AIAA Guidance, Navigation and Control Conference and Exhibit
18 - 21 August 2008, Honolulu, Hawaii

AIAA 2008-7147

Future Advanced System Collaborative T&E Requirements, Methods and Practices

Catherine Jaggard[*]*
APPTIS Corp, Atlantic City Airport, NJ 08405, USA

W. Clifton Baldwin[†]
WJHTC FAA, Atlantic City Airport, NJ 08405, USA

Mike Paglione[‡]
WJHTC FAA, Atlantic City Airport, NJ 08405, USA

This white paper explores the development of a new Verification and Validation (V&V) Methodology. The Federal Aviation Administration (FAA) has assigned the William J. Hughes Technical Center (WJHTC) to standardize the agency's testing procedures. This paper will discuss the multiple phases of the development of the V&V methodology: which includes the development and defining the Metrics-Based Approach for evaluating Air Traffic Control automation. The metrics discussed in this paper looks at historical data from legacy system and does a comparison using the data from the ERAM project. Three examples of metric studies are presented that evaluate the performance of the legacy systems versus ERAM. The first study utilizes global positioning satellite and performs a simulation analysis to determine the surveillance tracker accuracy of the legacy versus replacement ERAM system. The next two studies include a comparison of converted route processing and an analysis of the automatic handoff algorithms.

I. Introduction

The Federal Aviation Administration (FAA) has assigned the William J. Hughes Technical Center (WJHTC) to standardize the agency's testing procedures. The Verification and Validation (V&V) Protocol of Operations Project team is responsible for identifying and defining V&V methods that will be standardized at the WJHTC to assure the quality of V&V products and services, including test plans, test conduct, and test reports, as well as enabling risk management activities throughout the V&V process. These V&V methods are being documented in a handbook and provide accountability and consistency in the engineering and test of FAA systems/programs. The handbook will be used in conjunction with other FAA standards and policies to solidify organizational commitment to quality testing methods.

Testing of a complex computer system is an iterative process, which embraces all phases of the software development, maintenance and hardware upgrades. Continuous change to any system is a constant and if test and evaluation methods and measurements are not adequately scoped to meet the changing system then testing can never be complete. As in any system, the test environment and activities should always be in constant flux as the system matures. System life span activities change, but in most instances, how testing is conducted and metrics are collected have not changed, which was the first misconception of what works now will always be acceptable.

This paper explores the evolution of existing methodologies from lessons learned and creating a living test methodology to follow any system from cradle to grave. The methodology can then be reborn and used with the new system from the cradle, through maturity and then to the grave benefiting from the lessons from the previous system.

---

[*] Software Engineer/ Lead Test Engineer, Operational System 2nd Level Engineering, Associate Fellow
[†] System Engineer, Systems Engineering Group
[‡] Project Lead/Systems Engineer, Simulation & Analysis Group, AIAA Member

Copyright © 2008 by the American Institute of Aeronautics and Astronautics, Inc.
The U.S. Government has a royalty-free license to exercise all rights under the copyright claimed herein for Governmental purposes.
All other rights are reserved by the copyright owner.

II. Development of the New Methodology

By definition, a complex system has many interrelated parts. New systems destined for the National Airspace System (NAS) must be autonomous in fulfilling their goal while contributing to the mission of the Next Generation Airspace (NextGen). Therefore any testing strategy of such a complex system needs to consider all the elements including the component parts of the system as well as the system's place as a whole in NextGen. To accomplish this daunting task, a systems thinking approach is needed, which should consider the various perspectives throughout the system's life span.

All new systems start at the concept stage, then to development, testing, deploying and maintenance stage. The test group should be included at the concept stage of any new system where they can consider approaches to test the system alone as well as part of the NAS. The methodology discussed in this paper will start at the concept stage where the requirements begin their life.

Through research, the test team defined several characteristics for every system/software regardless of its form or level of refinement. The areas identified for the test team are as follow:

- Tests development
- Testing Considerations
- Security Considerations
- Operational Considerations
- Related System/Software Patterns
- Integration into legacy system

The test team must validate the test results and identify test metrics that will take place in the beginning of project. Historically the test team would identify metrics based on the system's objectives, but the new approach needs to include metrics from a NAS perspective as well across the life span. The metric requirements should be quantifiable, non-threatening, meaningful, and easy to collect and always keep it simple. Always remember if the statistics to be gathered become overbearing the metric will no longer serve a purpose of properly reporting the information it was intended to capture. There are five defined phases of software development: Requirements, Design, Coding, Integration and Test and Delivery. These phase identify the development of a system but should not be used to identify when to start or conduct V&V. V&V must be performed continuously throughout development. What happens to the ATC System starting at the inception to delivery of a full operational system? It all starts with the analysis and support of the quantitative management, configuration management/Quality Assurance, defect tracking and analysis and the stage of involvement for On-Site audits. First to understand what metrics that produces from the expected results, four phases of V&V should be discussed.

The V &V starts with conception of the system and development of the functional requirements and their activities. Process improvement starts in this phase by looking at the revised or new tools, revising the analysis, and revised processes. When the team looks at existing tools, analysis and processes they are modifying what has worked in the past and utilizing lessons learned gathered in past projects. The past project history assists in lowering the risk assessment of the project. It is like having your own crystal ball, if used properly the new project can save on leaving the past failures behind and bringing to the table what works. The recipe for a project success is the fine lines between the four basic phases of software development that applies in all methodologies.

Accordingly first phase of V&V is the concept development phase. This is where the requirement analysis begins and tracing of the requirements. Simultaneously the test planning phase should commence. The legacy systems now are an intricate part of the planning. Since the ATC system is an open system and interconnects to several platforms the developer of the code (new or reused) needs to understand the hierarchy of the existing systems (i.e., radar, existing ATC systems). The design analysis will embrace the legacy systems, the different codes (i.e., C++, Java, Oracle, SQL Basic, different flavors of UNIX operating systems, etc).

Second phase of V&V coding begins. On site audit are conducted. Regression test for the legacy systems should be already in place and note as we add the new system new regression test will be developed to test each addition. Any time code is introduced to a system all parts of its operation is tested to ensure that the improvement or enhancement has not change any part of the system intended functionality.

Third phase of V&V, any test metric collected if used properly, can aid in software development process improvement by providing pragmatic, objective evidence of process change initiatives. Metrics are defined as "standards of measurement" and are used as a method of gauging the effectiveness and efficiency of a particular activity within a project. Although test metrics are gathered during the test effort, they can provide measurements of many different activities performed throughout a project. In conjunction with root cause analysis, test metrics can be used to quantitatively track issues from points of occurrence throughout the development process. In addition, when

test metrics information is accumulated, updated and reported on a consistent and regular basis, it ensures that trends can be promptly captured and evaluated.

Fourth phase of V&V is the delivery of the system to the field and then it begins its life in 2nd level support arena. Test metrics exist in a variety of forms, and the question is not whether metrics should be used, but which ones would give us the true picture.

All systems being deployed to the field must have a signed National Change Proposal (NCP). This is the beginning phase where the security issues must be addressed. All systems have their own SCAP prepared to address security issues with the system. Note: Without a SCAP package the NCP will not be approved.

III. Changing the Face of V&V

The main task from V&V is to emphasize the modeling task to allow a more dynamic testing of systems. Testing is moving into a new arena where simulation modeling is being developed as a discrete event simulation model on the ATC Systems. The findings resulted in more meaningful results and it gives the ability to develop algorithm to simulate memory utilization. Reviews of the requirements traceability that helped identified gaps n parent (SSS) and child (SRS) relationships and then could be fixed with ECP. Identifying those test items and then gapping the results in an early stage of development will cut down on cost.

Risk assessment for each software function the V&V introduced the CARA scoring Methodology which rates the criticality and risk and then takes the average of the criticality and risk and then is placed in analysis (IAL) thresholds. Under criticality performance and operations, safety and cost/schedule is rated, and for Risk the complexity, technology maturity, requirements definition and stability, testability and developers experience is rated.

Defining relationships, roles, and responsibilities for the elements (such as subsystems and components) is a key aspect of a testing methodology. Conducting the requirements traceability analysis is to identify critical requirements identified in the NASA's CARA methodology. Plus, identify missing, incorrect, or inconsistent requirements using the CORE repository. In addition is it necessary to review all major CDRLS to validate their current performance requirements with the new ATC system.

V&V testing strategy is risk reduction in testing. Verify that the critical function implementation and the collection of statistics about component and subsystem behavior are analyzed. Conduct orthogonal testing which observes system behavior under realistic but atypical conditions. Another words, break the system to make it better! Tests are run to detect defects when the ATC capabilities are degraded. The identification of code that is inefficient or incorrectly implemented for example: algorithms. Plus the V&V collects date to estimate performance over time, e.g., tracker accuracy, conflict alert, trajectory modeling, conflict probe, and posting.

As noted, the V&V activities start early as possible in the development life-cycle. It establish continual feedback loop among the V&V, sponsor organization and utilizes the V&V products throughout the software development life-cycle. Metrics are used to determine whether the functional aspects of ERAM and how it performs against the legacy systems requirements.

IV. Metrics-Based Approach to Testing

A proposed metrics driven process is partitioned into a two-part process, but each part is independent enough to add value on its own. In the first part, Key Performance Parameters (KPP) is identified using operational use cases. These parameters are systemic and often based on the system's operational concept and functional requirements. The established FAA National Airspace System capabilities and services are correlated to the KPPs to facilitate pinpointing relevant metrics that will measure the system's effectiveness.

Once a list of metrics has been identified, they must be prioritized based on needs, criticality, or risks to the program. While rating the metrics, gaps may be identified and highlighted to management. More value can then be added by completing the second part of the proposed process.

Part two of the process involves actual studies and reports of the system under development against the defined metrics. Multiple metrics can be derived from each KPP. Each KPP should have a corresponding metrics report, which is released in stages. As the first stage defines the metrics for the KPP and documents a plan to proceed, the second stage validates the metrics. During this validation step, a metric may be revised or dropped as it is tested in a realistic scenario. The metric may be implemented in a legacy system, if one exists, or a model and/or simulation may be used. In any case, the results will include a quantified expected value for the new system that measures the degree in which it provides a set of services and capabilities. Once the new system is available for testing, the

metric can be quantified and a comparison made. The following sub-sections present this test metric activity applied to the ERAM Testing Program.

A. Metrics Methodology Implementation in the ERAM Testing Program

In 2004 the Federal Aviation Administration's (FAA's) En Route Automation Modernization (ERAM) Test Group formed the Automation Metrics Test Working Group (AMTWG). The team's charter was to support the developmental and operational testing of ERAM by developing a set of metrics that quantify the effectiveness of key system functions in ERAM. The targeted system functions were Surveillance Data Processing (SDP), Flight Data Processing (FDP), Conflict Probe Tool (CPT), and the Display System (DS) modules. The metrics were designed to measure the performance of ERAM. In addition they were designed to measure the performance of the legacy En Route automation systems in operation.

The project was divided into key phases: first a *metrics identification* process was performed. Lists of approximately one hundred metrics were generated by the AMTWG and mapped to the Air Traffic services and capabilities found in the Blueprint for the National Airspace System Modernization 2002 Update. This took place most of fiscal year 2004 and initial metrics results were published in a June 2004 whitepaper. Next, an *implementation-planning* phase was performed. In this step, the identified metrics were prioritized for more detailed refinement during 2005. Next, an implementation-planning phase was performed. In this step, the identified metrics were prioritized for more detailed refinement during 2005. A plan documented the implementation-planning phase. It listed these metrics, gave the rational for selecting them, and provided a high level description on how the highest priority metrics would be measured. It also provided the metric's traceability to the basic controller decisions, ERAM Critical Operational Issues (COIs), and the development contractor's technical performance measurements (TPMs).

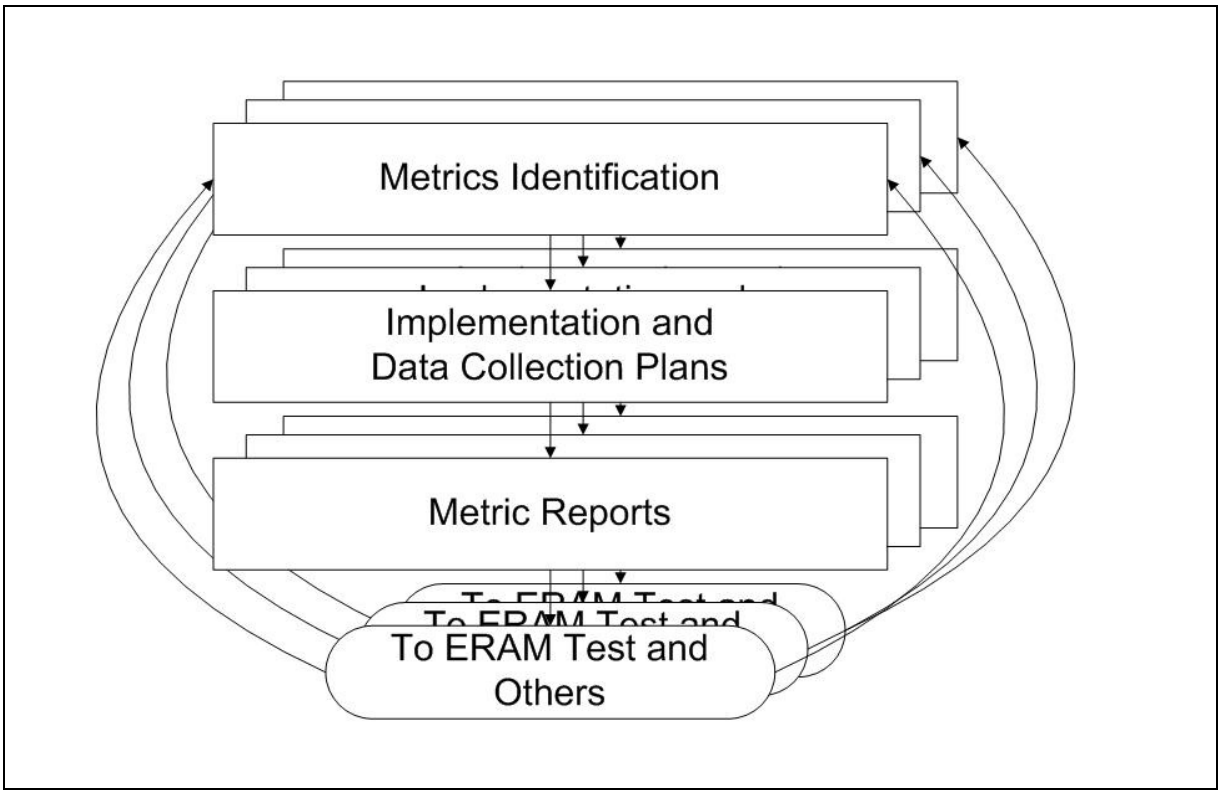

Figure 1: Iterative Process of Data Collection/Analysis Phase (adapted from Ref. 1)

The final project phase was the *data collection and analysis* phase. In this step, AMTWG documented the further refinement and application of these metrics on the legacy systems in a series of Metric Reports. AMTWG delivered four of these Metric Reports covering the different system functions.

As illustrated in the following Fig. 1, the Metric Reports were designed to generate feedback from the ERAM Test Team and other sources (e.g. subject matter experts) to identify more metrics or expand upon existing ones.

The implementation planning would then be revised, and new Metric Reports would be generated. Iterative with frequent output, applied immediately on legacy NAS to flush out and "accredit" metrics focus on validation during DT and OT but also supports verification by providing, valid test cases, expertise and tools when applicable.

Overall, this activity provided several significant benefits to the ERAM program. As support to the ERAM Test Team, it provided additional data points that helped develop test cases and measurements to supplement and enhance the requirements verification. It served as a risk reduction activity for the entire program by evaluating key subsystems and functions for effectiveness. The metrics provided the ERAM Program Office with supporting data on ERAM benefits. The activity allowed key ERAM Test Team personnel to receive in-depth experience on ERAM subsystems and similar functions in the legacy automation. This experience appeared to increase their effectiveness in reviewing contractor test plans and procedures. Finally, the tools, metrics, and traffic scenarios were incorporated, where appropriate, in the formal ERAM Test Program. The following sub-sections provide concrete examples of this process.

B. Example of Metrics Study for ERAM Surveillance Data Processing

One sample application of the test metrics approach was the implementation of metrics for evaluation of the ERAM surveillance tracking algorithm. At the highest level, the FAA's air traffic control system relies directly on aircraft locations provided by the long range en route surveillance radars. The accuracy of the radars is an important factor in determining the overall performance of the system. To support the planned modernization of the air traffic control system a metrics study was conducted to measure the accuracy of the radar tracking function of the legacy HCS. This was first done by comparing aircraft radar tracks produced by the existing system with the tracks for the same aircraft produced by the Global Positioning Satellite System (GPS) position reports. It was assumed that the GPS data was the ground truth. The GPS data was available from the FAA's Reduced Vertical Separation Minimum (RVSM) Certification Program. The Host Air Traffic Management Data Distribution System (HADDS) at each Air Route Traffic Control Center (ARTCC) captures the radar track data. This data is then archived at the William J. Hughes Technical Center (WJHTC) for a period of time. The radar tracks for 265 flights were compared to their GPS "tracks". Three distance metrics were used: horizontal track error and its two components cross track error and along track error. A total of 54170 pairs of position reports were compared. The distributions of the errors were plotted and basic descriptive statistics were determined. It was found that the average horizontal error was 0.69 nautical miles (nm), the root mean square value of the horizontal error was 0.78 nm, the average cross track error was 0.12, and the average along track error was 0.67 nm.[2] The complete results are summarized in Table 1.

Table 1: HCS Tracker Results from Ref. 2

		Horizontal Error		Cross Track Error		Along Track Error	
Type	Sample Size	Mean	RMS	Mean	RMS	Mean	RMS
Signed	54170	0.69	0.78	0.00	0.16	-0.67	0.77
Unsigned				0.12		0.67	

Once the GPS analysis was implemented, a second study repeated this approach but using a simulation environment that could be repeated later on the analogous ERAM tracking functions. The metrics study began with a recording of approximately four hours of air traffic data from Washington D. C. ARTCC (ZDC) on March 17, 2005. The Automation Metrics Test Working Group supplied this data to the Integration & Interoperability Facility that ran a simulation using the FAA's Graphical Simulation Generation Tool (GSGT) simulator. The simulation produced the HCS tracks reports and GSGT positions. The GSGT positions were considered the actual path the aircraft flew in which the HCS track positions were measured against. A series of data processing steps were run comparing the GSGT positions to the HCS track reports. Four error metrics were applied including horizontal error, its orthogonal components cross and along track error, and altitude error. For this study, the mean horizontal error was 0.85 nautical miles. The cross track error distribution is symmetrical about zero nautical miles and a root mean square value of 0.14 nautical miles. However, the along track error distribution is strongly skewed in the negative direction with an average error of -0.83 nautical miles. This represents an uncompensated time error, confirming the previous result using GPS aircraft positions.[3]

As described in the overview of metrics methodology is to repeat the experiment on the ERAM functions and compare the legacy results. A repeat run of the same scenario as described above was formatted for running into

ERAM. The cross track error is reasonably comparable to the HCS.[§] The experiment of 1100 flights from the same ZDC data produced an average unsigned cross track error of 0.04 nautical miles and a root mean square error of 0.0025 nautical miles. The cross track error in ERAM is a third the size of the HCS results. As illustrated in the histograms in Fig. 2-3, it is clear that the ERAM tracker has improved upon the tracking problem in the lateral or cross track dimension.[2,4]

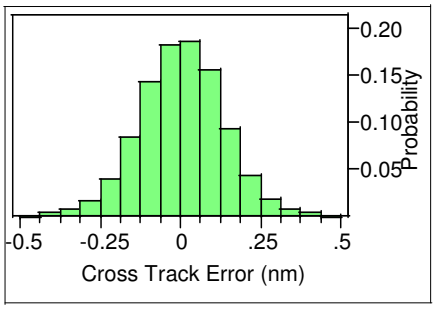

Figure 2: HCS Cross Track Error[1]

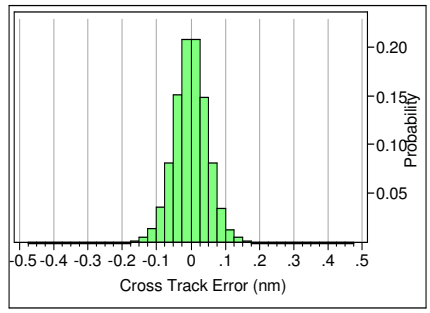

Figure 3: ERAM Cross Track Error[4]

C. Example of Metrics Study for ERAM Flight Data Processing – Converted Route Logic

A key function of most flight data processing systems, and certainly for the ERAM system, is the processing of the flight plan. The flight plan represents the current air traffic control's cleared 2-dimensional path of the aircraft's flight path. It represents the best intent information available to the automation and is used for many critical air traffic functions from hand-off coordination and flight strip printing to aircraft trajectory prediction. The flight plan is converted to a 2-dimensional set of positions, typically referred to as the converted route. Thus, a metric study was performed first on the legacy automation and then repeated on the ERAM replacement to measure this critical function. The primary metric is defined as the lateral deviation or distance from the current aircraft position to the converted route. This is illustrated in Fig. 4 below.

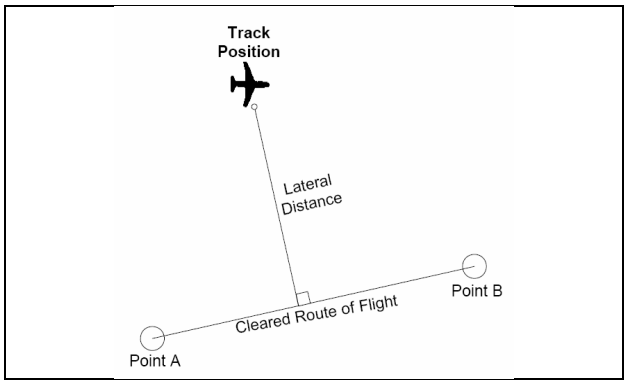

Figure 4: Lateral Distance (adapted from Ref. 5)

In the initial study performed in Ref. 5, data from ZDC was collected and applied to the legacy systems of URET and the HCS, the strategy was applied to determine which system had superior performance. The results of the tests indicate that URET performance exceeds the HCS performance with better than a 99.9% probability. The study was later repeated and presented in Ref. 4 using ERAM results. The average lateral deviations for the HCS, URET, and ERAM systems were 1.7, 0.9, and 0.9 nautical miles, respectively. The distributions of these deviations are illustrated in the following Figures 5, 6, and 7 for each of the systems HCS, URET, and ERAM, respectively. The HCS is significantly skewed to the right consistent with its mean value, while the URET and ERAM distributions are practically identical. This is as expected since the logic for both URET and ERAM flight plan route conversion

[§] Due to a problem with one of the interfaces still being repaired at the time of this publication, only the cross track error can be compared in this paper as presented in Ref. 4.

was the same. Thus, the test metrics analysis confirmed that the ERAM system would indeed provide enhanced performance over the legacy HCS system and matched the URET system's accuracy.

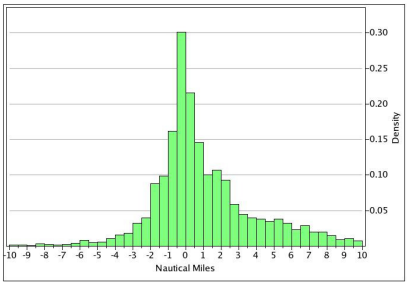

Figure 5: HCS Lateral Distances[4]

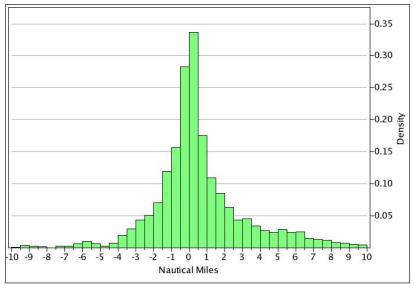

Figure 6: URET Lateral Distances[4]

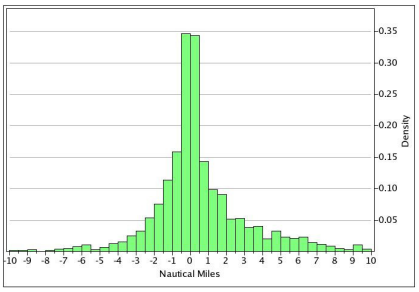

Figure 7: ERAM Lateral Distances[4]

D. Example of Metrics Study for ERAM Flight Data Processing – Automatic Handoff Function

The metrics overview discussed the process of identifying the test metrics by first evaluating their context to Air Traffic services and capabilities. It also indicated that some metrics were identified by subject matter experts. Unlike the previous two examples that are linked to critical air traffic controller decisions, the Automatic Handoff Test Metrics presented in this example is the later. The metric was generated by direct request by testing and air traffic specialists within the ERAM Test Group. In response, a study was employed to measure the performance of an automated process of initial handoff of flights between sectors from an objective viewpoint. These statistics are based on an unbiased sample of flights from ZDC. The study developed two methods of predicting the handoff sector: (1) based on the flight plan converted route, referred to as the Flight Plan Trajectory and (2) based on a full 4-D trajectory prediction, referred to as the Aircraft Trajectory in ERAM documentation. As one might expect intuitively, the aircraft trajectory-based predictions performed at a higher accuracy rate than the flight path predictions. Since a flight path prediction assumes the aircraft is on its flight path, there is an additional assumption not present with trajectory-based predictions. In particular, the automation checks the conformance of the prediction against the current aircraft prediction and builds a new trajectory. When the assumption is incorrect (aircraft are deviating from the known flight plan), the prediction should have a lower probability of success.

Besides the type of prediction, the performance of the predictions of the next controlling sector is calculated based on two reference positions: at the actual recorded operational initialization of handoff to the next sector and at a hypothetical predicted distance to the next sector. The predicted distance initializations had a higher accuracy than the actual initializations. This result is attributed to the large variation in the actual distance at which handoffs were initialized to the next sector. Hence, the trajectory-based predicted-distance initializations of handoff had the highest probability of success in this study and were usually physically closer to the next sector than the recorded operational distance.

To approximate the overall success rate of an auto-init handoff function, results of this study can be applied to a basic calculation. Of all initialization-handoff events, 93% could be identified by current automation. Of the resulting sample space, a 78.5% success rate was determined for trajectory-based predicted-distance initializations. Therefore an auto-init handoff function should be at least 73% accurate for all events. Alterations to the algorithms used and choosing alternate distances for the initializations could impact the accuracy of an auto-init function. To compensate, the software tools developed for this study are flexible and can be easily modified for future studies.[6]

Besides the detailed flight examples that did provide insights into the sources of error, statistical methods were applied as well. Correlations were calculated between successfully predicted sectors and the prediction error at the prediction location (e.g. sector handoff initialization time) in the horizontal and vertical dimensions. Although there was a significant correlation between the prediction accuracy and the vertical error, more analysis is required to establish cause and effect. It is possible that both effects are the result of an unknown factor. Additional metrics were calculated at the location of the predicted sector that may further explain the error sources of the prediction.

As planned the metrics were repeated on ERAM flight plan trajectories and presented in July 2008.[4] The Ref. 6 results for the HCS flight path predictions reported that 65% of the hand-off predictions were correct (contrasted to the higher rate accuracy for trajectory-based predictions). The same predictions using ERAM's flight plan trajectory produced almost identical results. This is not surprising. An analysis of the longitudinal error differences between the HCS flight path predictions against ERAM was performed. The results indicated that on average ERAM had 4.8 nautical miles less along track error or 42 seconds more timely than HCS predictions. The two resulting

distributions based of differences are illustrated in Fig. 8 and 9, respectively, for several thousand flights. Since the difference is calculated by subtracting the HCS along track or time error per flight by the same aircraft prediction in ERAM, a positively skewed distribution of the differences indicates significantly larger errors in the HCS. Thus, the test metrics and study provided confidence that the ERAM replacement of the HCS would indeed perform at least as good at the legacy system. Furthermore, for the longitudinal dimension of flight plan predictions, ERAM should perform significantly better.

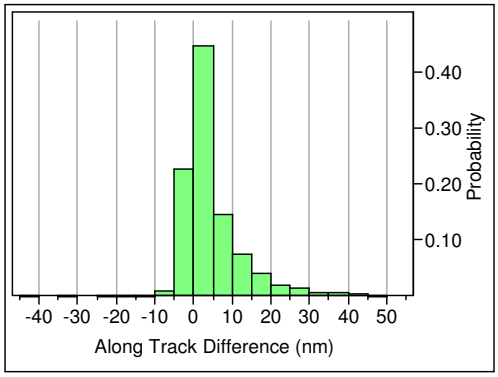

Figure 8: HCS-ERAM Along Track Error[4]

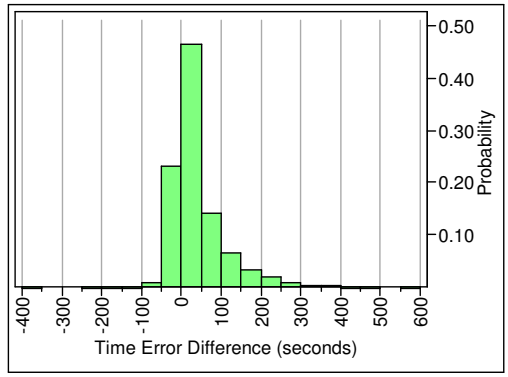

Figure 9: HCS-ERAM Time Error[4]

V. Conclusion

The use of tools, and metrics which included traffic scenarios were evaluated and compared against the existing legacy system and new ERAM system. The metrics testing proved to be helpful revealing that the ERAM not only performed to the standards of the legacy system but show enhanced performance in some areas. The samples used in this study were the surveillance data processing which proved that ERAM tracker improved upon the tracking problem in the lateral and cross track dimension. The ERAM Flight data processing analysis revealed enhanced performance over the legacy systems, plus ERAM would perform significantly better in the longitudinal dimension of flight plan predictions. Overall, development of the V&V Methodology using the Metrics-Based Approach proved to be a valuable access to the program in determining that ERAM can perform as well or even significantly better then its predecessor.

The overall objective is to apply the Metrics-Based Approach to future FAA projects. Through the example implementation on ERAM, the approach has been proven useful in providing measures of the performance on key functions needed at critical decision points. It can buy down FAA risk if integrated into the program's vendor contracts. The use of a cross organizational metrics team with varied skilled sets and expertise had proven to be very constructive as well. It allowed the testing personnel to bring larger sets of FAA resources onto complex problems within their program. With the lessons learned on ERAM, the improved V&V techniques of the Metrics-Based Approach can certainly be applied with even greater success to future FAA programs.

References

[1]Paglione, M., R. Oaks, H. Ryan, "Metrics-based Approach for Evaluating Air Traffic Control Automation of the Future," *51st Air Traffic Control Association Annual Conference Proceedings*, October 29, 2006.

[2]Ryan, H., M. Paglione, "Comparison of Host Radar Tracks to Aircraft Positions from the Global Positioning Satellite System" DOT/FAA/TC-TN05/30, Federal Aviation Administration, William J. Hughes Technical Center, August 2005.

[3]Paglione, M., W. C. Baldwin, S. Putney, "Host Radar Tracking Simulation and Performance Analysis" DOT/FAA/CT-TN05/31, Federal Aviation Administration, William J. Hughes Technical Center, August 2005.

[4]Paglione, M., "En Route Automation Modernization Testing Program: Minimum Operational Performance Standards from Test Metrics Activity," Internal Presentation, Federal Aviation Administration, William J. Hughes Technical Center, July 11, 2008 (unpublished).

[5]Baldwin, W. C., "Comparison of Converted Route Processing by Existing Versus Future En Route Automation," DOT/FAA/CT-TN05/29, Federal Aviation Administration, William J. Hughes Technical Center, August 2005.

[6]Baldwin, W. C., M. M. Paglione, J. E. Rhodes, S. R. Larimer, "Analysis of Handoffs for Future En Route Automation," DOT/FAA/TC-TN07/15, Federal Aviation Administration, William J. Hughes Technical Center, February 28, 2007.