS), the conduct of the operational evaluation was an important step towards better understanding both the challenges and potential benefits of a Level 4 ATM-weather integrated DST.

The benefits measured during the 2012 use of GPSM are summarized as follows:

- There were over 1,600 fewer minutes of initial delay per GDP (a 20% reduction) when GPSM recommendations were followed than when they were not.
- Planned use of arrival slots post-clearing improved by 29% over 2009-2011 levels when GPSM recommendations were followed.
- When adjusted for changes in forecast errors, planned use of arrival slots was improved by 62% relative to 2009-2011, and unnecessary delay was at its lowest levels since 2007 on days where GPSM recommendations were followed.
- Even when GPSM recommendations were not strictly followed, adjusted missed planned slots were 42% lower than the 2009-2011 average, and unnecessary delay was comparable to recent years even though 2012 traffic levels were substantially higher.
- GPSM benefits were achieved with negligible increase in risk.

The 2012 results continue to indicate that GPSM can provide benefits in terms of reduced delays and increased capacity utilization at SFO during typical summer stratus events. The data showed that the benefits could have been even greater if GPSM was used to plan revisions in the 15Z hour – on days that the initial GPSM recommendation was followed, GPSM-recommended revisions would have eliminated an additional 1,500 minutes of issued delay and 1,000 minutes of absorbed delay per GDP.

NextGen DSTs like GPSM do not negate the importance of improving the accuracy of weather forecasts. An analysis that manipulated clearing time forecast accuracy over the 2011 stratus season showed that a 10% reduction in forecast error would have allowed GPSM to reduce unnecessary delay by an additional 20%. However, error is an inherent part of weather forecasting, and GPSM and similar Level 4 ATM-weather integrated tools can reduce the negative impact of these errors on operational efficiency.

The operational evaluation also brought to light areas where GPSM's use can continue be improved. Better modeling of transition rates around the stratus clearing time would lead to more accurate metrics and more efficient recommendations. Improved identification of days where GPSM use is appropriate can expand the use of the tool and increase the overall benefit. And working with users to develop appropriate methods for reducing the frequency of changes in GPSM's recommendations will increase user acceptance. These changes will only add to the benefits observed during the 2011 and 2012 evaluation.

The GPSM concept is flexible and requires only that forecasts be translated into probabilistic capacity estimates for an airport. Thus, as new WTMs for other airports are developed, GPSM can be ported to other airports. Further research is required to determine if modifications to GPSM's model would be necessary to support specific airport/TRACON operations and concerns. It is our belief that GPSM's benefits can be realized at other airports in the NAS.

ACKNOWLEDGMENT

Mosaic ATM would like to acknowledge NASA Ames for providing the funding for the 2011 and 2012 GPSM field evaluations and the initial weather integration research starting in 2007 that led to the development of GPSM. Mosaic ATM would also like to acknowledge Midori Tanino of the FAA for her recognition of the potential of the GPSM concept while it was still a research effort at NASA and for providing the initial funding to develop a prototype version of the tool. Her interest was the result of the GPSM paper presented at the 2009 ATM Seminar. The GPSM operational evaluation required the involvement of many individuals across multiple organizations. We appreciate the collaborative spirit that existed amongst many personnel at NASA, FAA, NWS, and the technical team of Mosaic ATM and MIT Lincoln Laboratory. Lastly, we acknowledge the significant work done by many individuals at the FAA ATCSCC to support and conduct the evaluation, expertly led by Ed Corcoran and Daniel Horton.

REFERENCES

[1] Clark, D. A. and Wilson, F. W., "The San Francisco marine stratus initiative," 7th Conference on Aviation, Range and Aerospace Meteorology, Long Beach, CA, pp. 384-389, 2003.

[2] Clark, D., "Investigating a new ground delay program strategy for coping with SFO stratus," Aviation, Range, and Aerospace Meteorology Special Symposium on Weather - Air Traffic Management Integration, 89th AMS Annual Meeting, Phoenix, AZ, 11-15 January, 2009.

[3] Joint Planning and Development Office, "Concept of operations for the next generation air transportation system v 2.0," June 2007.

[4] Cook, L. and Wood, B., "A model for determining ground delay program parameters using a probabilistic forecast of stratus clearing," Eighth USA/Europe Air Traffic Management Research and Development Seminar, Napa, CA, June 2009.

[5] Joint Economic Committee Majority Staff, "Your flight has been delayed again; flight delays cost passengers, airlines, and the U.S. economy billions," May 2008.

[6] Weather-ATM Integration Working Group, "Research, engineering and development advisory committee," October 2007.

[7] Weather Integrated Product Team, Joint Planning and Development Office, "Weather concept of operations, v 1.0," May 2006.

[8] Joint Planning and Development Office: ATM-Weather integration plan, "Where we are and where we are going, v2.0," September 2010.

[9] Cook, L. and Wood, B., "A model for determining ground delay program parameters using a probabilistic forecast of stratus clearing," Air Traffic Control Quarterly, Volume 18 (1), 2010.

AUTHOR BIOGRAPHIES

Lara Shisler received an M.S. in 1991 in operations research and management science from George Mason University in Fairfax, VA. Her B.S. in mathematics was received in 1989 from The College of William and Mary in Williamsburg, VA.

She is a Principal Analyst with Mosaic ATM in Leesburg, VA, managing projects related to TFM and the integration of weather with ATM decision making. She has over 15 years' experience supporting research and development activities for the FAA and NASA in ATM, both at Mosaic ATM and Metron Aviation. Prior to that, Ms. Cook worked as an Operations Research Analyst for two major air carriers, passenger and cargo.

Christopher Provan received his M.S. in operations research from Cornell University in Ithaca, NY in 2008. He received a B.S. in mathematics and secondary education in 2003 from Vanderbilt University in Nashville, TN.

He is a Senior Analyst with Mosaic ATM in Leesburg, VA. He has supported multiple field evaluations of TFM decision support tools, and his research has addressed ATM topics including automation of TFM decision making, optimization of surface and terminal area operations, and probabilistic capacity prediction.

Dave Clark is a technical staff member in the Weather Sensing Group at MIT Lincoln Laboratory. He received degrees in meteorology from the University of Massachusetts at Lowell (B.S., 1981) and M.I.T. (S.M., 1983). He worked at Raytheon Company on the Next Generation Weather Radar system (WSR-88D) prior to joining Lincoln Laboratory in 1987. His early work within the Weather Sensing Group was associated with hazardous wind shear detection, making contributions to the Terminal Doppler Weather Radar (TDWR), Low Level Wind Shear Alert System (LLWAS), and the Integrated Terminal Weather System (ITWS). He served for five years as the lead of the Terminal Ceiling & Visibility Product Development Team within the FAA's Aviation Weather Research Program. He now serves as the Lincoln Lab technical lead in support of FAA programs for wake turbulence and for aviation environmental impact modeling. Dave is a member of the American Meteorological Society.

Shon Grabbe received a Ph.D. in 1997 in theoretical atomic physics from Kansas State University. He has over 15-years of experience in the air traffic management domain, and is the lead of NASA's Traffic Flow Management research area. Dr. Grabbe has over 60 publications in the fields of physics and air traffic management, and is an Associate Fellow of the AIAA.

William N. Chan is the Chief of the Systems Modeling and Optimization Branch at NASA Ames Research Center. Prior to joining NASA, Mr. Chan served as a lecturer in the Meteorology department at San Jose State University and was a C-17 flight test engineer at the Air Force Flight Test Center at Edwards Air Force Base in California. He earned a MS degree in Meteorology from San Jose State University in California and holds BS degrees in Aeronautical Engineering and Physics from the California Polytechnic State University in San Luis Obispo, California and is an Associate Fellow of the AIAA.

Daniel Gilani received an MBA in 2012 from the University of Maryland in College Park, MD. His B.S. in aerospace engineering was received in 2007 from California Polytechnic State University in San Luis Obispo, CA.

He is a general engineer with the FAA's Program Management Organization, leading multiple concept-level engineering projects to define Mid-Term (2017-2020) software enhancements for the FAA's Traffic Flow Management System. Prior to that, Mr. Gilani worked as a systems engineer for TASC Inc. and Northrop Grumman Information Systems. He has 5 years of experience in aviation system development.

Ed Corcoran graduated from San Jose State University with a B.S. in Aeronautical Flight Operations and a minor in meteorology.

He has been employed by the FAA since 1981, starting as an Air Traffic Controller at Oakland Enroute Air Traffic Control Center (ZOA), followed by serving as a specialist in the Traffic Management Unit. Since 1993, he has worked as a Traffic Management Specialist (TMS) at the Air Traffic Control System Command Center (ATCSCC), serving in many capacities to include TMS in both the Severe Weather and Terminal Areas, National Air Traffic Control Association Representative, Collaborative Decision Making participant and group lead, Automation Department, and Training Department. Mr. Corcoran was the GPSM lead at the ATCSCC.

Kenneth C. Venzke received a M.S. in 2000 in meteorology from the Air Force Institute of Technology at Wright-Patterson AFB, OH and a B.S. in atmospheric science in 1993 from the University of Kansas in Lawrence, KS.

Since 2004, he is the Meteorologist In Charge of the Center Weather Service Unit (CWSU) and manages a team of experienced National Weather Service aviation weather forecasters at the FAA ZOA Oakland Air Route Traffic Control Center in Fremont, CA. He has over 22 years' experience in aviation weather and research and has over 1,500 hours combined flight time in USAF EC-135 "Looking Glass" and Navy E6-B TACAMO aircraft while a Captain in the USAF from 1989-2004.

Christine Riley received a B.S. in 2008 in atmospheric science from the University of California Davis in Davis, California.

She is a Meteorologist at the National Weather Service office in Monterey, CA. She developed a training program for staff at the National Weather Service office in Monterey, CA with regards to the MSFS and GPSM. She also worked with the ZOA CWSU in developing new terminology for GPSM. She is the leader of the student mentoring program within the Monterey office. Prior to that, Ms. Riley worked at the National Weather Service office in Norman, Oklahoma where she lead the student volunteer program and studied severe weather.