technical note technical note

Evaluation of Parameter Adjustments to the En Route Automation Modernization's Conflict Probe

Andrew Crowell, FAA ANG-C41

Andrew Fabian, FAA ANG-C41

Christina Young, Ph.D., CSSI Inc.

Ben Musialek, General Dynamics Information Technology

Mike Paglione, FAA ANG-C41

December 2011

DOT.FAA/TC-TN12/2

Document is available to the public
through the National Technical Information
Service, Springfield, Virginia 22161

U.S. Department of Transportation

Federal Aviation Administration

William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

[THIS PAGE IS INTENTIONALLY LEFT BLANK]

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the objective of this report. This document does not constitute FAA certification policy. Consult your local FAA aircraft certification office as to its use.

This report is available at the Federal Aviation Administration William J. Hughes Technical Center's Full-Text Technical Reports page: actlibrary.tc.faa.gov in Adobe Acrobat portable document format (PDF).

[THIS PAGE IS INTENTIONALLY LEFT BLANK]

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT.FAA/TC-TN12/2		

4. Title and Subtitle	5. Report Date
Evaluation of Parameter Adjustments to the En Route Automation Modernization's Conflict Probe	September 2011
	6. Performing Organization Code ANG-C41
7. Author(s) Andrew Crowell, Andrew Fabian, Christina Young, Ph.D., Ben Musialek, and Mike Paglione	8. Performing Organization Report No. DOT.FAA/TC-TN12/2
9. Performing Organization Name and Address U. S. Department of Transportation Federal Aviation Administration, William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address U. S. Department of Transportation Separation Management Modern Procedures Program Office, AJE Conflict Probe/URET Product Team Washington, D. C. 20590	13. Type of Report and Period Covered Technical Note 14. Sponsoring Agency Code

15. Supplementary Notes

The authors identified above represent the following organizations: Andrew Crowell, Andrew Fabian, and Mike Paglione with FAA ANG-C41, Christina Young, Ph.D. with CSSI Inc, and Ben Musialek with General Dynamics Information Technology.

16. Abstract

The Federal Aviation Administration (FAA) is currently implementing a number of improvements to the National Airspace System (NAS) in the United States under a multi-agency initiative called the Next Generation Air Transportation System (NextGen) Program. The Separation Management and Modern Procedures Project is one of these NextGen initiatives. The FAA's Air Traffic Organization's En Route Program Office (ATO-E) has employed the FAA's Concept Analysis Branch (ANG-C41) to conduct a series of independent evaluations on the performance enhancements to the En Route Automation Modernization (ERAM) Trajectory Modeling (TM) and Conflict Probe (CP) sub-systems. ATO-E contracted the prime contractor of ERAM, Lockheed Martin under FAA Task Orders 45 and 51, to develop these prototypes within the actual ERAM architecture, so the FAA could evaluate their efficacy. Before analyzing the initial prototype enhancements being developed, an initial experiment was performed that altered parameters of ERAM to determine if improvement can be achieved from the current version of ERAM without additional upgrades. This report describes the metrics and methods developed and results of this initial experiment.

17. Key Words	18. Distribution Statement
Separation Management, Conflict Probe, ERAM, En Route, Automation	This report is approved for public release and is on file at the William J. Hughes Technical Center, Aviation Security Research and Development Library, Atlantic City International Airport, New Jersey 08405. This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	125	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

[THIS PAGE IS INTENTIONALLY LEFT BLANK]

Executive Summary

The Federal Aviation Administration (FAA) is currently implementing a number of improvements to the National Airspace System (NAS) in the United States under a multi-agency initiative called the Next Generation Air Transportation System (NextGen) Program. The Separation Management and Modern Procedures Project is one of these NextGen initiatives. The FAA's Air Traffic Organization's En Route Program Office (ATO-E) has employed the FAA's Concept Analysis Branch (ANG-C41) to conduct a series of independent evaluations on performance enhancements to the En Route Automation Modernization (ERAM) Trajectory Modeling (TM) and Conflict Probe (CP) sub-systems. This work is motivated by a Separation Management and Modern Procedures Project's objective of implementing the ERAM strategic conflict probe on the radar controller's display. The strategic conflict probe utilizes the TM and CP sub-systems to notify air traffic controllers when aircraft will violate separation standards as much as 20 minutes in the future. Furthermore, NextGen operational concept envisions a future air traffic environment managed by aircraft trajectory with advances in ground automation like the conflict probe. Thus, ATO-E contracted the ERAM prime contractor under FAA Task Orders 45 and 51 to develop these prototypes within the actual ERAM architecture, so the FAA could evaluate their efficacy.

This report describes two key evaluations. First, it describes an initial improvement to the ERAM TM function. This evaluation found that the initial trajectory prototype improved both the trajectory position accuracy and conflict probe performance. Second, it presents an initial experiment that altered configurable parameters in ERAM. This determines if improvement can be achieved from the current version of ERAM without additional upgrades. In addition, automation experts expect the prototype enhancements will require these same altered parameters to realize their improvements. This report describes the metrics and methods developed and results of this initial experiment.

The three parameters changed in this experiment are the lateral and longitudinal adherence bounds and the likelihood function. The adherence bounds are a buffer zone surrounding the aircraft and the aircraft prediction. The lateral bound is the distance of this buffer to the left and right of the aircraft, whereas the longitudinal adherence bound is the buffer to the front and back. This buffer is used for two purposes. If the reported position of the aircraft is outside of the buffer zone surrounding the predicted position of the aircraft, it triggers a new trajectory prediction. Also, if the buffer zones surrounding the predicted positions of two aircraft come within the minimum separation requirement (traditionally 5 nautical miles in En Route airspace), then this triggers the generation of an alert by the Conflict Probe. Lastly, the likelihood function assesses each conflict prediction and depending on its confidence estimate, may delay the alert notification until the estimate is improved or prediction is more imminent.

Longitudinal and lateral adherence bound thresholds were altered from the current operational levels of 1.5 nautical miles and 2.5 nautical miles, respectively, to 0.5 nautical miles. The current ERAM operational system bypasses the likelihood calculation for all alerts predicted to begin within the next 10 minutes and applies the likelihood function for conflict detections beyond 10 minutes only. The two other levels examined in this experiment reduced this minimum threshold from 10 minutes to 3 minutes and changed other parameters associated with the function. For one level, it reduced the upper bound to 10 minutes and for the other left this setting at 20 minutes. As a result, these three factors, including longitudinal adherence bounds, lateral adherence bounds, and likelihood function, were simultaneously altered using experimental design techniques and ERAM was run at twelve different treatment and two baseline runs.

Even though this initial study was limited to one air traffic scenario, the experiment produced a number of interesting results and motivated ANG-C41 to develop a number of new metrics and evaluation techniques. The longitudinal adherence bound was found to significantly increase the Late Alerts (i.e. ERAM conflict notifications presented with less than the required strategic minimum warning time, 3 minutes) when reduced too far from its original setting. The lateral adherence bound was determined to cause a significant decrease in False Alerts (i.e. ERAM conflict notifications that do not manifest into actual violation of separations, a.k.a. nuisance alerts) without causing a major negative impact on Late Alerts. This occurred when the lateral bound was reduced to the minimum parameter setting of 0.5 nm. Though there is some interaction of these two settings with the likelihood, the interactions seen were minimal, so recommendations can be made independent of the likelihood function.

Likelihood was determined to have a major effect on False Alerts, Late Alerts, and warning time. Unfortunately, it is not as easy to make a recommendation for a setting of likelihood. The three different settings of likelihood used in this study provided some insight to how the function affects the performance of the probe, but it is still not completely understood. There were two main findings in this study. The first is that likelihood alone can have a major effect on the performance of the probe. The second is that, because of the longitudinal inaccuracy of the probe, setting the minimum time of the likelihood function to the same as the minimum warning time requirement causes extra Late Alerts that should not have been. In order to further understand the likelihood and additional experiments are recommended.

Overall, the study provides important insights into the three factors examined. It developed a number of metrics, including traditional counts and rates of false and late detections as well as implementation of fuzzy detection theory that extracted even more information out of the data. The study also implemented advanced experimental design techniques that both maximized use of the ERAM runs and examined the potential of interactions between factors. The study provides a strong foundation from which a number of additional experiments are being planned and underway.

Table of Contents

EXECUTIVE SUMMARY ..... **VII**

TABLE OF CONTENTS ..... **IX**

LIST OF FIGURES ..... **XI**

LIST OF TABLES ..... **XIII**

1 INTRODUCTION **1**

1.1 BACKGROUND TO STUDY 1
1.2 SCOPE 4
1.3 DOCUMENT ORGANIZATION 4

2 ANALYSIS METRICS AND METHODOLOGY **5**

2.1 EVALUATION METRICS 8
2.1.1 Trajectory Modeling Metrics *8*
2.1.1.1 Trajectory Sampling 8
2.1.1.2 Definition of Metrics 9
2.1.1.3 Statistical Techniques 9
2.1.2 Conflict Probe Metrics *10*
2.1.2.1 Conflict Prediction "Crisp" Metrics 10
2.1.2.1.1 Definition of Basic Conflict Prediction Metrics 10
2.1.2.1.2 Comparison of Conflict Prediction Results 12
2.1.2.1.3 Statistical Approach of Comparing Conflict Predictions 16
2.1.2.1.4 Additional Crisp Metrics 19
2.1.2.2 Conflict Prediction "Fuzzy" Metrics 22
2.2 DESCRIPTION OF THE EXPERIMENTAL PLAN 29
2.2.1 Definition of the Problem Statement *31*
2.2.2 Design of the Experiment *31*
2.2.2.1 Factors 31
2.2.2.2 Model 33

3 PERFORMANCE EVALUATION **35**

3.1 TRAJECTORY MODELING ENHANCEMENTS 35
3.1.1 Analysis Description *36*
3.1.2 Results *36*
3.1.2.1 Trajectory Accuracy 36
3.1.2.2 Conflict Probe Accuracy 39
3.2 INTEGRATED EXPERIMENT ANALYSIS 41
3.2.1 Trajectory Modeling Analysis *42*
3.2.2 Conflict Probe Analysis *45*
3.2.2.1 Fuzzy Analysis Results 53
3.2.2.2 Model Analysis 56
3.2.2.2.1 Implementation of Statistical Model 56
3.2.2.2.2 Examine Model Results 64
3.2.2.2.3 Factor Effects 76
3.3 EXAMPLE FLIGHTS 77
3.3.1 Flight Example 1 ..... *77*
3.3.2 Flight Example 2 ..... *82*

4 RECOMMENDATIONS AND FUTURE WORK 86
LIST OF ACRONYMS AND ABBREVIATIONS 88
REFERENCES 90
APPENDIX A 93
A.1 FUZZY ALERT LOGIC 93
A.2 ADDITIONAL COMPUTATIONS FOR FUZZY ALERT LOGIC 100
APPENDIX B 102
APPENDIX C 107
APPENDIX D 109

List of Figures

Figure 1. Illustrative Example of a Conflict 2
Figure 2. Black Box – Crisp Metrics 7
Figure 3. White Box – Fuzzy Metrics 7
Figure 4. Diagram of Trajectory Errors 9
Figure 5. Example Alert Probability Illustrating Sharpness and Sharpness Bias 21
Figure 6. Function of vertical inverse closure time. 25
Figure 7. Function of horizontal separation re-centered so that zero represents the minimum horizontal separation standard. 26
Figure 8. Function of horizontal inverse closure time. 26
Figure 9. Top: Gaussian distribution altered by *Head-on* horz. geometry weighting. Bot: Same altered by *Left-To-Right Perp. Crossing* horz. geometry weighting instead. 27
Figure 10. Function of predicted horizontal separation. 29
Figure 11. Linear likelihood function for 10/20 setting 32
Figure 12. Piecewise linear likelihood function for 3/8/20 setting 33
Figure 13. Matched Pair Analyses for Trajectory Error Metrics. 37
Figure 14. Matched Pair Analyses for Trajectory Error Metrics for In-Adherence Data. 39
Figure 15. Prediction Profiler of TM Model with Baseline Parameter Settings 44
Figure 16. Hit Rate vs. False Alert Rate colored by Warning Time. The top-left corner of the graph is the most desirable. 48
Figure 17. Graph of % FA Improvement vs. % LA Improvement compared to the FA32 Baseline. Top right corner is the most improved. 51
Figure 18. Graph of % FA Improvement vs. % LA Improvement compared to the Initial Baseline. Top right corner is the most improved. 53
Figure 19. Fuzzy Hit Rate vs. Fuzzy False Alert Rate. 55
Figure 20. Traditional FA %, LA %, and WT % metrics leverage plots 56
Figure 21. Traditional FA %, LA %, and WT % metrics in the refined model 58
Figure 22. Residual Error Distributions for the Response Variables from the Refined Model 59
Figure 23. Fuzzy FA %, LA %, and WT % metrics leverage plots 61
Figure 24. Fuzzy FA %, LA %, and WT % metrics in the refined model 62
Figure 25. Residual Error Distributions for the Fuzzy Response Variables from the Refined Model 63
Figure 26. Prediction profiler of the model with baseline parameter settings. 64
Figure 27. Optimal settings of the conflict probe to minimize the LA count ignoring all other response variables. 65
Figure 28. Optimal settings of the conflict probe to minimize the FA count ignoring all other response variables. 66
Figure 29. Optimal settings of the conflict probe to maximize the warning time. 67
Figure 30. Interaction profile of the factors effects on the FA % response variable. 68
Figure 31. Interaction profile of the factors' effects on the LA % response variable. 69
Figure 32. Interaction profile of the factors' effects on the WT % response variable 70
Figure 33. Prediction profiler with settings that are a combination of the optimal values for each response. 71
Figure 34. Prediction profiler with a combination of optimal settings of the responses. 72
Figure 35. Prediction profiler of the model with baseline parameter settings. 73
Figure 36. Optimal settings of the conflict probe to minimize the LA count ignoring all other response variables. 73

Figure 37. Optimal settings of the conflict probe to minimize the FA count ignoring all other response variables. 74
Figure 38. Optimal settings of the conflict probe to maximize the warning time. 75
Figure 39. Prediction profiler with settings that are a combination of the optimal values for each response. 76
Figure 40. Baseline conflict detection (2.5 by 1.5 conformance bounds). 78
Figure 41. Reducing only longitudinal bounds to 0.5. 79
Figure 42. Reducing only lateral bounds to 0.5. 80
Figure 43. Mid-level conformance bounds (1.5 by 1.0). 81
Figure 44. Low-level conformance bounds (0.5 by 0.5). 82
Figure 45. Baseline settings showing a valid conflict detection. 83
Figure 46. Lateral conformance bounds only are reduced to minimum. 84
Figure 47. Longitudinal conformance bounds only are reduced to minimum. 85
Figure 48. High-level FARE Process Flow 94
Figure 49. Fuzzy Response Computation Process Flow 95
Figure 50. High-level Fuzzy Signal Computation Process Flow 98
Figure 51. Process Flow for Computation of Fuzzy Signal at a Point in Time 99
Figure 52. Determining Fuzzy Alert Values 100
Figure 53. Examples 1 and 2 of Matched Pairs Analysis from JMP 107
Figure 54. Examples 3 and 4 of Matched Pairs Analysis from JMP 108
Figure 55. StrategicAlertEvaluator Process A – Establishing Notification Sets from Alerts 110
Figure 56. StrategicAlertEvaluator Process B – Evaluating Actual Conflict Notification Set 111
Figure 57. StrategicAlertEvaluator Process C – Evaluating False Notification Set 112

List of Tables

Table 1. Summary of Prototyping Activities 5
Table 2. CP Alert and Conflict Event Combinations (adapted from [Paglione, 1999]) 11
Table 3. Conflict Prediction Result's Main Reason Codes 13
Table 4. Comparison of Two Run's Resulting Alert and Conflict Event Combinations 14
Table 5. Conflict Prediction Comparison Program Evaluation Codes 15
Table 6. Example Application of the McNemar's test 18
Table 7. Example of Sharpness Metric Summary 21
Table 8. Processing Steps for the Experimental Analysis 30
Table 9. Continuous factors of the Integrated Experiment 31
Table 10. Nominal factors of the Integrated Experiment 32
Table 11. Twelve runs for the integrated experiment and the baseline run 33
Table 12. Statistical Results for Trajectory Error. 36
Table 13. Statistical Results for Trajectory Error of In-Adherence Data. 38
Table 14. Conflict Probe Comparison Results. 40
Table 15. Adherence Conflict Probe Comparison Results. 41
Table 16. Statistical Results for Trajectory Cross Track Error 42
Table 17. Statistical Results for Trajectory Along Track Error 43
Table 18. Statistical Results for Trajectory Vertical Error 43
Table 19. Alert type counts for each of the treatment runs and the two baselines. 46
Table 20. Late Alert plus Missed Alert count and Adjusted Late Alert value for each of the treatment runs and the two baselines. 47
Table 21. Overall Alert Rates and False Alert Rates by separation category. 47
Table 22. Median, inter-quartile range, and 25th percentile of conflict warning time. 49
Table 23. The runs that can be eliminated based on Hit Rate, FA Rate and 25th percentile Warning Time 49
Table 24. Results of comparisons of each of the treatment runs to the FA32 Baseline run. 50
Table 25. Results of comparisons of each of the treatment runs to the Initial Baseline run. 52
Table 26. Fuzzy Alert values for each of the treatment runs and the baselines. 54
Table 27. Overall Fuzzy Alert Rates and Fuzzy False Alert Rates by separation category. 54
Table 28. Initial Model Effect Test for the FA % Response Variable 57
Table 29. Initial Model Effect Test for the LA % Response Variable 57
Table 30. Initial Model Effect Test for the WT % Response Variable 58
Table 31. Summary of Refined Model Effect Tests for Response Variables. 60
Table 32. Initial Model Effect Test for the Fuzzy FA % Response Variable. 61
Table 33. Initial Model Effect Test for the Fuzzy LA % Response Variable. 62
Table 34. Summary of Refined Fuzzy Model Effect Tests for Fuzzy Response Variables. 63
Table 35. Alert type counts for each of the twelve runs and the two baselines. 102
Table 36. Late Alert plus Missed Alert count and Adjusted Late Alert value for each of the twelve runs and the two baselines. 103
Table 37. Overall Alert Rates and False Alert Rates by separation category. 103
Table 38. Median, inter-quartile range, and 25th percentile of conflict warning time. 104
Table 39. Results of comparisons of each of the twelve runs to the FA32 Baseline run. 105
Table 40. Results of comparisons of each of the twelve runs to the Initial Baseline run. 105
Table 41. Fuzzy Alert values for each of the 12 runs and the baseline. 106
Table 42. Overall Fuzzy Alert Rates and Fuzzy False Alert Rates by separation category. 106

1 Introduction

The Federal Aviation Administration (FAA) is currently implementing a number of improvements to the National Airspace System (NAS) in the United States under a multi-agency initiative called the Next Generation Air Transportation System (NextGen) Program. The Separation Management and Modern Procedures Project is one of these NextGen initiatives. The FAA's Air Traffic Organization's En Route Program Office (ATO-E) has employed the FAA's Concept Analysis Branch (ANG-C41) to conduct a series of independent evaluations on the performance enhancements to the En Route Automation Modernization (ERAM) Trajectory Modeling (TM) and Conflict Probe (CP) sub-systems. These ERAM enhancements are required by the *Separation Management and Modern Procedures Project*. ATO-E contracted the prime contractor of ERAM, Lockheed Martin under FAA Task Orders 45 and 51, to develop these prototypes within the actual ERAM architecture so the FAA could evaluate their efficacy.

Before analyzing the initial prototype enhancements being developed, an initial experiment was performed that altered parameters of ERAM to determine if improvement can be achieved from the current version of ERAM without additional upgrades. This report describes this experiment and results.

1.1 Background to Study

The FAA created the NAS to provide a safe and efficient airspace environment for the air transportation system in the United States. This includes all commercial, general civilian and military aviation. The NAS is composed of a network of air navigation facilities, air traffic control facilities, and airports, along with the technologies and the rules and regulations to operate the system. As the air transportation system in the United States has grown, the NAS has evolved by incorporating new procedures and new technologies. The projected increases in demand could lead to a greater stress and perhaps to decreased quality of service for NAS users. In response to this the United States Congress created the multi-agency Joint Planning and Development Office (JPDO) in 2003 as a part of the "Vision-100" legislation (Public Law 108-176). The mission of the JPDO is to design and deploy an air transportation system meeting the nation's air transportation system's anticipated needs in 2025. Since its creation the JPDO has published an integrated plan [Federal Aviation Administration, 2004] and documented a concept of operations [Federal Aviation Administration, 2007] that establish a vision for the Next Generation Air Transportation System (NextGen). An integral part of this vision is Trajectory Based Operations (TBO), which represents a paradigm shift from clearance-based air traffic control to trajectory-based air traffic control. With TBO it is envisioned aircraft will fly negotiated trajectories and the air traffic control functions will move to trajectory management.

In the United States, the FAA is the single Air Navigation Service Provider (ANSP) that operates the key components of the NAS, such as Air Route Traffic Control Centers (ARTCCs) and Terminal Radar Approach Control (TRACON) centers. A key function provided by an ANSP is to ensure the safe separation of aircraft within the air transportation system. This separation management function remains much as it was when radar was first introduced to civil aviation in the late 1950s. Aircraft-to-aircraft separation is managed by human air traffic controllers who make strategic and tactical decisions using radar displays to visualize aircraft positions and flight paths. These decisions are then provided to the pilots through voice communications via very high frequency (VHF) radio. Although more recent automated decision support tools (DSTs) have been made available to controllers that predict potential loss of separation (i.e., conflicts) and aid in their evaluation, the effectiveness of the DSTs is limited by airspace complexity,

controller workload, and the use of voice communications.

NextGen envisions trajectory-based separation management that will provide precise management of the current and future positions of all controlled aircraft in the air transportation system. This will require enhanced DSTs that not only predict future conflicts, but also provide conflict resolution that is communicated directly to the aircraft from the ANSP through digital Data Comm*. This planned separation management capability will be able to handle the anticipated increase in traffic demand and aircraft diversity with minimal impact to user-desired performance profiles and to the environment, while retaining the existing strict safety standards.

A specific separation management problem is an aircraft-to-aircraft conflict, which can be defined as a situation where two or more aircraft violate minimum separation criteria. The separation criteria depends on the aircraft's navigational equipment and the airspace (e.g., En Route, Terminal, or Oceanic) in which they are flying. For example, in En Route airspace the separation criteria may be five nautical miles horizontally and 1,000 feet vertically. A Conflict Probe (CP) is a DST that predicts when conflict situations may occur by continuously comparing projected aircraft positions over a user-specified look-ahead time.

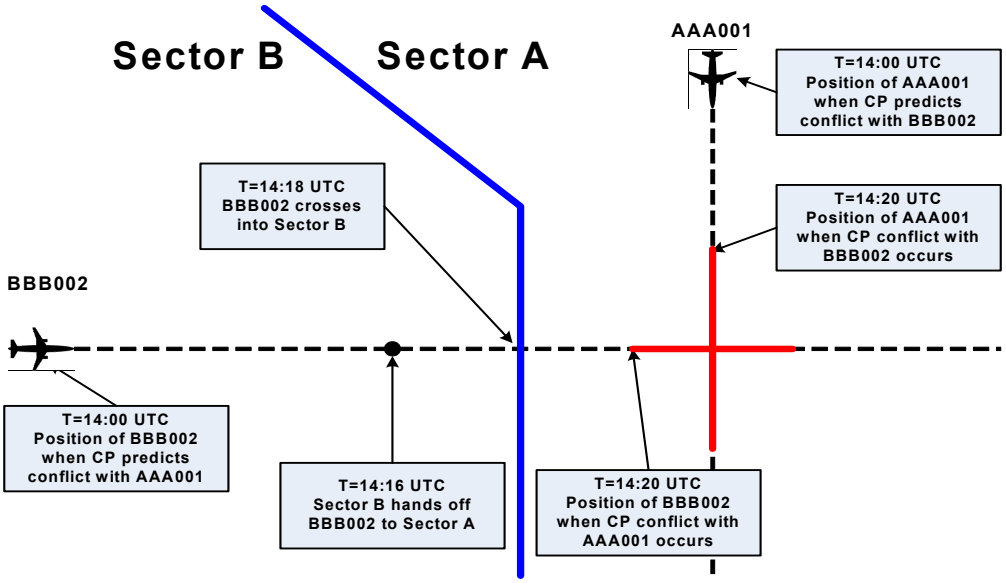


Figure 1. Illustrative Example of a Conflict

Figure 1 presents a horizontal view of an example conflict situation in which an aircraft, identified as AAA001, is flying southbound under the control of the Sector A controller. Another aircraft, identified as BBB002, is flying eastbound at the same altitude under the control of the Sector B controller. In this figure the airplane icons represent the positions of these aircraft at 14:00 UTC†, when a CP might predict a conflict in Sector A where their paths cross. The two aircraft are predicted to violate their separation criteria at 14:20 UTC, indicating a look-ahead time of 20 minutes. The situation is further complicated by the proximity of the aircraft to the sector boundary, in that the anticipated hand-off between the air traffic controllers would occur at about 14:16 UTC. This figure shows a typical separation management problem in which the flight path of the affected aircraft must be changed by heading, altitude, or speed changes.

* Data Comm refers to the program and technology required to digitally communicate air traffic and weather information between the individual aircraft and the responsible ground air traffic control automation.

† UTC refers to Coordinated Universal Time, which within this study is equivalent to Greenwich Mean Time

It must be emphasized, that this figure is presented for illustrative purposes only in order to provide a touchstone for describing how conflicts are resolved. The separation of aircraft is complex because each situation presents unique circumstances, where numerous variables such as other air traffic, special use airspace that must be avoided, proximity of one of the aircraft to its destination, weather, and controller workload influence the action selected by the air traffic controllers. For this reason a number of assumptions are made to simplify this example. First, it is assumed both AAA001 and BBB002 are mid-flight and at their cruise altitude and AAA001 is flying at a slower rate of speed than BBB002. Secondly, it is assumed there are no thunderstorms or other abnormal weather conditions nor any other aircraft in the area.

Lastly, it is assumed the CP presents the potential conflict to the Sector A controller. Given these assumptions, this is a simple conflict with few complicating factors that can be used in this Study Plan to illustrate approaches to conflict resolution. With this in mind, here is a comparison of the current and envisioned future ways that this conflict might be resolved:

- **Current Resolution of Conflicts.** In today's NAS there are a number of alternative ways to solve this illustrative conflict; none of which would be considered the "right" way to resolve the conflict. For example, since the Sector A controller is notified of the potential conflict by the CP he/she would typically call the Sector B controller and request an early hand-off of aircraft BBB002. Once the hand-off has occurred and BBB002 has established voice communications with the Sector A controller, the Sector A controller would notify the pilot of the heading change and the conflict would be avoided. Alternatively, the Sector A controller could contact the Sector B air traffic controller and have him/her initiate a heading change to BBB002 to resolve the conflict in a timelier manner. On the other hand, if the Sector A controller's workload was heavy, he/she might request AAA001 to change their altitude, thus avoiding the time consuming coordination with Sector B. With each of these alternatives there is much time lost due to coordination between sector controllers and voice communications with the affected pilots.
- **Future Resolution of Conflicts.** Since NextGen is currently in the process of defining the alternatives that will exist in the future NAS, it cannot be stated with certainty how this example conflict will be handled in the future. But it is anticipated the function of the CP will be enhanced in several fundamental ways. First, it will not only exist at the associate controllers display like today but will be integrated within the radar controller's display as well. The data block will contain alert data and additional pullout menus will contain time ordered listings of these notifications. Next, these enhanced alerts will automatically take into account user preferences and aircraft capabilities to generate a rank-ordered set of resolutions that both resolve the conflict and meet metering constraints. The enhanced CP will then notify the Sector A controller of the predicted conflict and present a menu of recommended resolutions. The Sector A controller will select the best resolution to the DST. If the best resolution involved maneuvering AAA001, automation would send the clearance directly to AAA001 via Data Comm. However, if the best resolution involved maneuvering BBB002, the Sector B controller would be notified of the conflict and recommended resolution, and if the Sector B controller concurs, automation would send the clearance directly to BBB002 via Data Comm.

The current CP used in the NAS is very effective at detecting potential aircraft-to-aircraft conflicts that are of real concern and providing controllers with enough time to properly separate the aircraft. However, the current CP of ERAM is not effective at filtering out those events that will not become conflicts without any controller interaction. These additional alerts that are not of interest to the controller are called nuisance alerts, because they detract from the events that are of

real concern. Before this CP can be used effectively by the radar controller or even by conflict resolution automation, the nuisance alert rate must be reduced to a level that is acceptable by Air Traffic Control (ATC) subject matter experts (SME). Likewise, in the process of reducing the nuisance alert rate, the rate of detecting the events that are of a real concern should not be adversely affected.

Thus, the motivation of this work is in support of the NextGen project titled, *Separation Management and Modern Procedures*. This particular project is charged with enhancing and then implementing a strategic conflict probe into the radar controller's display, required for the future NAS envisioned by NextGen. The prototype performance enhancements are required to present accurate and timely conflict predictions to the radar controller and support other NextGen advanced functions as they are implemented later. This paper presents results of an initial experiment limited to adjusting a few key parameters within the TM and CP functions. The details of this experiment and the analysis methodology will be presented in the Section 2.

1.2 Scope

This document reports on the results of an initial experiment limited to one large six hour traffic sample collected in May 2005 from the Washington Air Route Traffic Control Center (ARTCC). To induce conflicts between aircraft for evaluation purposes, the data sample was time shifted using a methodology documented in [Paglione, 2003]. This same scenario was originally developed and used for the formal testing of the En Route Automation Modernization (ERAM) system [Ryan, 2008]. It is envisioned that this study is an initial analysis with initial recommendations, but further experiments and analysis will follow that will confirm the results presented. Additional scenarios from other airspace regions and traffic days will expand the scope of these results. Being the first in a series of experiments, the methodology and metrics are explained in detail in this document. The descriptions provided will serve as a reference for subsequent experiments.

1.3 Document organization

Section 2 presents the analysis metrics and methodology employed to perform this experiment, including a description of the experiment itself. Section 3 presents the evaluation results of the experiment. Section 4 presents final conclusions and recommendations.

2 Analysis Metrics and Methodology

This section provides a detailed description of the metrics and methodology employed in this study. As introduced in Section 1.1, a key outcome of the authors' support for the *Separation Management and Modern Procedures Project* is to provide a detailed list of recommended enhancements to ERAM to support the subsequent implementation of a strategic conflict probe in the en route radar controller's display. Table 1 provides a comprehensive list of the various prototyping activities as defined by current ERAM development contractor's delivery schedule (i.e. Lockheed Martin). The prototypes are delineated by functional areas. The two key functional areas are:

1. Functional Area 32 which targets improvements to the aircraft trajectory modeling sub-systems and algorithms, and
2. Functional Area 18 which is focused on improvements of the conflict prediction algorithms within the ERAM Conflict Probe Tool (CPT) sub-system.

The FAA's Concept Analysis Branch (ANG-C41) is charged with the task of providing an independent engineering evaluation of these prototypes. The evaluation is limited to the laboratory performance of the ERAM's trajectory and conflict probe's prediction capability. It expands on the accuracy requirements used in the formal ERAM testing program as described in detail in [Ryan, 2008].

Table 1. Summary of Prototyping Activities

#	Functional Area	Description of Prototyping Effort	Estimated LM Delivery[‡]	Evaluation Strategy
1	FA32	Enhance initial point and lateral rejoin logic	6/17/2011	Integrated Experiment
2	FA32	Enhance turn modeling	2/14/2011	Individual Evaluation
3	FA32	Vertical Profile Modeling Enhancement	1/28/2011	Individual Evaluation
4	FA32	Enhance Restriction Modeling	8/31/2011	Individual Evaluation
5	FA32	Enhance modeling in terminal area (RNAV STARs to runway)	6/21/2011	N/A
6	FA18	Trajectory Monitoring (TM) and Conflict Detection (CD) adherence bounds (functionally separate adherence bounds)	9/30/2011	Integrated Experiment
7	FA18	Application of CD forced trajectory re-builds (FTR)	9/30/2011	Integrated Experiment
8	FA18	Application of growth adherence bounds (GAB) for "near-time" encounters	9/30/2011	Integrated Experiment
9	FA18	Application of conflict post-processing based on conflict geometry parameters	9/30/2011	Integrated Experiment
10	FA18	Application of tailored warning times based on conflict parameters	9/30/2011	Integrated Experiment
11	FA18	Application of new red/yellow criticality concept	9/30/2011	Integrated Experiment
12	FA18	Application of specific designated notification sector changes	9/30/2011	Individual Evaluation
13	FA18a	3nm strategic separation	6/30/2011	Individual Evaluation

[‡] Based on slide # 6 from the presentation listed in

However, the evaluation is not limited to these accuracy requirements but significantly expands upon them by modifying and adding dozens of additional metrics. The objective of the evaluation is to answer the following question with a certain level of confidence based on the sample traffic data utilized.

> *Has the prototype software listed in Table 1 in red and labeled "Integrated Experiment" installed in the ERAM subsystem, the Conflict Probe Tool (CPT), improved or degraded its trajectory (FA32) and conflict (FA18) prediction's performance?*

The hypothesis therefore is that the upgraded CPT software performs equivalent to baseline CPT in terms of the trajectory and conflict prediction metrics. These metrics are defined by the authors in collaboration with system engineers from the ERAM development contractor and MITRE's Center for Advanced Aviation System Development (CAASD). A set of experiments test whether this claim can be rejected and whether the performance improved. Since the CPT installed with the prototypes cannot be accepted if they degrade the system, the performance must be greater than the existing baseline performance. The amount of improvement required to be operationally significant will remain unknown for this phase of analyses. Only with the later consultation with subject matter experts (e.g. air traffic personnel) or through further evaluation using a human in the loop experiment proposed in [Willems, 2011] can this determination be made. The analysis objective is to first determine if the hypothesis can be rejected statistically (i.e. CPT improved) for each key metric and then synthesizes the results from the multitude of competing metrics. For example, the three key metrics for the conflict predictions are the:

- Measure of aircraft-to-aircraft conflict events that indeed should have been predicted in a timely manner and were not – missed alerts,
- Measure of nuisance aircraft-to-aircraft conflict notifications that should not have been presented but were – false alerts,
- Measure of lead time provided when aircraft-to-aircraft conflict event is presented when it should have been – warning time.

One challenge is these metrics could be inversely proportional. In particular, as missed alerts decrease false alerts tend to increase, thus the evaluation will need to balance these competing response variables and assess the overall performance of the CPT sub-system under various conditions. Another challenge is the various prototypes may interact with one another so these interactions must be taken into account. For example, if a reduction in the conflict detection bounds protects for less airspace inducing less alerts overall, it could leave in-trail overtake situations with less warning time. However, the Conflict Geometric Separation (CGS) prototype[§] that inhibits in-trail alerts but predicts overtake cases would address this issue, allowing the overall benefit of the reduced conflict detection bound to prevail.

Several methods will be brought to bear to address these challenges. Experimental design is well suited to help integrate the need to assess all the factors involved and their interactions, while examining the multitude of competing response metrics. It is defined as a "test or series of tests in which purposeful changes are made to input variables of a process or system so that we may observe and identify the reasons for changes in the output response."[**] Initially two experiments are being implemented. The first test examines the factors involved that simply change CPT's parameters (e.g. TM and CD adherence bounds and predicted notification time thresholds). The

[§]CGS prototype is an implementation of the software described as "conflict post-processing based on conflict geometry parameters" listed in Table 1.
[**]See introductory chapter in [Montgomery, 2009]

second larger experiment is designed to assess the performance of several CPT prototypes (e.g. growth adherence bounds or GAB, forced reconformance or FTR). By utilizing experimental design techniques, an optimal design algorithm calculates the minimal number of ERAM CPT runs required to extract the maximum amount of information. These techniques were applied by the authors in similar contexts as described in both [Paglione, 2008] and [Santiago, 2010].

A second advanced analysis method employed was development of the metrics themselves by applying what is referred to as black box and white box testing. Black box testing is employed when applying the author's traditional crisp metrics considering minimal details inside the process under study but more on its input and output. For example, the system generates N alerts and it is partitioned as correct predictions (valid) and nuisance or false alerts. This is illustrated in the general model of a process in Figure 2, adapted from [Montgomery, 2009].

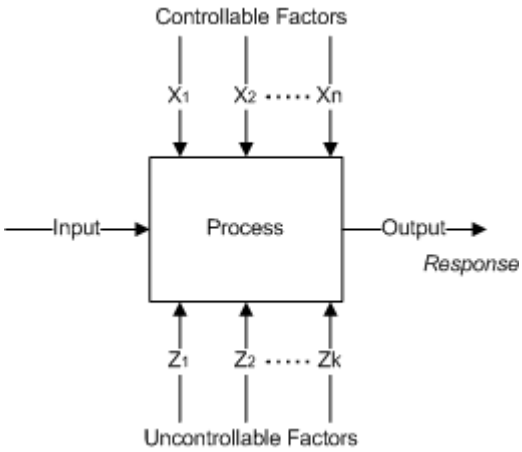


Figure 2. Black Box – Crisp Metrics

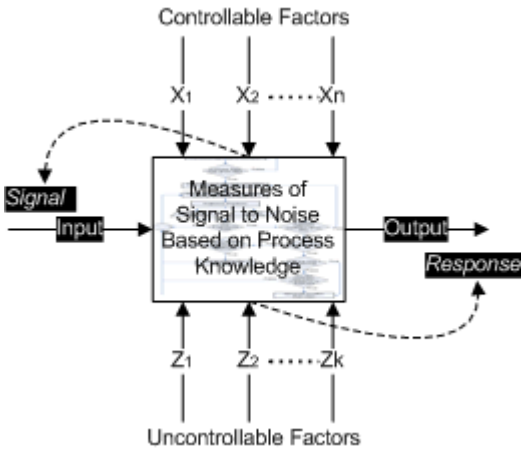


Figure 3. White Box – Fuzzy Metrics

Illustrated in Figure 3, the white box view of the same process takes additional knowledge from the process itself and calculates fuzzy or more sensitive weighted metrics for the same events [Parasuraman, 2000]. Thus, the same nuisance or false alert event from the crisp analysis could be partially valid and partially false depending on additional factors related to the signal (i.e. the encounter event itself) and the response (i.e. alert color or warning time). For example, an encounter between two aircraft that is 5.1 nautical miles is a near conflict and typically of interest to the air traffic controller if formed from the crossing paths of the flights. Under the white box – fuzzy metric approach, an alert predicted for this event would receive a high valid assessment, while the black box – crisp metric would determine it is not a legal conflict based on its separation so label it as a false alert.

In summary, the experimental design technique is being employed to plan and implement a set of experiments to evaluate the prototypes listed in Table 1. Extended from the ERAM Formal Test Program, a set of advanced metrics are being employed using both crisp and fuzzy measurement methods to fully test the hypothesis that these prototypes improved the ERAM CPT compared to a baseline without them. The evaluation will provide a statistical assessment that can be handed-off to the next level of evaluation. This next phase is the operational assessment focused at answering the question: is the prototypes' performance improvement significant enough to be operationally acceptable for the intended function, namely implementation of CPT functionality on the radar controller's console.

The metrics and methodology will be applied to all the designated prototypes in Table 1 (i.e. in red). However, this particular study reports on the initial experiment and focused only item number 6 in Table 1. This includes evaluating the impact of changing the TM and CP bounds[††]. It also includes changes to the CPT conflict likelihood parameters. Thus, the experiment evaluates the impact of altering existing configurable parameters within ERAM, offering NextGen the option of first enhancing the performance of the ERAM strategic conflict probe without actually making costly software changes. The study includes results of the first trajectory modeling enhancement, item number 1 from Table 1. It is expected that the other prototypes will require both changes in these parameters as well as the FA 32 (item number 1) enhancements, so logically the initial experiment focused on these items first.

The following sub-sections will first present the evaluation metrics for both trajectory and conflict predictions in Section 2.1. Next in Section 2.2, the methodology of the designed experiment will be described including details on the model used to fit the experimental results.

2.1 Evaluation Metrics

The metrics used for evaluating the performance of the conflict probe in this experiment differed slightly from those used in previous studies. While most of the metrics used in previous studies are also used in this study, some additional metrics have been implemented as well. As stated above, two basic grouping of conflict prediction metrics are employed, including the traditional crisp and newer fuzzy metrics. The trajectory prediction accuracy metrics are also utilized and presented first.

2.1.1 Trajectory Modeling Metrics

As noted in previous work [Crowell, 2011, A], trajectory error metrics provide a statistical method to measure the accuracy of trajectories in multiple dimensions with respect to actual flight position. Correct implementation of these metrics allows for comparison and validation of trajectory predictors. Sampling methods and definitions of these metrics along with related analysis techniques are presented in the following sections.

2.1.1.1 Trajectory Sampling

The Interval Based Sampling Technique (IBST) is the trajectory accuracy sampling method developed by the FAA's Concept Analysis Branch. It has been previously documented in [Paglione, 2007] and has been used in a number of FAA studies and test programs. As early as 1999, the Trajectory Predictors (TP) within the operational DSTs of the User Request Evaluation Tool (URET) and the Center TRACON Automation System were evaluated using this technique. More recently, it is being employed to evaluate the TP in the En Route Automation Modernization (ERAM) system, which will replace the en route operational systems such as URET and the Host Computer System. IBST is a two-step process that pairs the track and trajectory points to measure the prediction errors for an entire flight. This sampling technique takes the perspective of the DST user, the air traffic controller. The active trajectory at the time the controller is looking at the display may be several minutes old and in error. Consequently, in the IBST the trajectories are sampled at the current time for a look-ahead time of 0 seconds and at a number of parameter times in the future (e.g., 300, 900, and 1200 seconds). This is contrasted

[††] The prototype enhancement listed in Table 1 item number 6 refers to the software change of decoupling the TM and CP adherence bounds. The experiment described in this report examines the impact of reducing the bounds without decoupling them. Future experiments will need to test the impact of setting the TM and CP bounds at different thresholds.

with a sampling technique that uses the internal build time of the trajectory to start the sampling [Brudniki, 1998] [Paielli, 1999].

2.1.1.2 Definition of Metrics

The four basic metrics defined in [Paglione, 2007] are horizontal error, vertical error, along track error, and cross track error. Figure 4 provides a notional illustration of these errors. The horizontal error is the time coincident difference in nm between the predicted position on the trajectory and the actual position calculated from surveillance radar reports. Cross track error (nm) is the perpendicular distance between the actual position of an aircraft and its projection onto the trajectory. Along track error (nm) is the longitudinal distance along the trajectory between the same projection and the time coincident predicted position of the aircraft. The vertical error (not illustrated in the figure) is the altitude difference in feet between the predicted trajectory position and the time coincident actual position. For further details on these definitions and how they are calculated see [Paglione, 2007].

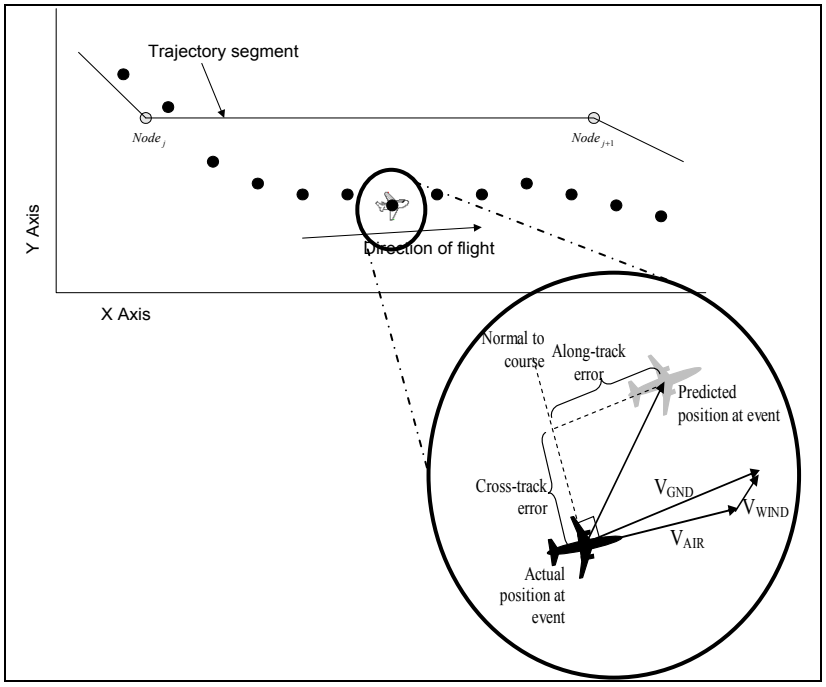


Figure 4. Diagram of Trajectory Errors

Following these definitions, horizontal error is unsigned while the other three are signed. For most analyses involving these metrics, it is desirable to consider absolute values because the distance from zero is of interest. Therefore the absolute value is taken before calculating average values. Finally, the average metrics comprise only relevant data points by considering error values for trajectory points where a clearance or route amendment has not been received within a specified time period (as that may have altered the trajectory) and the flight remains within control of the center.

2.1.1.3 Statistical Techniques

A common application of trajectory accuracy error metrics is in evaluating flight scenarios with both trajectory data and aircraft position data, often simulated. In the case where two scenarios containing data for the same set of flights are to be compared, it is possible to match up the performance of trajectories for the same flight in each scenario and get more information about

how the two scenarios compare. For example, the average error value is calculated for the same flight in two scenario runs; these values may be collected and compared for many flights to illustrate any underlying differences between the runs. This is similar to taking multiple measurements from the same test subject in different treatments, and is referred to as a "matched pairs" or "paired difference" test in statistics.

A matched pair analysis is performed for each of the four metrics using the average absolute error per flight. The analysis includes a paired t-test, which examines the distribution of differences in error between the two scenarios and tests if the mean of the differences is statistically different from zero. The results from the paired t-tests are provided in tables. The mean differences indicate which run has less error on average, and the p-value indicates whether or not this difference is statistically significant. The t-ratio is a statistical test parameter which is calculated as the mean difference over the standard error. Large negative values of the t-ratio produce small p-values and indicate that the mean difference is of statistical significance. The presented p-value is the probability of observing a discrepancy in means at least as large as that observed in the data set, even if there is no underlying difference in the means. A p-value less than 0.05 is typically considered to indicate statistical significance.

The results of the matched pair analysis may be presented graphically by plotting the difference in average trajectory error per flight between two runs against the average over these runs, where each data point represents one flight. The flight's average error in the first run minus the respective mean in the second run is plotted on the vertical axis, while the horizontal axis represents the average of these two measurements (calculated by taking the sum of the mean error in each run and dividing this sum by 2 for each matched flight). The resulting plot indicates net trends between runs. For example, if the errors are predominantly above a zero line in the middle of the plot, this indicates the error is larger in the first run more often. If the errors are below this line, they indicate the error is larger in the second run. Appendix C provides a detailed description of the matched pair analysis and graphical output from the commercial statistical software package JMP® , which is used for the analyses presented in this document.

2.1.2 Conflict Probe Metrics

This section will describe the conflict prediction metrics that gauge the performance of an aircraft conflict probe directly. There are two sub-sections – Section 2.1.2.1 presenting the "crisp" conflict prediction metrics and Section 0 the "fuzzy" conflict prediction metrics.

2.1.2.1 Conflict Prediction "Crisp" Metrics

This section provides a detailed description of the traditional crisp conflict prediction metrics.

2.1.2.1.1 Definition of Basic Conflict Prediction Metrics

A conflict probe predicts when two aircraft will violate separation standards some time in the future. A violation of separation standards is typically called a conflict. An event where two aircraft pass near each other but not close enough to violate separation standards is labeled an encounter in this study. As documented in [Paglione, 1999], [Bilimoria, 2001], [Brudnicki, 1998], and [Cale, 1998], the conflict probe is not perfect and does make mistakes. For example, it can miss a conflict (Missed Alert) or it can predict a conflict that never occurs (False or Nuisance Alert). The four possible situations are shown in Table 5.

For a real time system, it is important that an alert be given sufficiently earlier in time of the actual conflict so corrective action can be taken. In other words, an alert must be timely as well as accurate. To ensure timeliness in conflict predictions, a conflict probe is often required to have

some lead-time or actual warning time. This Minimum Warning Time (MWT) ranges from 1 to 5 minutes depending on the particular type of conflict probe being evaluated. For this study, specified notification lead time is required unless the conflict was determined to be a pop-up event. A pop-up conflict occurs when the probe is not provided with MWT threshold of continuous surveillance data or prediction for either of the associated flights. Detailed descriptions of the different situations that cause this to occur are described in [Paglione, 1999].

Table 2. CP Alert and Conflict Event Combinations (adapted from [Paglione, 1999])

	CONFLICT OCCURS	**CONFLICT DOES NOT OCCUR**
ALERT	CP predicts conflict and it occurs *(VA – valid alert)*	CP predicts conflict and it does not occur *(FA -- false alert)*
NO ALERT	CP does not predict conflict and it occurs *(MA -- missed alert)*	CP does not predict conflict and it does not occur *(NC -- correct no-calls)*
Total Number of Alerts	**Total Number of Conflicts**	**Total Number of Non-Conflicts (i.e. encounters without conflicts)**

For this study, the conflict prediction accuracy crisp metrics consist of counts of the error events, including the false alerts (FA) and missed alerts (MA) in context of the correctly predicted events of valid alerts (VA) and correct no-calls (NC). The Conflict Probe metrics used in this study were described in detail in [Paglione, 2004] and [Crowell, 2009]. Thus, three main metrics used in the past are the VA, MA, and FA counts. There are also Missed Alert Rate (MAR) and False Alert Rate (FAR) that use ratios of the main metrics counts. The following Eq. 1 and Eq. 2 describe these ratios as the number of missed alerts over the total missed and valid alert counts and number of false alerts over the number of false alerts and correct no-calls, respectively.

$$MAR = \frac{(MA)}{(MA + VA)} \qquad \textbf{Eq. 1}$$

where,
MA is total number of missed alerts
VA is the total number of valid alerts

$$FAR = \frac{(FA)}{(FA + NC)} \qquad \textbf{Eq. 2}$$

where,
FA is total number of false alerts
NC is the total number of non-conflict encounters without associated alerts

In this study, we split what previously were called Missed Alerts into two categories called Late Alerts (LA) and Missed Alerts (MA). Late Alerts occur when an alert is not posted within the minimum warning time of a conflict, but is still posted earlier than 40 seconds prior to the start of the actual conflict. Missed Alerts are all conflicts in which an alert is not posted until within 40 seconds of the start of the conflict. This includes conflicts which have no alert posting prior to the start of the conflict, sometimes referred to as no-call missed alerts. In the past, the minimum warning time requirement used for analysis of a strategic conflict probe was typically five minutes. However, after discussions with Air Traffic Controller (ATC) Subject Matter Experts

(SMEs), it was determined the use of a three minute minimum warning time provides a better measure of what should be considered a Late Alert. Thus, another rate is required and used in the analysis that focuses only on these Late Alerts and does not include the excused popup conflict events. It is summarized in the following Eq. 3.

$$LAR = \frac{(LA + MA)}{(LA + MA + VA^{*})} \qquad \textbf{Eq. 3}$$

where,

LA is total number of late alerts, that are missed due to warning time less than MWT but greater than 40 seconds warning time,

MA is the total number of missed alerts that are missing completely or due to a warning time less than 40 seconds, and

VA^{*} is the total number of standard valid alert conflicts (excludes valid alerts[‡‡] with less than threshold MWT of warning time associated to pop-up conflicts)

2.1.2.1.2 Comparison of Conflict Prediction Results

The typical approach of comparing an improvement or change in the CP is to assume the baseline performance as the role of the requirement and test the improved system against the baseline's performance. There are two main limitations to this approach. If the MAR and FAR of the legacy system is compared to the new "prototype" system, these metrics are themselves random variables. The test tends to under estimate this random variation. Second and more importantly, this approach summarizes the errors for both systems into a ratio and only net effects are compared. For example, if the ERAM system had a total of two more missed alerts than the legacy URET system; it is only considering the net difference. In reality, ERAM had four missed alerts that the legacy system had correctly predicted, yet two more missed alerts were generated by the legacy system that ERAM correctly predicted. The test only compared the net quantities of missed alert events. A more sensitive test would compare the same conflict and alert events, reporting all mismatches. Furthermore, it is necessary to identify all the specific error events from a practical standpoint, so software corrections can be made. This section will present a method to identify and statistically compare these events.

Before discussing this comparison, Table 3 lists the individual reason codes for each run's conflict prediction results. The alert type falls into four categories. First, the missed alert and false alerts are the two errors being measured. Next, the valid alerts are the correct prediction of a conflict and discards are the events excused due to out of adherence situations or other artifacts of the traffic sample being used. The alert types and reason codes are generated by a software program written by ANG-C41. It matches the ground truth conflict and non-conflict encounter events and produces a data base table with these codes. It is titled the *StrategicAlertEvaluator* or SAE.

[‡‡] This will be defined as late valid alerts defined in the next section and listed in Table 3 as LATE_VA.

Table 3. Conflict Prediction Result's Main Reason Codes[§§]

REASON CODE	ALERT TYPE	REASON DESCRIPTION
STD_VA	Valid Alert	Standard Valid Alert
LATE_VA	Valid Alert	Late Valid Alert, Valid since conflict was determined a pop-up
NO_CALL_MA	Missed Alert	Missed Alert due to no call (no alert at all before the actual conflict start time)
LATE_MA	Missed Alert	Late alert – alert presented with less than the minimum required warning time
SHRT_NO_CALL_DISCARD	Discard Alert	Missed Alert no call discarded because conflict duration below a threshold time
NO_CALL_DISCARD	Discard Alert	Missed Alert no call discarded since out of adherence
SHRT_LATE_DISCARD	Discard Alert	Late alert discard because conflict duration below a threshold time
LATE_DISCARD	Discard Alert	Late alert discard since out of adherence
NO_TRK_FA_DISCARD	Discard Alert	No post processed track at predicted conflict start time so discard
NO_ADHER_FA_DISCARD	Discard Alert	Out of adherence at predicted conflict start time so discard
CLR_FA_DISCARD	Discard Alert	Retracted False Alert assigned by an ATC clearance so discard
CFL_FA_DISCARD	Discard Alert	False Alert notified beyond last conflict actual start time so discard
STD_FA	False Alert	Standard False Alert
RETRACT_FA	False Alert	Retracted False Alert, notification end time earlier than predicted conflict start time
IN_APDIA_FA	False Alert	False alert generated but predicted conflict start time determined to be inside an automated problem detection inhibited area

The sets of conflict predictions generated by the legacy system run (e.g. Run A) and a new system (e.g. Run B) are first evaluated separately. The analysis results in a database table with records labeled with the reason codes defined in Table 3. The resulting paired evaluations of the two runs are listed in Table 4, but for this version it is assumed both runs are provided the same input traffic scenario. The first column in Table 4 lists all combinations of intersection and union of the events from Table 3. As in Table 4 and throughout this section, the first run compared will be referred to as Run A and the second as Run B. For example, Run A may generate a missed alert that is either a missed alert or valid alert in Run B and vice versa. Of particular interest is when Run A makes an incorrect prediction and Run B makes a correct prediction and vice versa. This is highlighted by the light orange shaded rows. It indicates one run has a VA and the other an MA for the same conflict event. The light green shading highlights where one run has an FA and the other a correct no-call for the same encounter.

[§§] This table summarizes the reason codes into 15 types that capture the essence of the processing involved. The actual processing software produces a total of 19 reason codes.

Table 4. Comparison of Two Run's Resulting Alert and Conflict Event Combinations*

	CONFLICT OCCURS	CONFLICT DOES NOT OCCUR
ALERT by both Runs A and B	Both predicts conflict and it occurs (V_A & V_B -- valid alerts both)	Both predicts conflict and it does not occur (F_A & F_B -- false alert both)
ALERT by A and not B	A predicts conflict and it occurs (V_A -- valid alerts by A only)	A predicts conflict and it does not occur (F_A -- false alert by A only)
	B does not predict conflict and it occurs (M_B -- missed alert by B only)	B does not predict conflict and it does not occur (NC_B – B correct no-call)
ALERT by A and B ALERT or non-ALERT is discarded	A predicts conflict and it occurs (V_A -- valid alerts by A only)	A predicts conflict and it does not occur (** F_A Continued **)
	B does not predict conflict correctly but is discarded ($DiscardM_B$ -- B discard an MA)	B does not predict conflict correctly but is discarded ($DiscardF_B$ – B discard an FA)
ALERT by B and not A	B predicts conflict and it occurs (V_B -- valid alerts by B only)	B predicts conflict and it does not occur (F_B -- false alert by B only)
	A does not predict conflict and it occurs (M_A -- missed alert by A only)	A does not predict conflict and it does not occur (NC_A – A correct no-call)
ALERT by B and A ALERT or non-ALERT is discarded	B predicts conflict and it occurs (V_B -- valid alerts by B only)	B predicts conflict and it does not occur (** F_B Continued **)
	A does not predict conflict correctly but is discarded ($DiscardM_A$ -- A discard an MA)	A does not predict conflict correctly but is discarded ($DiscardF_A$ – A discard an FA)
NO ALERT by both Runs A and B	Both do not predict conflict and it occurs (M_A& M_B -- missed alert by both)	Both do not predict conflict and it does not occur (NC -- correct no-calls by both)
Total Number of Alerts for each/both	Total Number of Conflicts *(Same for both Runs!)*	Total Number of Non-Conflicts (Encounters that did not have conflicts; *Same for both Runs!*)

To determine the various combinations of comparative events as defined in Table 4, ANG-C41 wrote a software tool to identify them from the conflict prediction results for a pair of conflict probe runs. The program is titled the *StrategicAlertComparer (SAC)* or newer version called *AlertComparerEnhanced*. The program produces a database table of entries with evaluation codes for each of these events. A listing of these combinations and their corresponding codes are listed in the following Table 5. Because of limitations in the fidelity of the ERAM simulation platform, there are small differences in the input scenarios from run to run (e.g. slight time deviations in the radar track reports). Thus, an advanced matching algorithm was developed and presented in [Crowell, 2009] that matches the events in an efficient and fair manner. Table 5 delineates the events for equivalent ground truth input scenarios or when the scenarios are slightly different. Thus, if the codes are labeled false in the column "Same Ground Truth Only," the event will only occur if the scenarios were different in some way.

*** These events are for comparison of Run A and Run B input with the same exact ground truth scenario.

Table 5. Conflict Prediction Comparison Program Evaluation Codes[†††]

Event (Labels from Table 4)	Evaluation Code	Same Ground Truth Only	Description
V_A and V_B	SAME_VA	True	Both runs have valid alerts for the same conflict
M_A and M_A	SAME_MA	True	Both runs have missed alerts for the same conflict
F_A and F_B	SAME_FA	True	Both runs have false alerts for the same encounter
V_A and M_B	VA_MA	True	Run A has a valid alert and Run B has a missed alert for the same conflict
M_A and V_B	MA_VA	True	Run A has a missed alert and Run B has a valid alert for the same conflict
V_A and DiscardM_B	VA_DISCARD	True	Run A has a valid alert while Run B discards the conflict
DiscardM_A and V_B	DISCARD_VA	True	Run A discards the conflict while Run B has a valid alert
DiscardM_A and M_B	DISCARD_MA	True	Run A discards the conflict while Run B has a missed alert
M_A and DiscardM_B	MA_DISCARD	True	Run A has a missed alert while Run B discards the conflict
F_A and NC_B	FA_NC	True	Run A has a false alert while Run B has no prediction to match
NC_A and F_B	NC_FA	True	Run A has no prediction to match while Run B has a false alert for the same encounter
F_A and V_B	FA_VA	False	Run A has a false alert while Run B has a valid alert
V_A and F_B	VA_FA	False	Run A has a valid alert while Run B has a false alert
F_A and M_B	FA_MA	False	Run A has a false alert while Run B has a missed alert
M_A and F_B	MA_FA	False	Run A has a missed alert while Run B has a false alert
V_A and no match in B	VA_NOMATCH	False	Run A has a valid alert while Run B has no event to match
No match in A and V_B	NOMATCH_VA	False	Run A has no event to match while Run B has a valid alert
M_A and no match in B	MA_NOMATCH	False	Run A has a missed alert while Run B has no event to match
No match in A and M_B	NOMATCH_MA	False	Run A has no event to match while Run B has a missed alert
F_A and DiscardF_B	FA_DISCARD	True	Run A has a false alert while Run B discards the event
DiscardF_A and F_B	DISCARD_FA	True	Run A discard the event while Run B has a false alert
Discard$_A$ and no match B	DISCARD_NOMATCH	False	Run A discard the conflict while Run B has no event to match
No match A and Discard$_B$	NOMATCH_DISCARD	False	Run A has no event to match while Run B discards the conflict
Discard$_A$ and correct no call B	DISCARD_NC	True	Run A discards the event while Run B has no prediction to match
Correct no call A and Discard$_B$	NC_DISCARD	True	Run A has no prediction to match while Run B discards the event

[†††] These are a roll-up of the evaluation codes from SAC for when the scenario input of the Run A and B is the same and is not exactly the same, delineated by "Same Ground Truth Only" column. Also, assume MA is the sum of LA and MA events.

There are two additional comparison metrics used in this document that help to summarize the results of the comparison. These are % FA Improvement and % LA Improvement. These metrics should not be confused with FA% or LA%, which are not comparison metrics. Eq. 4 shows the formula for calculating % FA Improvement. Note that only the FA_NC and NC_FA counts are used. Any False Alerts in either run that match to Discards in the other run are intentionally not used in this metric. This metric is intended to show the differences between the two runs when both runs had the best chances of being correct. Since Discards usually mean something was wrong with the track or the trajectory, these counts are not included in this metric.

$$\%FA_{IMP} = \frac{FA_NC - NC_FA}{FA_A} \qquad \textbf{Eq. 4}$$

where,

FA_NC is the count of FA_NC events in the comparison
NC_FA is the count of NC_FA events in the comparison
FA_A is the count of FA events in the Run A of the comparison
$\%FA_{IMP}$ is the percentage of improvement of Run B over Run A

Eq. 5 shows the formula for calculating % LA Improvement. Once again, no Discard events are used because the metric is intended to show the differences between the two runs when they both had the best chance of correctly predicting the alert.

$$\%LA_{IMP} = \frac{LA_VA - VA_LA}{LA_A} \qquad \textbf{Eq. 5}$$

where,

LA_VA[‡‡‡] is the count of LA_VA events in the comparison
VA_LA[§§§] is the count of VA_LA events in the comparison
LA_A is the count of LA events in the Run A of the comparison
$\%LA_{IMP}$ is the percentage of improvement of Run B over Run A

2.1.2.1.3 Statistical Approach of Comparing Conflict Predictions

As implied in equations presented in Section 2.1.2.1.1 and the shading in Table 4, the most critical quantities to detect a statistical difference between runs are when one run correctly predicts an event and the other does not. Comparison of the MAR, LAR, and the FAR metrics will indicate the net magnitude of these differences. One approach to determine if the difference is statistically significance is to utilize a binomial distribution and perform a hypothesis test concerning the difference between population proportions [Devore, 2000]. However, this technique assumes that the respective runs are independent. For this study, each run is not independent, since the system is run with the same air traffic scenario and weather files.

An alternative technique is presented in [Kachigan, 1986], utilizing categorical data analysis techniques. For categorical data analysis, we examine the difference in frequencies not proportions. For this study, the frequencies directly relating the missed and false alert rates include the counts of these events. Paired counts that are mutually exclusive and exhaustive, which is required for this test, occur when the error event occurs in one run and the correct event occurs in the other.

[‡‡‡]Eq. 5 is assuming LA is the sum of LA and MA events. In Table 5, MA is used to refer to this same sum, both the LA and MA events. Thus, LA_VA from Eq. 5 is equivalent to MA_VA from Table 5.
[§§§]Same as previous footnote, VA_LA is equivalent to VA_MA and refers to all LA and MA events.

For the missed alert analysis, the count of interest is the missed alert count in Run A when simultaneously getting a valid alert in Run B or vice versa for the opposite case. These include the counts V_{A2} or M_{B2} compared to the V_{B2} or M_{A2}. Therefore, the count of valid alerts in Run A and simultaneous missed alerts in Run B is statistically compared to the count of valid alerts in Run B and simultaneous missed alerts in Run A. These counts should be equally likely if the two runs are statistically equivalent. Calculating the ratio of the squared difference between the expected value of each run and the observed value can test this hypothesis. If the hypothesis is true, this ratio will follow a chi-squared distribution or χ^2 with one degree of freedom. The test statistic is as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad \textbf{Eq. 6}$$

where

O_i is the observed frequency in category i

E_i is the expected frequency in category i

k is the total number of categories

For this study, k is always two, since only paired runs are compared. For example, the observed frequencies are the extracted V_{A2}/M_{B2} and V_{B2}/M_{A2} counts for the two runs. Since the null hypothesis assumes both events are equally likely, both expected frequencies are equal and calculated from the following equation:

$$E_i = \frac{\sum_{j=1}^{k} O_j}{k} \qquad \textbf{Eq. 7}$$

The resulting test statistic in Eq. 6 can be expressed as a probability or P-value[****] by assuming a chi-squared distribution with one degree of freedom. For example, let's say we observe a $V_{A2}/M_{B2} = 8$ and a $V_{B2}/M_{A2} = 22$. The expected frequency from Eq. 7 is 15 for both values, and the resulting test statistic from Eq. 6 is 6.53. Therefore for this example exercise, the P-value is 0.011. This expresses that the hypothesis that these runs have equivalent missed alerts is only about one percent likely and provides evidence to reject the null hypothesis. For this test in the study, a P-value, which is less than 0.10, is considered sufficient to reject the hypothesis.

False alert probabilities can be analyzed in an analogous way. For the false alert counts, the observed frequency of F_{A2}/NC_B and F_{B2}/NC_A are compared.

The preferred method and further simplification to the test above was presented in [Agresti, 2002]. The test is referred to as the McNemar's test and is specifically designed for testing two

[****] In [Devore, 2000], the P-value is defined as the "smallest level of significance at which the null hypothesis would be rejected when a specified test procedure is used on a given data set." Thus, the P-value is the probability of the null hypothesis has occurred, so a small P-value (less than 0.10) would indicate the null hypothesis unlikely and should be rejected and if large should be assumed correct.

data sets that are not independent. This is clearly the case in this study where the same flights are examined between two runs of ERAM. An example is illustrated in the following Table 6.

Table 6. Example Application of the McNemar's test

Encounters in Treatment ERAM System	Encounters in Legacy ERAM System		
	With False Alert	Without False Alert	Total
With False Alert	47	11	58
Without False Alert	37	63	100
Total	84	74	158
	χ^2=14.083, *df*=1; p-value=0.000		

The test statistic, χ^2, is defined generically as follows:

$$\chi^2 = \frac{(n_{21} - n_{12})^2}{(n_{21} + n_{12})} \qquad \textbf{Eq. 8}$$

where,
n21 is the quantity of flights in the second row, first column of the table
n12 is the quantity of flights in the first row, second column of the table

Under assumed conditions, the test statistic assumes a Chi-Squared Distribution. The test statistic can be applied directly to the evaluation code quantities listed in Table 5. The following Eq. 9 and Eq. 10 both apply the generic Eq. 8 to these evaluation codes first for missed alert processing and then false alert, respectively.

$$\chi^2_{MA} = \frac{(MA_VA - VA_MA)^2}{(MA_VA + VA_MA)} \qquad \textbf{Eq. 9}$$

where,
MA_VA is quantity of Run A missed alerts with matching Run B valid alerts
VA_MA is quantity of Run A valid alerts with matching Run B missed alerts

$$\chi^2_{FA} = \frac{(FA_NC - NC_FA)^2}{(FA_NC + NC_FA)} \qquad \textbf{Eq. 10}$$

where,
FA_NC is quantity of Run A false alerts with matching Run B correct no-calls
NC_FA is quantity of Run A correct no-calls with matching Run B false alerts

Note, it can be shown that the methods in Eq. 6 and Eq. 7 will produce equivalent results as those in Eq. 8 through Eq. 10. Also, for the test statistics above to assume a Chi-Squared Distribution the sum of $n_{21} + n_{12}$ in Eq. 8 or equivalently (MA_VA + VA_MA) and (FA_NC + NC_FA) must all be greater than 25. If their sum is less, an exact test can be used to utilize the Binomial Distribution with $n_{21} + n_{12}$ size parameter and 0.5 for the probability of success. Details are provided in [Agresti, 2002].

2.1.2.1.4 Additional Crisp Metrics

Other than the counts and rates discussed above, the single scenario metrics include the Sharpness and Sharpness Bias metrics described in [Paglione, 1999]. Also, point statistics describing the distribution of the warning time (WT) of the predicted conflicts provide an overview of the timeliness of the conflict notifications. Warning time is the lead time provided by the predicted notification. It is measured by the following Eq. 11.

$$WT_i = ACST_i - NST_i \qquad \textbf{Eq. 11}$$

where,

WT_i is warning time for i^{th} valid alert conflict prediction

$ACST_i$ is actual conflict start time for i^{th} valid alert conflict prediction

NST_i is the notification start time for i^{th} valid alert conflict prediction

The warning time is calculated on all the valid alerts and late alerts not associated to a popup conflict event (i.e. STD_VA and LATE_MA events from Table 3). There are a number of point statistics calculated including average, median, maximum, minimum, standard deviation, 25th percentile, 75th percentile, and inter-quartile range which is the difference between the 75th and 25th percentiles. The warning time statistic that exhibits an important insight into the warning time distribution is the 25th percentile. It provides reasonably sensitive measure of the lower end of the distribution indicating how close this tail is to the MWT threshold (3 minutes in this study) and the tactical threshold of 40 seconds. The greater the 25th percentile more warning time is provided by the CP overall. The smaller the value and closer to the tactical threshold indicates the particular run is less suitable for strategic operations.

Another useful CP performance metric is called sharpness which involves two metrics including a measure of sensitivity and bias. It was developed by the authors in 1999 and is described in detail in [Paglione, 1999]. The sharpness metric indicates the sensitivity of the probe as a function of the unit-less metric min-max-ratio. Min-max-ratio is the measure of the separation between a pair of aircraft. It is calculated for all aircraft with encounters of a large separation distance simultaneously in both vertical and horizontal dimensions. For this study, it is aircraft that simultaneously penetrate the horizontal separation distance of 30 nautical miles and vertical separation of 4000 feet. The min-max-ratio is defined by the following Eq. 12.

$$\rho = \min_i^k \left[\max\left(\lambda_i, \pi_i\right)\right] \qquad \textbf{Eq. 12}$$

Where

i = current i^{th} radar track position report

k = total number of track points

λ_i = horizontal separation ratio of ith radar track position report

π_i = vertical separation ratio of ith radar track position report

The λ_i and π_i sub-ratios are expressed in the following two equations. First for horizontal separation Eq. 13 expresses the λ_i as the instantaneous ratio of horizontal separation versus the horizontal separation standard. Next, vertical separation is illustrated in Eq. 14, which is the vertical separation distance versus the vertical separation standard. For this study, the horizontal separation standard is 5 nautical miles and vertical is 1000 feet.

$$\lambda_i = \frac{\sqrt{\left(\left(x_i^a - x_i^b\right)^2 + \left(y_i^a - y_i^b\right)^2\right)}}{\delta_i} \qquad \textbf{Eq. 13}$$

Where

δ_i = horizontal separation standard for the ith synchronized track point

x_i^a = x position of the ith track point of aircraft a in nautical miles

x_i^b = x position of the ith track point of aircraft b in nautical miles

and y_i^a , y_i^b are the corresponding y positions

$$\pi_i = \frac{\left|z_i^a - z_i^b\right|}{\upsilon_i} \qquad \textbf{Eq. 14}$$

Where

υ_i = vertical separation standard for the ith synchronized track point

z_i^a = altitude position of the ith track point of aircraft a in feet

z_i^b = altitude position of the ith track point of aircraft b in feet

The maximum of these two ratios is calculated for each synchronized track position and minimum of this value for the entire overlapping flight times is expressed in Eq. 12.

The sharpness metric is calculated by measuring the min-max-ratio distance from the fitted alert probability (quartic fit) from 0.99 to 0.1. The smaller the measured sharpness metric the more sensitive the CP is as a function of separation distance. In other words if the sharpness is smaller, the probe is more sensitive to detecting aircraft-to-aircraft conflicts with smaller separations than a probe that has a larger sharpness.

The sharpness bias metric is the distance (positive/negative) that the alert probability of 0.5 is from a min-max-ratio of one (at target separation, e.g. legal separation 5nm/1000ft). A CP could be perfect in terms of sharpness with a zero value, but if the sharpness bias is 1.0 it will have 100% false alerts (see Fig. 7 in [Paglione, 1999]).

An excellent way to illustrate the mechanics of sharpness is by example. For illustrative purposes three runs were compared, including runs 3, 7, and 11 from the FA18 Interim Report 2. It compares three runs of different adherence bounds including 2.5, 3.5, and 1.5 nautical miles (nm), respectively. Figure 5 illustrates the three alert probability curves as a function of the min-max-ratio metric and Table 7 summarizes the metric results. Therefore, the 1.5 nm adherence bound run produces the best results in terms of the smallest sharpness metric and smallest sharpness bias. The 3.5 nm probed adherence box produces the next best performance, while 2.5 nm run produces the least performance.

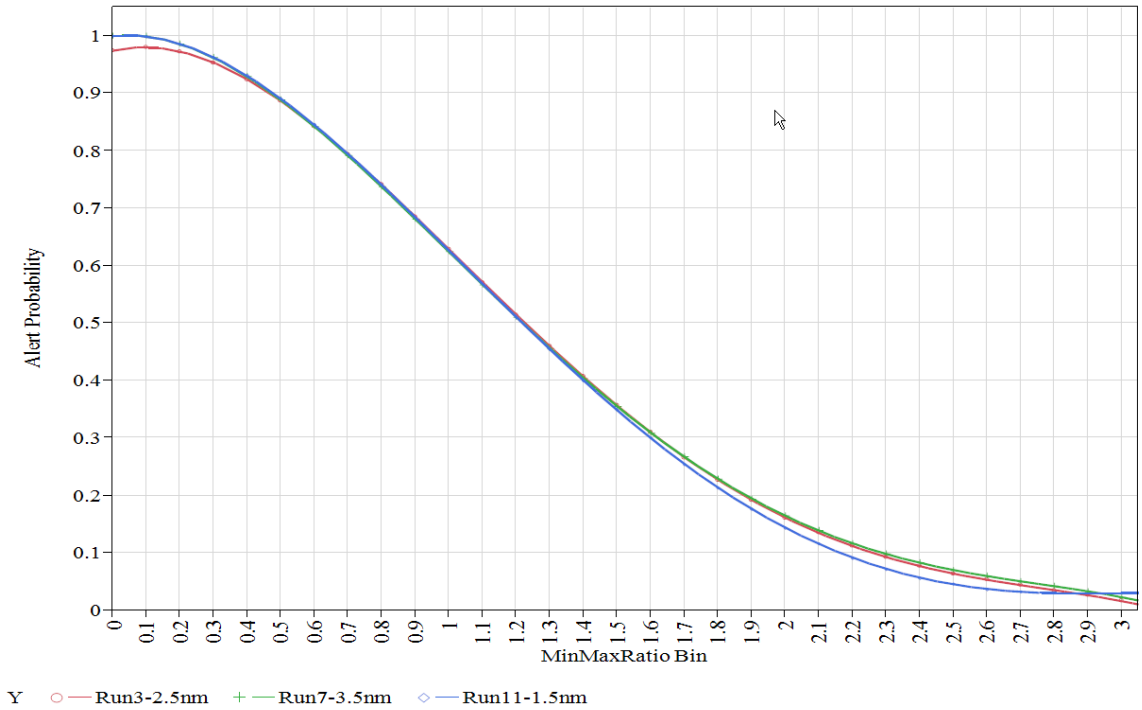


Figure 5. Example Alert Probability Illustrating Sharpness and Sharpness Bias

From the sharpness metrics listed in
Table 7, even though the best results are as expected with run 1.5 nm, the results also indicate that the 3.5 nm adherence bound run outperforms the 2.5 nm run. This occurs because one missed alert reduces the 2.5 nm curve to a max alert probability of 0.98 and thus calculation of sharpness defaults to 2.5's max (never reaches 0.99) and thus increases the sharpness curve. This is reasonable and makes sense. An effect to the left of 1.0 min-max-ratio has a larger effect on the sharpness metric (missed/valid alert impacts) versus the area to the right of 1.0 (false alert impacts) which is stretched out to 3.0 min-max-ratio (has double the relative distance). Finally, it is not clear why missed alerts are being generated under the 2.5 nm run and correctly detected for the 1.5 and 3.5 runs. This warrants further investigation. However, the example illustrates how sharpness and sharpness bias are calculated, and how they model the sensitivity and bias of the conflict probe's prediction process.

Table 7. Example of Sharpness Metric Summary

Adherence Bound Distance	Sharpness Metric	Sharpness Bias Metric
1.5	1.98528	0.21912
2.5	2.16854	0.22647
3.5	2.11261	0.21996

There are three main metrics used throughout this technical note that are a variation of those described in the paragraphs above. These are labeled in the figures as FA%, LA%, and WT%. These metrics correspond to the False Alert, Late Alert, and Warning Time metrics discussed previously, but normalizes them by the results of the baseline for easier comparison of the performance among different runs. These three metrics are summarized in Equations 15, 16, and 17.

$$FA\% = \left(\frac{FA_R - FA_{BL}}{FA_{BL}} \right) \cdot 100 \qquad \textbf{Eq. 15}$$

where,

FA_R is the number of False Alerts in the treatment run.
FA_{BL} is the number of False Alerts in the baseline run.

$$LA\% = \left(\frac{LA_R - LA_{BL}}{LA_{BL}} \right) \cdot 100 \qquad \textbf{Eq. 16}$$

where,

LA_R is the total number of Missed and Late Alerts in the treatment run.
LA_{BL} is the total number of Missed and Late Alerts in the baseline run.

$$WT\% = \left(\frac{WT_R - WT_{BL}}{WT_{BL}} \right) \cdot 100 \qquad \textbf{Eq. 17}$$

where,

WT_R is the 25th percentile of Warning Time in the treatment run.
WT_{BL} is the 25th percentile of Warning Time in the baseline run.

2.1.2.2 Conflict Prediction "Fuzzy" Metrics

The fuzzy metrics offer a way to re-evaluate the classifications made by the SAE by accounting for uncertainties when classifying alerts. This uncertainty allows for any event, being an alert posting, encounter with no alerts, conflict with no alerts, or early alert deletion, to represent several event categories. For example, an encounter with no alerts would traditionally be counted as a no-call, but one would consider a well-separated encounter to be better than one that came close to minimum separation and generated no alert. The fuzzy metrics attempt to represent this subjective nature. A well-separated encounter will score very high for a no-call and next to nothing in the other categories, but as the encounter gets closer to being a conflict the score for no-call will decrease as the score for a late alert increases. A score of 0.2 for no-call and 0.8 for late alert would be a possible score for such an encounter, encoding more information than simply being a no-call event. Here the situation may have warranted attention since the no-call event in the new metric scored much closer to being a late alert.

The fuzzy metric is an application of the fuzzy signal detection theory discussed in [Parasuraman, 2000]. Given that the metric here is used to rank conflict prediction performance, the natural source of signal is the true aircraft-pair geometry during the duration of the event. This gives a ranking from 0 (certainly not conflict-like) to 1 (certainly conflict-like) of how conflict-like the situation actually became while the event occurred. The response is a measure of perceived utility of the event from 0 (no perceived utility) to 1 (perceived to be of the utmost importance). The signal and response are designed to scale to each other, so that low-signal merits low-response and vice versa. After the involved task of finding appropriate functions for signal and response, they are easily combined to produce scores for the four classic event ratings:

True Positive = Valid Alert	VA = min (r, s)
False Negative = Late Alert	LA = max (s-r, 0)
False Positive = False Alert	FA = max (r-s, 0)
True Negative = No-Call	NC = min (1-r, 1-s)

In fact, the classic system can be derived for these equations given that the response and signal are strictly binary:

True Positive = Valid Alert	VA = min (r, s)	= 1, $r = s = 1$
False Negative = Late Alert	LA = max (s-r, 0)	= 1, $r = 0$ and $s = 1$
False Positive = False Alert	FA = max (r-s, 0)	= 1, $r = 1$ and $s = 0$
True Negative = No-Call	NC = min (1-r, 1-s)	= 1, $r = s = 0$

The metric's computation is described in detail in the next several paragraphs.

Given the collection of alerts generated by the conflict prediction process, if they have no associated flight pair, a track is missing during the duration of an alert, or the entirety of a track occurred within an Automated Problem Detection Inhibited Area (APDIA) during this time, the signal is zero. At any time there is missing data or a point is inside an APDIA, that data is excluded. If any point occurred within an APDIA, a flag is set to alert that at least one of the flights entered an APDIA for at least a moment while the signal was computed. This flag informs an analyst that some care is needed in determining performance for these events. Also, if an alert was originally discarded, was not late, and not a no-call, it is re-included as a false alert.
If the alert is a false alert (whether originally false or a re-included discard), a time buffer is computed after the alert duration. This buffer t_b is the minimum of some buffer time t and the time between the end of the current alert and the beginning of the next alert for a flight pair. If this alert is the last, $t_b = t$. The min-max ratio for the flight pair is calculated for the duration of the event and the duration of the buffer. To calculate the min-max ratio, the maximum of the horizontal and vertical separations in ratio to their minimum separation standards are measured, then the minimum of these values over the duration of the event is returned. This value indicates whether a flight pair violated minimum separation standards by being less than unity for in conflict, unity for at minimum separation, and greater than unity for not conflicted. For example, a value of 2 would represent aircraft separated in one of their separation standards by twice the minimum whereas 0.5 would represent aircraft separated by half of their separation standards.

Given that the min-max ratio is smaller inside the alert or $t_b = 0$, it is assumed that the alert was not an early-deletion event. The signal is computed for some time window about the min-max ratio, and allowed to extend outside the bounds of the event duration. The signal for the buffer region s_b is also calculated. *early_del* = $1 - s_b$ is reported only for this type of event. *early_del* gives an indication of the stability of the alert when it was removed. A high *early_del* indicates to an analyst that the end time was very stable, as very little signal actually occurred for some time after the end time. A low *early_del* indicates to an analyst that the end time was unstable. There was a high signal after the alert was removed. *early_del* = 1 when $t_b = 0$ since another action was immediately taken.

If the min-max ratio is smaller in the buffer zone ($t_b > 0$), it is assumed that the alert was an early-deletion event. The signal is computed for some time window about the min-max ratio, and allowed to extend outside the bounds of the event duration. A second event is created to account for the alert being removed even though the signal increases past the deletion time. This event is marked as an early-deletion miss event and the signal for the buffer region is set to its signal.

All other alerts require no special treatment. They and all encounters with no associated alerts have a signal computed around their min-max ratio times.

To get a value for signal around the min-max ratio time, the following values are computed for an aircraft pair:

- *hmin* = minimum horizontal separation standard
- *vmin* = minimum vertical separation standard
- *h* = horizontal separation at the min-max ratio time
- *v* = signed vertical separation from cleaned vertical profile at the min-max ratio time, where negative denotes object below subject and vice versa
- $\overline{\varphi}$ = average encounter angle about the min-max ratio time
- σ_{φ} = standard deviation of the encounter angle about the min-max ratio time
- $\overline{w}$ = average horizontal geometry weighting about the min-max ratio time, discussed in detail later
- $\overline{n}$ = average number of incidents about the min-max ratio time in which the object had a relative angle to the subject within 15 degrees of 0 or 360
- $\overline{vcr}$ = average vertical closure rate about the min-max ratio time
- $\overline{hcr}$ = average horizontal closure rate before the min-max ratio time
- *clr_sep* = flag stating whether the aircraft are both within 300ft of their clearances and are separated by at least the minimum vertical separation standard at min-max ratio time

If *clr_sep* is true, the aircraft are considered well-separated and the signal is zero, or else the vertical component of signal is calculated as follows.

$$f_V = \begin{cases} 1 & v < vmin \\ f_{vict} & otherwise \end{cases} \qquad \textbf{Eq.18}$$

$$f_{vict} = \begin{cases} 0 & vict \le -0.5 \\ spline_A & -0.5 < vict < 2 \\ 1 & otherwise \end{cases} \qquad \text{[Figure 6]} \qquad \textbf{Eq.19}$$

$$vict = \frac{\overline{vcr}}{|v|} \qquad \textbf{Eq.20}$$

vict is the vertical inverse closure time, which is the inverse of the time elapsed before the aircraft would meet vertically if continuing at the same vertical closure rate. Negative values indicate divergence. Large values are desirable for closure time meaning small values are desirable in the inverse. Since the function of *vict* needs to be close to zero for desirable values the inverse of the vertical closure time is used. The same argument is made for *hict*, the horizontal inverse closure time used later.

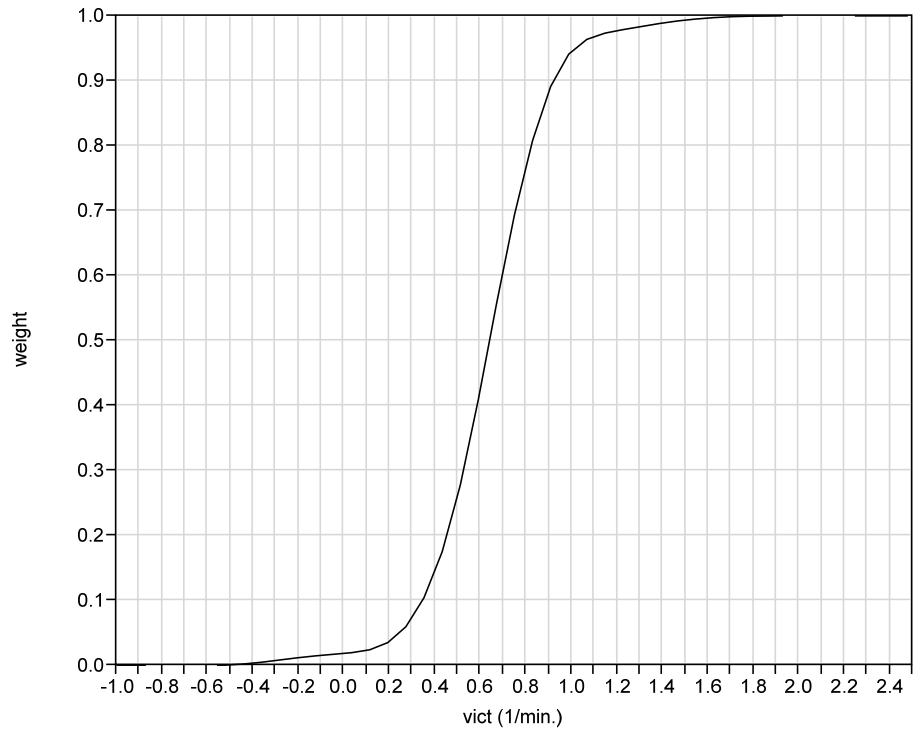


Figure 6. Function of vertical inverse closure time.

If $\overline{n} \geq 0.7$, $\sigma_{\varphi} < 45$, and $\overline{\varphi} < 15$, then the subject and object were consistently in-line with a persistent and similar heading. The flights are categorized as in-trail and the horizontal component of the signal comprises horizontal separation and closure rate only. If this is not true, the horizontal component of the signal comprises horizontal separation and average encounter angle weighting.

$$f_H = \begin{cases} f_{hict} f_h & intrail \\ \overline{w} f_h & otherwise \end{cases} \qquad \textbf{Eq.21}$$

$$f_h = \begin{cases} 0 & h \geq hmin + 4 \\ spline_B & hmin - 1 < h < hmin + 4 \\ 1 & otherwise \end{cases} \quad \text{[Figure 7]} \qquad \textbf{Eq.22}$$

$$f_{hict} = \begin{cases} 0 & hict \leq -0.5 \\ spline_C & -0.5 < hict < 0.7 \\ 1 & otherwise \end{cases} \quad \text{[Figure 8]} \qquad \textbf{Eq.23}$$

$$hict = \frac{\overline{hcr}}{h - hmin} \qquad \textbf{Eq.24}$$

See the earlier definition for *vict*, the vertical inverse closure time, to understand the rationale for *hict*, the horizontal inverse closure time.

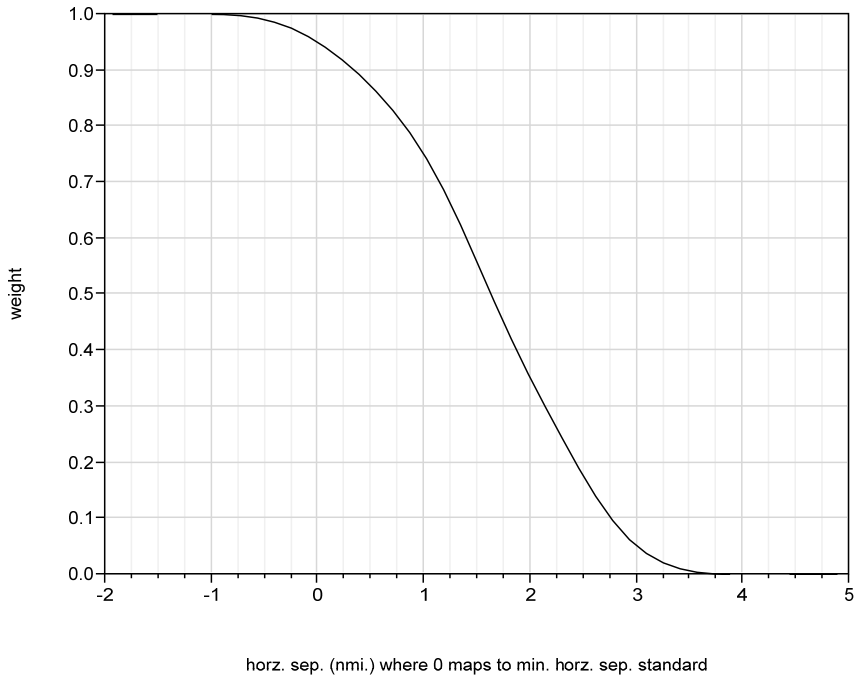


Figure 7. Function of horizontal separation re-centered so that zero represents the minimum horizontal separation standard.

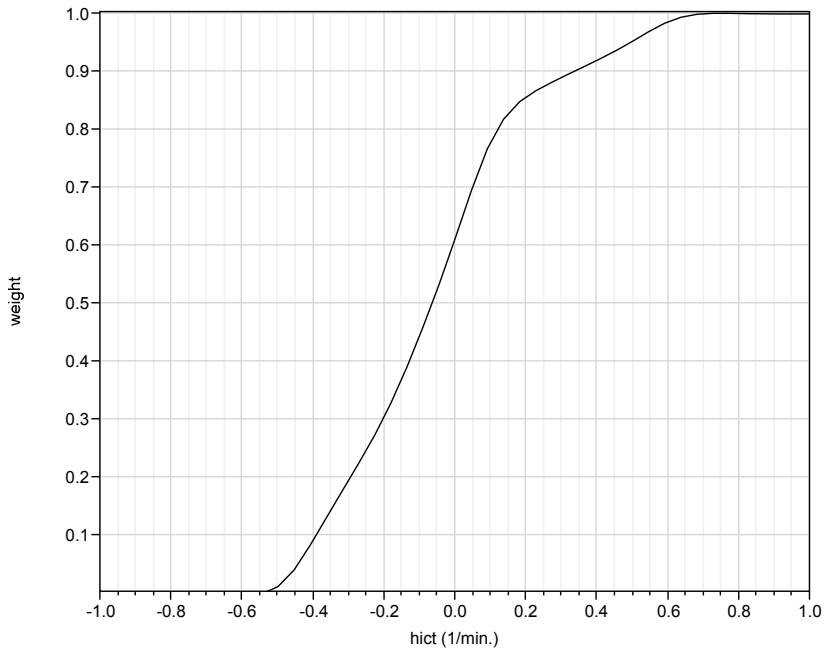


Figure 8. Function of horizontal inverse closure time.

$$w = 1 - \frac{1}{2}\left(1 + \frac{\alpha(\cos\varphi' - 1) + \beta\sin\varphi'}{2\sqrt{\alpha^2 + \beta^2}}\right)e^{-hmin^2/(\alpha^2+\beta^2)} \qquad \textbf{Eq.25}$$

Where:

$\varphi' =$ difference in object's heading from subject's heading
$\alpha =$ longitudinal distance to object from subject
$\beta =$ lateral distance to object from subject

See Appendix A for additional information on the geometric weighting function *w*.

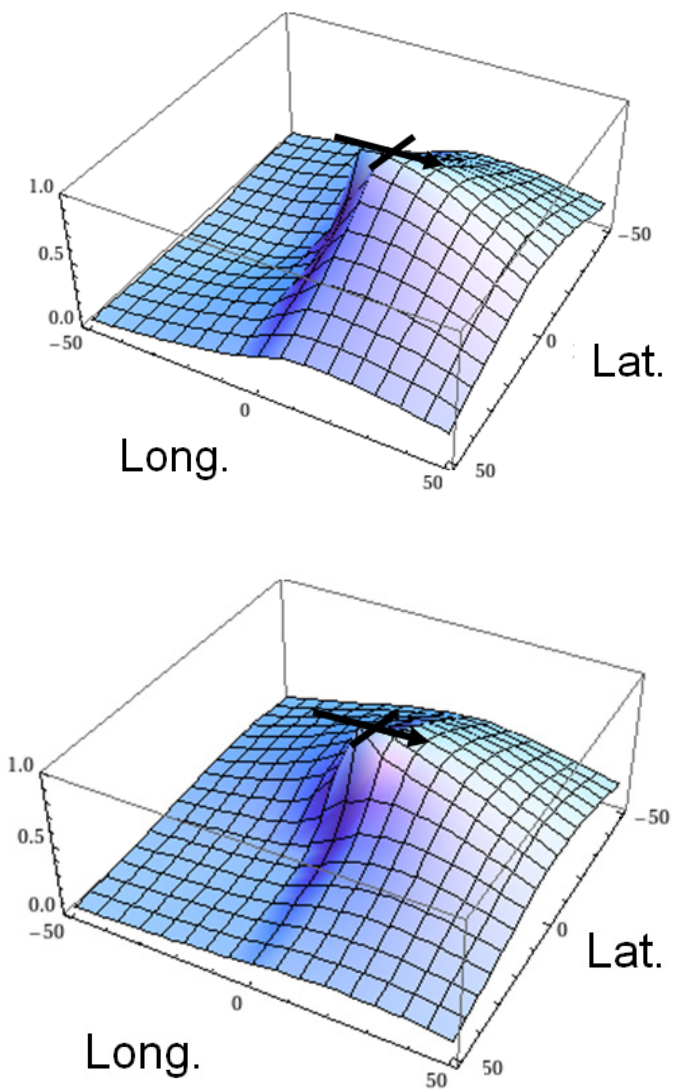


Figure 9. Top: Gaussian distribution altered by *Head-on* horz. geometry weighting. Bot: Same altered by *Left-To-Right Perp. Crossing* horz. geometry weighting instead.

In Figure 9, the signal is represented by a Gaussian distribution for illustrative purposes. Let the arrow in the middle indicate the subject aircraft. The longitudinal direction and lateral direction are labeled where negative indicates behind and left respectively. Choosing a position on the grid, imagine the object aircraft is located there. In the top situation, the object is facing the completely opposite direction as the subject. In the bottom situation, the object is facing perpendicular to the subject traveling toward the right. In the top situation of Figure 9, if the object is behind the subject, they are diverging and the signal is sharply decreased. In the bottom situation, if the object is already to the left of the subject, they are diverging, more so if the object is also behind the subject, which is why the signal is most impacted in the (-long., +lat.) region. The head-on situation is symmetric on the lateral axis since a parallel case is identical on either side of the subject. The highest signal is when the object is directly in front of the subject, the highest chance of a conflict occurring. The crossing case has largest signal to the left and in front of the subject since the aircraft will be most rapidly converging in this case if they stay to course.

The splines are created using the Fritsch-Carlson method to ensure weak monotonicity. The ends of the splines are required to be flat so that they smoothly connect to the areas where the functions are constantly either zero or one. The splines are defined by their control points and can be seen in Figure 6, Figure 7, and Figure 8.

$spline_A$:

$(-0.50, 0.00),\quad (-0.25, 0.01),\quad (0.00, 0.02),\quad (0.25, 0.05),$

$(0.50, 0.25),\quad (0.75, 0.70),\quad (1.00, 0.95),\quad (2.00, 1.00)$

$spline_B$:

$(hmin-1, 1.00),\quad (hmin, 0.95),\quad (hmin+1, 0.75)$

$(hmin+2, 0.35),\quad (hmin+3, 0.05),\quad (hmin+4, 0.00)$

$spline_C$:

$(-0.50, 0.00),\quad (-0.40, 0.10),\quad (-0.30, 0.20),\quad (-0.20, 0.30),\quad (-0.10, 0.45),$

$(0.00, 0.60),\quad (0.10, 0.80),\quad (0.20, 0.85),\quad (0.30, 0.89),$

$(0.40, 0.92),\quad (0.50, 0.95),\quad (0.60, 0.99),\quad (0.70, 1.00)$

Finally, the signal is found to be the product of the horizontal and vertical components.

$$s = f_H f_V \qquad \textbf{Eq.26}$$

As for the response, if the event was classified as an early-deletion miss event, the response is automatically zero, since this is a no-action event. Also, if the encounter has no associated alerts, no action was taken so the response is also zero in this case. The actual warning time is the first part of the response. It is used to devalue alerts that are presented after the minimum warning time, which are late alerts and no-calls (since the alert was never made). If the alert was an FA, including the discards now marked as FAs, then the actual warning time is not computable, though this part of the response is set to 1 since there is no computable warning time to devalue the event. The warning time for computable events is the actual conflict start time minus the event start time. If for some reason the warning time is less than zero or the event is a no-call then the warning time is set to zero, having the effect of setting this portion of the response to zero. After this, the warning time is divided by the minimum warning time standard to get a unitless

value t. If $t < 1$, the warning time did not meet the standard, and this part of the response takes the value $r_1 = t^2$.

The second part of the response is based on predicted horizontal separation of the flights (Figure 10). The predicted horizontal separation is divided by the minimum horizontal separation standard to get a unitless ratio q. If this value is less than 1, the perceived utility of the response is high and this part of the response takes the value of 1. If q is at least 13, then the perceived utility is very low and this part of the response takes the value of 0. The values in-between are computed by a spline of the same type as in the signal. This spline is defined by its control points as follows:

$spline_D$:

(1.00, 1.00), (1.40, 0.85), (2.00, 0.25), (3.00, 0.05), (13.00, 0.00)

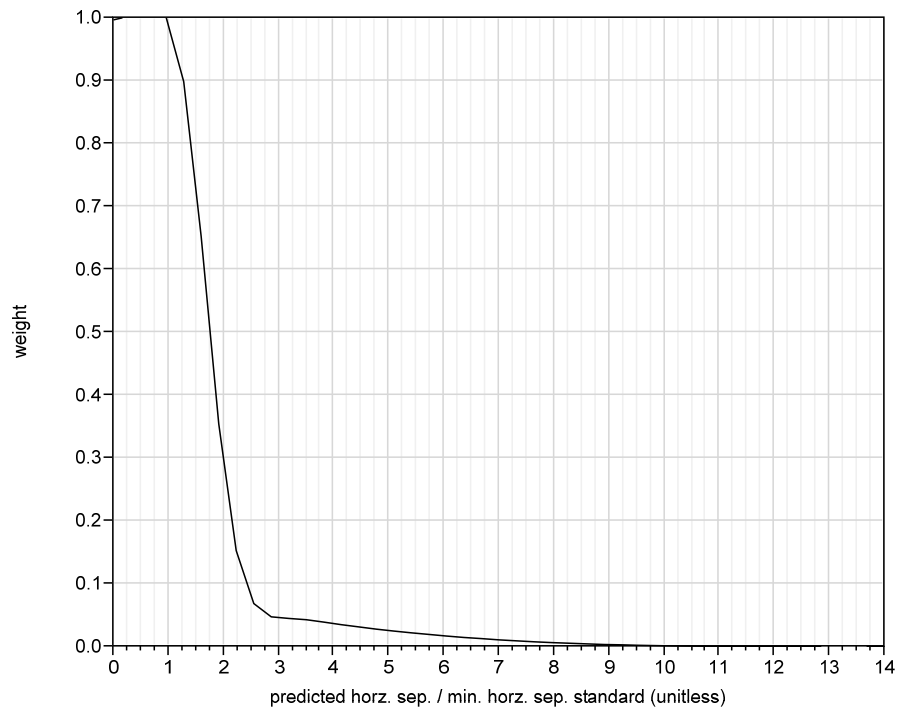


Figure 10. Function of predicted horizontal separation.

The final value of the response is

$$r = r_1 \, r_2 \qquad \textbf{Eq.27}$$

2.2 Description of the Experimental Plan

One of the most powerful inferential statistical approaches is the design, implementation, and synthesis of experiments. Experiments are performed by most researchers and scientists in practically all disciplines. An input stimulus is entered into a process with a set of controllable factors. These processes were already illustrated in the Figure 2 and Figure 3, which contrasted the black and white box testing approaches. For both processes and for application of both the

crisp and fuzzy metrics, these are the factors or independent variables in the experiment are manipulated to study the output or response variables. The uncontrollable factors are not easily manipulated, but through experimental design techniques such as blocking and randomization can be removed from the experiment. The output response variables are the dependent variables of the experiment. They are often determined by application of a metric or measured by a sensor device.

Table 8. Processing Steps for the Experimental Analysis

Step	Description	Section
1 – Problem Definition	Define the problem statement	2.2.1
2 – Design of Experiment	Design the experiment – The factors, levels of the factors, response variables to be run, and the model to be used for analysis are defined.	2.2.2
3 – Execute Experiment	Execute the experiment and prepare output data – The system is configured for the experimental runs defined by the design, runs executed, and resulting output data is processed for input into model	3.2
4 – Implement Model	Implement statistical model defined by the experiment.	3.2.2.2.1
5 – Model Results	Examine the results of the model and discuss factor effects	3.2.2.2.2 and 3.2.2.2.3
6 – Synthesize Impact	Synthesize overall results from the model and publish conclusions.	4

There are many purposes for performing an experiment. For this study, the objective of designing and executing an experiment is to determine (1) which pre-determined factors and interactions of these factors show a statistically significant effect on the ERAM system's performance, and (2) the relative sizes of the determined significant effects. From designing the experiment to concluding on its results, a series of processing steps should be performed as identified in Table 8. The first two steps presented in Table 8 are described in this section, which documents the plan for the experimental analysis. The last four steps are described in Section 3.2 and Section 4, which presents the results by documenting the actual execution and analysis of the experiment.

This initial integrated experiment is the main focus of this study. Various runs at different settings allowed a model to be developed that can estimate performance of the conflict probe at any settings between those chosen for the experimental levels. This will be discussed in detail in the subsequent Section 2.2.2.

In the introduction to Section 2 and listed in Table 1 (item #1), an enhanced trajectory modeling feature was listed that improved the initial point and lateral rejoin logic. This prototype activity is organized under Functional Area 32 (FA32) Trajectory Modeling (TM) Enhancements. In Section 3.1, this prototype will be compared against the initial-baseline (IBL) ERAM run with no enhancements and parameter levels set to operational levels. For the remainder of this study, this FA32 scenario, referred to as baseline run (BL), with parameters equal to IBL settings is treated as the baseline scenario against which the scenarios with varying parameter settings (also referred to as treatment runs) are compared. This experimental plan applies only to the second task of the two presented in this paper. Thus, the first task compares the FA32 prototype run to the IBL ERAM run and subject of the analysis presented in Section 3.1. The second task and main focus of this paper uses the FA32 run as its baseline comparing it to the various parameter treatment runs, which are presented in the next sub-sections.

2.2.1 Definition of the Problem Statement

As stated in the introduction to Section 2, the objective of this study is to examine prototype enhancements and parameter changes to ERAM to recommend these changes for implementation in the operational ERAM trajectory modeling (TM) and CPT sub-systems. The motivation is to provide air traffic control users a conflict probe functionality on the radar display and without the improvements envisioned the current instantiation will not be suitable. Thus, this study documents an initial experiment. Several other experiments are either implemented or planned for the near future that expands upon this work. This initial experiment is limited to the parameter changes of the TM and CD adherence bounds and settings of the conflict likelihood function. The following problem statement captures the objective of this initial experiment.

> *Through a set of purposeful runs of ERAM, input with the legacy ZDC time-shifted test traffic scenario, the experiment shall determine the statistically significant impact that the TM/CD adherence bound changes longitudinally and laterally, and likelihood function parameter changes has in terms of trajectory and conflict prediction accuracy performance.*

2.2.2 Design of the Experiment

In order to develop the model without running a full factorial, which would have been 27 runs, a fractional design of experiment (DOE) was done using the JMP [SAS, 2010, B] software tool, which resulted in an experiment with only 12 runs.

2.2.2.1 Factors

The factors used in the experiment included only settings of ERAM that can be changed in the current version without any prototype upgrades or code changes. The factors of this particular experiment are simply parameters within ERAM's configuration files and do not require any new software enhancements.

The lateral and longitudinal bounds of the conformance box were varied independently from each other. This variance did not include the prototype changes in FA18 Interim 2 [Crowell, 2011, C] that decoupled the TM bounds from the CP bounds. Instead all changes to the conformance bounds affected both the TM and CP bounds. Both of these factors are continuous factors, modeled using a quadratic equation. The ranges of two continuous factors are listed in Table 9.

Table 9. Continuous factors of the Integrated Experiment

Factor	Min	Max
Lateral Bound	0.5 nm	2.5 nm
Longitudinal Bound	0.5 nm	1.5 nm

In some initial experiments, likelihood appeared to be a significant factor. In order to further understand the effects of likelihood, it was varied among three discrete values. Effects of likelihood cannot easily be modeled as a continuous function because the likelihood parameter is a function in itself. The functions used for likelihood contain either two or three parameters. When two parameters are used, the first one is the maximum time in minutes at which a likelihood value of 0.0 will generate an alert. The second parameter is the minimum at which a likelihood value of 1.0 is required in order to generate an alert. This creates a linear function similar to that shown in Figure 11. The white area above the line is where the likelihood must fall in order for an alert to be generated. When three parameters are used, it becomes a piecewise linear function, with the first parameter being the maximum time at which a likelihood value of 0.0 will generate an alert. The last parameter is the minimum time in minutes at which a

likelihood value of 1.0 is required in order to generate an alert, and the center parameter is the time in minutes at which a chosen value is required in order to generate an alert. In the case of the 3-parameter settings used in this experiment, this chosen value is set to 0.8. This results in a function like the one shown in Figure 12. The settings used for likelihood are shown in Table 10.

Table 10. Nominal factors of the Integrated Experiment

Factor	Settings		
Likelihood Function	10/20	3/8/20	3/8/10

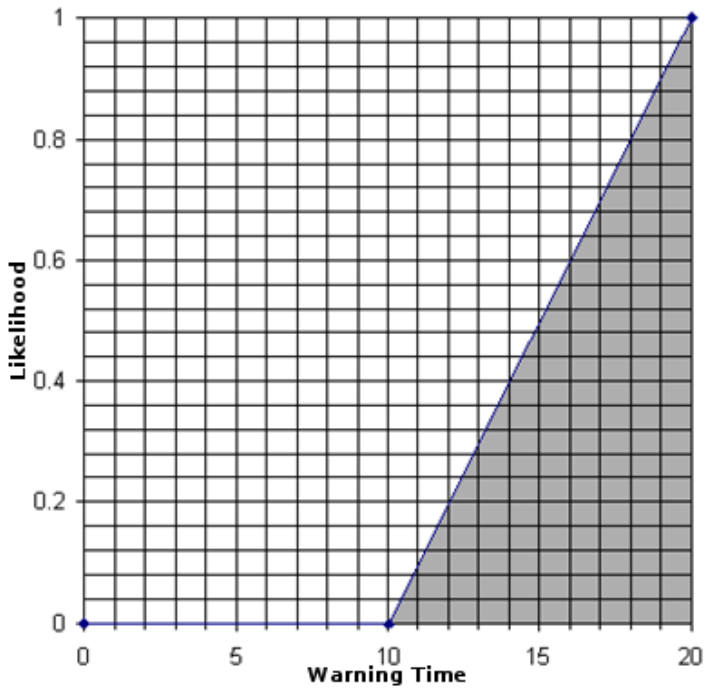


Figure 11. Linear likelihood function for 10/20 setting

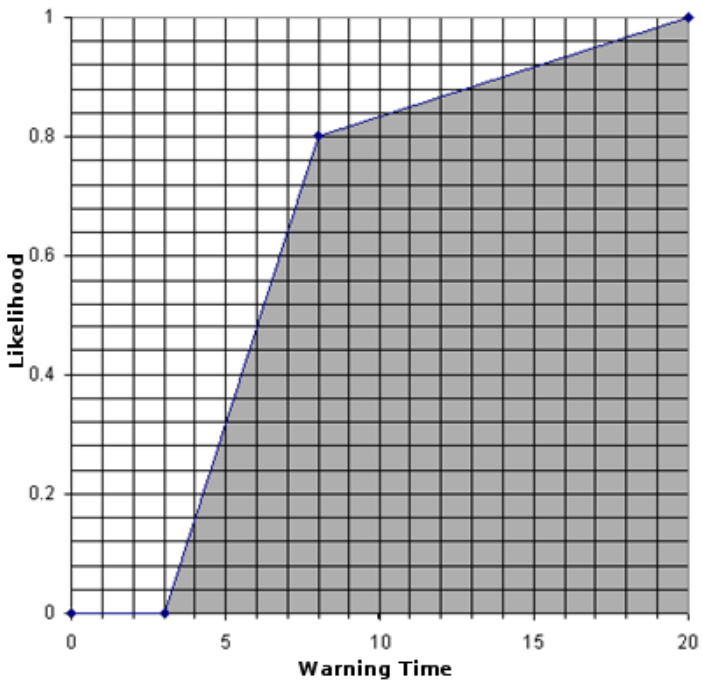


Figure 12. Piecewise linear likelihood function for 3/8/20 setting

These settings resulted in the 12 runs shown in Table 11. Also shown in this table are the settings used currently in the deployed version of ERAM. This run is referred to as the baseline (BL) run or in some cases the FA32 Baseline (32BL).

Table 11. Twelve runs for the integrated experiment and the baseline run

Run	Lateral	Longitudinal	Likelihood
1	0.5	0.5	10/20
2	2.5	1.5	3/8/10
3	0.5	1.5	10/20
4	1.5	0.5	3/8/20
5	1.5	1.0	10/20
6	0.5	1.5	3/8/10
7	1.5	1.5	3/8/20
8	2.5	0.5	10/20
9	2.5	1.0	3/8/20
10	2.5	0.5	3/8/10
11	0.5	0.5	3/8/10
12	0.5	1.0	3/8/20
BL	2.5	1.5	10/20

2.2.2.2 Model

The initial model allowed both continuous factors to have at most a quadratic effect. It was assumed all factors could interact with each other only in pairs (two-way interactions only). The constant or overall mean effect is represented in the model as μ, and $\varepsilon_{n(ijk)}$ represents the assumption of independently normally distributed random error with a zero mean. All factors are assumed additive. The model is defined as in Eq.28.

Response:

$$R_{ijk} = \mu + Lat_i^2 + Lat_i + Long_j^2 + Long_j + Lat_i Long_j + Like_k + Lat_i Like_k + Long_i Like_k + \varepsilon_{n(ijk)} \quad \textbf{Eq.28}$$

Where:

Lat_i = lateral conformance bounds in nautical miles, i = 0.5, 1.5, 2.5
$Long_j$ = longitudinal conformance bounds in nautical miles, j = 0.5, 1.0, 1.5
$Like_k$ = likelihood, k = "10/20", "3/8/10", "3/8/20"
$\varepsilon_{n(ijk)}$ = random error, n = 1, 2, … for all i, j, k

There are five response variables, all of which are addressed by this model. R_{ijk} will represent the percentage improvement in traditional and fuzzy false alert counts, traditional and fuzzy late alert counts, or the 25th percentile of warning time from the baseline depending on context.

3 Performance Evaluation

There were two major evaluation tasks of this study. The first task was the analysis of Function Area 32 (FA32) Trajectory Modeling (TM) Enhancements. This analysis was presented in [Young, 2011] and compares the performance of the IBL scenario to a scenario including the trajectory modeling enhancements defined in FA32. This comparison was done independent of and prior to other analyses because all of the following runs include the enhancements.

The next task was an integrated experiment that uses 12 runs of ERAM at different CP and TM settings from which a model is built that allows estimations to be performed within a continuous range of CP and TM settings. The factors defined for this experiment are parameters in the currently deployed version of ERAM whose settings can be adjusted without requiring modifications to the system.

In all the analyses included in this document, the performance evaluations are performed on a test scenario and should not be taken out of context. The scenario used for the evaluation has had conflicts injected into it intentionally, some of which are particularly difficult for ERAM to detect. The numbers presented in this document are for purposes of comparison to other numbers within this document only. The purpose of this analysis is not to determine if the CP meets some numerical performance requirement, but rather to determine if an improvement can be gained when compared to the same scenario run through the CP with different settings.

3.1 Trajectory Modeling Enhancements

This chapter provides the analytical results from the evaluation of algorithmic enhancements to the aircraft trajectory modeling that has been implemented in the En Route Automation Modernization (ERAM) prototyping effort that was presented in [Young, 2011]. The algorithmic details of the planned prototype effort can be found in the Lockheed Martin report [McKay, 2011] delivered as part of the Separation Management Task Order 51 activity for Functional Area 32 (FA32) –trajectory modeling improvement.

In earlier ERAM releases, trajectory modeling only begins at the track position on a lateral re-adherence of the trajectory, or change of track control. Otherwise, the initial point is taken to be the track projection onto the previous trajectory. On some trajectory rebuilds, the trajectory initial point could be up to 2.5nm from the latest track reported position.

This prototype effort changes the aircraft trajectory algorithm to always start at the latest track position, and investigates different lateral rejoin enhancements. It provides for risk reduction, firming up of algorithmic changes, an associated accuracy benefit, and accelerated software development and implementation strategy.

These changes to the trajectory modeling were included in all the treatment runs for this experiment. Since this technical note is concerned only with the effects of altering the parameters of the ERAM Conflict Probe, a new baseline is needed to compare the treatment runs to. The new baseline must be run at the baseline settings, but contain the FA32 trajectory modeling enhancements. This chapter provides a comparison of this new FA32 Baseline (32BL) to the Initial Baseline (IBL) run, to provide a perspective for the treatment run analyses.

3.1.1 Analysis Description

The Initial Baseline (IBL) is detailed in [Crowell, 2011, A] and used to gain an understanding of the performance differences of the FA32 Baseline (32BL). The FA32 update introduces an enhanced method of trajectory modeling, and a new scenario was generated using this method. The purpose of this chapter is to analyze the effect of the trajectory modeling enhancements by comparing the 32BL to the IBL in terms of trajectory accuracy and conflict probe performance. Specifically, the accuracy of the trajectories generated in each scenario is calculated using the simulated track data as a baseline.

FA32 scenario data was provided by LM and run through CPAT tools, using the same process as in the initial baseline scenario. The conformance bounds and likelihood parameter were set to match the values in the IBL run, which were 2.5 nm lateral, 1.5 nm longitudinal and 10/20 likelihood. The only difference between the two runs is the FA32 trajectory modeling enhancement.

3.1.2 Results

The two parts of the analysis results—accuracy of trajectory prediction and performance of conflict probe—are presented in the following two sections respectively.

3.1.2.1 Trajectory Accuracy

Average error values are calculated for the same flight in the IBL and 32BL runs; these values are collected and compared over many flights to illustrate any underlying differences between the runs. A matched pair analysis is performed for each of the four metrics using the average error per flight between the IBL and 32BL with the trajectory modeling enhancements. The analysis includes a paired *t*-test, which examines the distribution of differences in error between the two scenarios and tests if the mean of the differences is statistically different from zero. The results from the paired *t*-tests are provided in Table 12. In each of the four graphs in Figure 13 the difference in average trajectory error per flight from the IBL to the 32BL is plotted against the mean of the two errors.

Appendix C provides a detailed description of the matched pair analysis and graphical output. Positive mean differences indicate the baseline has more error on average than the prototype run.

Table 12. Statistical Results for Trajectory Error.

Error	Mean Diff	Std Error	*p*-value
Horizontal (nm)	0.1153	0.0082	< 0.0001
Abs. Vertical (ft)	2.1481	1.5954	0.1783
Abs. Cross Track (nm)	0.1237	0.0066	< 0.0001
Abs. Along Track (nm)	0.0268	0.0058	< 0.0001

From the results in Table 12, the 32BL has less average error per flight for all four types of error. Each *p*-value presented in Table 12 is the probability of observing a discrepancy in means at least as large as that observed, even if there is no underlying difference in the means. A *p*-value less than 0.05 is typically considered statistically significant. For this analysis the mean difference is statistically significant for horizontal, absolute cross track, and absolute along track errors. For absolute vertical error, the new run has less mean error but the difference is not statistically significant.

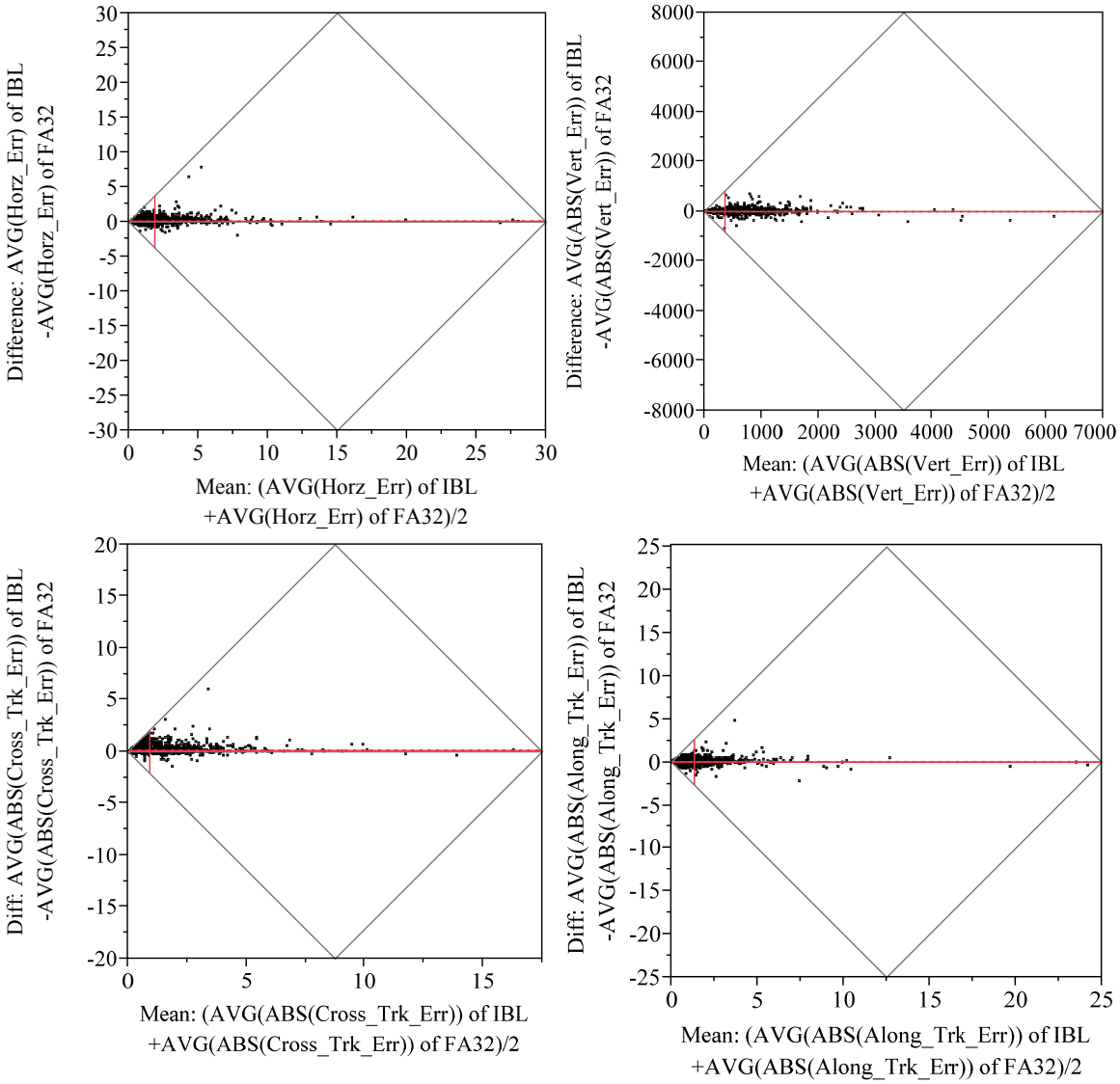


Figure 13. Matched Pair Analyses for Trajectory Error Metrics.

Figure 13 illustrates the same results of the matched pair analysis graphically. It presents the measurements for each metric using a special plot from the commercial statistical software package called JMP®. It plots each paired difference, with the vertical axis being the difference of the flight's average error in the IBL run minus the respective mean in the treatment run. The horizontal axis is the average of these two measurements[††††]. The resulting plot normalizes the error differences indicating net trends between runs. For example, if the errors are predominantly above the red line at zero in the middle of the plot, this indicates the error is larger in the baseline run more often. If the errors are below this line, they indicate the error is larger in the treatment run (in this case the 32BL run). Supporting the same conclusion as drawn from the results of Table 12, Figure 13 illustrates a trend of most error differences being above the zero line, indicating the net benefit of FA32.

[††††] The average of the two metrics used on the horizontal axis of the matched pair plot is calculated by taking the sum of the mean error of the IBL and mean of the FA32 treatment run and dividing this sum by 2 for each matched flight.

It is helpful to know how accurate the trajectories are under certain conditions, such as when the flight is considered to be adhering to its known flight path (including clearances entered into the trajectory predictor automation) within specified parameters[‡‡‡‡]. This is the focus of the second stage of trajectory analysis. The data is filtered to include only data from a sample of points along the trajectory where the flight is in adherence. See [Pankok, 2011] for further details of the parameters determining adherence status. The matched pair analysis is repeated on this filtered data set. The results from the paired *t*-tests are provided in Table 13, while the graphical results are presented in Figure 14.

Table 13. Statistical Results for Trajectory Error of In-Adherence Data.

Error	Mean Diff	Std Error	*p*-value
Horizontal (nm)	0.0899	0.0072	< 0.0001
Abs. Vertical (ft)	5.3179	1.9219	0.0057
Abs. Cross Track (nm)	0.0991	0.0062	< 0.0001
Abs. Along Track (nm)	0.0197	0.0057	0.0005

From the results in Table 13, the FA32 baseline run has less average error per flight for all four types of error when only data from in adherence track points are considered. This difference is statistically significant for all error types: horizontal, absolute vertical, absolute cross track and absolute along track errors. However, the mean differences are less than in Table 12 for overall data, thus the positive impact from the trajectory modeling enhancements on trajectory accuracy is observed in all error types when considering flights that are adhering to the known flight path. Compared to the unfiltered data results, this impact is more consistent for vertical errors but overall is reduced in magnitude. This is expected because the prototype affects re-adherence trajectories for flights out of lateral route adherence, and the filtered data only considers flights in lateral route adherence.

[‡‡‡‡] Adherence for this study is measured at each radar track position. Position reports are flagged as out or in a maximum threshold of 2.5 nautical miles simultaneously returning to route by 30 degrees track course heading and within 300 feet of the flight's vertical clearance during cruise.

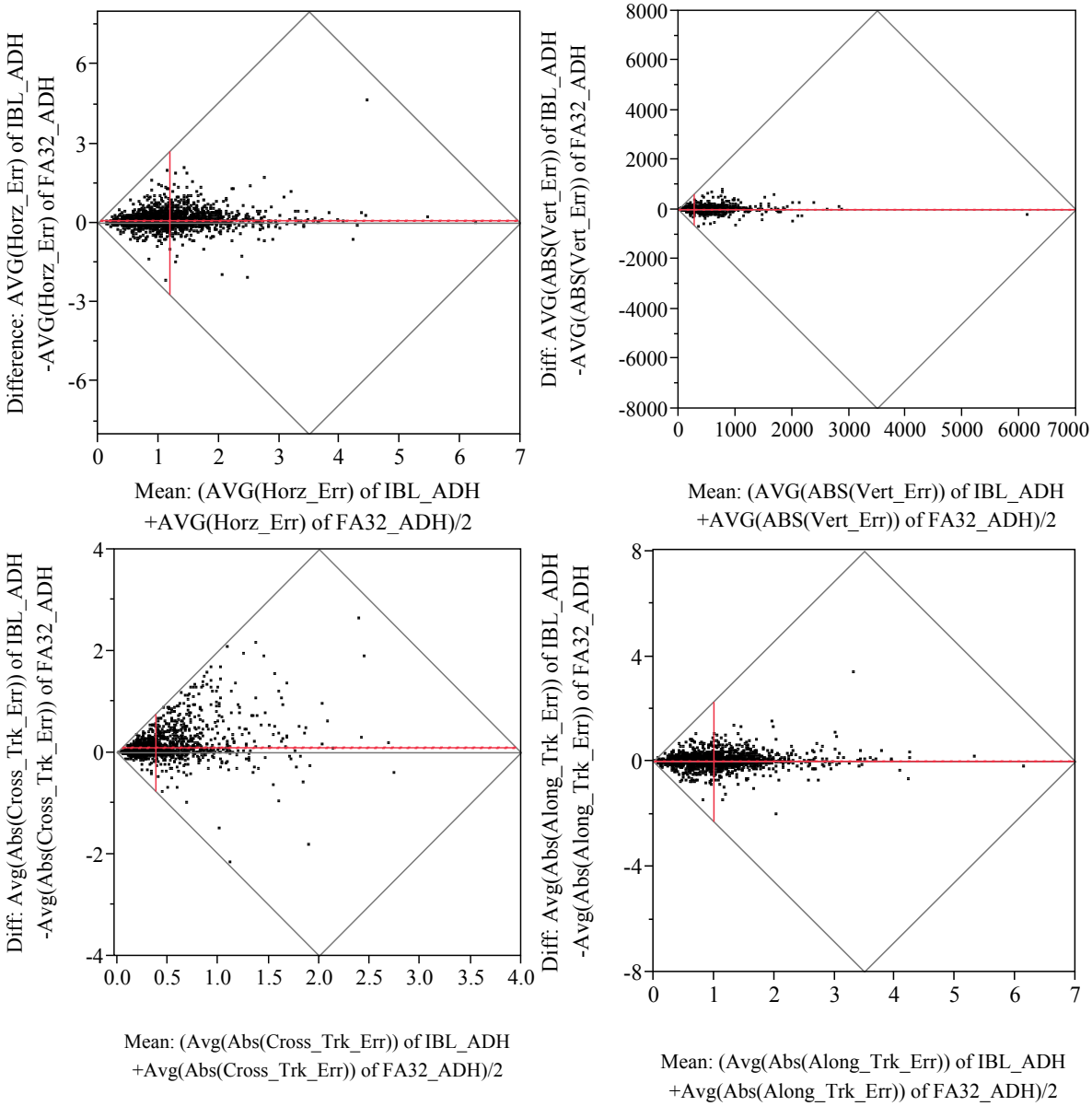


Figure 14. Matched Pair Analyses for Trajectory Error Metrics for In-Adherence Data.

3.1.2.2 Conflict Probe Accuracy

A simulated conflict probe is run on both baselines. Since the underlying ground truth is theoretically the same, individual conflicts and encounters can be matched between the two scenarios and the conflict probe response in each scenario can be compared to demonstrate the impact of FA32 changes. It should be noted that there may be some small differences in track position data due to randomness in the radar signal generation process. These differences may affect the trajectory prediction process and even the designation of conflicts, so much so that a pair of flights designated in-conflict in one scenario may exhibit track positions in the other that are separated just enough to avoid being in conflict.

Unlike the other analyses performed in this study, all alerts generated by ERAM were considered in this conflict probe analysis. No filtering by predicted separation or conflict duration was done,

whereas the other analyses in this study filtered out conflicts with less than 10 second duration. One exception is that this analysis, like the other analyses, does not consider muted alerts, which are alerts for conflicts that are predicted to occur outside of the current cleared altitude. An interim clearance, for example, may be assigned to a flight during a climb to its intended cruise level. Meanwhile, the simulated trajectory prediction engine uses the flight plan clearance as input, which may generate muted alerts. The decision to exclude muted alerts was made based on input from ATC SMEs.

Software applications developed by the Conflict Probe Assessment Team (CPAT) were used to evaluate the results of the conflict probe and compare the performance of the FA32 run to the baseline by matching encounter pairs. The *Strategic Alert Evaluator* and *Strategic Alert Comparer* applications are described in [Paglione, 1999] and [Crowell, 2009] respectively. Table 14 presents the results of the comparison. The shaded rows represent cases where the slight differences in ground truth between the two scenarios caused some events that were near the threshold of being defined as a conflict to be recorded as conflicts in one scenario and not the other. The last three rows represent cases where the conflict probe's performance in two scenarios matched perfectly. The first three rows represent the more interesting cases, and they are discussed here.

Table 14. Conflict Probe Comparison Results.

IBL	FA32	Count
LA	VA	3
FA	N/A	220
N/A	FA	140
LA	N/A	1
N/A	VA	9
VA	N/A	3
FA	FA	1167
LA	LA	26
VA	VA	160

From Table 14, there are three cases where a Late Alert in the IBL changed to a Valid Alert in the 32BL, which demonstrates that the trajectory modeling enhancements had a positive impact on the performance of the conflict probe. These cases will be explored in further detail and the results reported in a final version of this analysis. There are no instances of a Valid Alert in the IBL changing to a Late Alert in the 32BL. Therefore, the improvement in Late Alerts can be stated as the count of Late Alerts that changed to Valid Alerts divided by the number of Late Alerts common to both. This is calculated as LA_VA / LA_LA = 3/26 = 11.5%. These alerts do not comprise a large enough sample to test for statistical significance in this improvement.

There are 220 cases where a False Alert in the IBL changed to a correct No Call in the 32BL, and only 140 cases where the opposite occurred. Decreasing the false alert rate this way is an important objective for trajectory prediction improvements. The improvement in False Alerts can be stated as the count of False Alerts that changed to correct No Calls (FA_NC) minus the count of correct No Calls in the IBL that changed to False Alerts (NC_FA), divided by the number of False Alerts common to both (FA_FA). This is calculated as (FA_NC – NC_FA)/FA_FA = (220-140) / 1167 = 6.85%.

The false alert results were tested for significance using the chi-squared test presented in [Vivona, 2010]. The null hypothesis being tested is that the incidence of false alerts is statistically

equivalent for the two runs. The test calculates the squared difference between the counts of FA_NC and NC_FA and divides by their sum. The test statistic is defined as:

$$\chi^2 = \frac{(FA_NC - NC_FA)^2}{(FA_NC + NC_FA)}$$

and the calculated value is 17.77. With one degree of freedom the corresponding probability is less than 3E-5, so the null hypothesis can be rejected [Agresti, 2002]. This indicates that the reduction in false alerts in the FA32 scenario is statistically significant.

It is useful to know how the conflict probe performs when the adherence of flights to their known flight paths is considered as a factor in categorizing conflicts and alert responses. For more information on how adherence age is used in the evaluation of the conflict probe please refer to [Crowell, 2011, A]. When the length of time in adherence is considered, many more alerts will be categorized as discards. This can be seen in Table 15, which presents the results of the conflict probe on the same two scenarios when considering adherence age in the evaluation of alert responses. Similar to Table 14, the first two rows represent the cases of most interest to this analysis.

Table 15. Adherence Conflict Probe Comparison Results.

IBL	FA32	Count
FA	N/A	59
N/A	FA	45
N/A	VA	9
VA	N/A	3
FA	FA	429
LA	LA	1
VA	VA	160

When adherence is used, there are 59 cases where a False Alert in the IBL changed to a correct No Call in the 32BL, and 45 cases where the opposite happened. The improvement in False Alerts can be stated as above, (FA_NC – NC_FA)/FA_FA = (59-45) / 429 = 3.26%.

Another chi-squared test is formulated to test whether the incidence of false alerts is statistically equivalent for these results. The calculated value of the test statistic is 1.88. With one degree of freedom the corresponding probability is 0.17, so the null hypothesis cannot be rejected. This indicates that the reduction in false alarms in the FA32 scenario cannot be said to be statistically significant. The reason for this may be similar to the reason that the observed improvement in trajectory metric accuracy was less for in-adherence data.

The analysis shows that the FA32 trajectory modeling enhancements did provide a net improvement in both trajectory and conflict prediction accuracy. Most importantly, FA32 prototype was observed to have an overall net improvement in reducing both late and false alerts. Typically it is difficult to reduce both simultaneously; generally one can be reduced with degradation to the other. This was not the case for the FA32 trajectory modeling enhancements. The evidence showed that the false alert rate was reduced while the late alert rate was not degraded.

3.2 Integrated Experiment Analysis

The main experiment of this study was the integrated experiment that is defined in Section 2.2. Trajectory modeling and conflict probe performance were analyzed in detail using several

approaches, including data exploration using discrete results from the treatment runs and analysis of effects using the least squares fit model created with the d-optimal design.

3.2.1 Trajectory Modeling Analysis

This section analyzes the effect of the three defined factors on the trajectory accuracy in the scenario runs. The trajectories generated in each scenario are compared to the simulated track data—which should be very similar but not identical across runs—and three trajectory metrics are calculated as described in Section 2.1.2. The null hypothesis for this trajectory modeling analysis is the following:

> *A significant decrease in trajectory modeling errors cannot be gained by altering the lateral or longitudinal adherence bounds or the likelihood function of the current probe.*

To test this hypothesis, the trajectory error in a treatment run is compared to the error in the FA32 baseline scenario (32BL) using a matched pair test. This statistical test is carried out for all three error types and for each of the 12 treatment runs. For the matched pair analysis, the average error is calculated for each individual flight in the baseline and treatment run scenarios. The values are compared for every flight, and the mean difference is calculated to illustrate any underlying differences between the scenarios. The matched pair analysis includes a paired *t*-test, which examines the distribution of differences in per flight error between the two scenarios and tests if the mean of the differences is statistically different from zero. The results from the paired *t*-tests for cross track, along track, and vertical errors are provided in Table 16, Table 17, and Table 18 respectively. The likelihood parameter has no effect on the trajectory modeling, so it is left out of these tables.

Table 16. Statistical Results for Trajectory Cross Track Error

Run	Lat bound	Long bound	Mean by flight (nm)	*t*-ratio vs. 32BL	*p*-value vs. 32BL
1	0.5	0.5	0.72941	-23.0753	<.0001
2	2.5	1.5	0.85348	0.449675	0.653
3	0.5	1.5	0.73571	-22.0115	<.0001
4	1.5	0.5	0.78731	-14.5896	<.0001
5	1.5	1.0	0.79885	-11.9374	<.0001
6	0.5	1.5	0.73486	-22.2937	<.0001
7	1.5	1.5	0.80568	-10.5545	<.0001
8	2.5	0.5	0.81791	-16.0585	<.0001
9	2.5	1.0	0.84138	-7.23434	<.0001
10	2.5	0.5	0.81917	-15.6695	<.0001
11	0.5	0.5	0.72927	-23.2975	<.0001
12	0.5	1.0	0.73328	-22.5304	<.0001
32BL	2.5	1.5	0.85316	-	-

Table 17. Statistical Results for Trajectory Along Track Error

Run	Lat bound	Long bound	Mean by flight (nm)	*t*-ratio vs. 32BL	*p*-value vs. 32BL
1	0.5	0.5	1.03443	-37.7576	<.0001
2	2.5	1.5	1.32823	1.154313	0.2485
3	0.5	1.5	1.24105	-9.89563	<.0001
4	1.5	0.5	1.05273	-34.3245	<.0001
5	1.5	1.0	1.17805	-19.1516	<.0001
6	0.5	1.5	1.2426	-9.74754	<.0001
7	1.5	1.5	1.29028	-1.95322	0.0509
8	2.5	0.5	1.05834	-38.539	<.0001
9	2.5	1.0	1.19206	-21.4522	<.0001
10	2.5	0.5	1.05846	-38.5739	<.0001
11	0.5	0.5	1.03259	-38.6468	<.0001
12	0.5	1.0	1.14615	-23.7055	<.0001
32BL	2.5	1.5	1.3008	-	-

Table 18. Statistical Results for Trajectory Vertical Error

Run	Lat bound	Long bound	Mean by flight (ft)	*t*-ratio vs. 32BL	*p*-value vs. 32BL
1	0.5	0.5	340.985	-7.77013	<.0001
2	2.5	1.5	359.383	0.419611	0.6748
3	0.5	1.5	350.159	-4.52418	<.0001
4	1.5	0.5	343.959	-7.53102	<.0001
5	1.5	1.0	352.234	-4.06503	<.0001
6	0.5	1.5	349.732	-4.75879	<.0001
7	1.5	1.5	354.995	-2.79515	0.0052
8	2.5	0.5	346.388	-7.9294	<.0001
9	2.5	1.0	355.486	-2.9365	0.0034
10	2.5	0.5	346.958	-6.94365	<.0001
11	0.5	0.5	341.066	-7.85201	<.0001
12	0.5	1.0	349.432	-4.75783	<.0001
32BL	2.5	1.5	358.964	-	-

When the mean error in the baseline is greater than that in the treatment run, the mean difference is negative. The *t*-ratio is a statistical test parameter which is calculated as the mean difference over the standard error. Large negative values of the *t*-ratio produce small *p*-values and indicate that the mean difference is of statistical significance.

From the results in Table 16 for cross track error, the low *p*-values for almost all runs indicate there is a statistically significant difference in error for every treatment run except one. Run 2 uses the same lateral and longitudinal bounds as the baseline and it is intuitive that it would be an exception, because the trajectory modeling process in this treatment run would behave similar to that in the baseline. In all the other treatment runs, it can be observed that reducing the lateral or longitudinal bounds resulted in lower average error.

In Table 17, the results for along track error also include low *p*-values for many runs, with exceptions in Run 2 and Run 7. The reasoning for Run 2 is the same as above. Noting that the lateral and longitudinal bounds for Run 7 are 1.5 each (compared to 2.5 and 1.5, respectively, in the baseline) it can be surmised that the slightly smaller lateral bound resulted in decreased errors, similar to the other treatment runs, and that in this case the effect was just under the threshold of statistical significance.

The results for vertical error in Table 18 also include statistically[§§§§] significant *p*-values for all runs except Run 2. However, the *t*-ratio values overall are less negative, which implies less drastic effects than in the previous two tables. This may be explained by the fact that while changes in lateral and longitudinal bounds directly affect cross and along track errors, respectively, their impact on vertical error is much less straightforward.

These impacts are further explored using a statistical model of the factors' effects on the error metrics. Average error values were taken over all valid data points and entered into a model of treatment effects. The prediction profiler is a tool from JMP [SAS, 2010, B] that graphically presents the results of the model by plotting average error versus the lateral, longitudinal, and likelihood factors. A screenshot of this tool is shown in Figure 15. Note that for the matched pair tests, the average error value was calculated by individual flights first and then the average over these per flight values was taken. Those values differ from the overall average values of all sampled error data points, which is what was used in the model.

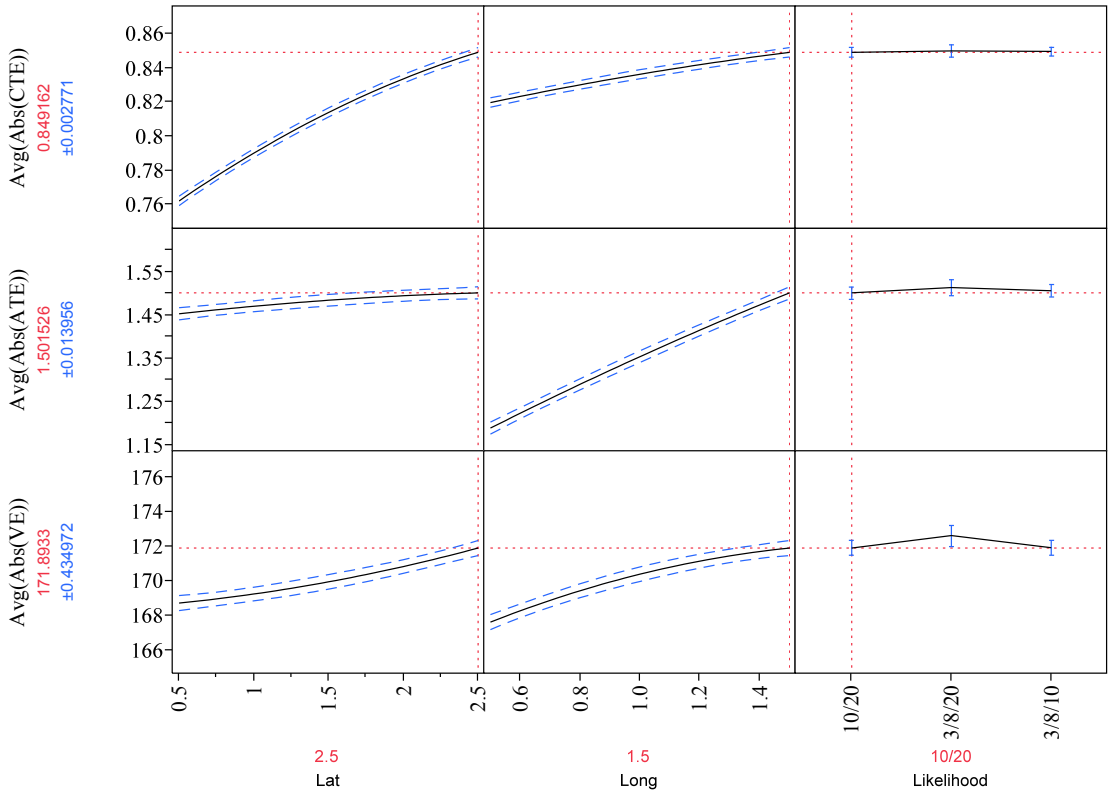


Figure 15. Prediction Profiler of TM Model with Baseline Parameter Settings

[§§§§] It should be noted that the maximum decrease in error is 18ft, which is a practically insignificant change for vertical error.

Figure 15 profiles some of the treatment effects from the statistical model, and a few important patterns can be discerned. As the lateral bound is reduced, the average cross track error decreases sharply, thereby increasing the trajectory accuracy in this dimension. Next, as the longitudinal bound is reduced, the average along track error decreases sharply, increasing trajectory accuracy in the longitudinal dimension as well. Average vertical error is also affected by reducing either bound, however the error is measured in feet so the practical effect is much less in comparison to the other errors in nm. In other words, the effect is statistically significant but not practically significant.

3.2.2 Conflict Probe Analysis

Using the integrated experiment, each effect of the conflict probe performance can be evaluated at many different settings. This allows the analyst to understand the effects of each of the three factors on the conflict probe performance. The purpose of this integrated experiment is not to determine the optimal settings, but rather to understand the effects of each of the settings and determine if a significant improvement to performance can be gained by altering these settings of the conflict probe. The null hypothesis for the conflict probe evaluation is the following:

> *A significant Conflict Probe performance improvement cannot be gained by altering the lateral or longitudinal adherence bounds or the likelihood function of the current probe.*

The purpose of this section is to prove this null hypothesis false. In order to do this we are looking for a significant decrease in False Alert Rate as well as either a decrease in Late Alert Rate or no significant increase in Late Alert Rate. Furthermore, a desirable result would include no significant decrease in warning time.

In the following analysis the 32BL is the baseline scenario that includes the FA32 trajectory modeling upgrades described in Section 3.1. IBL is the initial baseline that was described in [Crowell, 2011, A], and does not include the FA32 upgrades. In this chapter, the main analysis performed compares the results of the twelve treatment runs to the 32BL, since the purpose of this technical note is to determine the performance that can be gained by altering the three settings described in Section 2.2.2.1, and these three settings are the only differences between 32BL and each of the treatment runs. The comparison to the IBL is included since the conclusion of the FA32 analysis was that the upgrades should be implemented, so this information shows the overall performance gained through just that upgrade with several different settings.

In Table 19, the alert type counts of each of the treatment runs and the two baseline runs are shown. It can be seen from this table, that although all runs decrease the FA count from both baselines, they also increase the LA count. This table also shows that all runs do not significantly affect MA count. Although there are four runs that decrease the MA count from 4 to 3, this decrease is not significant and can likely be attributed to the noise inherent among runs of ERAM. Likewise, the three that show an increase to 5 MAs do not show a significant change.

Table 19. Alert type counts for each of the treatment runs and the two baselines.

Run	VA	STD VA	LA	MA	FA
1	155	125	19	4	878
2	159	131	15	4	1148
3	160	131	15	3	1013
4	151	120	24	3	814
5	160	130	15	3	1066
6	161	132	13	4	774
7	156	127	18	4	994
8	156	126	17	5	1288
9	157	128	17	3	1092
10	154	123	19	5	981
11	150	119	23	5	633
12	158	129	16	4	718
32BL	163	135	11	4	1394
IBL	158	134	12	4	1470

During the development of the fuzzy metrics defined in Section 0, a new metric was discovered that weights the value of each Late Alert or Missed Alert by the amount of time by which it was late. This metric is less susceptible to the noise inherent among runs, since a small shift in time will be represented by a small difference in the value, instead of a difference between counting as a LA or not. The values of each alert are summed for the run and this value is called Adjusted LA. In Table 20 this value is shown for each run and compared to the sum of Late Alerts and Missed Alerts. An example of the differences between the two metrics can be seen by comparing Runs 6 and 7. Using the LA+MA metric, Run 7 seems to have a significant degradation in performance, with a value of 22 compared to the 17 of Run 6. However, looking at the Adjusted LA metric, it can be seen that there is no significant difference between the two runs, since their scores differ by a value of only 0.04.

Similar to the LA metric, the Adjusted LA metric shows a significant degradation in late alert performance for all of the treatment runs, when compared to either of the two baseline scenarios. Those that have the highest values of LA count, particularly Runs 4 and 11, also have the highest values of Adjusted LA. The lowest values of Adjusted LA are from Runs 2 and 3, with Run 5 close behind. Runs 6, 7, and 9 are only slightly further degraded in performance. These initial findings are interesting in that Runs 2, 3, 6, and 7 are run with a longitudinal setting of 1.5 nm, whereas Runs 5 and 9 are run with a longitudinal setting of 1.0 nm with lateral settings of 1.5 and 2.5 nm respectively. The only run with a longitudinal bound of 1.0 nm or greater not included in this list is Run 12, with a longitudinal of 1.0 and lateral of 0.5 nm. These initial findings indicate that longitudinal conformance bound may be the most significant factor affecting LAs, and there is a small interaction with the lateral conformance bound that becomes most prominent when the lateral bounds are set to the smallest value. This is an important finding and will be analyzed further later.

Table 20. Late Alert plus Missed Alert count and Adjusted Late Alert value for each of the treatment runs and the two baselines.

Run	LA+MA	Adj LA
1	23	14.58
2	19	10.46
3	18	10.64
4	27	16.45
5	18	11.13
6	17	11.78
7	22	11.74
8	22	12.90
9	20	11.64
10	24	12.57
11	28	16.32
12	20	13.26
32BL	15	8.95
IBL	16	8.79

Next, we look at the LA and FA Rates in Table 21. LA Rate is calculated by summing the number of LAs and MAs and dividing by the total number of conflict events. FA Rate is calculated by dividing the number of FAs by the number of non-conflict events. Non-conflict events are encounters that do not result in a conflict. The overall rates are shown on the left, and then the FA Rate is broken down into bins based on the horizontal and vertical separation of the encounter event at the minimum max-ratio time.

Table 21. Overall Alert Rates and False Alert Rates by separation category.

			FA Rate								
			Horz	<5	<5	5 ≤ h < 8	5 ≤ h < 8	8 ≤ h ≤ 13	8 ≤ h ≤ 13	>13	>13
Run	LA Rate	FA Rate	Vert	<1000	≥1000	<1000	≥1000	<1000	≥1000	<1000	≥1000
1	0.155	0.051		1.000	0.119	0.805	0.244	0.225	0.036	0.076	0.024
2	0.127	0.068		1.000	0.107	1.000	0.365	0.600	0.185	0.037	0.013
3	0.121	0.059		1.000	0.109	0.921	0.347	0.319	0.060	0.061	0.024
4	0.184	0.048		1.000	0.122	0.931	0.328	0.241	0.059	0.046	0.011
5	0.122	0.063		1.000	0.119	0.978	0.372	0.338	0.085	0.045	0.024
6	0.114	0.045		1.000	0.102	0.916	0.334	0.292	0.055	0.036	0.009
7	0.148	0.058		1.000	0.117	0.977	0.378	0.413	0.120	0.038	0.011
8	0.149	0.076		1.000	0.123	0.973	0.365	0.494	0.135	0.069	0.029
9	0.135	0.064		1.000	0.106	0.967	0.364	0.523	0.152	0.044	0.014
10	0.163	0.058		1.000	0.121	0.951	0.346	0.416	0.126	0.038	0.011
11	0.190	0.037		1.000	0.119	0.759	0.234	0.168	0.033	0.033	0.011
12	0.134	0.042		1.000	0.118	0.862	0.307	0.177	0.034	0.043	0.010
32BL	0.100	0.082		1.000	0.113	0.994	0.365	0.648	0.192	0.058	0.031
IBL	0.107	0.087		1.000	0.114	1.000	0.381	0.650	0.206	0.053	0.033

The first bin is a horizontal separation of less than 5 nm and a vertical separation of less than 1000 ft. This column includes all events that are conflicts, and since there are no non-conflict events in this bin, all rates are a value of 1, because the denominator becomes only the number of FAs. Looking at these false alert rates, it can clearly be seen that false alerts are far less likely

when the flights remain separated vertically. Every column shows a higher FA Rate in both baseline scenarios than in all runs, which is exactly what was expected, so there are no real surprises to be investigated in this table.

In Figure 16 the Hit Rate is graphed versus the False Alert Rate for each of the treatment runs and the two baselines. The FA Rates are the same values from Table 21, whereas Hit Rate is 1 – *LA Rate* from the same table. The top-left corner of the plot is the most desirable, since the most desired performance is high Hit Rate and low FA Rate. From this plot we can see that both baselines have a high FA Rate but a desirable LA Rate, whereas all of the treatment runs sacrifice Hit Rate in order to reduce the FA Rate. Looking at this plot, of the twelve treatment runs, Run 6 seems to provide the best balance of Hit Rate versus FA Rate, since it greatly reduces the FA Rate when compared to the baselines, but barely reduces the Hit Rate. Assuming no other metrics, all runs to the right of and below Run 6 can be eliminated, since they degrade in performance of both FA Rate and Hit Rate. This leaves only Runs 6, 11, and 12 as the top contenders, but another important metric to observe is Warning Time.

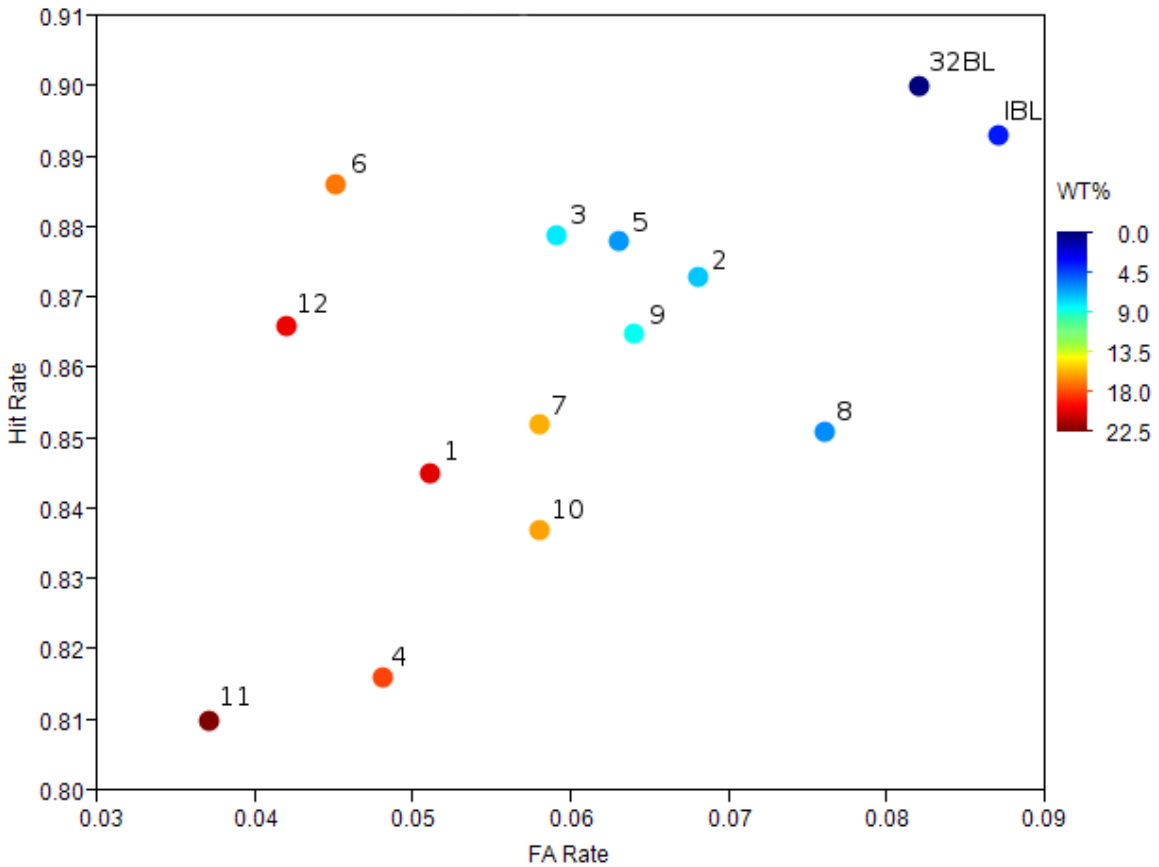


Figure 16. Hit Rate vs. False Alert Rate colored by Warning Time. The top-left corner of the graph is the most desirable.

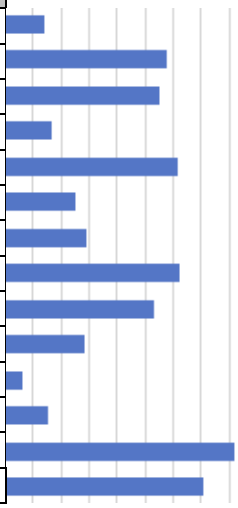
The warning time of alerts is an important metric because it tells not only how correct the probe is, but how much workload it puts on the controller. Less warning time before the conflict requires quicker reaction times from the controllers and therefore increases controller workload.

Table 22 shows several statistics of warning time for each of the treatment runs and the two baselines. The statistics used are the median, inter-quartile range (IQR), and 25th percentile. The

graph shown next to the columns is the 25th percentile. The 25th percentile is the main metric we want to look at, since this tells us the low end of the warning times of alerts. Raising this low end is more desirable than raising the median, since the median is almost always well above the required warning time, which in this case is 180 seconds. Back in Figure 16, each point representing a run has been colored based on its warning time value. The exact metric used for this coloring is the percentage difference of the 25th percentile of warning time from the FA32 Baseline, where deep blue is no reduction and deep red is the maximum reduction of 22.5%. Looking at the graph, we can see that warning time is consistently reduced as the FA Rate is reduced.

Table 22. Median, inter-quartile range, and 25th percentile of conflict warning time.

Run	Median	IQR	25th %
1	397.0	374.0	264.0
2	409.0	264.0	307.5
3	445.5	399.0	305.0
4	387.5	306.0	266.0
5	443.0	414.0	311.0
6	379.0	232.5	275.0
7	389.0	292.0	279.0
8	436.0	432.0	312.0
9	406.0	271.0	303.0
10	391.5	273.0	278.0
11	361.5	212.0	256.0
12	376.5	241.0	265.0
32BL	492.0	440.0	331.5
IBL	460.0	378.5	320.5



Run 6 that was previously deemed as the best run, can clearly be seen to have a significant reduction in warning time. Now, from the graph, we can eliminate 4 runs: Runs 1, 2, 4, and 9. Runs 1, and 4 can be eliminated since they are to the right and below Run 6, which they are also darker than. Runs 2 and 9 can be eliminated because it is to the right, below, and darker than Run 5. Table 23 shows the settings of these four runs, but nothing immediately stands out as a common factor. However, if they are split into the two sets of eliminations, as signified by the red and green color coding, we can see that the two reds were eliminated because of their longitudinal conformance bounds which we have earlier determined to be a major factor in LA Rate. The blue runs have the default 2.5 nm lateral conformance bound, with no change or little change to longitudinal bound. These two runs performed well with warning time but had little reduction to FA Rate when compared to other runs. This is another important finding that will continue to be investigated further in this paper.

Table 23. The runs that can be eliminated based on Hit Rate, FA Rate and 25th percentile Warning Time.

Run	Lateral	Longitudinal	Likelihood
1	0.5	0.5	10/20
2	2.5	1.5	3/8/10
4	1.5	0.5	3/8/20
9	2.5	1.0	3/8/20

Another analysis used to help validate our previous findings is the Conflict Probe Comparison that compares the results of each of the treatment runs' conflict probes to the baseline conflict

probes. Table 24 shows the results of each comparison to the FA32 Baseline run. The table contains a lot of information, but the main focus points are the VA-LA, LA-VA, FA-N/A, and N/A-FA. These fields are wrapped into the two fields at the bottom of the table: % FA Improvement and % LA Improvement. The percent FA improvement metric is the N/A-FA count subtracted from the FA-N/A count, divided by the number of FAs in the baseline run. Similarly percent LA improvement is the VA-LA count subtracted from the LA-VA count, divided by the number of LAs in the baseline run. LAs in this table include Late Alerts as well as Missed Alerts. A positive percentage is an improvement over the baseline whereas a negative percentage is degradation.

Table 24. Results of comparisons of each of the treatment runs to the FA32 Baseline run.

	Lat	0.5	2.5	0.5	1.5	1.5	0.5	1.5	2.5	2.5	2.5	0.5	0.5
	Lon	0.5	1.5	1.5	0.5	1	1.5	1.5	0.5	1	0.5	0.5	1
	Like	10/20	3/8/10	10/20	3/8/20	10/20	3/8/10	3/8/20	10/20	3/8/20	3/8/10	3/8/10	3/8/20
BL	**Proto**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
FA	**FA**	597	1107	772	645	864	629	851	1010	948	813	473	551
FA	**N/A**	718	276	548	687	449	701	491	310	408	516	858	779
N/A	**FA**	189	24	166	111	131	95	92	194	92	116	99	104
LA	**LA**	13	15	13	14	13	12	14	15	14	14	14	13
LA	**VA**	2	0	2	1	2	3	1	0	1	1	1	2
VA	**LA**	10	4	5	13	5	5	8	7	6	10	14	7
VA	**VA**	153	159	158	150	158	158	155	156	156	153	149	156
VA	**N/A**	0	0	0	0	0	0	0	0	1	0	0	0
DSC	**DSC**	625	696	783	446	838	455	565	901	588	490	362	431
DSC	**FA**	92	17	75	58	71	50	51	84	52	52	61	63
FA	**DSC**	79	11	74	62	81	64	52	74	38	65	63	64
DSC	**N/A**	555	559	414	768	363	767	656	287	632	730	849	778
N/A	**DSC**	133	10	92	65	107	40	39	130	59	53	67	65
% FA Imp		38%	18%	27%	41%	23%	43%	29%	8%	23%	29%	54%	48%
% LA Imp		-53%	-27%	-20%	-80%	-20%	-13%	-47%	-47%	-33%	-60%	-87%	-33%

The results seen in this table are similar to those seen in the previous analyses performed in this technical note. Each of the treatment runs improves the FA performance but degrades in LA performance when compared to the baseline. Figure 17 shows the two improvement metrics plotted against each other. Once again the points are colored by the warning time. The plot is essentially the mirror image of Figure 16. It is mirrored because that figure plotted Hit Rate vs. FA Rate, where Hit Rate is 1 – LA Rate.

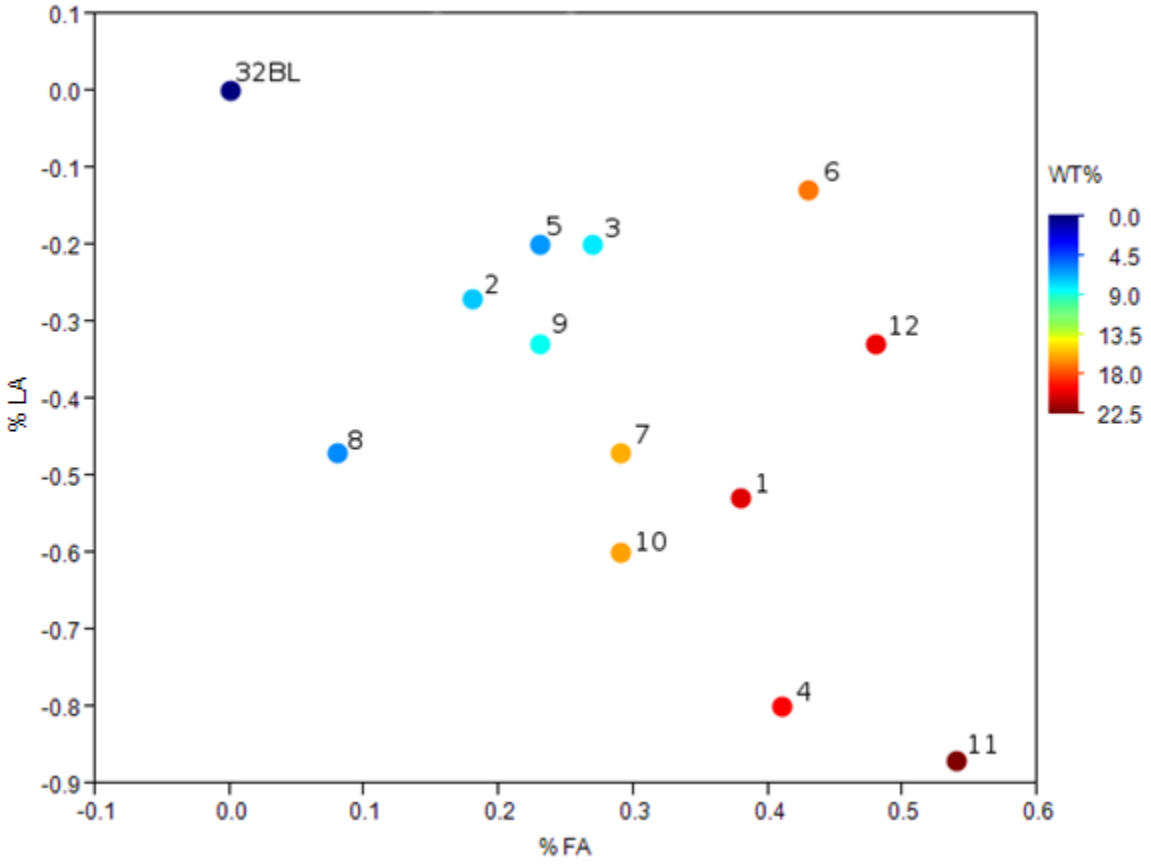


Figure 17. Graph of % FA Improvement vs. % LA Improvement compared to the FA32 Baseline. Top right corner is the most improved.

To give an idea of how these results compare to the Initial Baseline, Table 25 shows the results of the conflict probe comparison of each of the treatment runs to the IBL. Once again, looking at the percent improvements, we notice that all runs significantly improve upon the False Alert performance, but not in the LA performance. However, unlike in the comparison to the FA32 Baseline, Run 6 does not degrade in LA performance at all.

Figure 18 plots the percent improvement of the FA versus the LA. This figure looks almost exactly the same as Figure 17, only the LA degradation is smaller. This table and figure only confirm that any findings when analyzed against the FA32 Baseline will hold true against the Initial Baseline.

Table 25. Results of comparisons of each of the treatment runs to the Initial Baseline run.

	Lat	0.5	2.5	0.5	1.5	1.5	0.5	1.5	2.5	2.5	2.5	0.5	0.5
	Lon	0.5	1.5	1.5	0.5	1	1.5	1.5	0.5	1	0.5	0.5	1
	Like	10/20	3/8/10	10/20	3/8/20	10/20	3/8/10	3/8/20	10/20	3/8/20	3/8/10	3/8/10	3/8/20
BL	**Proto**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DSC	**DSC**	616	606	758	452	829	446	549	865	559	475	360	433
DSC	**FA**	100	66	92	71	88	61	68	110	74	76	69	70
DSC	**N/A**	599	643	465	792	398	808	698	340	682	764	886	812
FA	**DSC**	86	56	87	61	90	68	60	86	62	71	66	63
FA	**FA**	602	989	762	634	838	624	840	972	906	784	467	546
FA	**N/A**	782	425	621	775	542	778	570	412	502	615	937	861
LA	**LA**	13	15	12	14	13	12	14	15	14	14	14	13
LA	**VA**	3	1	4	2	3	4	2	1	2	2	2	3
N/A	**DSC**	135	55	104	60	107	45	47	154	64	62	66	64
N/A	**FA**	176	93	159	109	140	89	86	206	112	121	97	102
N.A	**LA**	1	1	1	1	1	1	1	1	1	1	1	1
N/A	**VA**	5	5	5	5	5	5	5	5	4	5	5	5
VA	**LA**	9	3	5	12	4	4	7	6	5	9	13	6
VA	**N/A**	2	2	2	2	2	2	2	2	2	2	2	2
VA	**VA**	147	153	151	144	152	152	149	150	151	147	143	150
% FA Imp		41%	23%	31%	45%	27%	47%	33%	14%	27%	34%	57%	52%
% LA Imp		-38%	-13%	-6%	-63%	-6%	0%	-31%	-31%	-19%	-44%	-69%	-19%

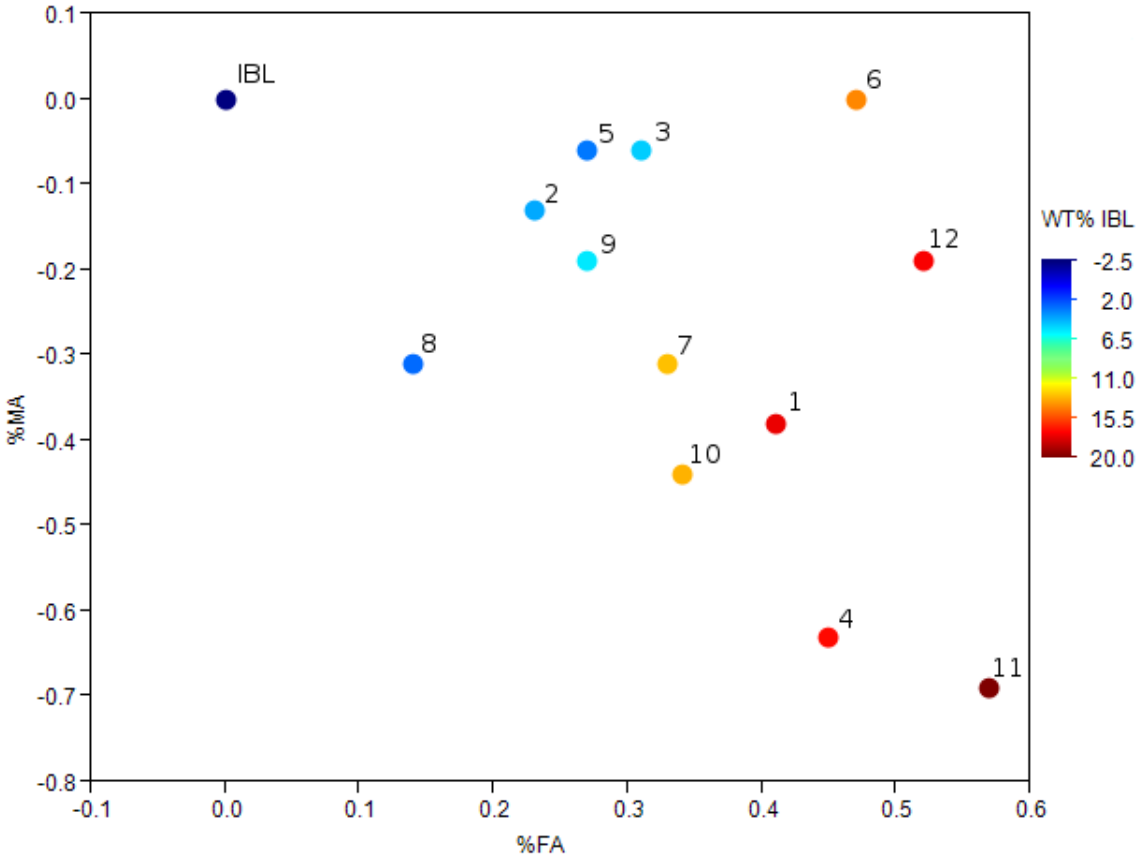


Figure 18. Graph of % FA Improvement vs. % LA Improvement compared to the Initial Baseline. Top right corner is the most improved.

3.2.2.1 Fuzzy Analysis Results

The fuzzy alert metrics are used in addition to the traditional metrics to provide some insight where the traditional metrics cannot. Since the fuzzy metrics use continuous values, they are less susceptible to the noise among runs of ERAM, but this is only one advantage provided by these metrics. The metrics were developed to capture the severity of an encounter event, rather than a simple yes or no of whether the event was a conflict or not, like the traditional metrics do. By capturing this severity, many more events are available for analysis that were previously simply encounters and are now very severe events. Many of these events were considered False Alerts using the traditional metrics and are captured as Valid in the fuzzy metrics. The fuzzy metrics attempt to capture the events that should be alerted regardless of whether or not they are an actual conflict.

In Table 26 we see the fuzzy alert values for each of the three main metrics. Though the measures provide similar numbers to the traditional metrics, it is important to note that, as with all fuzzy metrics, these numbers are calculated using a much different method than the traditional metrics and are only intended for comparison to fuzzy metrics of other runs. These fuzzy metrics should not be compared directly to the traditional metrics.

Table 26. Fuzzy Alert values for each of the treatment runs and the baselines.

Run	VA	LA	FA
1	261.05	29.57	740.02
2	260.04	20.42	763.46
3	267.02	20.39	815.86
4	266.14	24.01	629.73
5	271.44	18.33	837.26
6	264.99	22.47	594.09
7	268.33	17.31	730.34
8	266.51	23.12	954.01
9	267.22	20.90	748.33
10	264.45	23.13	694.92
11	253.96	33.76	502.16
12	262.68	25.03	570.33
32BL	260.31	20.55	949.11
IBL	257.33	19.08	994.56

Table 27. Overall Fuzzy Alert Rates and Fuzzy False Alert Rates by separation category.

			FA Rate									
			Horz	<5	<5	5 ≤ h < 8	5 ≤ h < 8	8 ≤ h ≤ 13	8 ≤ h ≤ 13	>13	>13
Run	LA Rate	FA Rate	Vert	<1000	≥1000	<1000	≥1000	<1000	≥1000	<1000	≥1000
1	0.102	0.044		0.832	0.342	0.686	0.210	0.224	0.035	0.075	0.023
2	0.073	0.045		0.822	0.197	0.815	0.300	0.343	0.118	0.030	0.011
3	0.071	0.048		0.837	0.334	0.775	0.295	0.276	0.055	0.057	0.023
4	0.083	0.037		0.823	0.210	0.800	0.276	0.216	0.052	0.043	0.011
5	0.063	0.050		0.809	0.338	0.830	0.316	0.283	0.071	0.041	0.022
6	0.078	0.035		0.843	0.172	0.772	0.283	0.250	0.051	0.034	0.009
7	0.061	0.043		0.805	0.202	0.809	0.313	0.314	0.091	0.034	0.010
8	0.080	0.056		0.755	0.353	0.824	0.302	0.344	0.097	0.058	0.025
9	0.073	0.044		0.825	0.204	0.803	0.301	0.325	0.101	0.035	0.012
10	0.080	0.041		0.791	0.201	0.785	0.287	0.278	0.089	0.032	0.010
11	0.117	0.030		0.815	0.203	0.628	0.198	0.165	0.033	0.033	0.011
12	0.087	0.034		0.836	0.201	0.740	0.266	0.170	0.033	0.042	0.010
32BL	0.073	0.057		0.761	0.330	0.798	0.298	0.389	0.124	0.045	0.024
IBL	0.069	0.059		0.771	0.334	0.817	0.312	0.367	0.130	0.039	0.026

The fuzzy alert values can be analyzed in the same way the traditional alert type counts are. Although all of the treatment runs improve the FA value over the Initial Baseline, Run 8 does not improve over the FA32 Baseline, indicating that the real improvement in that run was achieved by the FA32 prototypes and not by the altered settings. Another thing to note that is very different than the traditional metrics is that four of the runs actually reduce the LA value from both baseline values. Runs 2, 3, 5, and 7 all reduce the LA value. These are four of the runs that were noted to have a low LA count in the traditional metrics. The other two runs that were noted in the traditional metrics were Runs 6 and 9. Although these two runs increase the LA value over the

baselines, they only increase slightly and have considerably lower values than most other runs. These findings with the fuzzy metrics further verify our speculation that lowering the longitudinal conformance bounds has a negative impact on the LA performance of the conflict probe.

Table 27 shows the fuzzy alert rates. These are similar to the alert rates shown for the traditional metrics, but the numerator and denominator of these metrics use fuzzy values. The LA Rate is the LA value from the Table 26 divided by the sum of the LA value and the VA value. The FA Rate is the FA value from Table 26 divided by the sum of the FA value and the No-call (NC) value. NC Value is not shown in the table above, but the NC value for each event is defined as:

$$1 - \max(S, R)$$

The variables S and R are signal and response, respectively, as defined in Section 0. The overall NC value for a run is the sum of all NC values for events in that run.

The FA Rate columns on the right of the table are separated into categories by horizontal and vertical separation at the minimum max-ratio time.

Figure 19 shows the fuzzy Hit Rate versus FA Rate of each of the treatment runs and the baselines, once again colored by percent difference of the warning time from the FA32 Baseline. Interestingly, there are quite a few runs that improve over the baseline in fuzzy hit rate, whereas all runs degraded in the traditional hit rate.

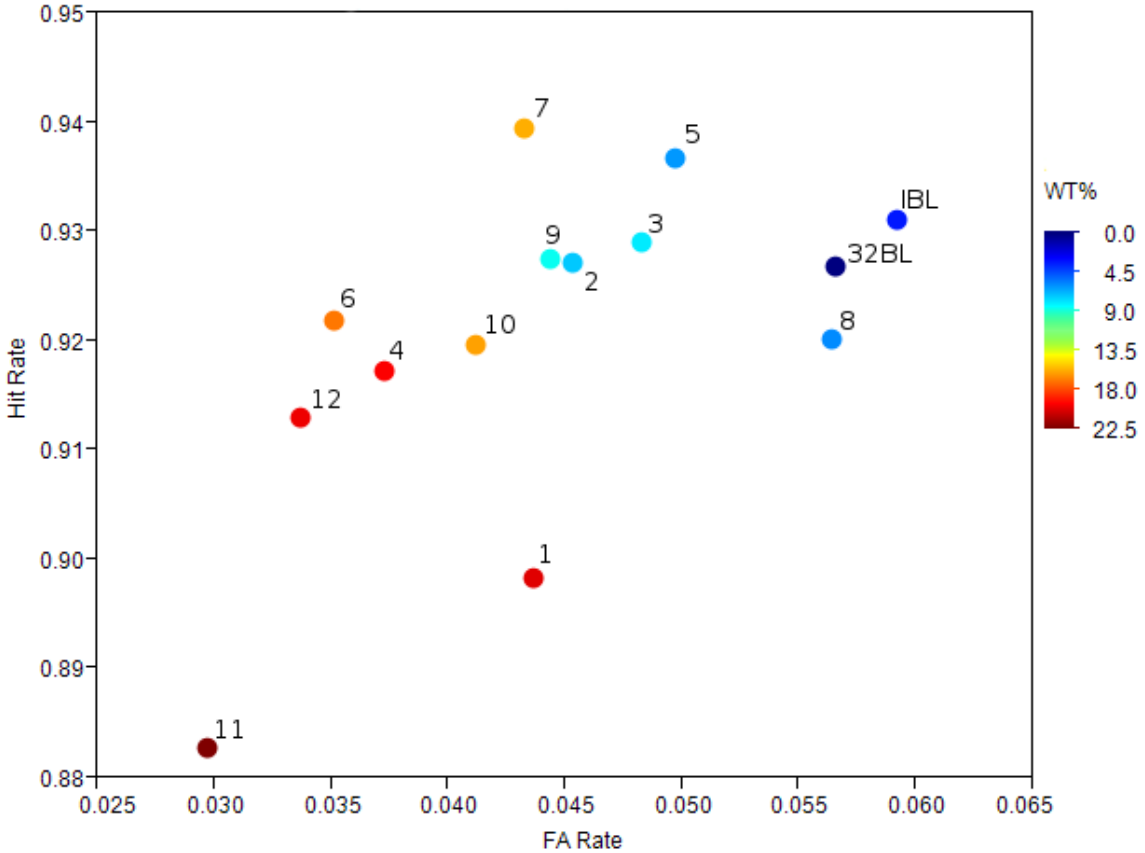


Figure 19. Fuzzy Hit Rate vs. Fuzzy False Alert Rate.

3.2.2.2 Model Analysis

The main analysis performed for this study used an integrated experiment with a D-optimal design that was described in Section 2.2.1. The design allows the analyst to model many different parameter settings with only the twelve experimental runs of ERAM and the FA32 Baseline. The model provides many options for analysis of the three factors. For initial discovery we look at a graph called the prediction profiler, which allows the analyst to observe effects of several factors on several response variables in one interactive graph. Since the model is a multi-dimensional model the graph must be interacted with in order to gain the full perspective of the data. In this document we provide several perspectives of the data.

3.2.2.2.1 Implementation of Statistical Model

The traditional metrics will be modeled separately from the fuzzy model (using the same warning time percentage in both models since it does not change). Using the data collected from the 12 runs and the baseline, the following leverage plots are obtained (Figure 20). In these plots the measured values (*y*-axis) are plotted against their modeled values (*x*-axis). Therefore, all points that fall on the diagonal are exactly modeled. The horizontal blue line represents the mean value of the samples and the red curves indicate the 95% confidence interval. Although the model captures between 95% and 100% of the variation in the study, which can be seen from the RSq value under each plot (RSq is the coefficient of determination R^2), none of the responses are statistically significant. The significance is quickly determined in a leverage plot by determining if the confidence interval intersects the mean, where no intersection indicates insignificance. There are likely more than one interaction terms in the model whose interactions have little correlation and can be removed from the model for improvement.

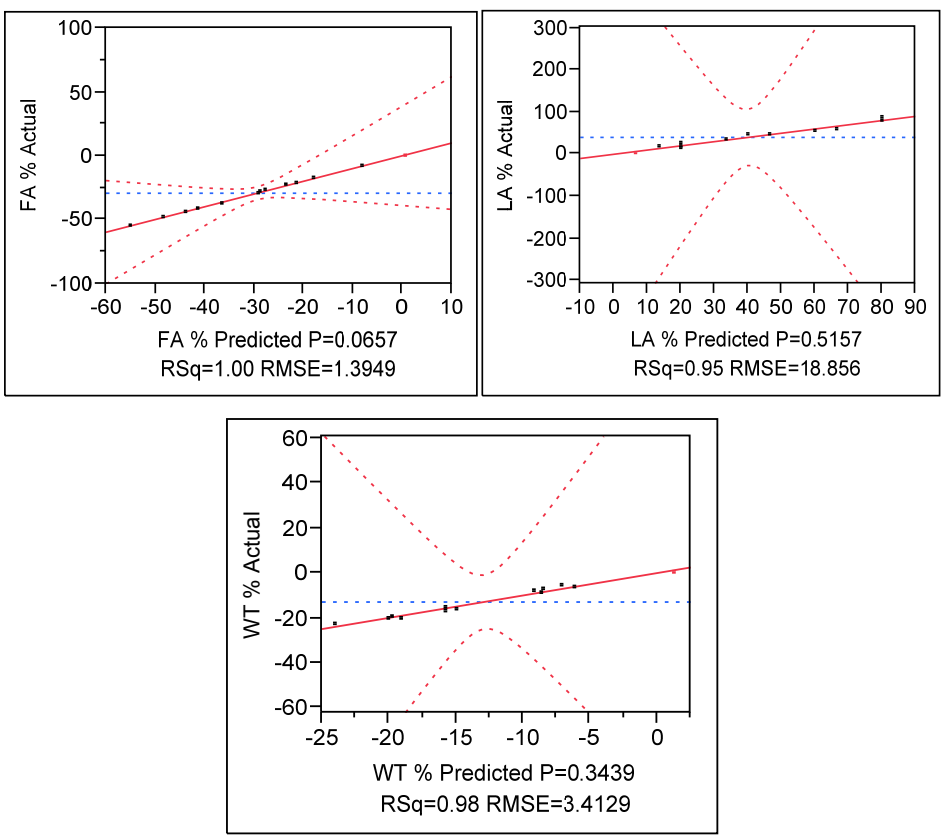


Figure 20. Traditional FA %, LA %, and WT % metrics leverage plots

The following tables (Table 28, Table 29,) list the effect tests for each interaction of the various factors for each of the three responses currently being investigated. Highlighted are those effects that have a *p*-value of at least 0.7000, where $p > 0.0500$ indicates statistical insignificance. The two-way interaction of lateral and longitudinal bounds is highly insignificant for all three response variables, indicating that it should be removed from the model. Also, the two-way interactions of likelihood are highly insignificant for LA % and WT %. These values are below the initial threshold of 0.7000 for FA %, though are still much higher than 0.0500. These three interactions will therefore be excluded for the next iteration of the model.

As an aside, removing only Lat and Lat*Likelihood interactions still produces a model with all three responses significant, although LA % only has a significance value of 0.0368, the R^2 values would be 94% for LA %, 97% for WT % and 100% for FA %. The confidence gain of removing all three outweighs the degradation of the model fit.

Table 28. Initial Model Effect Test for the FA % Response Variable

Source	DF	Sum of Squares	F Ratio	Prob > F
Lat(0.5,2.5)	1	1644.2036	844.9787	0.0219
Long(0.5,1.5)	1	265.7555	136.5754	0.0543
Likelihood	2	869.0968	223.3204	0.0473
Lat*Long	1	0.0058	0.0030	0.9653
Lat*Likelihood	2	3.0997	0.7965	0.6210
Long*Likelihood	2	6.6495	1.7086	0.4758
Lat*Lat	1	14.0043	7.1970	0.2271
Long*Long	1	13.2975	6.8338	0.2326

Table 29. Initial Model Effect Test for the LA % Response Variable

Source	DF	Sum of Squares	F Ratio	Prob > F
Lat(0.5,2.5)	1	100.0000	0.2813	0.6896
Long(0.5,1.5)	1	4011.1111	11.2813	0.1842
Likelihood	2	1247.0588	1.7537	0.4710
Lat*Long	1	88.8889	0.2500	0.7048
Lat*Likelihood	2	62.2222	0.0875	0.9225
Long*Likelihood	2	160.0000	0.2250	0.8305
Lat*Lat	1	177.7778	0.5000	0.6082
Long*Long	1	711.1111	2.0000	0.3918

Table 30. Initial Model Effect Test for the WT % Response Variable

Source	DF	Sum of Squares	F Ratio	Prob > F
Lat(0.5,2.5)	1	239.0125	20.5200	0.1383
Long(0.5,1.5)	1	103.6527	8.8989	0.2059
Likelihood	2	193.9714	8.3265	0.2380
Lat*Long	1	1.3764	0.1182	0.7892
Lat*Likelihood	2	5.7534	0.2470	0.8181
Long*Likelihood	2	9.0293	0.3876	0.7505
Lat*Lat	1	0.5315	0.0456	0.8660
Long*Long	1	15.1822	1.3034	0.4579

From Figure 21, it can be seen that although R^2 has degraded to the range of 91% to 100% (as opposed to 95% to 100%), the degradation is acceptable considering the huge impact on the confidence intervals of the three responses. Now all three response variables are significant. The new model is therefore Eq.29.

Response:

$$R_{ijk} = \mu + Lat_i^2 + Lat_i + Long_j^2 + Long_j + Like_k + \varepsilon_{n(ijk)} \qquad \textbf{Eq.29}$$

Where:

Lat_i = lateral conformance bounds in nautical miles, i = 0.5, 1.5, 2.5
$Long_j$ = longitudinal conformance bounds in nautical miles, j = 0.5, 1.0, 1.5
$Like_k$ = likelihood, k = "10/20", "3/8/10", "3/8/20"
$\varepsilon_{n(ijk)}$ = random error, n = 1, 2, … for all i, j, k

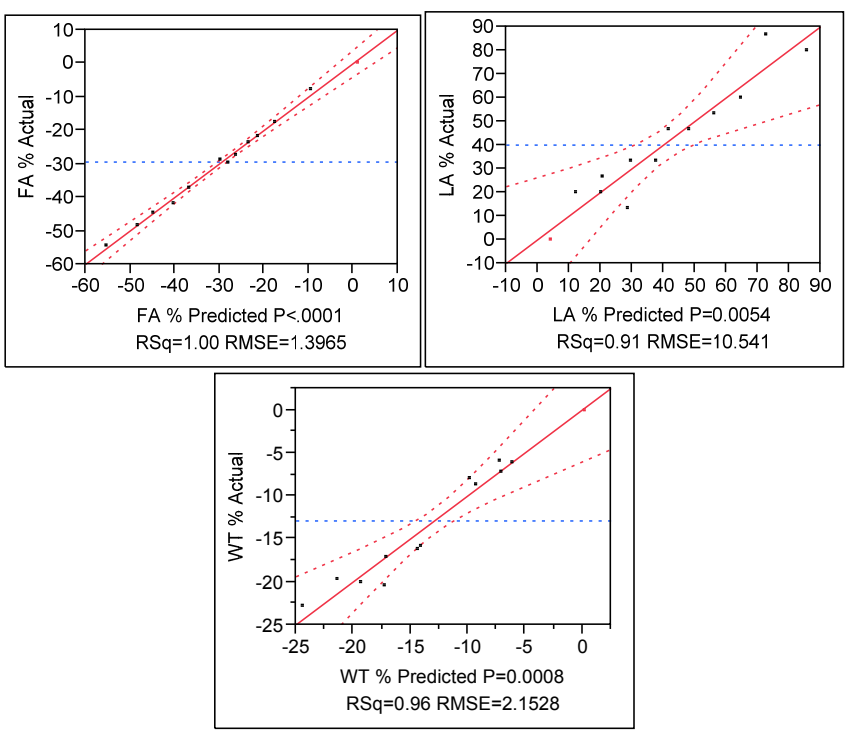


Figure 21. Traditional FA %, LA %, and WT % metrics in the refined model

The model also relies on the assumption that the random error $\varepsilon_{n(ijk)}$ is normally distributed. The residual errors should therefore be tested for normalcy. Figure 22 shows normal probability plots, box plots, and histograms fitted to a normal density curve. Although the data set is sparse, the histograms and box plots do trend toward a symmetric distribution.

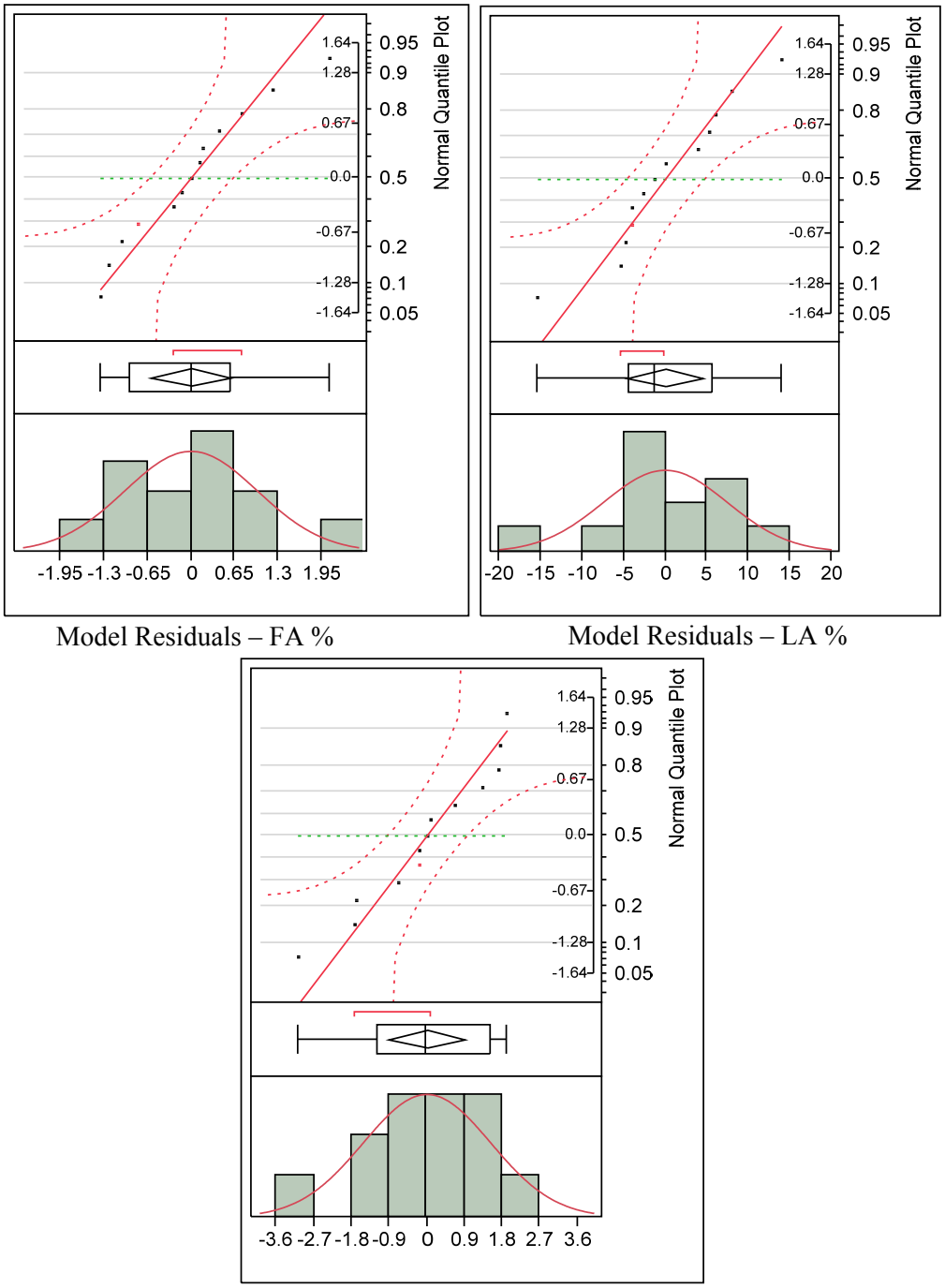


Figure 22. Residual Error Distributions for the Response Variables from the Refined Model.

Also, the normal probability plots illustrate that for each response, the model errors fall within the confidence interval along the diagonal line of the plot, indicating that each residual is at least approximately normally distributed. This provides evidence that the model is indeed appropriate.

Table 31 shows that all of the remaining interactions are significant to FA %, so the model cannot be further refined by removing interactions. Lateral conformance bounds and their squares are the only interactions that are not significant for LA %, although they have low enough *p*-values to be of interest. The two-way interactions are not significant for WT %, though the square of longitudinal conformance bounds has a small *p*-value and is of interest, while the square of the lateral conformance bounds is not of much interest for WT %.

Table 31. Summary of Refined Model Effect Tests for Response Variables.

Source	DF	FA % P-Value	LA % P-Value	WT % P-Value
Lat(0.5,2.5)	1	<0.0001*	0.2754	0.0003*
Long(0.5,1.5)	1	<0.0001*	0.0006*	0.0017*
Likelihood	2	<0.0001*	0.0423*	0.0020*
Lat*Lat	1	0.0365*	0.2528	0.7464
Long*Long	1	0.0401*	0.0447*	0.1203

Since it is unknown at this point how the fuzzy counts will interact with the model, the initial model will be used for the starting point again. Since WT % is the same as the first model, it is already known that it would be insignificant with the initial model, though the fuzzy variables are found to both be insignificant as well (Figure 23). Table 32 and Table 33 show the effect tests for the fuzzy responses and can be referred to for the WT % values. It seems as one of Lat*Like and Long*Like should be removed, which will actually be enough for all effects to be significant, though removing both effects greatly improves confidence in WT % for little degradation of the other responses, much like the reasoning with the traditional LA %, both interactions will be removed. The new model is therefore Eq.30.

Response:

$$R_{ijk} = \mu + Lat_i^2 + Lat_i + Long_j^2 + Long_j + Lat_i Long_j + Like_k + \varepsilon_{n(ijk)} \qquad \textbf{Eq.30}$$

Where:

Lat_i = lateral conformance bounds in nautical miles, i = 0.5, 1.5, 2.5
$Long_j$ = longitudinal conformance bounds in nautical miles, j = 0.5, 1.0, 1.5
$Like_k$ = likelihood, k = “10/20”, “3/8/10”, “3/8/20”
$\varepsilon_{n(ijk)}$ = random error, n = 1, 2, … for all i, j, k

In the fuzzy model, it is noticed that the interaction of the lateral and longitudinal conformance bounds must now be included in the model.

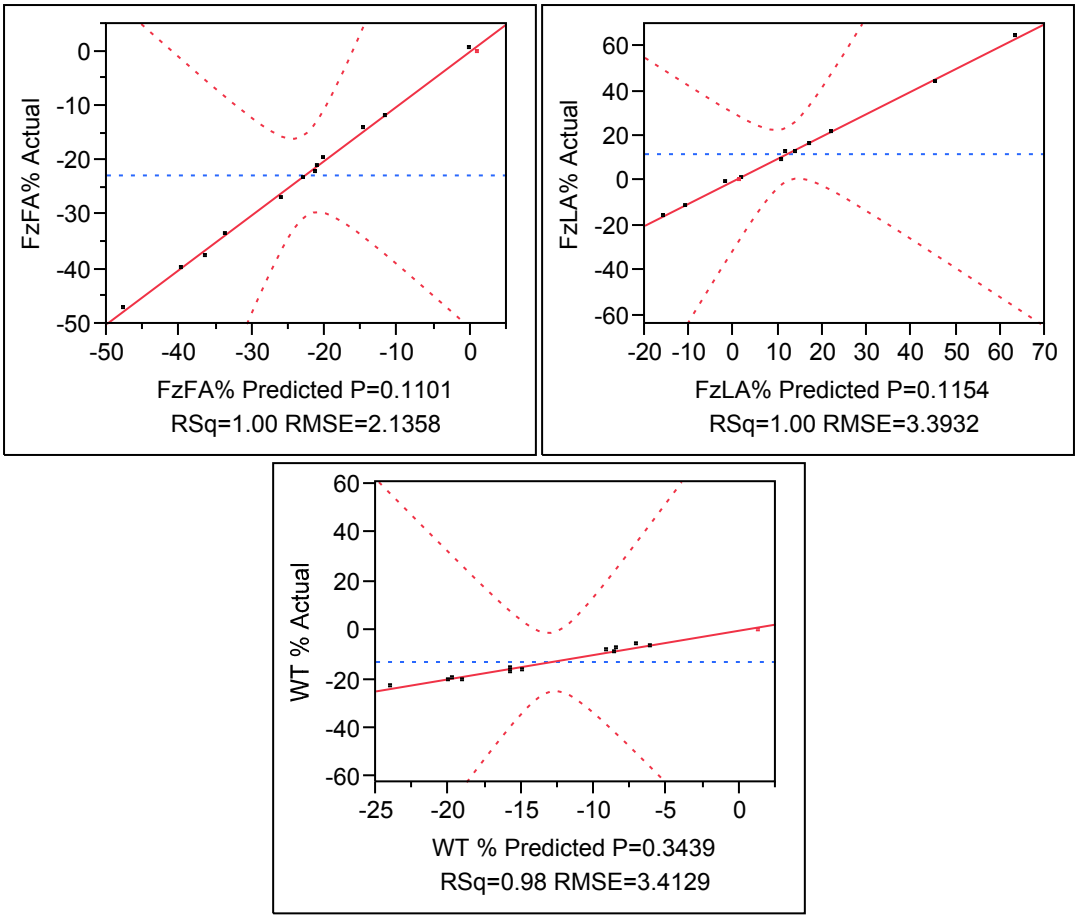


Figure 23. Fuzzy FA %, LA %, and WT % metrics leverage plots.

Table 32. Initial Model Effect Test for the Fuzzy FA % Response Variable.

Source	DF	Sum of Squares	F Ratio	Prob > F
Lat(0.5,2.5)	1	787.4953	172.627	0.0484
Long(0.5,1.5)	1	129.8698	28.4688	0.1179
Likelihood	2	1460.3960	160.0668	0.0558
Lat*Long	1	15.0463	3.2983	0.3204
Lat*Likelihood	2	0.3091	0.0339	0.9678
Long*Likelihood	2	19.2427	2.1091	0.4378
Lat*Lat	1	0.2271	0.0498	0.8602
Long*Long	1	11.4582	2.5117	0.3583

Table 33. Initial Model Effect Test for the Fuzzy LA % Response Variable.

Source	DF	Sum of Squares	F Ratio	Prob > F
Lat(0.5,2.5)	1	1098.7736	95.4297	0.0649
Long(0.5,1.5)	1	2273.4201	197.4491	0.0452
Likelihood	2	125.8989	5.4672	0.2895
Lat*Long	1	683.0919	59.3273	0.0822
Lat*Likelihood	2	124.6678	5.4138	0.2908
Long*Likelihood	2	15.4549	0.6711	0.6534
Lat*Lat	1	574.0277	49.8549	0.0896
Long*Long	1	81.3805	7.068	0.229

The refined fuzzy model gives the leverage plots in Figure 24.

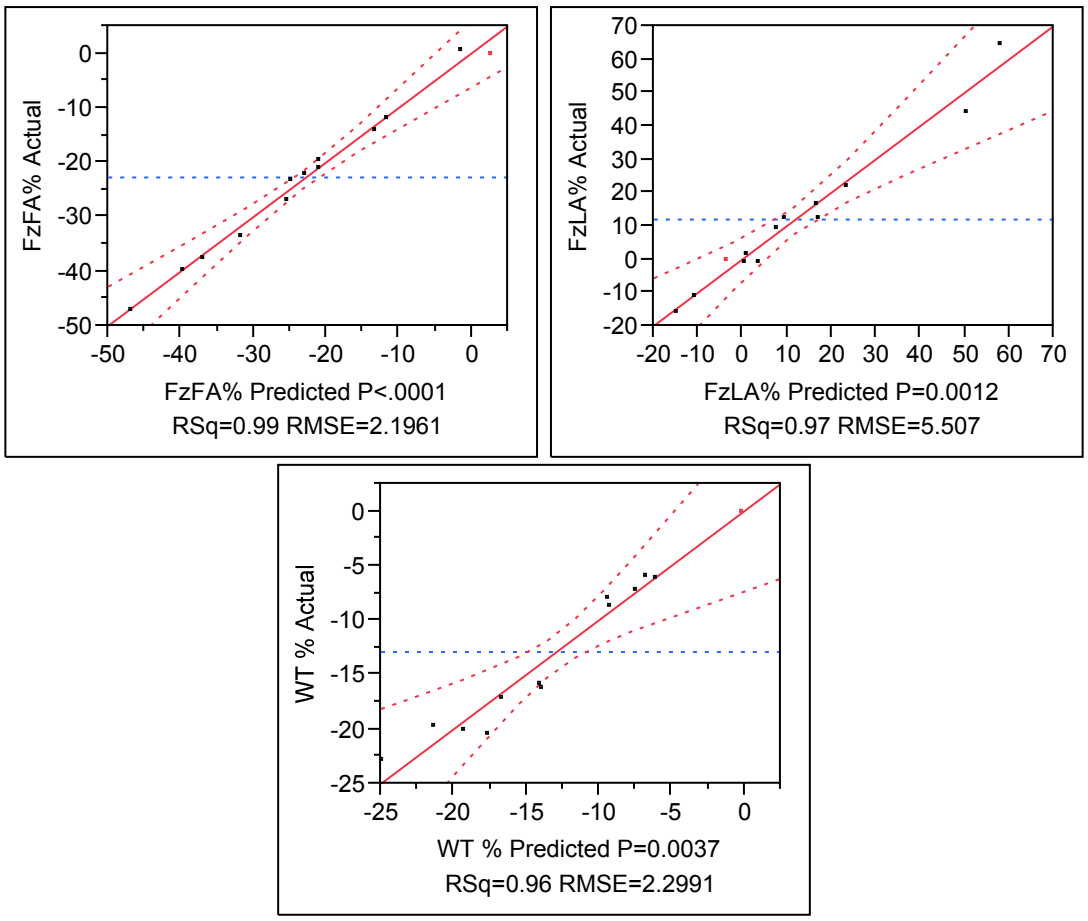


Figure 24. Fuzzy FA %, LA %, and WT % metrics in the refined model.

The fit on the fuzzy metrics is overall better than the fit on the traditional metrics, both with similar *p*-values. It can be seen that the range of the R^2 values in Figure 24 is 96% to 99%. Again, the random error is tested for normalcy, where it is already known from Figure 22 that the residual error for the WT % response has evidence for normalcy.

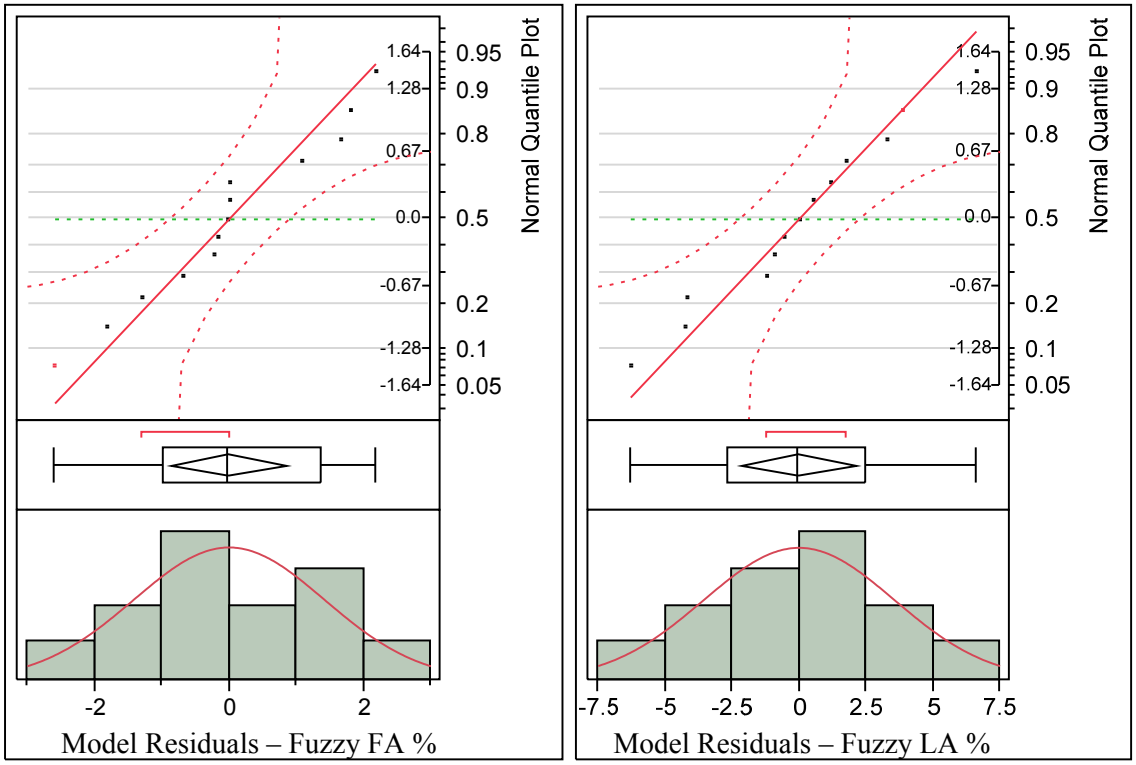


Figure 25. Residual Error Distributions for the Fuzzy Response Variables from the Refined Model.

Figure 25 shows nice distributions on the histograms, especially for Fuzzy LA % and the box plots show decent symmetry. The best evidence here is the normal probability plots, which show that the fuzzy response variables fit within the confidence interval around the diagonal suggesting they approximate normal error, thus the fuzzy responses have good evidence of having normally distributed error as well so the refined fuzzy model is claimed to be appropriate.

Table 34 summarizes the *p*-values for the significance of interaction effect tests on the factors of the refined fuzzy model. For fuzzy FA %, the effect of squared lateral conformance bounds is insignificant, also the rest of the two-way interactions are not significant but have small enough *p*-values to be interesting. The squared longitudinal conformance bounds and the main effect of likelihood are not significant for fuzzy LA % although have small enough *p*-values to be interesting. None of the two-way interactions are significant for WT % and only the square of longitudinal conformance bounds has a small enough *p*-value to be interesting. Long*Long is the only effect that is not significant for any of the responses, but has a small enough *p*-value that there is not enough justification to remove it from the model.

Table 34. Summary of Refined Fuzzy Model Effect Tests for Fuzzy Response Variables.

Source	DF	Fz FA % P-Value	Fz LA % P-Value	WT % P-Value
Lat(0.5,2.5)	1	<0.0001*	0.0013*	0.0010*
Long(0.5,1.5)	1	0.0040*	0.0003*	0.0039*
Likelihood	2	<0.0001*	0.2207	0.0050*
Lat*Long	1	0.1376	0.0051*	0.6316
Lat*Lat	1	0.8368	0.0074*	0.7640
Long*Long	1	0.1839	0.1623	0.1509

3.2.2.2.2 Examine Model Results

Figure 26 shows the prediction profiler with the parameters set to the baseline values. This figure provides an idea of the model accuracy and trends in responses from each factor. By investigating the tails of the histograms in Figure 22, it can be shown that the maximum unsigned residuals are 2.08 for FA %, 15.33 for LA %, and 3.08 for WT %. These were caused by Runs 8, 6, and 1 respectively. These percentages convert into approximately 29.0 FAs, 2.3 LAs, and 10.2 seconds of warning time. These bounds are considered acceptable since we are already aware of the large uncertainty associated with the LA % response and the most this error will cause is an error of 2.3 LAs. We are also willing to accept an uncertainty of 29 FAs and 10 seconds of warning time for the purposes of this study.

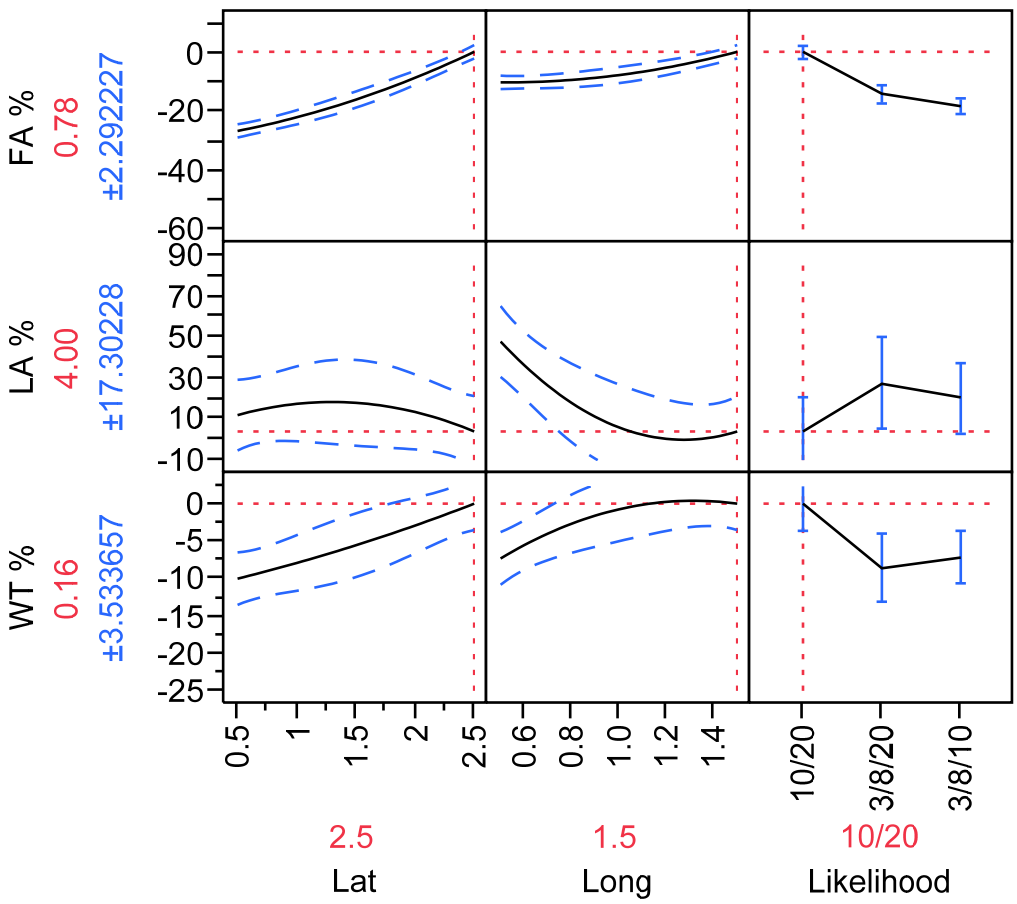


Figure 26. Prediction profiler of the model with baseline parameter settings.

The metrics used in the prediction profiler are the percent differences from the FA32 baseline of warning time (WT %), Late Alert count (LA %), and False Alert count (FA %). These should not be confused with the "%FA" and "%LA" metrics from Table 24 and Table 25. The metrics here are calculated with the following formula:

$$d = \left(\frac{r-b}{b}\right) \cdot 100$$

Where *r* is the FA count, LA count, or 25th percentile of warning time for one of the treatment runs, and *b* is the respective value for the FA32 baseline. The desire of the model is to decrease FA and LA count and increase warning time, so smaller values of FA % and LA % are desirable, whereas a larger value of WT % is desirable.

Typically in an integrated experiment, the model is used in operations research in order to determine the optimal settings of the factors defined in the model. In this experiment, there is no simple definition of what is optimal since all responses cannot be optimized together. We instead try to determine the effects of the different settings in order to make recommendations. The first step is to determine the optimal settings for each metric independent of the other metrics.

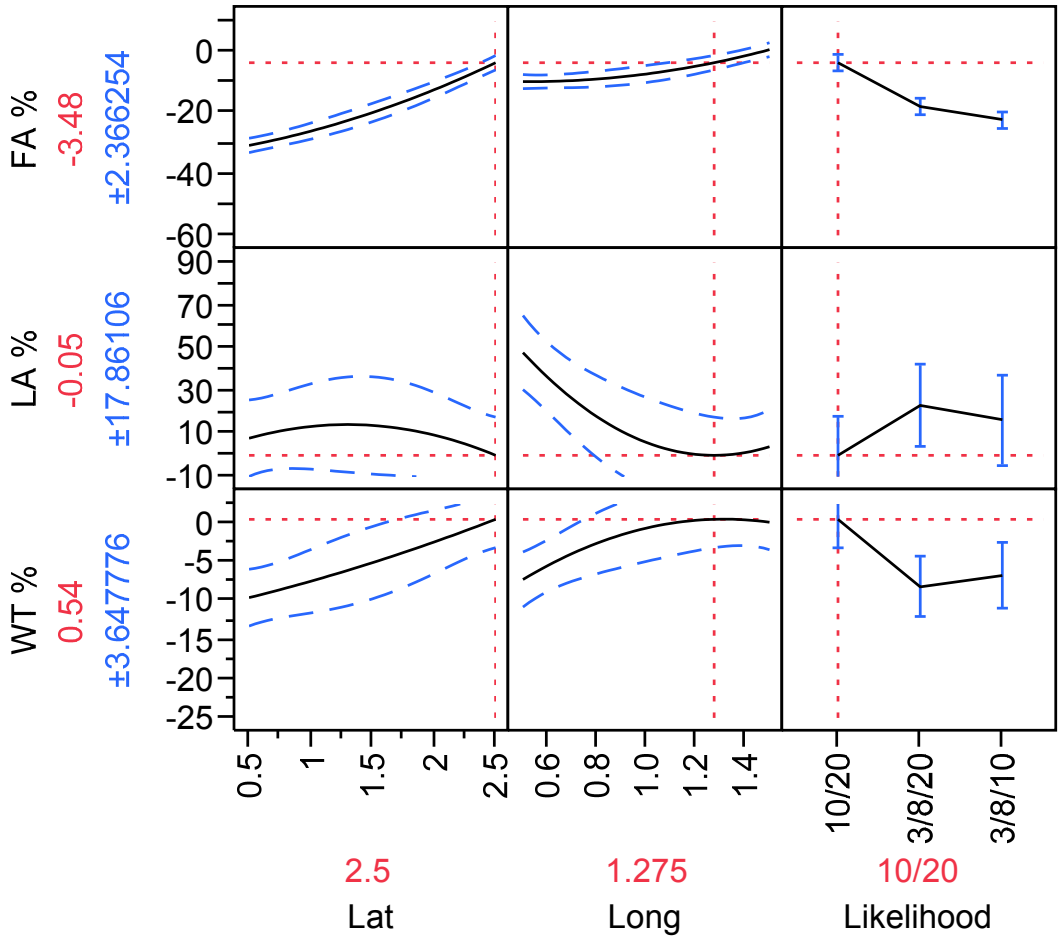


Figure 27. Optimal settings of the conflict probe to minimize the LA count ignoring all other response variables.

Figure 27 shows the optimal parameter settings of the conflict probe in order to minimize Late Alerts, without considering any other response variables. As expected, given the earlier analysis, there is very little change in the parameters from the baseline settings. The model does suggest a small reduction in the longitudinal conformance bounds to 1.275 nm. This may be due only to error in the model, since it is only a decrease of 0.0075, which is within the error of the model that we know to be at most 2.3. The part of the model between 1.0 and 1.5 longitudinal shows very little change in the effect on LAs and could use some further investigation to determine if that is truly the case.

The optimal settings to decrease the LA % show a decrease of -0.05%. Compare this value to the modeled baseline at 4.00%. There also is a slight decrease shown in the FA % by -3.48%. Finally, there is an increase in WT % by 0.54%. These parameter settings improve upon all responses, though the effects are very slight compared to those by other settings.

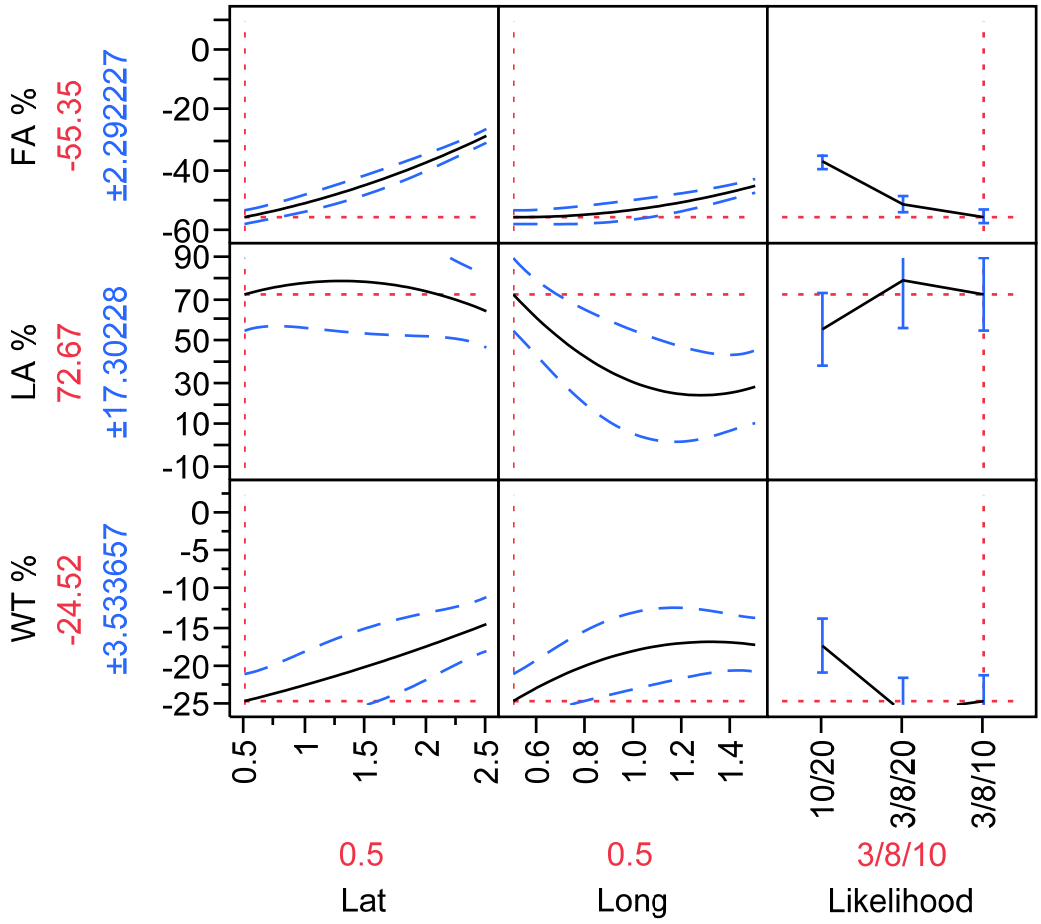


Figure 28. Optimal settings of the conflict probe to minimize the FA count ignoring all other response variables.

Figure 28 shows the optimal settings of the CP to minimize FA % ignoring the effects on all other response variables. These settings are the opposite extreme of what the original baseline settings are. Lateral and longitudinal conformance bounds are both set to 0.5 nm, and the likelihood is set to 3/8/10, which was considered to be the most extreme change from the default 10/20 setting.

In contrast to the settings for optimizing LA %, these settings for optimizing FA % have a major effect on all response variables, though a negative effect on LA % and WT %. False Alerts are decreased by a very significant 55% with these settings, but Late Alerts are increased by 73% and warning time is decreased by nearly 25%.

Figure 29 shows the optimal settings to increase the Warning Time. This is not the goal of these parameter changes, but rather the goal is to not adversely affect the warning time while improving the FA and LA performance. This figure is included simply to show the effects of the parameters on the warning time.

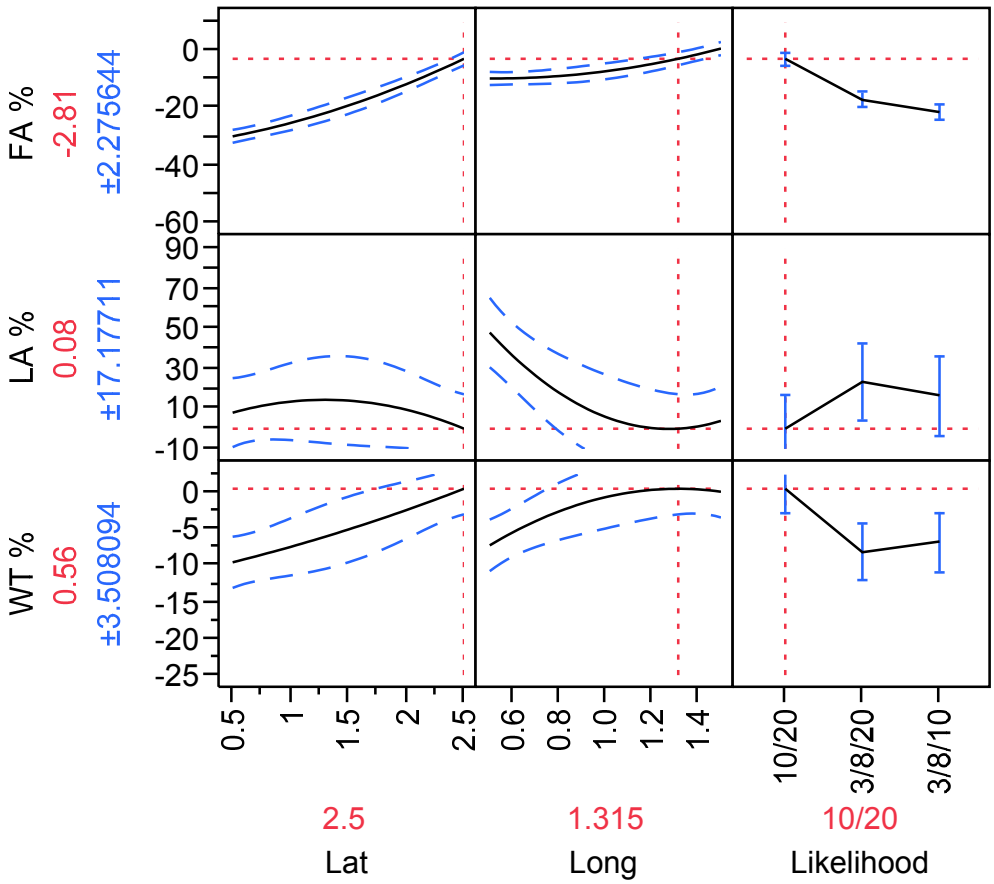


Figure 29. Optimal settings of the conflict probe to maximize the warning time.

The parameter settings for optimal warning time are not too different from the baseline settings nor are they very different from the settings for optimal LA %. This is understandable since a higher warning time usually coincides with less Late Alerts.

These three different settings give an idea of how the factors affect the three responses being analyzed. The False Alerts are affected positively by reducing the lateral or longitudinal, and by changing to either of the new likelihood settings. The Late Alerts and warning time follow the same trends and are affected negatively by reducing the lateral or by changing the likelihood. They are also affected negatively when the longitudinal is reduced too far.

Next we want to look at the interactions of the factors in regards to each of the response variables. The interaction profiler plot will give an idea of how changing the value of each factor affects how the other factors affect the response.

Figure 30 shows the interactions of the factors' effects on the FA % response variable. Each cell of this plot is a graph of the response by the factor listed in its respective column. The multiple curves in each cell represent the different settings of the factor listed in its respective row. Each of the curves is labeled by which setting it is. The two continuous factors have curves for the minimum and maximum settings used, whereas likelihood has a curve for each of the three nominal values it was set to. If there is no interaction between the factors listed in the respective row and column of the cell, then each of the curves will be the same shape, only shifted up or

down. If the curves cross or change shape, this indicates there is some interaction between the two factors.

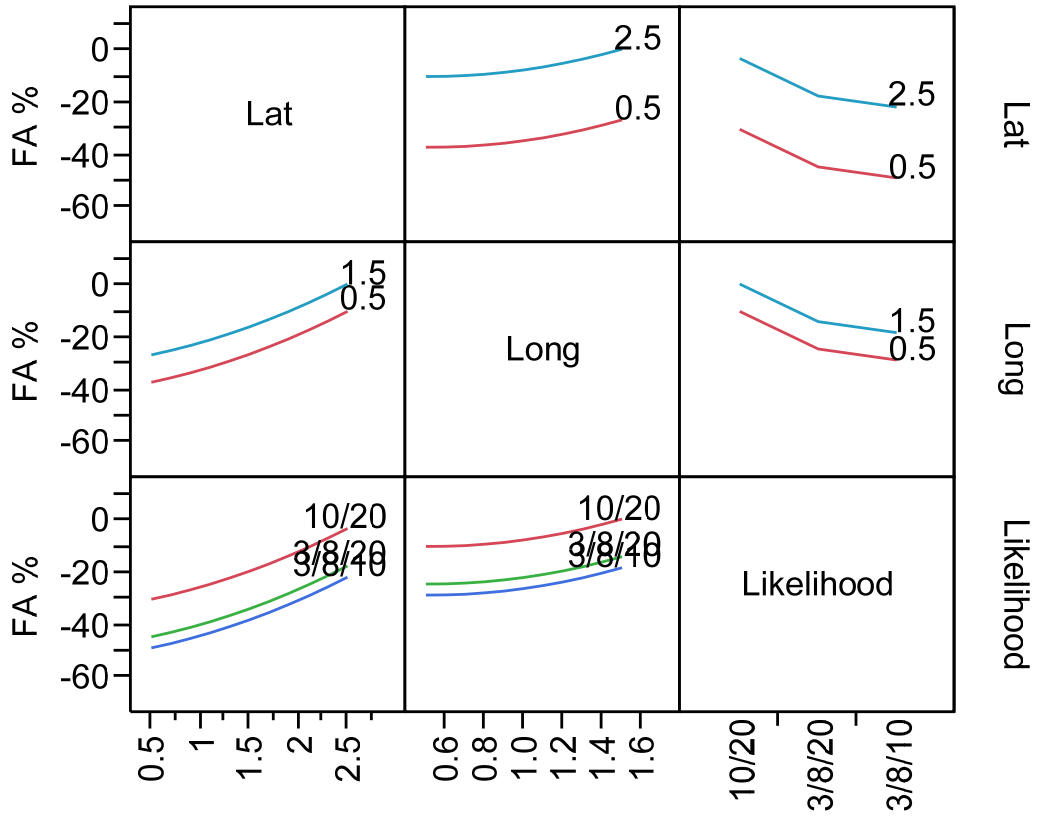


Figure 30. Interaction profile of the factors effects on the FA % response variable.

In Figure 30 we can see there is no interaction between any two factors. We expect this since we removed all two-way interactions from the model. Having no major interactions for this response gives the advantage of being able to change the setting of a single parameter to reduce the False Alerts without being concerned about the settings of the other parameters.

Figure 31 shows the interaction profile of the factors' effects on the LA % response variable. Again there is no interaction as expected.

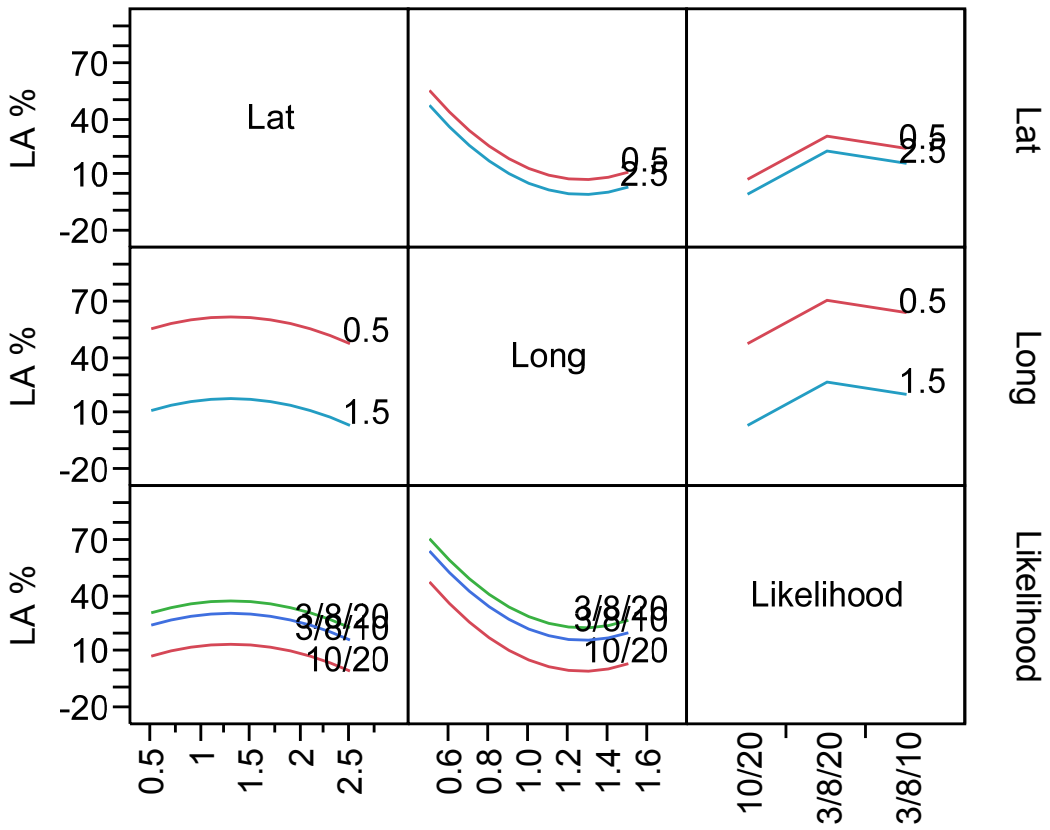


Figure 31. Interaction profile of the factors' effects on the LA % response variable.

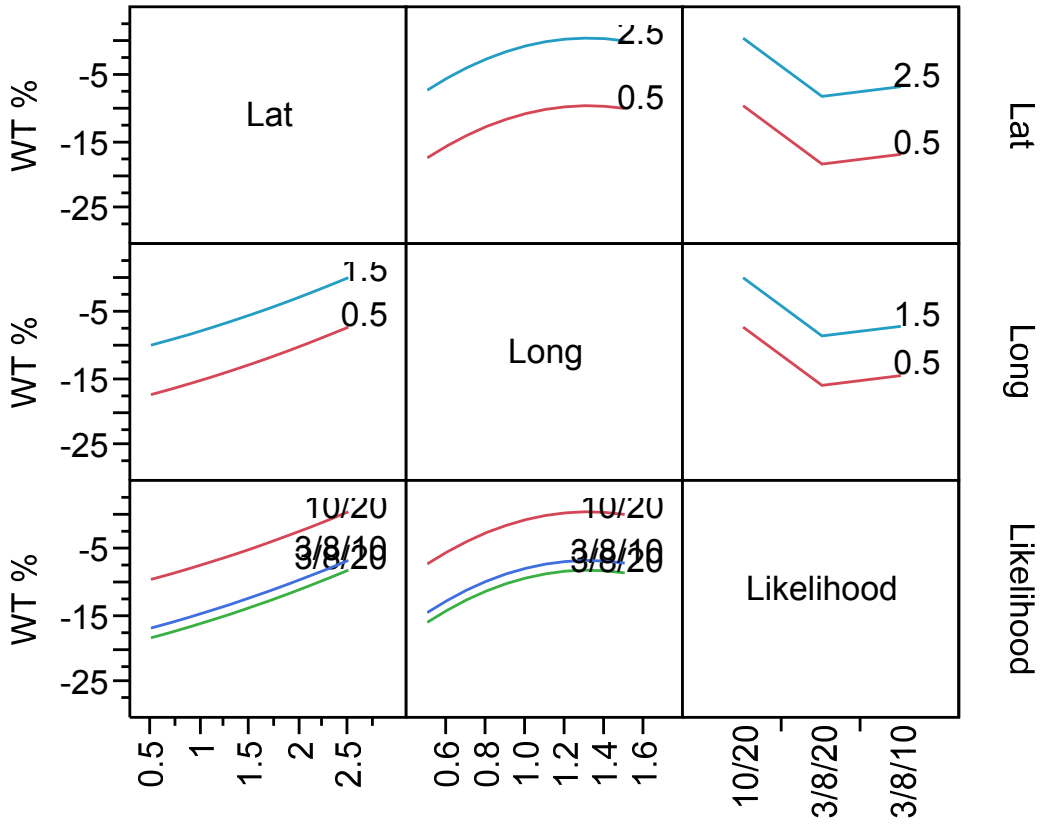


Figure 32. Interaction profile of the factors' effects on the WT % response variable.

Finally, Figure 32 shows the interaction profile of the factors' effects on the WT % response variable. Still the interactions are not existent as expected since no two-way interactions remain in our model.

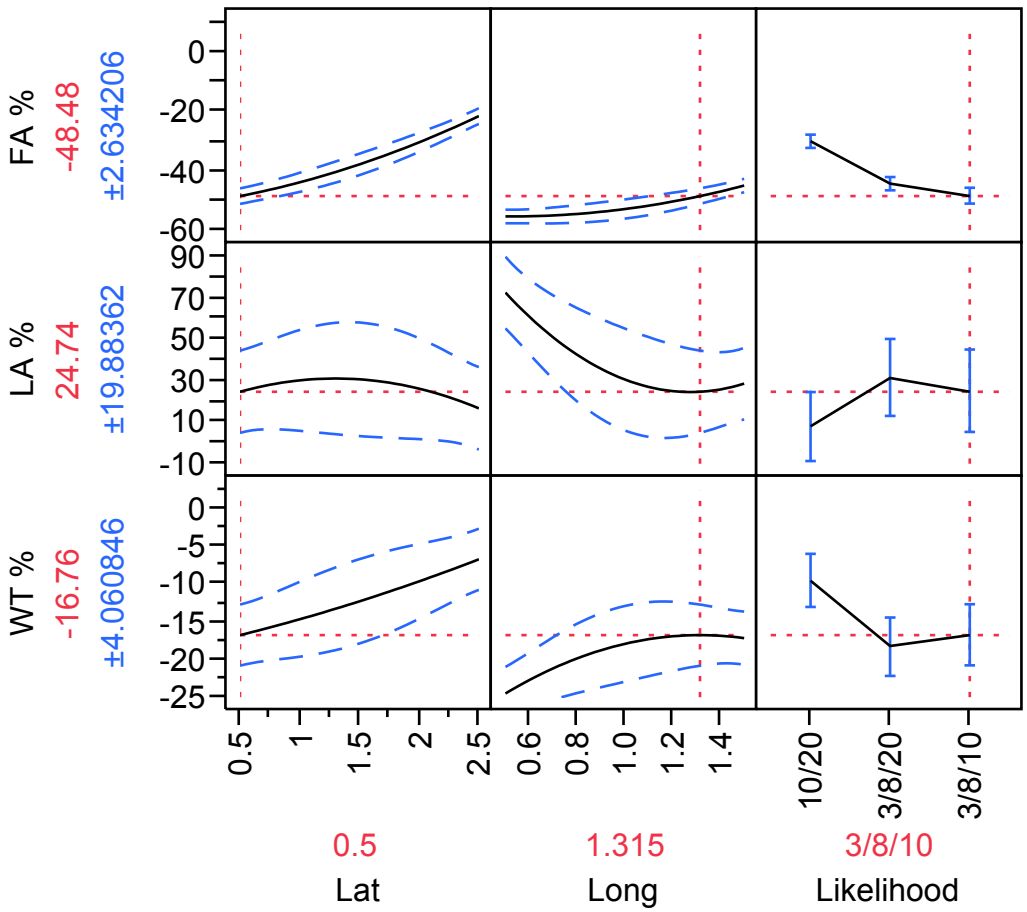


Figure 33. Prediction profiler with settings that are a combination of the optimal values for each response.

Now that the optimal settings for each of the responses have been found and determined to be very different from one another, the prediction profiler can be used to vary the settings based on what has been learned so far. Figure 33 shows one option of combining what has been learned to generate parameter settings. The lateral and likelihood were determined to have a major effect on the False Alerts whereas the longitudinal was determined to have a major effect on the Late Alerts and warning time. These settings combine the optimal longitudinal for warning time, which was also very close to the optimal for the Late Alerts, with the optimal lateral and likelihood for the False Alerts. The settings provide an over 48% decrease to false alerts, with a nearly 25% increase to Late Alerts and over 16% decrease in warning time. The absolute optimal settings for False Alerts only provided less than an 8% higher decrease to False Alerts, but cost 50% more Late Alerts and 8% less warning time.

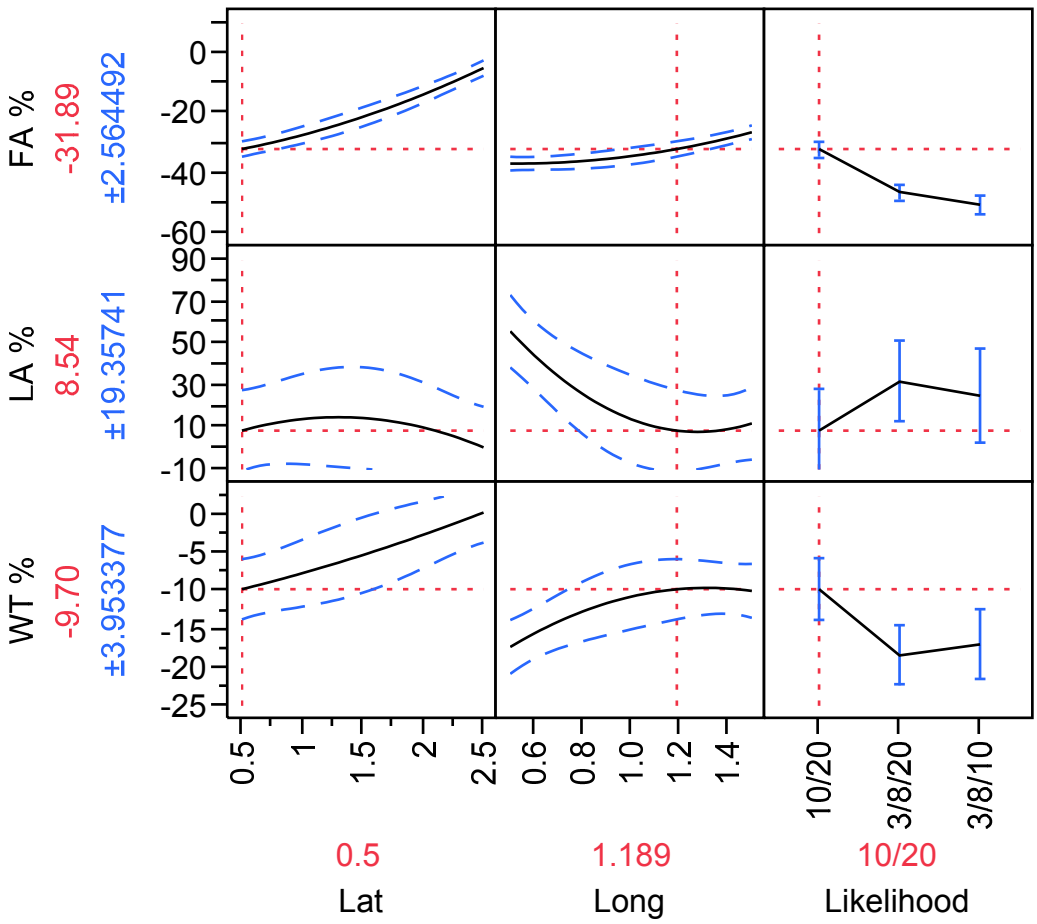


Figure 34. Prediction profiler with a combination of optimal settings of the responses.

Figure 34 shows the prediction profiler with changing the lateral and longitudinal based on what has been learned, but leaving the likelihood at the baseline setting of 10/20. These settings do not provide as large a decrease to the False Alerts, but also do not increase the Late Alerts or decrease the warning time as much. These settings also would provide a performance increase without changing the likelihood, which may be a desirable result, since the likelihood may need to be investigated further before it should be changed.

The prediction profiler is also applied to the fuzzy metrics with the expectation of similar findings to the prediction profiler of the traditional metrics. Figure 35 shows the prediction profiler at baseline configuration. Decreasing the fuzzy FA percentage (FzFA%) and fuzzy LA percentage (FzLA%) is desirable, whereas increasing the warning time percentage (WT%) is desirable.

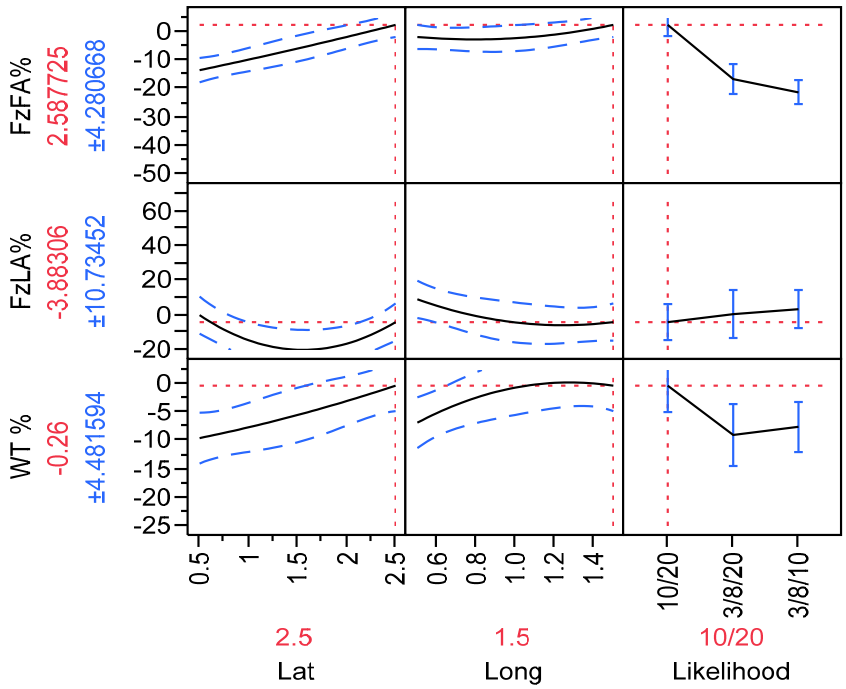


Figure 35. Prediction profiler of the model with baseline parameter settings.

Figure 36 gives optimal settings for late alert reduction only. These settings are different because the late alerts curve by lateral conformance bounds has changed in the fuzzy metrics significantly. The prediction does suggest that decreasing only the lateral conformance bounds will positively influence the conflict probe's late alerts count down to about 1.5 nautical miles.

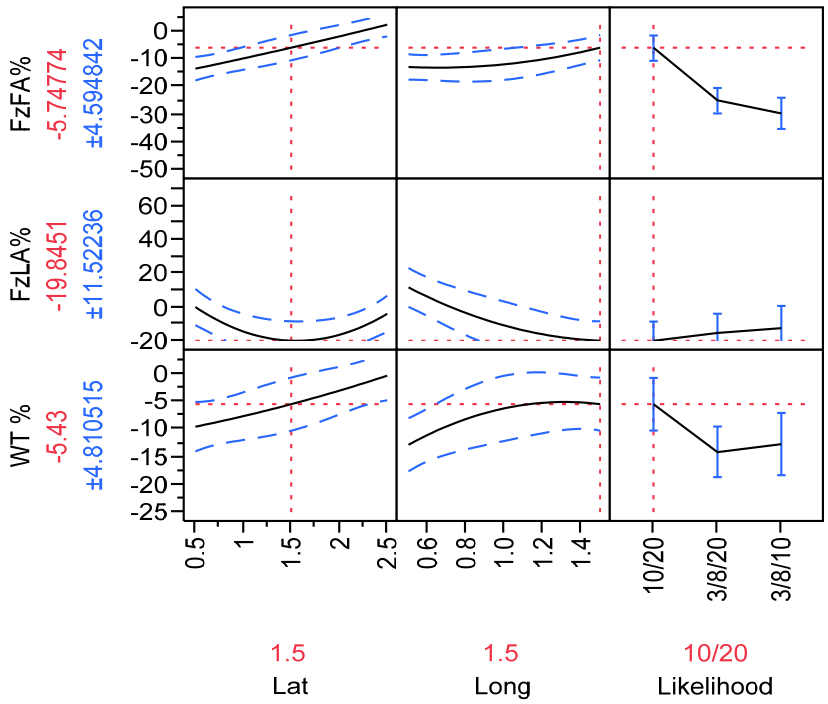


Figure 36. Optimal settings of the conflict probe to minimize the LA count ignoring all other response variables.

Figure 37 gives optimal settings for false alert reduction only. These settings are the same ones suggested by the traditional metrics, and comparing the trending of the false alert curves, it is apparent that the fuzzy metrics provide a similar prediction of false alerts compared to the traditional metrics. Figure 38 gives optimal settings for warning time maximization only. The same warning times are being used here (giving the same optimal settings). The optimal settings of warning time show very little change in fuzzy late alerts although the false alerts are still inversely correlated with the late alerts.

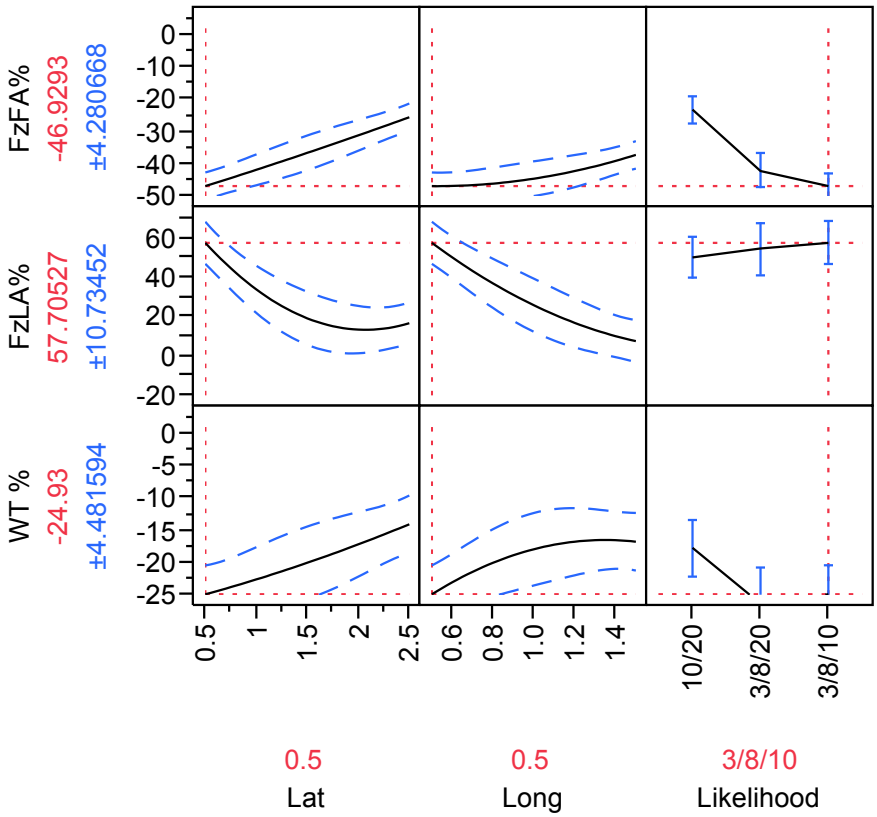


Figure 37. Optimal settings of the conflict probe to minimize the FA count ignoring all other response variables.

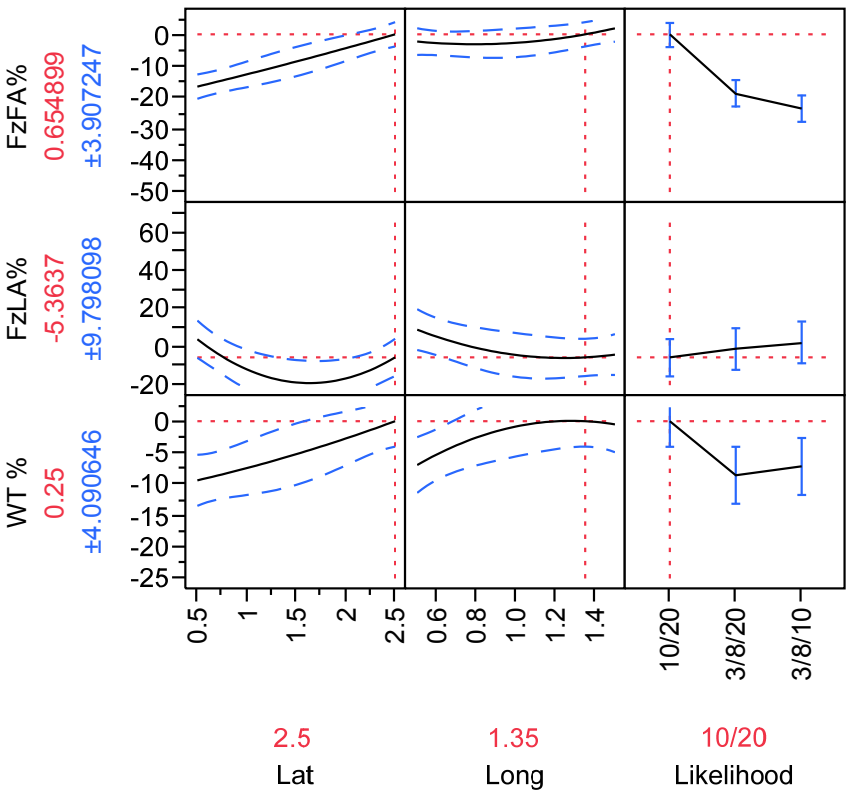


Figure 38. Optimal settings of the conflict probe to maximize the warning time.

Figure 39 still gives 3/8/10 as the best likelihood setting for reducing false alerts sacrificing warning, but not having as great an impact on late alerts. 1.35 nautical mile longitudinal conformance bounds are still optimal giving a benefit to both warning time percentage and fuzzy late alerts for a smaller reduction in fuzzy false alerts. Finally the 0.5 nautical mile setting for lateral conformance bounds now has the largest negative impact on fuzzy late alerts and warning time percentage, whereas 2.5 nautical mile conformance bounds have the best warning time percentage and very little impact on fuzzy late alerts. This setting was chosen since the total setting gives a 23% improvement to fuzzy false alerts for 6% loss of warning time and 2% loss of late alert performance. Sacrificing more warning time will improve both alert performances to 1.5 nautical miles (32% improvement to FA, 11% improvement to LA, and 12% loss of warning time) then late alert performance degrades quickly below 0.65 nautical miles (39% improvement to FA, 6% loss to LA, 16% loss of warning time).

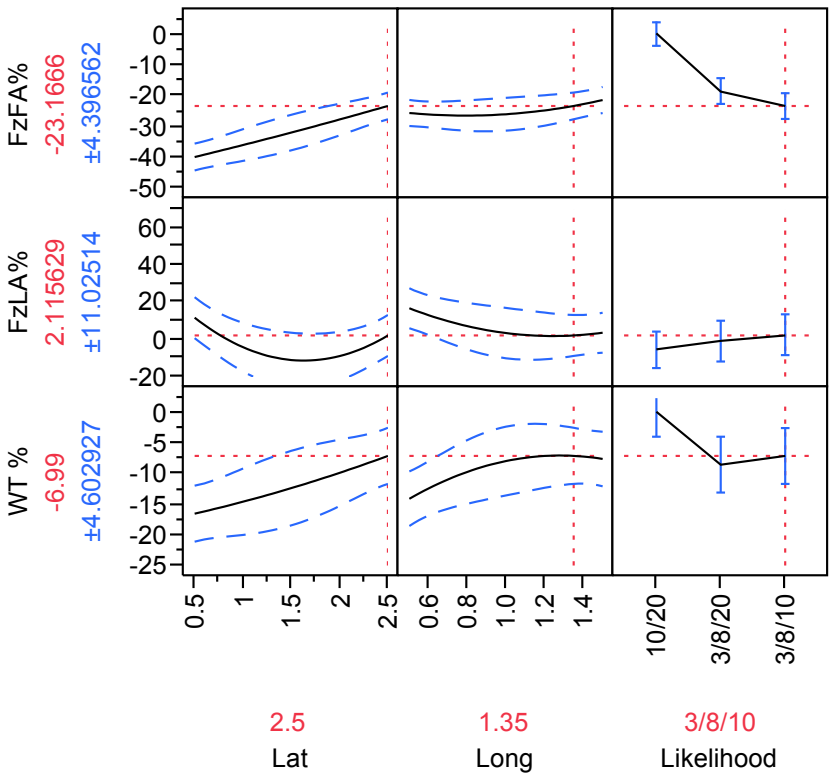


Figure 39. Prediction profiler with settings that are a combination of the optimal values for each response.

3.2.2.2.3 Factor Effects

In the previous sections, many discoveries were made as to the effects of each of the factors on each of the responses. Several of these effects were seen from many different perspectives through the model and through analysis of the discrete values.

A recurring theme through each analysis has been the effect of the longitudinal conformance bound. Early in the analysis it was determined that longitudinal had a major effect on the Late Alerts and warning time, and each subsequent analysis further confirmed this suspicion. Something that was discovered later, when looking at the results of the model, was that the longitudinal bound does not have nearly as large an effect on the False Alerts as it does the Late Alerts. Finally, the conformance bound has the largest effect on False Alerts between 1.5 and about 1.0 nm, and then the slope of the tangents to the curve become much flatter between 1.0 and 0.5 nm. Likewise, the effect of the longitudinal conformance bound on the False Alerts is not nearly as large as the lateral conformance bound's is.

The lateral conformance bound was determined to have a major effect on the False Alerts, decreasing them by over 26% when no changes are made to the other parameters. The lateral has almost no effect on the Late Alerts, increasing them by only 12% at the most, as opposed to the longitudinal that can increase them by up to 48%. Lateral does have a significant effect on the warning time, decreasing it by less than 10%, which is slightly higher than the 7% decrease caused by the longitudinal.

Finally, the likelihood can have a major effect on all responses. By changing only the likelihood the FAs can be decreased by 18%, the LAs can be increased by up to 27% and the warning time can be decreased by up to 9%. There are still many open questions with likelihood since only three different parameter settings were used. Both of the changes from the default decreased the minimum notification time to three minutes, which is also the required minimum warning time. Having this value right on the edge of the requirement may have caused many more Late Alerts than it reduced False Alerts, so it may be best to have a larger minimum in the likelihood function.

It is important to note the three factors' effects on the warning time. Although each of them significantly impact the warning time, the highest decrease by any single factor is only 10%, and the highest decrease overall is not more than 26%. This 26% is a significant decrease, but the baseline settings of ERAM were strategically chosen to increase the warning time above what was then the requirement of five minutes. So, although the warning time is decreased significantly at the lowered settings, the 25th percentile of warning time is still far above the three minute minimum warning time that was used as the requirement in this study.

3.3 Example Flights

A detailed statistical analysis has been provided in the previous sub-sections of Section 3 on the entire sample scenario of air traffic. This section is dedicated to an in depth description of a set of selected example flights, their flight plans in the NAS, their ERAM trajectory and conflict predictions, and their accuracy results. It illustrates how the performance evaluation is being applied on these individual flights and how the prediction errors manifest themselves in the metric calculations.

3.3.1 Flight Example 1

Flight A (red) is a McDonnell Douglas MD-83 equipped with RNP and RVSM (Code /Q) flying from Dallas/Fort Worth International Airport (KDFW) to Bradley International Airport (KBDL). In each figure, it is moving from lower left to upper right.

Flight B (green) is a Boeing 737-800 also with aircraft equipment code /Q flying from Fort Lauderdale-Hollywood International Airport (KFLL) to Newark Liberty International Airport (KEWR). In each figure it is moving from bottom left to top right.

These flights become sufficiently close to trigger a false alert with baseline settings while both are at the same altitude, level, and cruising. The dots represent the actual track points that each aircraft are flying. The disks are centered on each aircraft and have a 2.5nm radius (minimum separation requirement violated when disks intersect). The line path extending through each disk is the predicted trajectory of each flight created by the conflict probe. Finally the boxes (both sharp and rounded corners make a box) that are centered at a point along the trajectory, in this instance not far from the current time, represent the conformance boxes, which extend from the minimum separation box (5nm sides) by 2.5nm on either side laterally and 1.5 nm on either side longitudinally. The final dimensions of the boxes are therefore 10nm lateral x 8nm longitudinal. The intersection of a conformance box rectangle with sharp corners with a conformance box rectangle with rounded corners show pictorially what the conflict probe is measuring. The intersection as described gives the conflict probe's definition of separation violation as long as the flights have a horizontal separation not exceeding $2\sqrt{lat^2 + long^2} + hsep$ (e.g. $2\sqrt{2.5^2 + 1.5^2} + 5 \approx 10.83$ nautical miles). Keeping this slight limitation of the visualization in

mind, these boxes are very useful for understanding how the conformance boxes are interacting in each situation.

Figure 40 shows a conflict detection occurring 11 seconds after a trajectory rebuild by Flight B. The figure makes sense as being the point of first contact between the rectangle and the rounded-corner rectangle. The conflict detection ultimately creates a false alert, since Flight B passes safely behind Flight A.

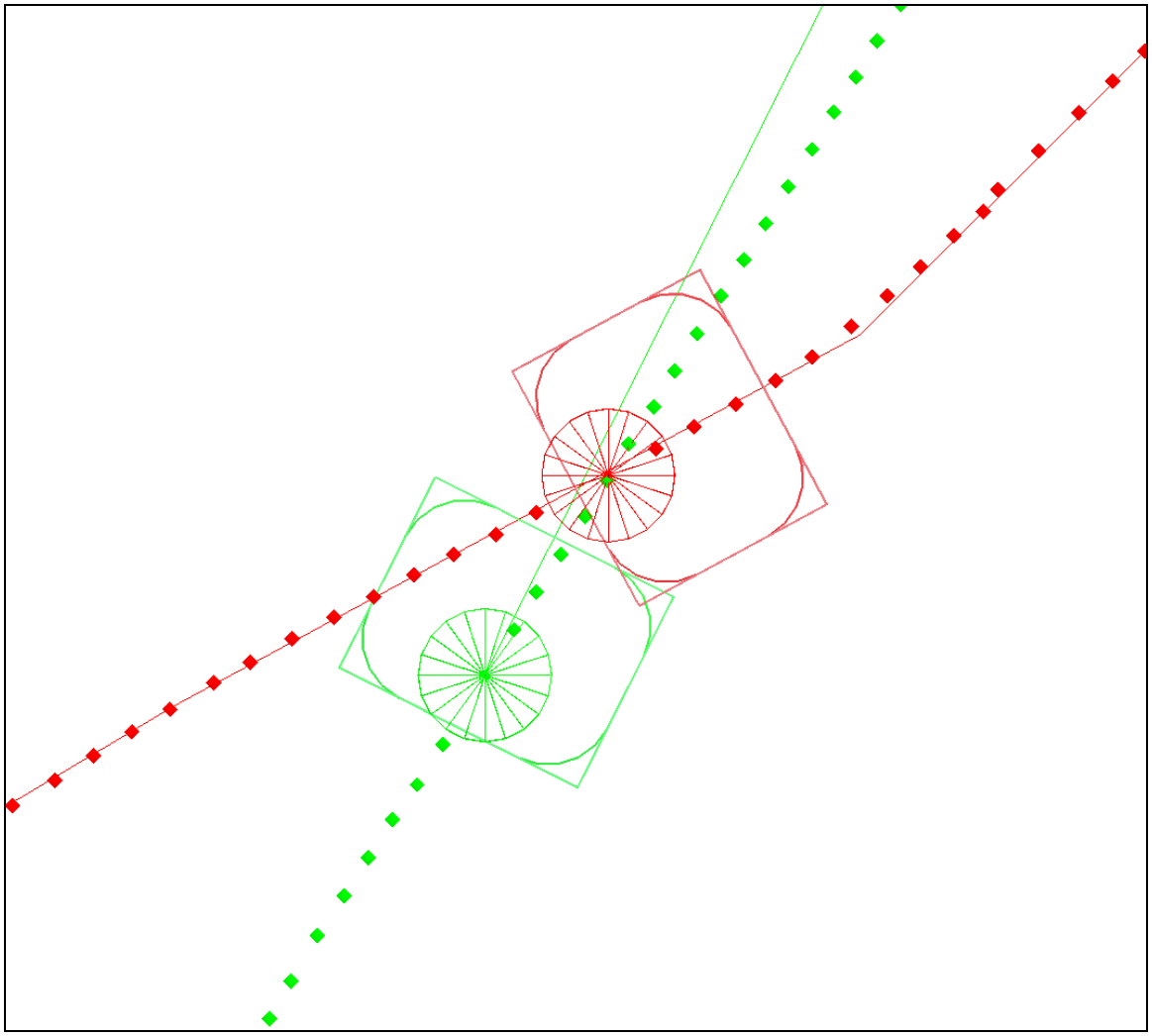


Figure 40. Baseline conflict detection (2.5 by 1.5 conformance bounds).

Figure 41 shows the effect on detection when the longitudinal bounds are reduced to the minimum tested value of 0.5nm. Here the estimated boxes do not even intersect at the same time. In fact no prediction was made because the conformance bounds are never violated here. The conflict prediction conformance bounds are coupled to the trajectory adherence bounds in these runs so the trajectories will potentially update more often, since speed differences will have a larger impact on the system with reduced longitudinal bounds. This can be seen by the more current trajectory being used by Flight A as compared to the baseline scenario.

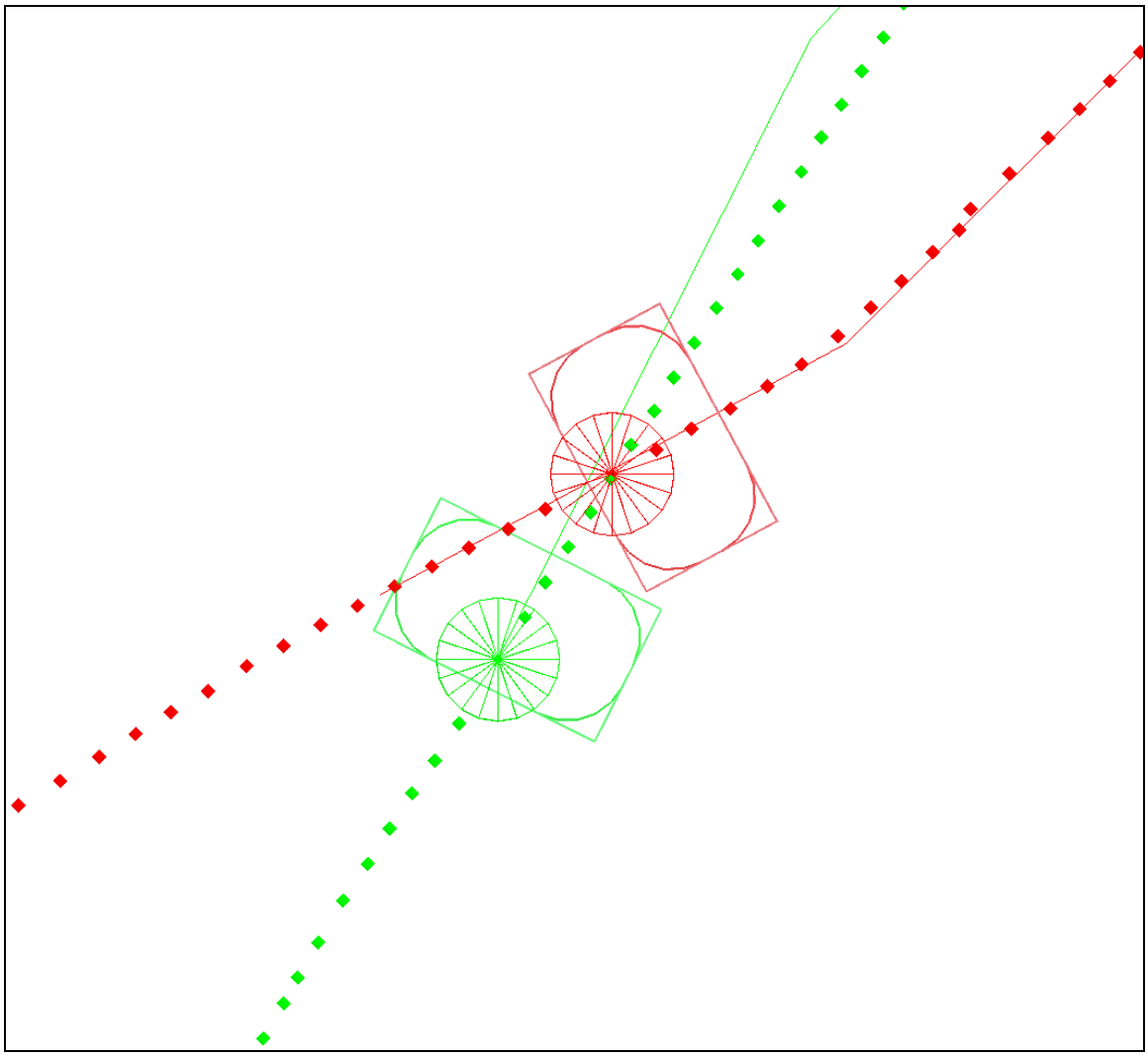


Figure 41. Reducing only longitudinal bounds to 0.5.

In Figure 42 only lateral conformance bounds were tightened, giving conformances boxes of 0.5 by 1.5 nautical miles. In this situation, the conflict prediction was not made, but it was close to being called, remembering that the pictorial conformance boxes intersecting must do so with a rectangle and a rounded-corner rectangle.

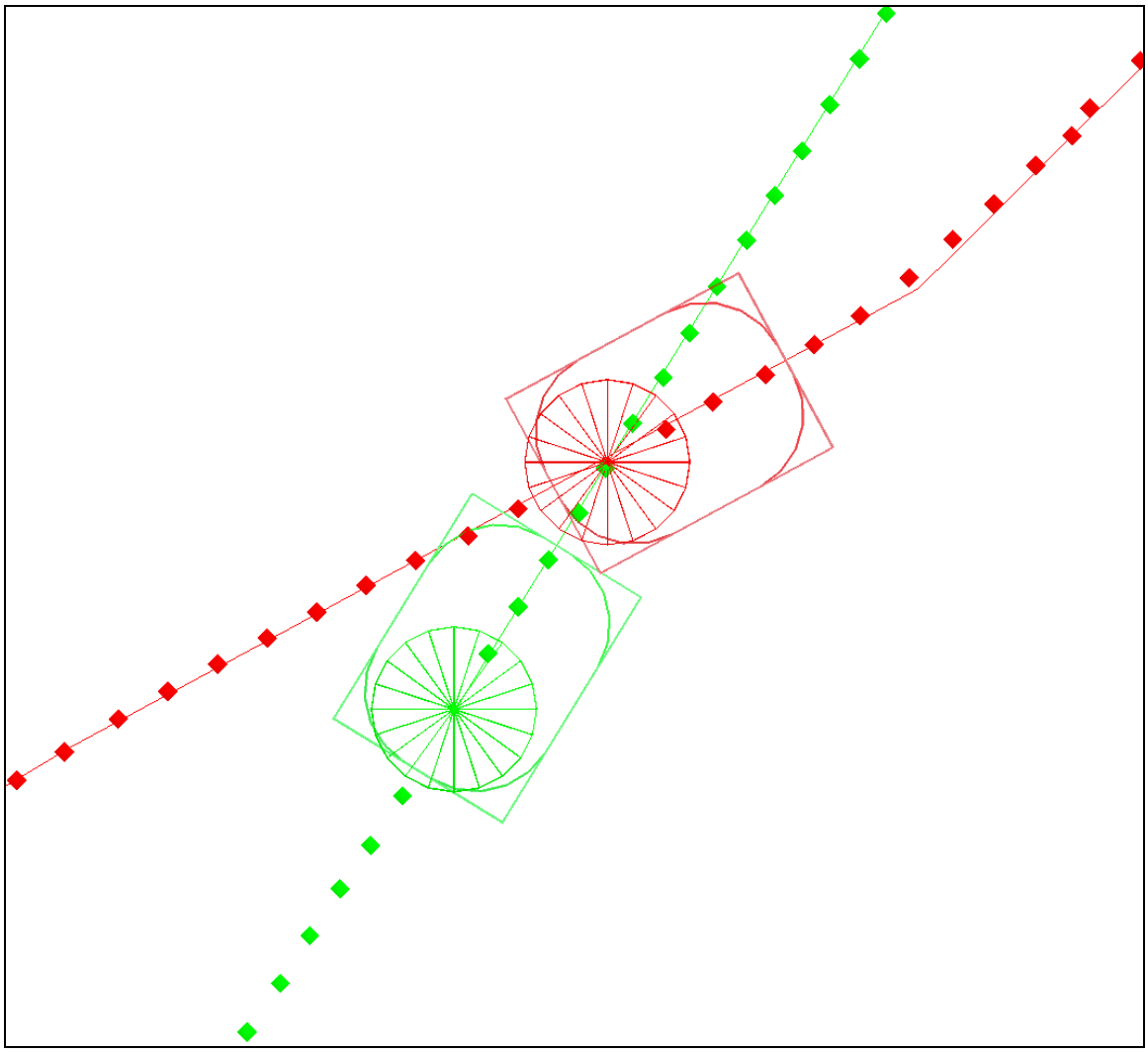


Figure 42. Reducing only lateral bounds to 0.5.

Figure 43 shows both conformance bounds reduced to mid-level values, being 1.5nm and 1.0nm for lateral and longitudinal conformance respectively. No prediction is made at these settings. The trajectory of Flight A is significantly older than that in Figure 41 due to the increase of longitudinal conformance bounds.

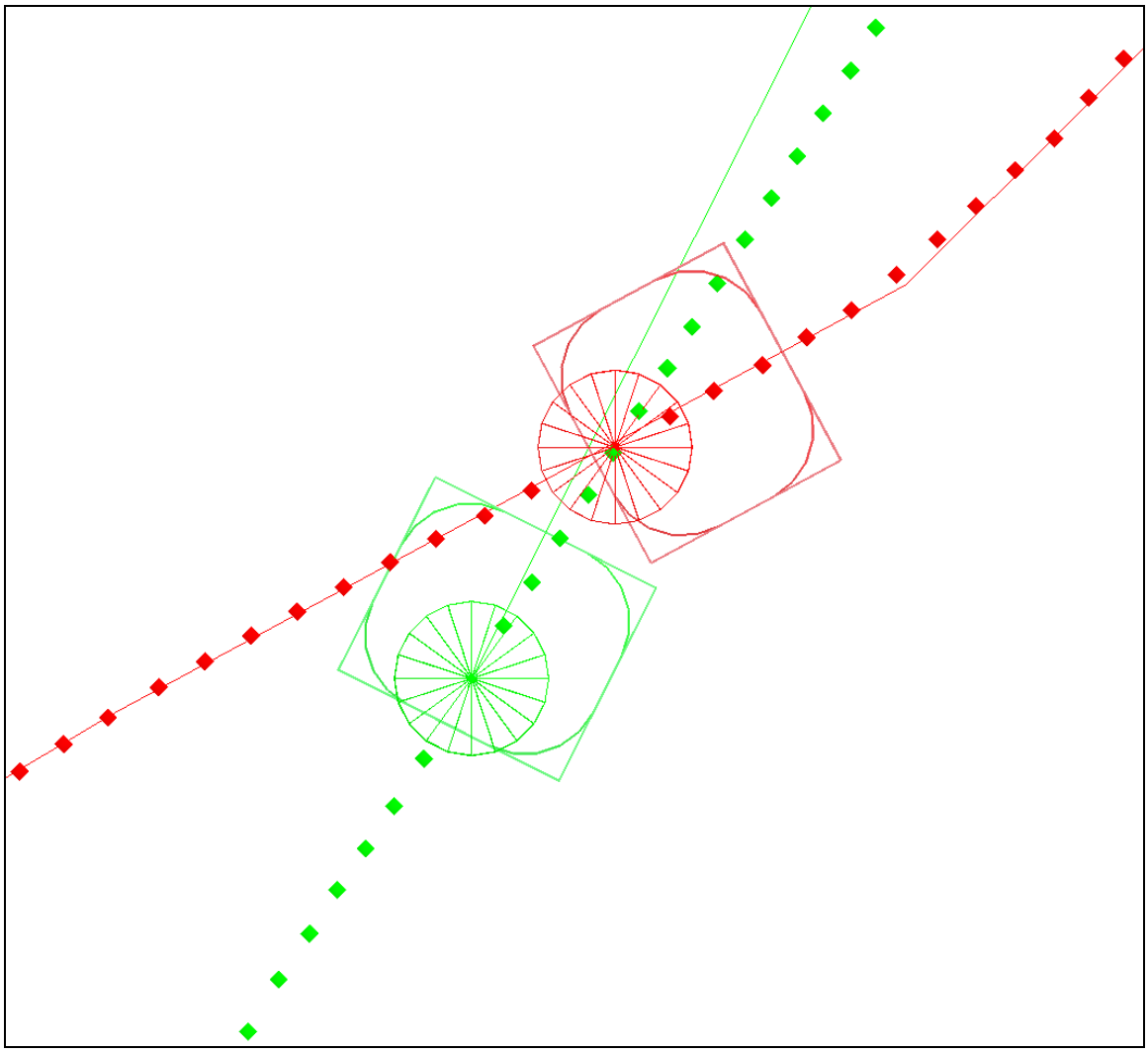


Figure 43. Mid-level conformance bounds (1.5 by 1.0).

Finally, Figure 44 shows the case that the conformance bounds are reduced to the minima being tested, 0.5 by 0.5 nautical miles. This of course offers the largest gap between the conformance boxes, but since these boxes are so small, one expects a reduction of false alerts to come at the price of loss of warning time and increase in missed alerts.

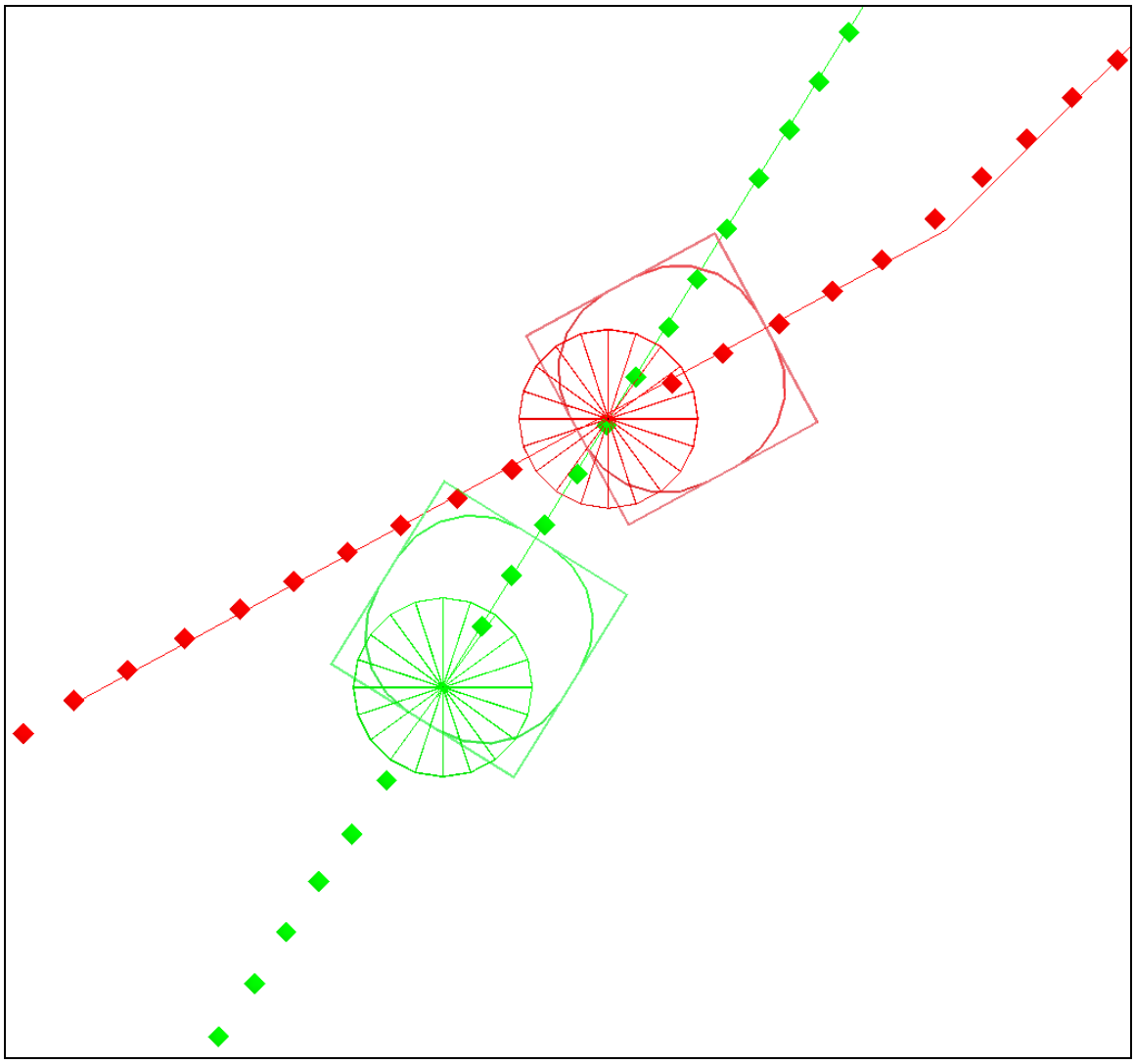


Figure 44. Low-level conformance bounds (0.5 by 0.5).

3.3.2 Flight Example 2

In the second example, reducing lateral conformance does not deter detection of a valid conflict, but reducing longitudinal conformance causes a missed alert.

Flight C (red) is an Airbus A320 with aircraft equipage \Q flying from Orlando International Airport (KMCO) to John F. Kennedy International Airport (KJFK). In each figure it is approaching the crossing point from the left.

Flight D (green) is an Airbus A310 with aircraft equipage \Q flying from Vilo Acuña Airport in Cuba (MUCL) to Montréal-Pierre Elliott Trudeau International Airport in Canada (CYUL). In each figure it is approaching the crossing point from the right.

These flights are at the same altitude, level, and cruising. Each figure shows the time of detection for each setting of the conflict probe. Figure 45 shows that the valid conflict detection was made with ample warning time, and this prediction must have persisted as it did not become a retracted false alert.

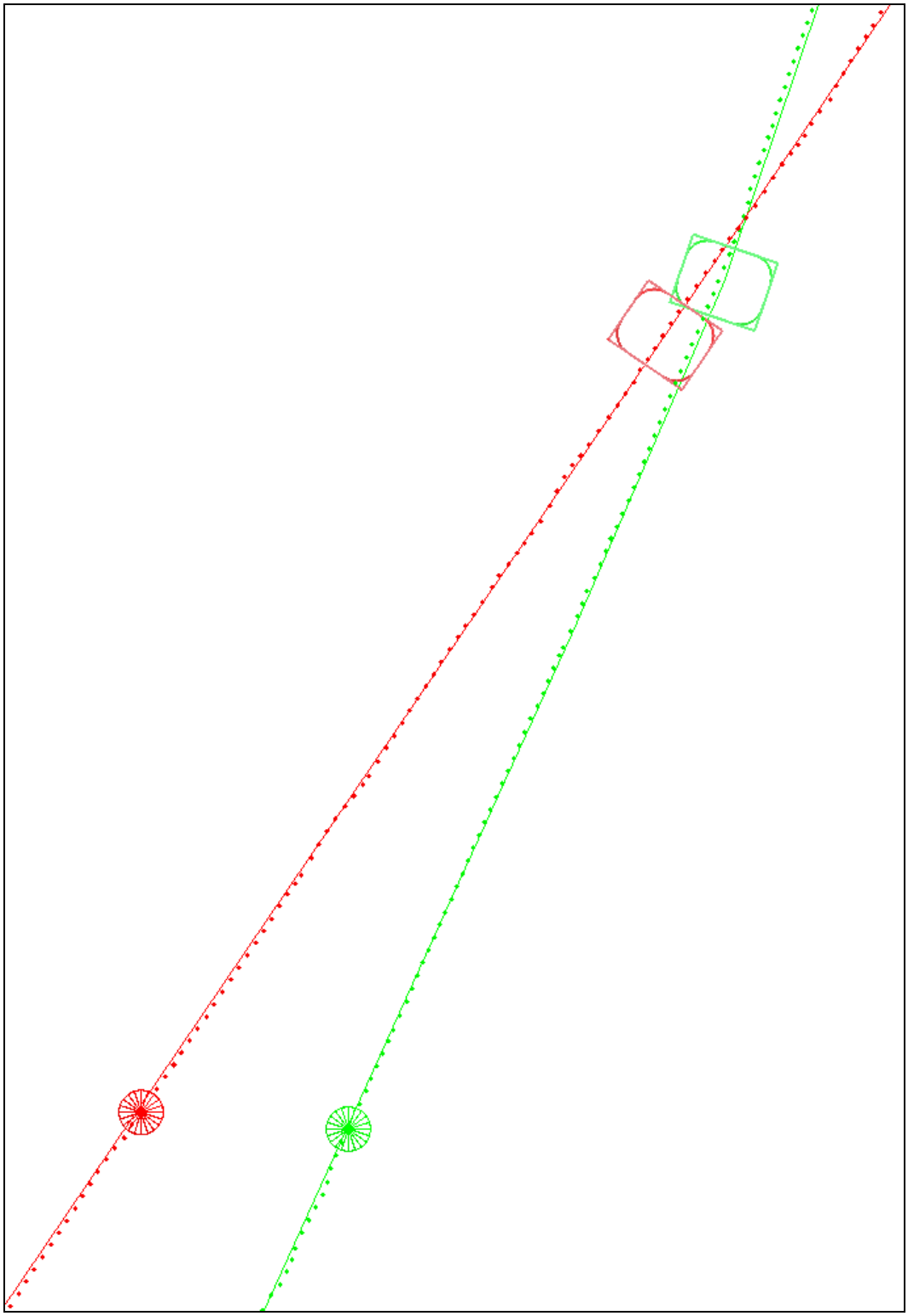


Figure 45. Baseline settings showing a valid conflict detection.

Figure 46 shows the effect of reducing the lateral conformance bounds to the minimum setting of 0.5nm. Here the detection cannot be made until the trajectories become much closer than needed in the baseline. However the prediction is made slightly earlier than the baseline. This is attributable to better quality trajectories. Here Flight C has just created a new trajectory and instead of Flight D being predicted to pass in front of Flight C, as in the baseline, now the trajectories predict that the flights will be side-to-side when entering conflict. This has to do with the delay caused by likelihood. The baseline actually detected the alert earlier than the case in Figure 46, though it waited to notify because of likelihood. In the reduced lateral conformance bounds treatment, the prediction was not able to be made until much later and then was immediately presented with improved accuracy instead of being held back by the likelihood function.

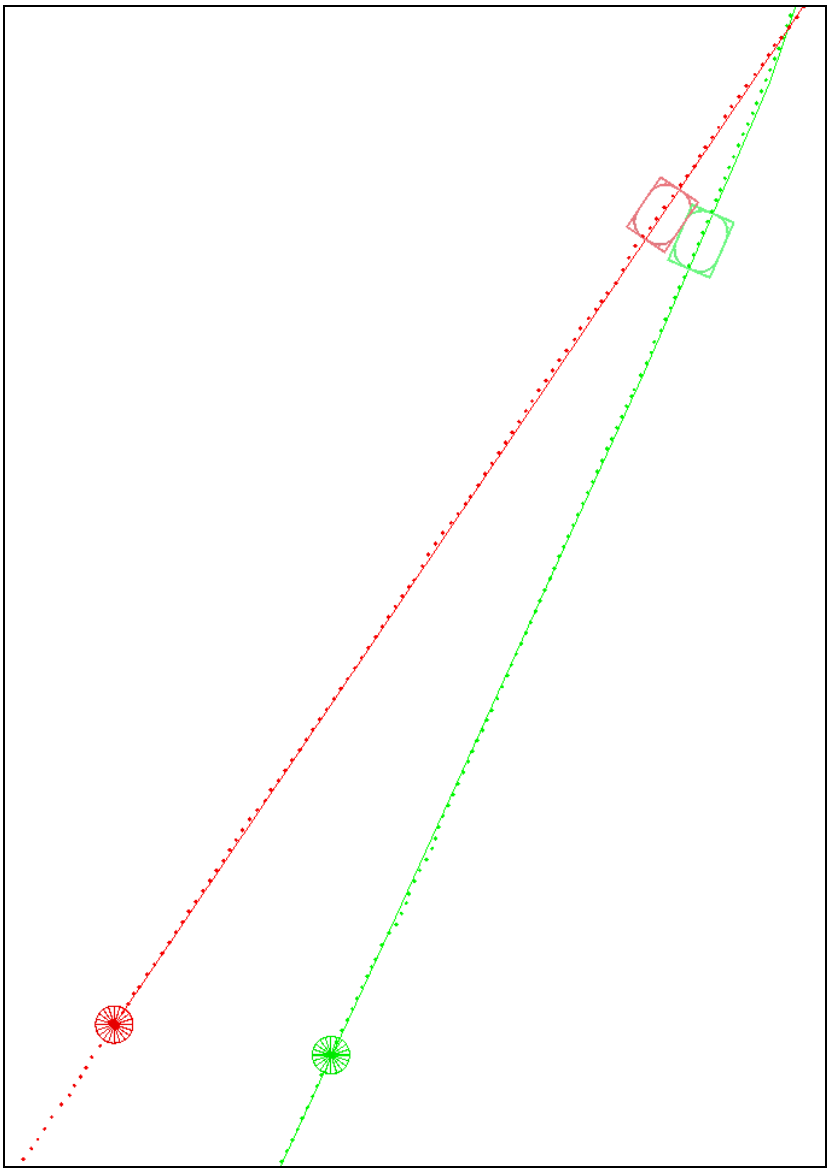


Figure 46. Lateral conformance bounds only are reduced to minimum.

Finally, Figure 47 shows the result of using conformance bounds where only the longitudinal conformance is reduced to minimum. Here the larger lateral conformance bounds allow the detection of the conflict in the same manner as in the baseline, but the predictions are being retracted by the probe until the final prediction is made below minimum warning time and retraction is no longer possible, turning the prediction into a missed alert. Tightening the longitudinal bounds translates to a greater confidence in speed profiling, though the conflict probe is less certain causing the predictions to be detected and then retracted due to the lack of confidence in speed profile.

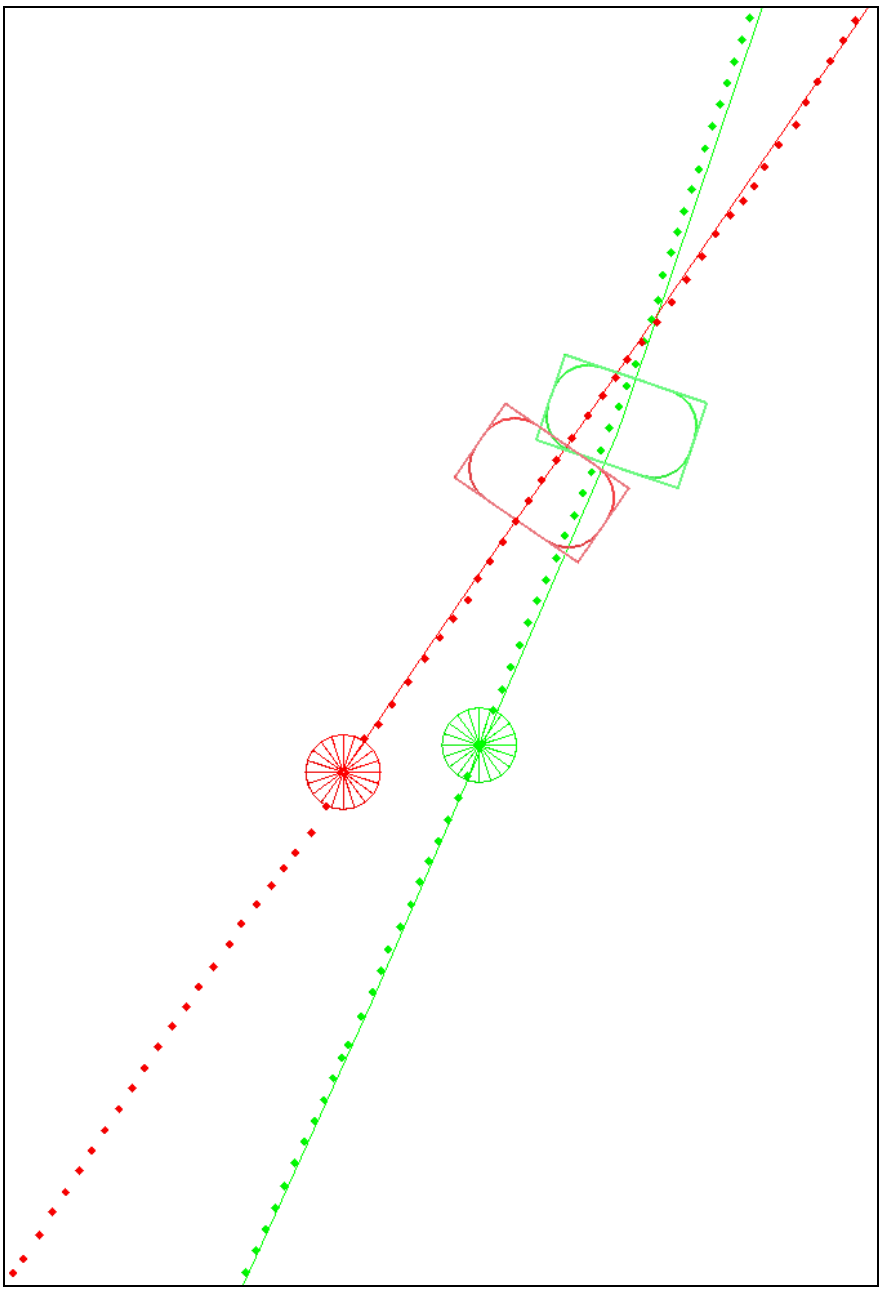


Figure 47. Longitudinal conformance bounds only are reduced to minimum.

4 Recommendations and Future Work

This report documents an initial study on the prototype enhancements required to advance the ERAM strategic conflict probe and allow it to be implemented within the radar controller's display. It is an initial study because it focuses mostly on altering parameter settings of functionality that already exists in ERAM. It does make use of one prototype enhancement on the trajectory modeling subsystem (see item #1, Table 1), but the experiment focuses on three factors that can be implemented in ERAM today without any software changes. Being an initial study, development of metrics and methodology was accomplished that can be utilized on all subsequent experiments. These metrics and methods are documented in Section 2 of this report. Furthermore, the study is limited in its use of a single data set of roughly 6-7 hours of traffic and about 2000 recorded flights. However, additional data sets have been collected from other facilities and traffic days. These scenarios will be used for follow-up experiments to both verify the results of this initial experiment and expand it by examining other factors and prototypes.

As presented in Section 3.1, task one of this report showed that the FA32 trajectory modeling prototype did have a significant improvement on both trajectory and conflict prediction performance. It was subsequently used for the experiment of task two, which is the main focus of this report. As outlined in the experimental plan in Section 2.2 and illustrated in the results in Section 3.2, the experiment focused on three key factors: longitudinal and lateral conformance or adherence bounds, reducing them for a range of values from operational settings of 1.5 and 2.5 nm, respectively, to as low as 0.5 nm. The third factor was changes to configurable parameters within the likelihood function. This algorithm assesses each conflict prediction and depending on its confidence estimate, may delay the alert notification until the estimate is improved or prediction is more imminent. The current ERAM operational system bypasses this calculation for all alerts predicted to begin within the next 10 minutes and applies the likelihood function for conflict detections beyond 10 minutes only. The two other levels examined in this experiment reduced this minimum threshold from 10 minutes to 3 minutes and changed other parameters associated with the function. For one level, it reduced the upper bound to 10 minutes and for the other left this setting at 20 minutes. As a result, these three factors, including longitudinal conformance bounds, lateral conformance bounds, and likelihood function, were simultaneously altered using experimental design techniques and ERAM run at twelve different treatment and two baseline runs.

The longitudinal conformance bound was found to significantly increase the Late Alerts when reduced too far from its original setting. The lateral conformance bound was determined to produce a significant decrease in False Alerts when reduced to the minimum parameter setting of 0.5 nm, without causing a major negative impact on Late Alerts. Though there is some interaction of these two settings with the likelihood, the interactions seen were minimal, so recommendations can be made independent of the likelihood function.

The longitudinal conformance bounds should not be lowered too far from the original setting of 1.5 nm. At a minimum, it should be set to 1.189 nm. However, to determine the best value of the longitudinal, a follow up study should be performed that varies the longitudinal conformance bounds between 1.0 and 1.5 nm. The study should be performed as an integrated experiment and contain center point runs at 1.25 nm longitudinal setting. Since no interactions were found between the lateral and longitudinal bounds, it is not necessary to include the lateral as a factor in this experiment.

The lateral conformance bounds were found to almost linearly decrease the False Alerts when lowered from 2.5 to 0.5 nm. Though the warning time was impacted, there was not a major

impact on Late Alerts when this change was made. It is recommended that the lateral conformance bound be lowered significantly from its original setting of 2.5 nm. A setting of 0.5 nm was found to provide the biggest impact of False Alerts, but more analysis is needed before a setting can be recommended. Once again, since no interaction was found with the longitudinal bound, it is not necessary to include it as a factor in this experiment.

Likelihood was determined to have a major effect on False Alerts, Late Alerts, and warning time. Unfortunately, it is not as easy to make a recommendation for a setting of likelihood. The three different settings of likelihood used in this study provided some insight to how the function affects the performance of the probe, but it is still not completely understood. There were two main findings in this study. The first is that likelihood alone can have a major effect on the performance of the probe. The second is that, because of the longitudinal inaccuracy of the probe, setting the minimum time of the likelihood function to the same as the minimum warning time requirement causes extra Late Alerts that should not have been. In order to further understand the likelihood and how the minimum, maximum, and middle setting of the probe affects the performance a separate experiment should be performed. This experiment should contain far more than only the three settings that were used in this experiment. An experiment is being planned in the year following this publication to precisely address these issues.

Overall, the study provides important insights into the three factors examined. It developed a number of metrics, including traditional counts and rates of false and late detections as well as implementation of fuzzy detection theory that extracted even more information out of the data. The study also implemented advanced experimental design techniques that both maximized use of the ERAM runs and examined the potential of interactions between factors. The study provides a strong foundation from which a number of additional experiments are being planned and underway.

List of Acronyms and Abbreviations

32BL	FA**32** **B**ase**l**ine Scenario
ANG-C41	FAA Concept Analysis Branch
ANSP	**A**ir **N**avigation **S**ervice **P**rovider
APDIA	**A**utomated **P**roblem **D**etection **I**nhibited **A**rea
ARTCC	**A**ir **R**oute **T**raffic **C**ontrol **C**enter
ATC	**A**ir **T**raffic **C**ontrol
ATO-E	Air Traffic Organization's En Route Program Office
BL	FA32 **B**ase**l**ine Scenario
CAASD	**C**enter for **A**dvanced **A**viation **S**ystem **D**evelopment
CD	**C**onflict **D**etection
CGS	**C**onflict **G**eometric **S**eparation
CP	**C**onflict **P**robe
CPAT	**C**onflict **P**robe **A**ssessment **T**eam
CPT	**C**onflict **P**robe **T**ool
DOE	**D**esign **o**f **E**xperiment
DST	**D**ecision **S**upport **T**ool
ERAM	**E**n **R**oute **A**utomation **M**odernization
FA	**F**alse **A**lert
FA18	**F**unctional **A**rea **18**
FA32	**F**unctional **A**rea **32**
FAA	**F**ederal **A**viation **A**dministration
FAR	**F**alse **A**lert **R**ate
ft	**F**ee**t**
FTR	**F**orced **T**rajectory **R**ebuild
GAB	**G**rowth **A**dherence **B**ounds
IBL	**I**nitial **B**ase**l**ine Scenario
IBST	**I**nterval **B**ased **S**ampling **T**echnique
IQR	**I**nter-**q**uartile **R**ange
JPDO	**J**oint **P**lanning and **D**evelopment **O**ffice
LA	**L**ate **A**lert
LAR	**L**ate **A**lert **R**ate
Llh	**L**ike**lih**ood
LM	**L**ockheed **M**artin
MA	**M**issed **A**lert
MAR	**M**issed **A**lert **R**ate
MITRE	The **MITRE** Corporation
MWT	**M**inimum **W**arning **T**ime
NAS	**N**ational **A**irspace **S**ystem
NC	Correct **N**o-**C**all
NextGen	**Next Gen**eration Air Transportation System

nm	**N**autical **M**iles
RNP	**R**equired **N**avigation **P**erformance
RVSM	**R**educed **V**ertical **S**eparation **M**inimum
SAC	**S**trategic **A**lert **C**omparer
SAE	**S**trategic **A**lert **E**valuator
SME	**S**ubject **M**atter **E**xpert
TBO	**T**rajectory **B**ased **O**perations
TM	**T**rajectory **M**odeling
TP	**T**rajectory **P**redictor
TRACON	**T**erminal **R**adar **A**pproach **Con**trol Center
URET	**U**ser **R**equest **E**valuation **T**ool
UTC	**C**oordinated **U**niversal **T**ime
VA	**V**alid **A**lert
VHF	**V**ery **H**igh **F**requency
WT	**W**arning **T**ime
ZDC	Washington, **DC** ARTCC

References

Agresti, A. 2002. *Categorical Data Analysis*, 2nd ed., John Wiley and Sons.

Bilimoria, Karl. May 2001. "A Methodology for the Performance Evaluation of a Conflict Probe," Journal of Guidance, Control, and Dynamics, Vol. 24 (3).

Box, George, J. Stuart Hunter, William G. Hunter. 2005. *Statistics for Experimenters, Design, Innovation, and Discovery*, Second Edition, Hoboken, NJ, John Wiley & Sons.

Brudnicki, D.J., W. C. Arthur, K. L. Lindsay. April 1998. "URET Scenario-based Functional Performance Requirements Document, MTR98W0000044, The MITRE Corporation, McLean, VA.

Cale, M. L., M. Paglione, H. Ryan, D. Timoteo, R. Oaks. April 1998. "URET Conflict Prediction Accuracy Report," DOT/FAA/CT-TN98/8, FAA WJHTC/ACT-250.

Crowell, Andrew, Confesor Santiago. August 2009. "An Algorithmic Method for Regression Analysis of Conflict Probe Accuracy," American Institute of Aeronautics and Astronautics (AIAA) Guidance, Navigation, and Control (GNC) Conference.

Crowell, Andrew, Christina Young, Ph.D. April 21, 2011. "Separation Management Modern Procedures Baseline Scenario Analysis Report," FAA Memorandum. Separation Management Modern Procedures.

Crowell, Andrew, Christina Young, Ph.D. May 4, 2011. "FA18 Interim 2 Prototype Baseline Scenario Analysis Report," FAA Memorandum, Separation Management Modern Procedures.

Crowell, Andrew, Christina Young, Ph. D., Dr. Hollis Ryan. June 13, 2011. "Response to FA18 Interim 2 Report – Conflict Probe Analysis," FAA Memorandum, Separation Management Modern Procedures.

Devore, Jay L. 2000. *Probability and Statistics for Engineering and the Sciences*, 5th Edition, ISBN 0-534-37281-3, Duxbury Press, Belmont, CA.

Federal Aviation Administration. December 12, 2004. "Next Generation Air Transportation System Integrated Plan," Joint Planning and Development Office, 1500 K Street, Suite 500, Washington, DC 20005.

Federal Aviation Administration. June 13, 2007. "Concept of Operations for the Next Generation Air Transportation System," Version 2.0, Joint Planning and Development Office, 1500 K Street, Suite 500, Washington, DC 20005.

Groupy, J., L. Creighton. 2007. *Introduction to Design of Experiments with JMP® Examples*, Third Edition, Cary, NC, SAS Press Series.

Hicks, Charles R. 1993. *Fundamental Concepts in the Design of Experiments*, Fourth Edition, New York, NY, Saunders College Publishing.

Kachigan, Sam Kash. 1986. *Statistical Analysis, An Interdisciplinary Introduction to Univariate and Multivariate Method*, Radius Press.

Lockheed Martin Corporation. May 19, 2011. "Separation Management Coordination Meeting Presentation, Lockheed Martin Status," briefing slides, Separation Management Modern Procedures.

McKay, Edward, Ph.D. February 7, 2011. "ERAM Aircraft Trajectory Lateral Modeling Enhancements, Prototype Effort," Lockheed Martin Report, Separation Management Modern Procedures.

Montgomery, Douglas C. 2009. *Design and Analysis of Experiments*, 7th Edition, New York, NY, John Wiley & Sons.

Oaks, Robert D., Mike M. Paglione. November 2001. "Generation of Realistic Air Traffic Scenarios Based on Recorded Field Data," Air Traffic Control Association (ATCA) 46th Annual Conference Proceedings.

Paglione, Mike M., Mary Lee Cale, Dr. Hollis F. Ryan. 1999. "Generic Metrics for the Estimation of the Prediction Accuracy of Aircraft to Aircraft Conflicts by a Strategic Conflict Probe Tool," *Air Traffic Quarterly*, Air Traffic Control Association, Volume 7(3) 147-165.

Paglione, Mike M., Robert D. Oaks, Karl D. Bilimoria. November 2003. "Methodology for Generating Conflict Scenarios by Time Shifting Recorded Traffic Data," Proceedings of the American Institute of Aeronautics and Astronautics (AIAA) Technology, Integration, and Operations (ATIO) Technical Forum.

Paglione, Mike M., Robert D. Oaks, Dr. Hollis Ryan. October 2004. "Methodology for Evaluating and Regression Testing a Conflict Probe," 23rd Digital Avionics Systems Conference (DASC), Salt Lake City, Utah.

Paglione, Mike M., Robert D. Oaks. August 2007. "Implementation and Metrics for a Trajectory Prediction Validation Methodology," American Institute of Aeronautics and Astronautics (AIAA) Guidance, Navigation, and Control Conference, Hilton Head, South Carolina.

Paglione, Mike M., Confesor Santiago, Andrew T. Crowell, Robert D. Oaks. August 2008. "Analysis of Aircraft to Aircraft Conflict Properties in the National Airspace System," Proceedings of the American Institute of Aeronautics and Astronautics (AIAA) Guidance, Navigation, and Control Conference.

Paielli, Russel A., Heinz Erzberger. October 1996. "Conflict Probability Estimation for Free Flight," NASA Technical Memorandum.

Paielli, Russel A. 1998. "Empirical Test of Conflict Probability Estimation," NASA Technical Report.

Paielli, Russel A., Heinz Erzberger. 1999. "Conflict Probability Estimation Generalization to Non-Level Flight," NASA Technical Report.

Pankok, Carl Jr. March 7, 2011. "Lateral Adherence Study," FAA Memorandum, Separation Management Modern Procedures..

Parasuraman, R., A. Masalonis, P. Hancock. Winter 2000. "Fuzzy Signal Detection Theory: Basic Postulates and Formulas for Analyzing Human and Machine Performance," Human Factors 42 (4) 636-659.

Ryan, Dr. Hollis F., George Chandler, Confesor Santiago, Mike M. Paglione, Shurong Liu. November 2008. "Evaluation of En Route Automation's Trajectory Generation and Strategic Alert Processing: Analysis of ERAM Performance," DOT/FAA/TC-TN08/10.

Sall, John. Fall 1997. "A New Angle on Matched Pairs," *JMPer Cable® Newsletter for JMP® Users*, SAS Institute Inc., Issue 5.

Santiago, Confesor, Jeffrey Lehman, Andrew Crowell. August 2010. "Accuracy Comparison of an Operational and Experimental Strategic Conflict Probe," Proceedings of the American

Institute of Aeronautics and Astronautics (AIAA) Guidance, Navigation, and Control Conference.

SAS Institute Inc. 2007. *JMP® Software: ANOVA and Regression Course Notes*, SAS Institute Inc., Cary, NC.

SAS Institute Inc. 2010. *JMP® 9 Basic Analysis and Graphing*, SAS Institute Inc., Cary, NC.

SAS Institute Inc. 2010. *JMP® 9 Design of Experiments Guide*, SAS Institute Inc., Cary, NC.

Vivona, Robert A., Mike M. Paglione, Karen T. Cate, Gabriele Enea. August 2010. "Definition and Demonstration of a Methodology for Validating Aircraft Trajectory Predictors," American Institute of Aeronautics and Astronautics (AIAA) Guidance, Navigation, and Control Conference, Toronto, Canada.

Willems, Ben. April 25, 2011. "Risk Reduction Task for Implementation of an R-side Conflict Probe," FAA Memorandum, Separation Management Modern Procedures.

Young, Christina, Ph. D. and Ben Musialek. September 30, 2011. "Initial Independent Evaluation of ERAM Aircraft Trajectory Lateral Modeling Enhancements," FAA Memorandum, Separation Management Modern Procedures.

Appendix A

This appendix provides more details of the Fuzzy Alert analysis. Section A.1 describes the logic used by the Fuzzy Alert Re-Evaluator software application, and Section A.2 provides details of computations performed for the Fuzzy Alert analysis.

A.1 Fuzzy Alert Logic

The Fuzzy Alert Re-Evaluator (FARE) program assigns fuzzy values to events from the Strategic Alert Evaluator (SAE) program. There are two types of events: encounter pairs with no associated conflict predictions presented, and encounter pairs with one or more conflict predictions presented (alerts). An encounter pair is defined to be a pair of aircraft that passed both spatial and temporal gross filters. If two aircraft are found to be an encounter pair more than once, then each event is treated as a separate encounter pair. For easier reference, each encounter pair is given a unique integer identification number greater than zero called the encounter index. Events marked with encounter indexes of zero or negative one are special case events in SAE that will be described later in this appendix.

Some information relevant to FARE is the data associated with an alert or encounter pair. For each encounter pair the time of closest approach, called the min-max ratio time, is recorded. Alerts have a predicted conflict start time (PCST), which is the time at which the conflict probe predicts an encounter pair will become a conflict pair (an encounter pair that also is violating minimum separation standards), and a predicted conflict end time (PCET). The other times recorded in a conflict prediction are the actual conflict start time (ACST) marking the start of the conflict duration (if it occurred) in the encounter pair, and the actual conflict end time (ACET) marking the end of the conflict duration (if it occurred) in the encounter pair, NOTIF_START_TIME marking the beginning time of the alert, and NOTIF_END_TIME marking the ending time of the alert. Also associated with alerts are predicted separation values at the predicted min-max ratio time.

User-defined parameters include the MAX_LOOK_AHEAD_TIME which defines the maximum duration of the BUFFER_TIME used in signal calculation of false alerts. Additionally, MIN_WARNING_TIME defines the minimum time before ACST in which the NOTIF_START_TIME can occur before becoming a late alert (which needs to match the value used by SAE). Finally, the minimum separation values MIN_HORZ_SEP and MIN_VERT_SEP are user-defined parameters in FARE which should match those used in SAE.

At the top-level of the flowchart, presented in Figure 48, the program begins by gathering all of the conflict predictions and encounter pairs. Each presented conflict prediction made by the conflict probe needs to map onto an encounter pair, however SAE does not provide mapping for all discarded alerts. In Figure 48, any conflict prediction that is not mapped to an encounter pair is matched by finding the encounter pair involving the same aircraft and with the closest min-max ratio time to the alert's PCST, though it is still possible to not have a mapping after this matching, which is handled in the subsequent flowcharts. The response for conflict predictions must be evaluated using the flowchart shown in Figure 49. Encounter pairs with no associated conflict predictions were never presented as potential conflict pairs by the conflict probe, and thus have no associated response. Both types of events then have their signal computed using the flowcharts in Figure 50 and Figure 51. With both signal and response calculated, the flowchart in Figure 52 provides the fuzzy classification of the event into the valid alert, late alert, false alert, and correct no-call categories.

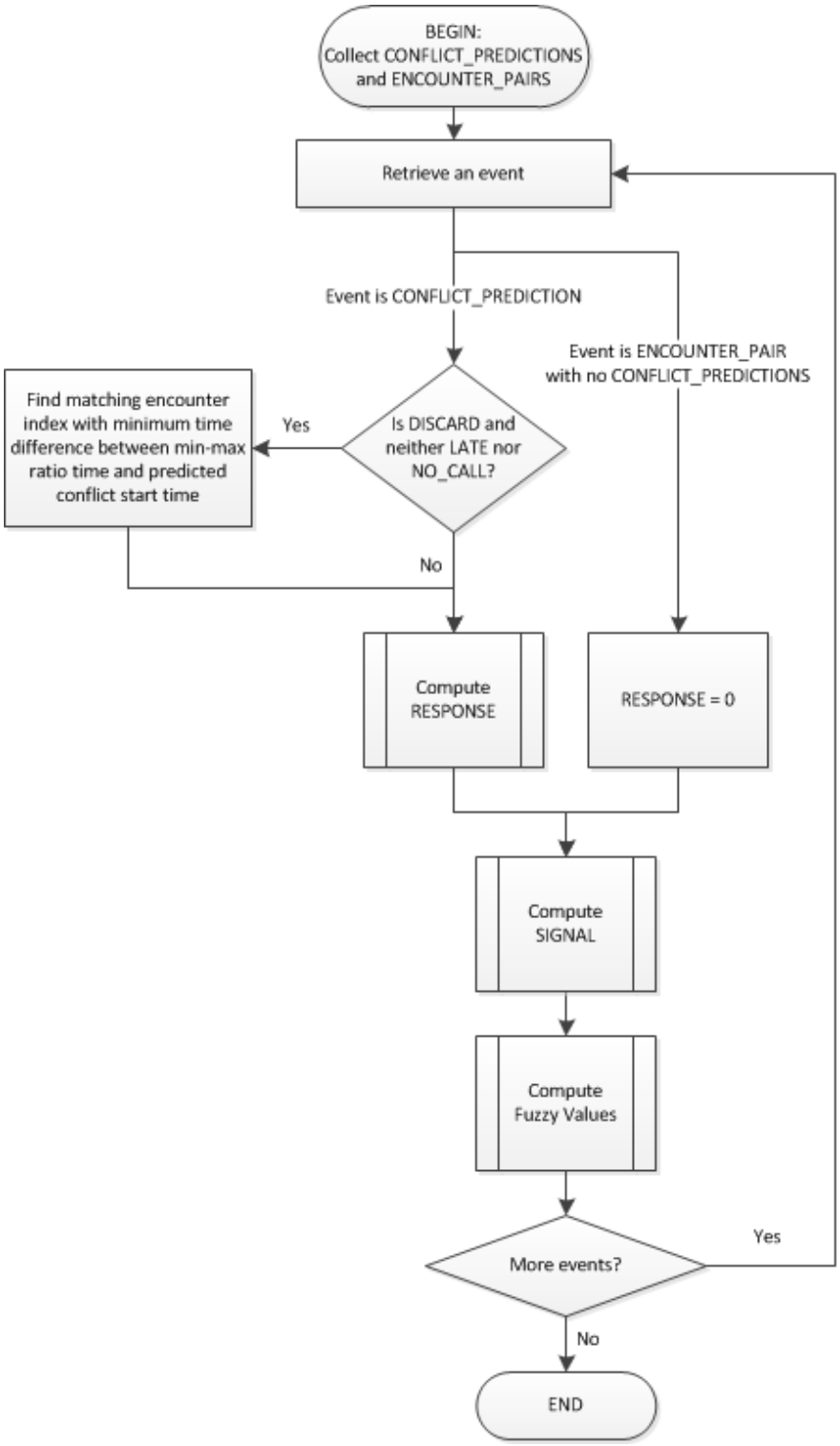


Figure 48. High-level FARE Process Flow

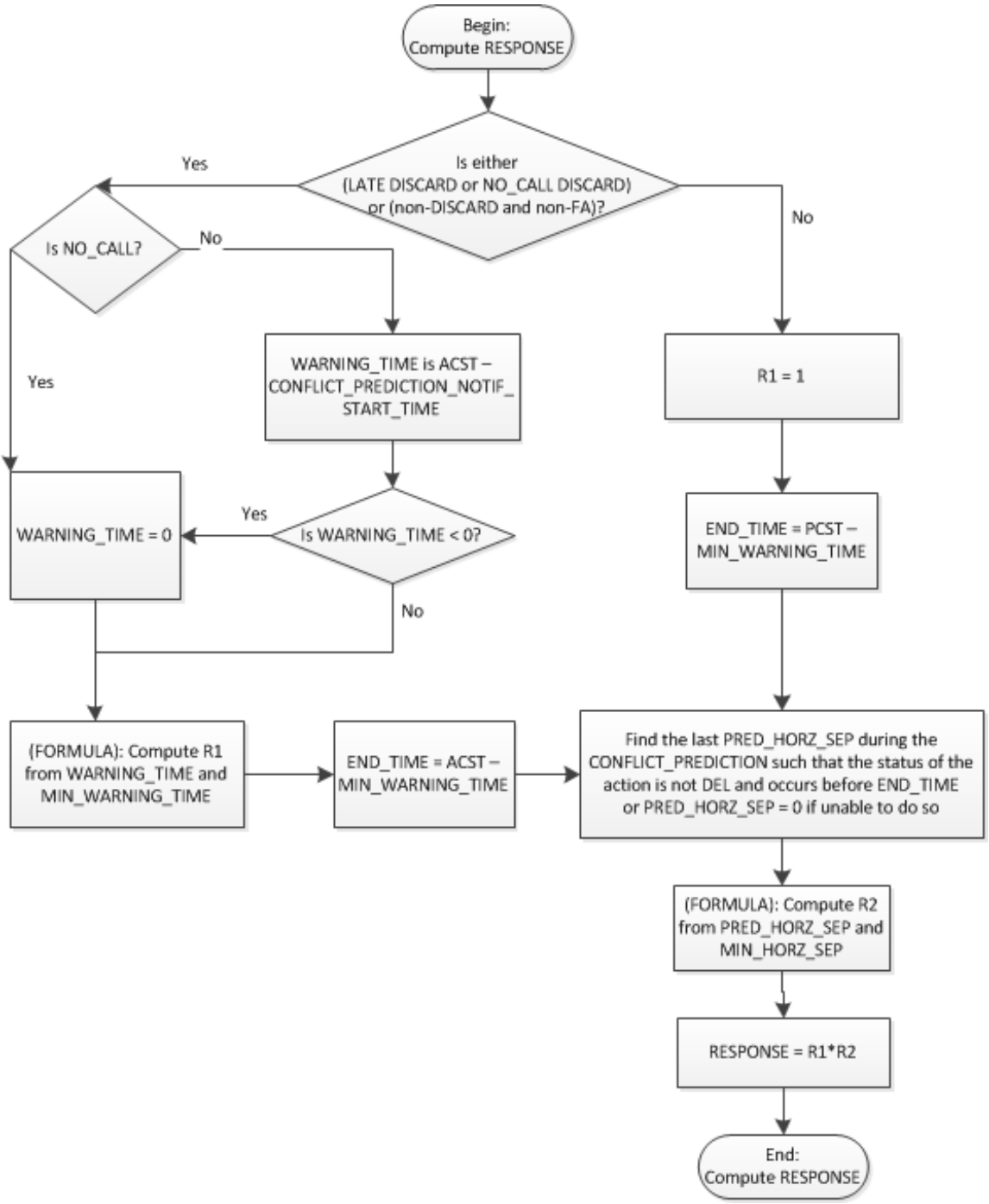


Figure 49. Fuzzy Response Computation Process Flow

Conflict predictions enter the flowchart presented in Figure 49, and are separated into two categories. The "yes" branch corresponds to conflict predictions that were determined to be late alerts discarded by SAE, or any non-discard that is not a false alert (excluding correct no-call events because they are not explicitly created by SAE). The "no" branch then corresponds to the remaining discards and false alerts. The reason this split is made is that the latter category does not have an intuitive notion of warning time, since no actual conflict occurred. In this case, R1 is set to one so that total response is reduced to R2. On the other hand, no-call events fail to generate an alert for an actual conflict so warning time is set to zero in this case. For the

remaining events, WARNING_TIME is simply ACST – NOTIF_START_TIME, the time between when the alert was made and the actual conflict began. WARNING_TIME is forced to be no less than zero to reflect the notion that making a prediction after the fact has the same perceived utility as making no prediction at all. The ratio of WARNING_TIME to MIN_WARNING_TIME is taken and if this ratio is greater than or equal to one, R1 is set to one; otherwise R1 is set to the square of the ratio to reflect a lateness penalty. END_TIME is the last possible time an alert could update before becoming late, so it is simply ACST – MIN_WARNING_TIME for the "yes" branch where a conflict actually occurred. For the "no" branch where no conflict actually occurred the END_TIME needs to be interpreted as a perceived threshold and is thus PCST-MIN_WARNING_TIME. Once END_TIME is found the last amendment to the predicted horizontal separation during the duration of the conflict prediction notification set is retrieved up to the END_TIME, after which no amendments could be usefully posted. The ratio of this predicted horizontal separation to the MIN_HORZ_SEP is taken and R2 is calculated by the spline as described in Section 2.1.2.2. Finally, the response is calculated as R1*R2.

In the flowchart shown in Figure 50, events are split into three groups: encounter pairs with no associated conflict predictions, conflict predictions, and error events. Encounter pairs with no associated conflict predictions are labeled with reason code = ENC and their signal is computed around their already-recorded min-max ratio time by using the flowchart in Figure 51. ENC becomes ENC_A if at any point along either track, an APDIA was encountered when computing the signal at min-max ratio time. It is noted that any time the flowchart in Figure 51 is used, '_A' will be appended to the reason code label to warn that the computed value may be skewed due to the presence of an APDIA. The second category includes any event that has an error code (0 or -1) encounter index from SAE. For these events, the reason code is set to SIGNAL_0 and the signal is set to 0, with warning values (-9999999) recorded instead of actual values such as min-max ratio, min-max ratio time, etc. The third category, conflict predictions, leads to more complex flowchart paths than the other two categories. The simplest path is when a conflict prediction cannot be matched to an encounter pair, and the event is processed as if it belonged to the SIGNAL_0 category previously discussed. Otherwise, the conflict predictions are again split into two categories exactly as in the response calculations. The "yes" branch matches to the "yes" branch of the response calculation and represents alerts made about actual conflicts, so these can be treated exactly the same as the ENC category. This leaves the "no" branch, corresponding to false predictions and the most involved branch. All conflict predictions associated with an encounter pair are first sorted by NOTIF_START_TIME so that the duration between the current false alert event and its successor can be calculated. BUFFER_TIME is taken to be the minimum of this duration and MAX_LOOK_AHEAD_TIME. The min-max ratio and min-max ratio time are then computed for the duration of the alert (between NOTIF_START_TIME and NOTIF_END_TIME) and for the duration of the buffer (between NOTIF_END_TIME and NOTIF_END_TIME+BUFFER_TIME). The encounter pair min-max ratio is not used here since the event does not necessarily align with the encounter pair min-max ratio time due to the event being a false alert. S1 is set to zero if the min-max calculations fail due to no track data or being within an APDIA for the duration of the event. The same is done for S2 if the calculations fail in the buffer or the buffer has zero duration. If both fail the reason code is IN_NOTIF_A and warning values are set as in SIGNAL_0 meaning the signal is ultimately zero here but is assumed to be within the event's duration (IN_NOTIF) and the assumed reason for failure is the presence of an APDIA (_A). Although not necessarily true, these IN_NOTIF_A events are easy to distinguish from actual IN_NOTIF_A events due to the warning values set and would require further investigation to determine actual cause of failure. If both S1 and S2 do not fail, then there are two possible cases. The first case is when S1 is not set, implying a min-max ratio and an associated min-max ratio time were successfully computed for the event, and also the calculated

min-max ratio is less than or equal to that of the buffer's, where this condition is assumed true if S2 was set, implying a min-max ratio and an associated min-max ratio time were not found in the buffer. This case can be restated as determining the aircraft were closer during the duration of the event than they were immediately after the event, therefore the signal is potentially stronger during the event than immediately after it. This case provides the reason code of IN_NOTIF. The second case is the reverse of the first case, where the aircraft were found to be closer after the event than during it implying the event would have potentially computed a higher signal if it had not been deleted, which means the false alert was most likely an early deletion event. This case therefore provides the reason code of ED_NOTIF. In either case, signals S1 and S2 are computed for the duration of the event and the buffer respectively if they have not already been set by failing to compute min-max ratio and its time. IN_NOTIF events are the only events assigned an early deletion value (1 – S2), which provides a score of how "stable" the situation was when the false alert was removed. ED_NOTIF events are paired with ED_MISS events, which are the only new events that this program creates. ED_MISS events are assigned a response of zero and a signal of S2, and these events denote the situation of missing potentially high signal by deleting a false alert. In either case the signal for the actual event is given as S1.

The flowchart shown in Figure 51 is used to compute signal at some time. The track data is retrieved for the aircraft, as well as their active clearance altitudes. Any track point with altitude within 300ft of the ACTIVE_CLEARANCE is said to be in vertical adherence. If the point at TIME, the time about which the signal is being calculated, is in vertical adherence all points within the computation region are set to their ACTIVE_CLEARANCE if they are within 300ft of it, called snapping altitude to clearance. If the point at TIME is out of vertical adherence, the path is more complex. First, the number of points out of vertical adherence to both their ACTIVE_CLEARANCE and PREVIOUS_CLEARANCE (immediately before active) is calculated. If the result is no more than 4 data points and does not represent more than half of the sample points, it is said that overall the aircraft is in vertical adherence and the outliers are snapped to their respective clearances (either ACTIVE or PREVIOUS). If it is instead found that most points are out of adherence, it is checked to see if the vertical separation at TIME is less than 1.1*MIN_VERT_SEP and the horizontal separation at TIME is less than MIN_HORZ_SEP, indicating the aircraft are either in conflict or near conflict due to vertical separation while out of adherence. In case this is due to a spike in the data, the altitude at TIME is set to the middle value of the average altitudes before and after the current TIME. After this is done, all points are snapped to clearance as was done when the current point was in vertical adherence, and the vertical profile is said to be cleaned and ready for use. If it is found that at TIME both aircraft are in vertical adherence and separated by at least MIN_VERT_SEP, then by the vertical adherence rule it is said that the aircraft are well separated and have signal 0. Otherwise, vertical and horizontal separation are calculated at TIME, then HORZ_CLOSURE_RATE (up to TIME), ENCOUNTER_ANGLE_WEIGHTING (up to TIME), VERT_CLOSURE_RATE, ENCOUNTER_ANGLE, and the number of points with ENCOUNTER_ANGLE within 15 degrees of 0 or 360 are computed for each point in the signal region. After this, HORZ_CLOSURE_RATE, VERT_CLOSURE_RATE, ENCOUNTER_ANGLE, and ENCOUNTER_ANGLE_WEIGHTING are averaged, the percentage of points with ENCOUNTER_ANGLE within 15 degrees of 0 or 360 is calculated, and the standard deviation of ENCOUNTER_ANGLE is calculated. From these values, the signal is computed using the formulas in Section 2.1.2.2.

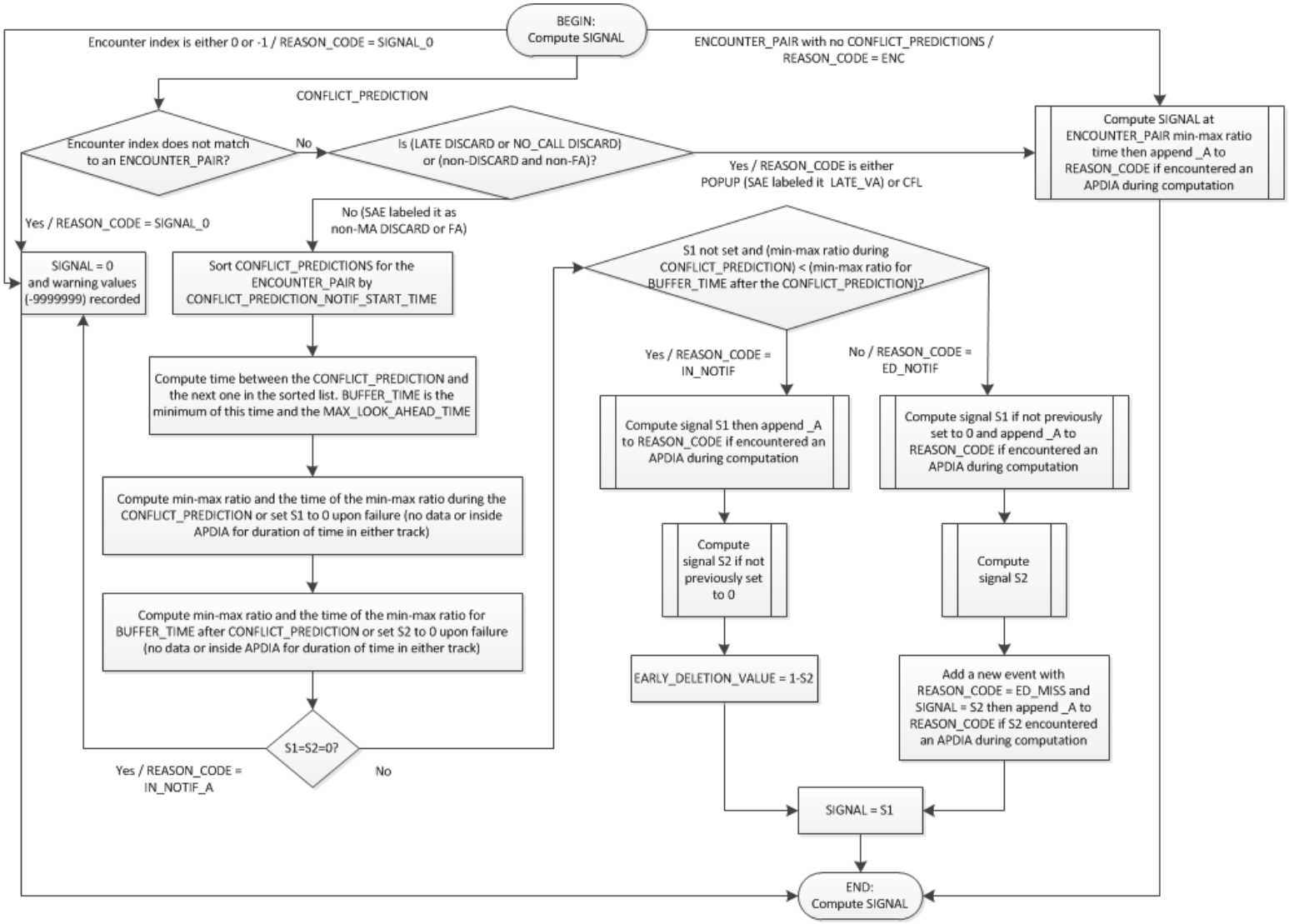


Figure 50. High-level Fuzzy Signal Computation Process Flow

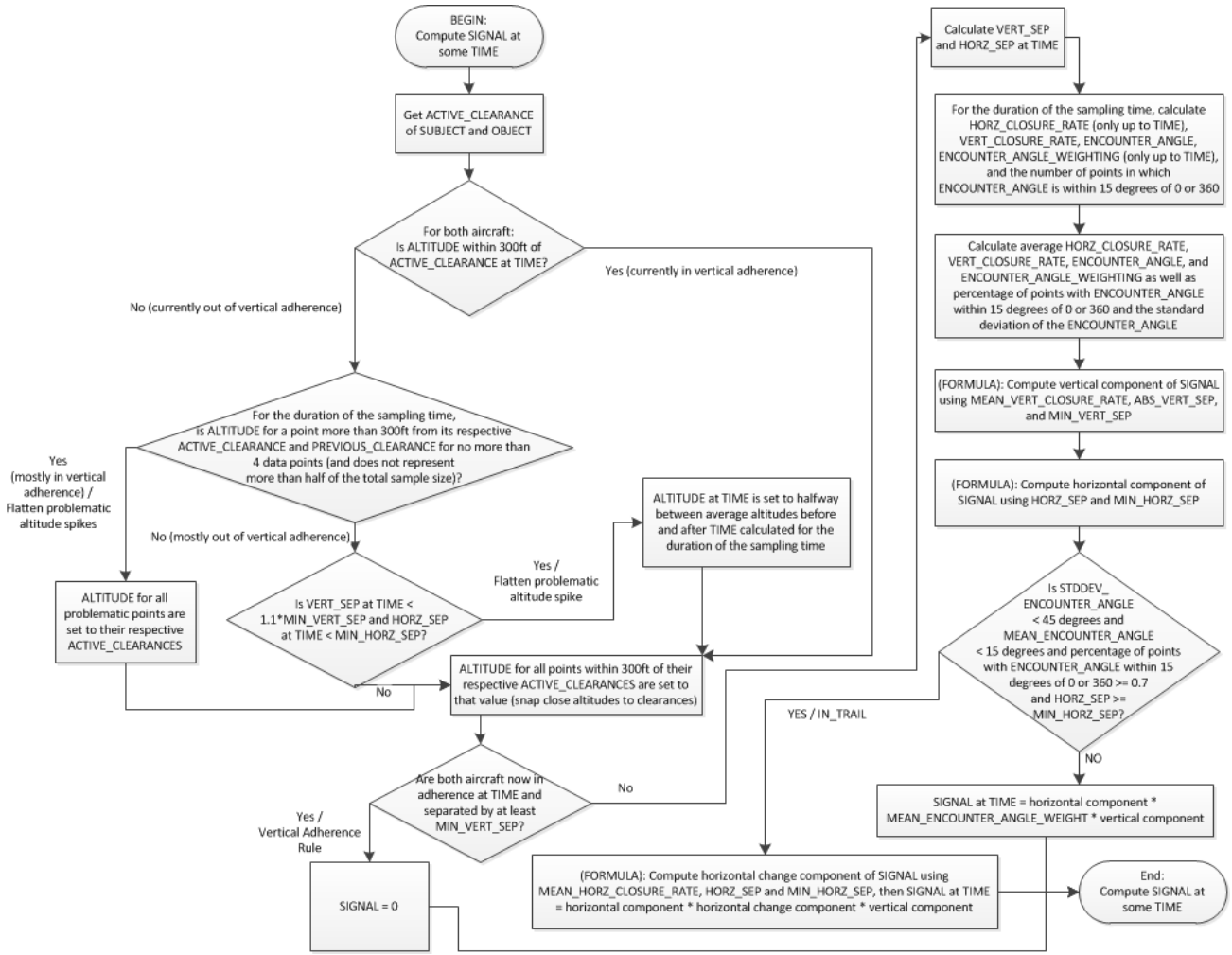


Figure 51. Process Flow for Computation of Fuzzy Signal at a Point in Time

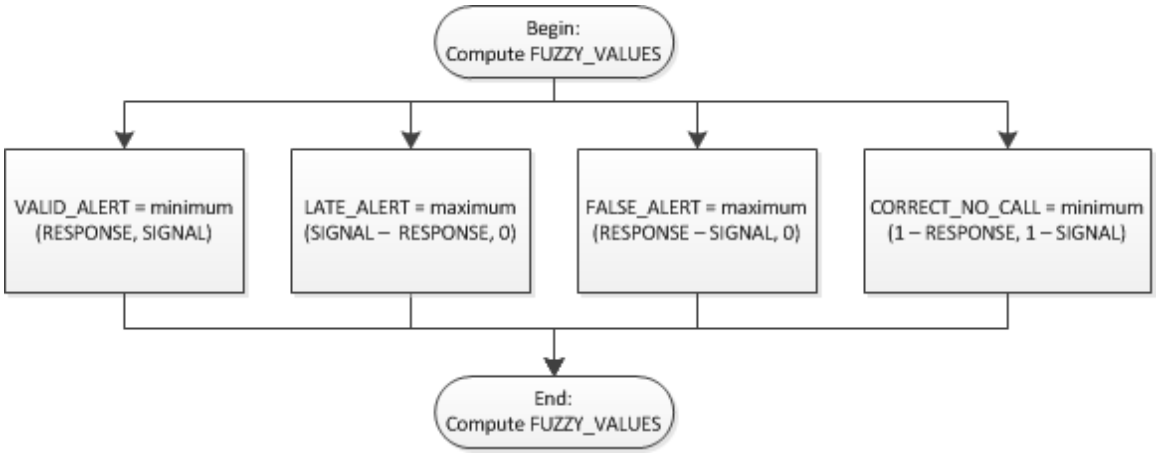


Figure 52. Determining Fuzzy Alert Values

Following the calculation of response and signal, the final values for VALID_ALERT, LATE_ALERT, FALSE_ALERT, and CORRECT_NO_CALL are computed in the flowchart, presented in Figure 52.

A.2 Additional Computations for Fuzzy Alert Logic

Eq.22 is a weighting function for flight pair geometry; as such its parameters are relative angle and relative position of the object of the pair to the subject of the pair (Figure 9). Certain cases of the weighting have been solved in the following to give a better idea as to how it performs.

Metal-to-metal at any relative angle is a certain conflict, thus $w_{\varphi,0,0} = 1$.

At *hmin* given any relative angle, the weighting is:

$$w_{\varphi,\alpha,\beta} = 1 - \frac{1}{2e}\left(1 + \frac{\alpha(\cos\varphi' - 1) + \beta\sin\varphi'}{2hmin}\right), \quad \alpha^2 + \beta^2 = hmin^2$$

The minimum value of the weighting is 0.5.

$$w_{\varphi,\alpha,\beta} = 0.5, \quad \alpha^2 + \beta^2 \to \infty$$

If the object has positive lateral distance from the subject, it is on the right of the subject. If the object has negative lateral distance from the subject, it is on the left of the subject. Approaching from the right is the mirror of approaching from the left, thus $w_{\varphi,\alpha,\beta} = w_{-\varphi,\alpha,-\beta}$.

Flights with the same heading are related by:

$$w_{0,\alpha,\beta} = 1 - \frac{1}{2}e^{-hmin^2/(\alpha^2+\beta^2)}$$

Flights with the same heading at *hmin* have exactly the weighting:

$$w_{0,\alpha,\beta} = 1 - \frac{1}{2e} \approx 0.816060, \quad \alpha^2 + \beta^2 = hmin^2$$

Flights that are at *hmin* with only lateral distance separating at 90 degrees have exactly the weighting:

$$w_{\pi/2,0,hmin} = w_{3\pi/2,0,-hmin} = 1 - \frac{3}{4e} \approx 0.724090$$

Flights that are at *hmin* with only lateral distance approaching at 90 degrees have exactly the weighting:

$$w_{\pi/2,0,-hmin} = w_{3\pi/2,0,hmin} = 1 - \frac{1}{4e} \approx 0.908030$$

Flights that are at *hmin* with only longitudinal distance with exactly opposite headings directed at each other are certainly in conflict.

$$w_{\pi,hmin,0} = 1$$

Flights that are at *hmin* with only longitudinal distance with exactly opposite headings separating from each other have a low chance of being in conflict.

$$w_{\pi,-hmin,0} = 1 - \frac{1}{e} \approx 0.632121$$

Appendix B

The following repeats the work done in Section 3.2.2, but now with adherence settings applied. It is expected that performance will improve on the whole from 3.2.2 though the same patterns are likely to emerge. In Table 35, the alert type counts of each of the twelve runs and the two baseline runs are shown, it is easy to see that the total number of FAs, LAs, and MAs have decreased significantly without affecting VAs (except for a slight impact in IBL). Except Run 11, it is difficult to say much about the LAs and MAs simply due to there being so few in the new data set. It is also seen here that every run has fewer false alerts than each baseline.

Table 35. Alert type counts for each of the twelve runs and the two baselines.

Run	VA	STD VA	LA	MA	FA
1	155	125	2	1	294
2	159	131	1	1	524
3	160	131	1	1	365
4	151	120	2	1	350
5	160	130	1	1	395
6	161	132	1	1	325
7	156	127	1	1	447
8	156	126	2	1	475
9	157	128	1	1	475
10	154	123	2	1	425
11	150	119	4	1	245
12	158	129	1	1	299
32BL	163	135	1	1	540
IBL	146	134	1	0	535

By only looking at the counts of LAs and MAs it is difficult to make much of a judgment about late alert performance, though the adjusted LA metric is of interest (Table 36). Again Run 11 has the worst performance by this metric, although now all other runs are comparable to 32BL. IBL has a superior score in this metric.

Table 36. Late Alert plus Missed Alert count and Adjusted Late Alert value for each of the twelve runs and the two baselines.

Run	LA+MA	Adj. LA
1	3	1.21
2	2	1.17
3	2	1.17
4	3	1.21
5	2	1.17
6	2	1.17
7	2	1.17
8	3	1.21
9	2	1.17
10	3	1.21
11	5	2.41
12	2	1.17
32BL	2	1.17
IBL	1	0.22

Table 37 also supports what was found in 3.2.2. There it was found that the baselines had the lowest LA Rate and the highest FA Rate and that FA Rate is greatly diminished when aircraft are well-separated vertically. It is not fair to focus on LA Rate heavily, though the same pattern has emerged, since there are so few samples there, however the FA Rates follow the same pattern, with the baseline runs having the highest rates, although now Run 2 is close to having the same rate.

Table 37. Overall Alert Rates and False Alert Rates by separation category.

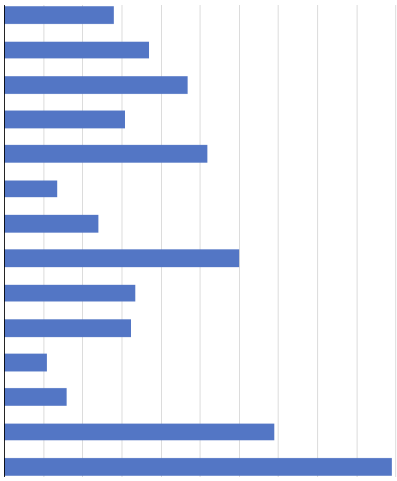
			FA Rate									
			Horz	< 5	< 5	5 ≤ h < 8	5 ≤ h < 8	8 ≤ h ≤ 13	8 ≤ h ≤ 13	> 13	> 13
Run	LA Rate	FA Rate	Vert	< 1000	≥ 1000	< 1000	≥ 1000	< 1000	≥ 1000	< 1000	≥ 1000
1	0.023	0.018		1.000	0.066	0.673	0.140	0.108	0.008	0.028	0.003
2	0.015	0.032		1.000	0.063	1.000	0.241	0.411	0.089	0.014	0.003
3	0.015	0.022		1.000	0.059	0.867	0.213	0.169	0.014	0.027	0.003
4	0.024	0.021		1.000	0.072	0.878	0.204	0.122	0.021	0.016	0.002
5	0.015	0.024		1.000	0.063	0.960	0.233	0.184	0.029	0.018	0.003
6	0.015	0.020		1.000	0.054	0.860	0.206	0.137	0.014	0.015	0.001
7	0.016	0.027		1.000	0.063	0.960	0.245	0.249	0.052	0.018	0.002
8	0.023	0.029		1.000	0.077	0.950	0.220	0.319	0.060	0.026	0.003
9	0.015	0.029		1.000	0.063	0.942	0.227	0.331	0.068	0.013	0.003
10	0.024	0.026		1.000	0.071	0.912	0.212	0.253	0.056	0.009	0.002
11	0.040	0.015		1.000	0.063	0.620	0.136	0.077	0.007	0.010	0.002
12	0.015	0.018		1.000	0.069	0.779	0.185	0.073	0.006	0.014	0.002
32BL	0.015	0.034		1.000	0.063	0.990	0.233	0.447	0.090	0.023	0.005
IBL	0.008	0.033		1.000	0.059	1.000	0.223	0.423	0.101	0.017	0.004

In Table 38, Runs 5 and 8 have the largest 25th percentile of the 12 runs, with the baselines outperforming all runs. This supports what was found before. Here Runs 6, 11, and 12 have the

worst performance in warning time, though Runs 1, 4, 11, and 12 were the worst in 3.2.2. Runs 11 and 12 therefore scored the worst in both analyses.

Table 38. Median, inter-quartile range, and 25th percentile of conflict warning time.

Run	Median	IQR	25th %
1	446.0	367.0	328.0
2	444.5	282.5	337.0
3	487.5	389.0	347.0
4	424.5	322.0	331.0
5	505.0	393.0	352.0
6	397.0	211.0	313.5
7	424.5	296.5	324.0
8	490.5	436.0	360.0
9	426.0	323.0	333.5
10	420.0	289.5	332.5
11	391.0	196.0	311.0
12	399.5	231.0	316.0
32BL	511.5	476.0	369.0
IBL	564.0	440.0	399.0



Runs 1, 4, 6, 11, and 12 show the most improvement in terms of FAs (Table 39, Table 40) when compared to either baseline, which is what was demonstrated in the analysis of 3.2.2 as well. Using this and the warning time, there is a significant effect between warning time and the FA improvement metric, where the runs with the most FA improvement have some of the worst scores for the 25th percentile of warning time between the two analyses. Also of note, except for Run 4, these runs all have a lateral conformance bound of 0.5. The classic metrics used here all support the conclusion that while 0.5 nautical mile lateral conformance bounds have the largest effect on reducing FAs, they adversely affect warning time, though it should be noted that the median and 25th percentile of warning time are well above the minimum warning time of 180 seconds.

Table 39. Results of comparisons of each of the twelve runs to the FA32 Baseline run.

	Lat	0.5	2.5	0.5	1.5	1.5	0.5	1.5	2.5	2.5	2.5	0.5	0.5
	Lon	0.5	1.5	1.5	0.5	1	1.5	1.5	0.5	1	0.5	0.5	1
	Like	10/20	3/8/10	10/20	3/8/20	10/20	3/8/10	3/8/20	10/20	3/8/20	3/8/10	3/8/10	3/8/20
BL	**Proto**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DSC	**DSC**	1114	1330	1360	877	1464	873	1093	1642	1177	1018	709	819
DSC	**FA**	43	19	26	39	35	23	34	38	40	38	36	35
DSC	**N/A**	980	790	751	1222	638	1240	1011	459	921	1082	1393	1283
DSC	**VA**	2	0	2	1	2	3	1	0	1	1	1	2
FA	**DSC**	50	6	34	41	42	37	29	41	28	39	50	38
FA	**FA**	197	489	295	266	324	275	375	361	393	337	176	228
FA	**N/A**	293	45	211	233	174	228	136	138	119	164	314	274
LA	**LA**	2	2	2	2	2	2	2	2	2	2	2	2
N/A	**DSC**	268	18	214	131	202	108	93	248	109	119	133	133
N/A	**FA**	54	16	44	45	36	27	38	76	42	50	33	36
VA	**DSC**	9	4	5	12	5	5	8	6	6	9	11	7
VA	**LA**	1	0	0	1	0	0	0	1	0	1	3	0
VA	**N/A**	0	0	0	0	0	0	0	0	1	0	0	0
VA	**VA**	153	159	158	150	158	158	155	156	156	153	149	156
% FA Imp		44%	5%	31%	35%	26%	37%	18%	11%	14%	21%	52%	44%
%LA Imp		-50%	0%	0%	-50%	0%	0%	0%	-50%	0%	-50%	-150%	0%

Table 40. Results of comparisons of each of the twelve runs to the Initial Baseline run.

	Lat	0.5	2.5	0.5	1.5	1.5	0.5	1.5	2.5	2.5	2.5	0.5	0.5
	Lon	0.5	1.5	1.5	0.5	1	1.5	1.5	0.5	1	0.5	0.5	1
	Like	10/20	3/8/10	10/20	3/8/20	10/20	3/8/10	3/8/20	10/20	3/8/20	3/8/10	3/8/10	3/8/20
BL	**Proto**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DSC	**DSC**	1134	1220	1352	886	1460	873	1086	1603	1157	1002	728	829
DSC	**FA**	45	63	44	53	53	40	63	58	63	61	40	48
DSC	**N/A**	1086	984	869	1329	754	1351	1119	606	1048	1205	1499	1389
DSC	**VA**	12	10	12	9	10	13	9	10	9	9	10	11
FA	**DSC**	44	32	46	44	48	43	36	42	31	41	41	36
FA	**FA**	195	418	271	252	300	256	349	346	367	319	169	214
FA	**N/A**	296	85	218	239	187	236	150	147	137	175	325	285
LA	**LA**	1	1	1	1	1	1	1	1	1	1	1	1
N/A	**DSC**	257	105	213	124	205	105	98	289	131	138	127	129
N/A	**FA**	54	43	50	45	42	29	35	71	45	45	36	37
N/A	**LA**	1	1	1	1	1	1	1	1	1	1	1	1
N/A	**VA**	5	5	5	5	5	5	5	5	4	5	5	5
VA	**DSC**	6	1	2	7	0	2	3	3	1	4	7	3
VA	**LA**	1	0	0	1	0	0	0	1	0	1	3	0
VA	**N/A**	1	1	1	1	1	1	1	1	1	1	1	1
VA	**VA**	138	144	143	137	145	143	142	141	144	140	135	142
% FA Imp		45%	8%	31%	36%	27%	39%	21%	14%	17%	24%	54%	46%
%LA Imp		-100%	0%	0%	-100%	0%	0%	0%	-100%	0%	-100%	-300%	0%

As for the fuzzy metrics (Table 41), the FA counts still agree that Runs 1, 4, 6, 11, and 12 all return the lowest FA counts, yet also return the highest LA counts.

Table 41. Fuzzy Alert values for each of the 12 runs and the baseline.

Run	VA	LA	FA
1	193.46	17.69	223.37
2	197.41	10.04	340.64
3	199.67	12.04	272.49
4	196.22	10.17	254.68
5	202.14	9.45	286.91
6	197.51	13.16	239.12
7	199.02	7.61	318.54
8	195.71	11.74	331.68
9	200.65	10.40	313.69
10	195.82	11.27	287.10
11	185.91	21.35	177.83
12	198.76	14.53	221.01
32BL	198.47	12.31	357.20
IBL	180.49	7.53	346.84

Again there is agreement in the data (Table 42) that vertically well-separated events have much lower FA rates. Here Runs 2 and 8 have similar FA rates to the baselines though all runs improve on the FA rates on the whole.

Table 42. Overall Fuzzy Alert Rates and Fuzzy False Alert Rates by separation category.

			FA Rate								
			Horz	< 5	< 5	5 ≤ h < 8	5 ≤ h < 8	8 ≤ h ≤ 13	8 ≤ h ≤ 13	> 13	> 13
Run	LA Rate	FA Rate	Vert	< 1000	≥ 1000	< 1000	≥ 1000	< 1000	≥ 1000	< 1000	≥ 1000
1	0.084	0.014		0.945	0.090	0.505	0.117	0.108	0.008	0.027	0.003
2	0.048	0.021		0.918	0.080	0.823	0.198	0.215	0.057	0.012	0.002
3	0.057	0.017		0.905	0.082	0.707	0.179	0.146	0.013	0.026	0.003
4	0.049	0.015		0.973	0.078	0.721	0.169	0.109	0.018	0.015	0.002
5	0.045	0.018		0.892	0.084	0.803	0.195	0.153	0.023	0.017	0.003
6	0.062	0.014		0.928	0.055	0.714	0.175	0.121	0.013	0.014	0.001
7	0.037	0.019		0.938	0.077	0.797	0.200	0.192	0.039	0.017	0.002
8	0.057	0.021		0.854	0.098	0.814	0.180	0.211	0.042	0.022	0.003
9	0.049	0.019		0.907	0.083	0.807	0.183	0.200	0.044	0.011	0.003
10	0.054	0.018		0.901	0.079	0.757	0.173	0.167	0.038	0.008	0.002
11	0.103	0.011		0.944	0.069	0.450	0.112	0.077	0.007	0.009	0.002
12	0.068	0.013		0.962	0.075	0.641	0.160	0.070	0.005	0.014	0.002
32BL	0.058	0.022		0.811	0.101	0.817	0.189	0.251	0.058	0.020	0.004
IBL	0.040	0.022		0.852	0.085	0.846	0.186	0.226	0.065	0.015	0.003

Appendix C

The statistical software program JMP, and specifically the Matched Pairs platform was used to compare response metrics and reveal any difference between scenarios. The graphical output from the Matched Pairs platform is called a Tukey mean-difference plot, which presents the paired differences by the paired means. In other words, the value from one set is subtracted from the value from the second set and this difference is plotted on the vertical axis, while the average of the same two values is plotted on the horizontal axis. If there are points that have no difference between the two sets, they will fall exactly on the zero horizontal axis. Points from values that are much greater in one set will appear far away from the zero axis, and if the majority of points follow this trend it becomes apparent that there is a difference between the two sets.

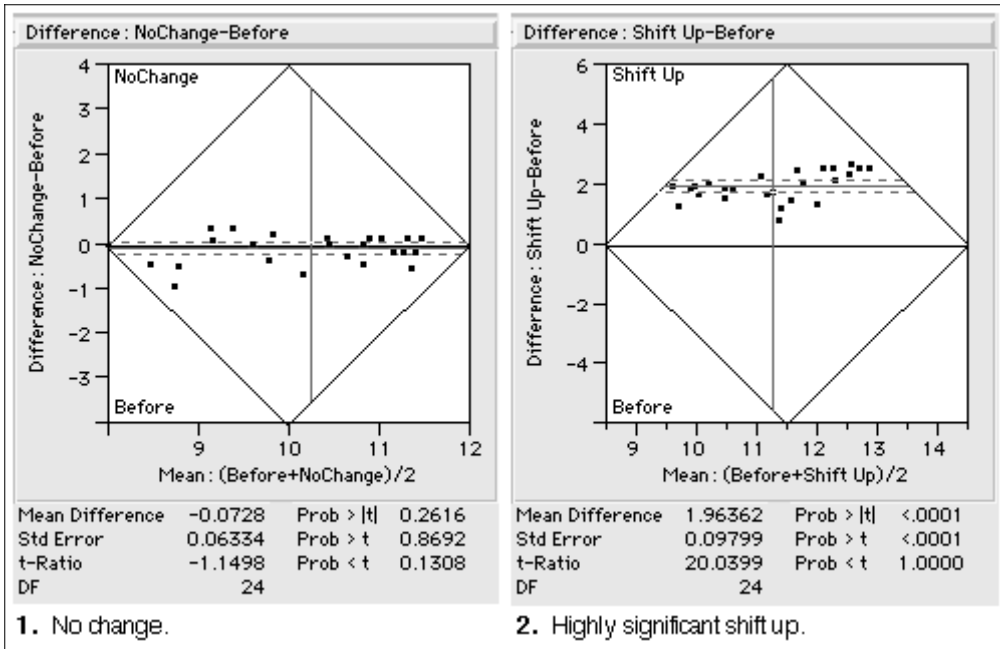
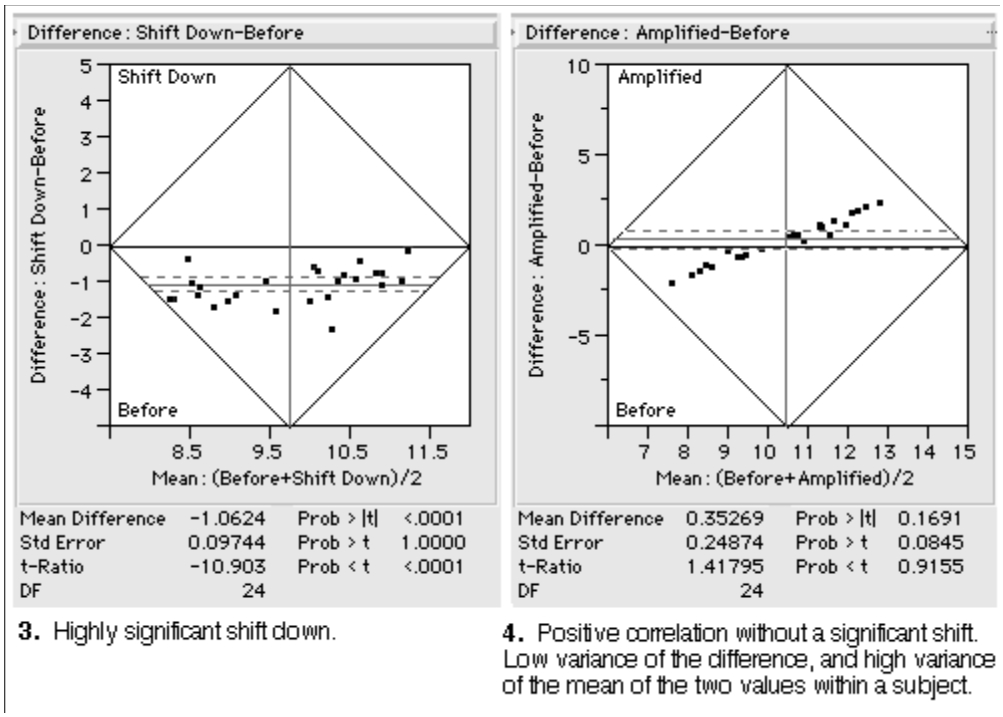


Figure 53. Examples 1 and 2 of Matched Pairs Analysis from JMP



3. Highly significant shift down.

4. Positive correlation without a significant shift. Low variance of the difference, and high variance of the mean of the two values within a subject.

Figure 54. Examples 3 and 4 of Matched Pairs Analysis from JMP

A JMP reference book [SAS, 2007] notes that this graph is the same as a scatterplot of the two original variables, with a 45 degree rotation and rescaling to turn the original coordinates into a difference and a mean. Red horizontal lines are added to the graphs to illustrate the mean of differences (solid line) and 95% confidence interval around that value (dotted lines above and below). If the confidence interval includes the zero axis then the means are not significantly different at the 0.05 level, and it is possible the two sample sets do not have any underlying difference that affects the metric.

The statistical method underlying the matched pair approach is the paired t-test, which examines the distribution of the differences between two sets and tests if the mean of the differences is statistically different from zero. The output *p*-value is the probability of observing a discrepancy in means as large as (or larger than) that observed, even if there is no underlying difference in the means. This *p*-value is compared against a given alpha level (0.05 is commonly used) and if the *p*-value is lower, the difference is determined to be statistically significant. For more information on the Matched Pairs platform including its output and statistical tests, please see [13].

The graphical plots of differences can reveal patterns or correlation in the data, while the statistical test will state the difference in means and whether that difference is statistically significant. It then remains to decide whether that difference is practically significant, or if it is too small to have any practical effect.

Appendix D

The *Strategic Alert Evaluator* (SAE) software evaluates the alerts generated by a conflict probe in order to determine how they match to actual conflicts that occurred in the scenario. SAE provides a result for each event in the scenario. The result is whether the event is evaluated as a Valid Alert (VA), False Alert (FA), Missed Alert (MA), or Discard. The current version of SAE does not support the new definition change to Missed Alerts or the new Late Alert (LA) evaluation. The LA/MA determination was performed by separate software for this analysis.

The following flowcharts describe the process flow of SAE and how each evaluation is assigned to the event.

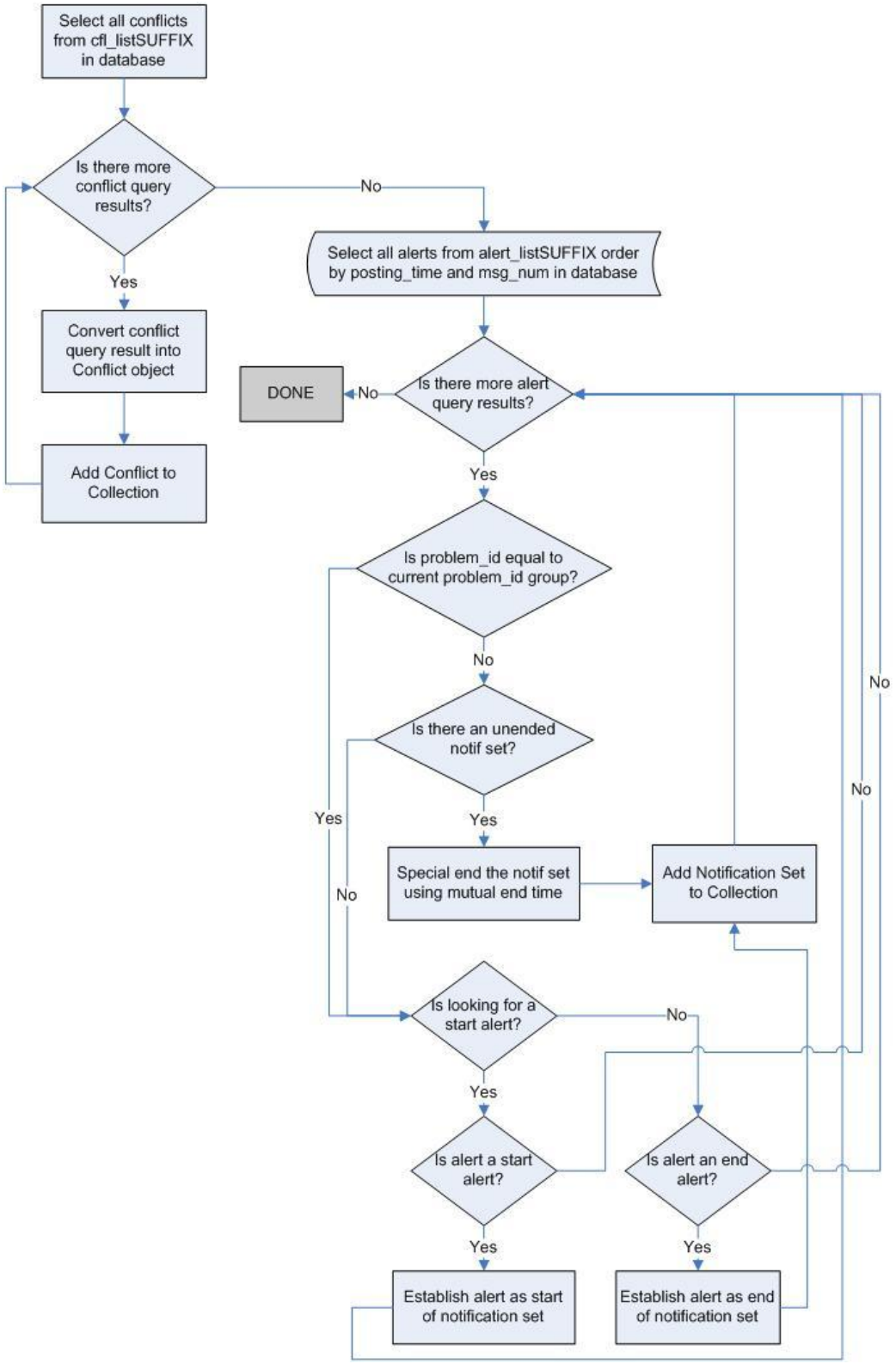


Figure 55. StrategicAlertEvaluator Process A – Establishing Notification Sets from Alerts

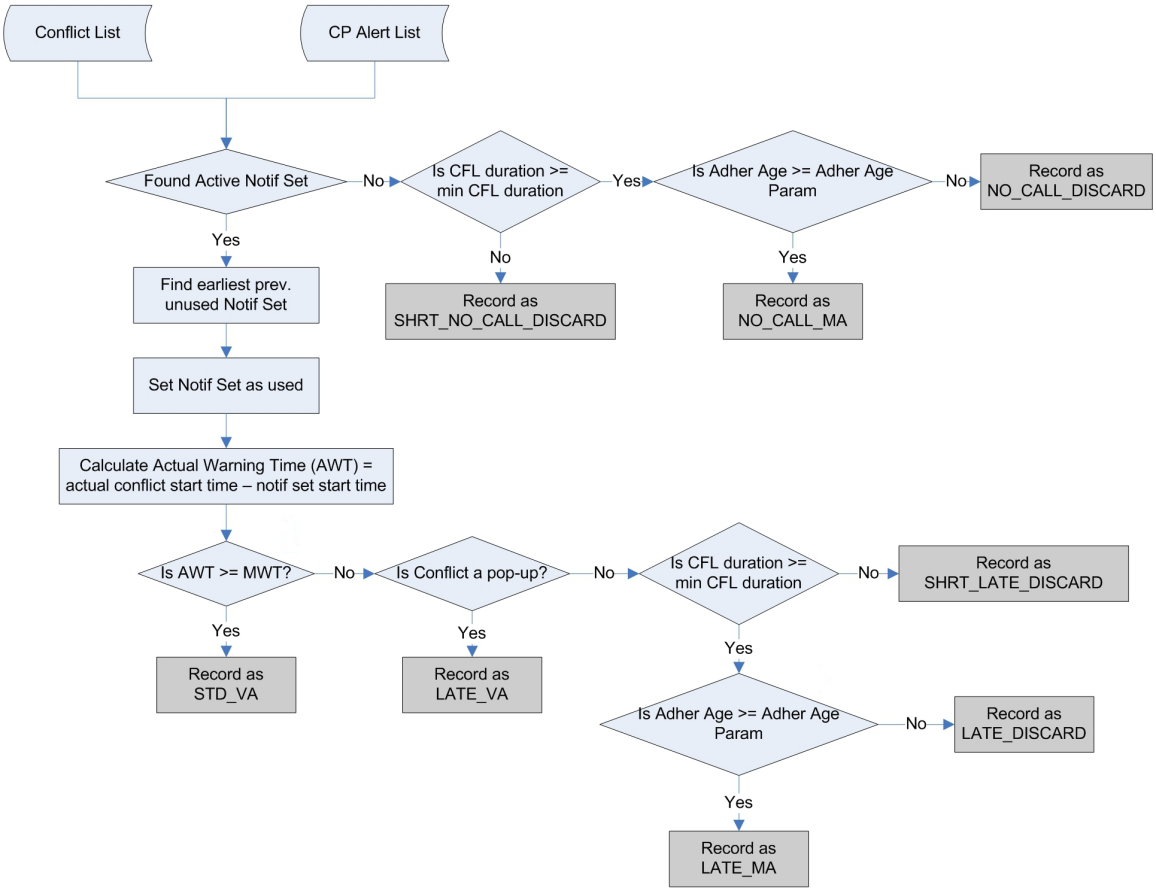


Figure 56. StrategicAlertEvaluator Process B – Evaluating Actual Conflict Notification Set

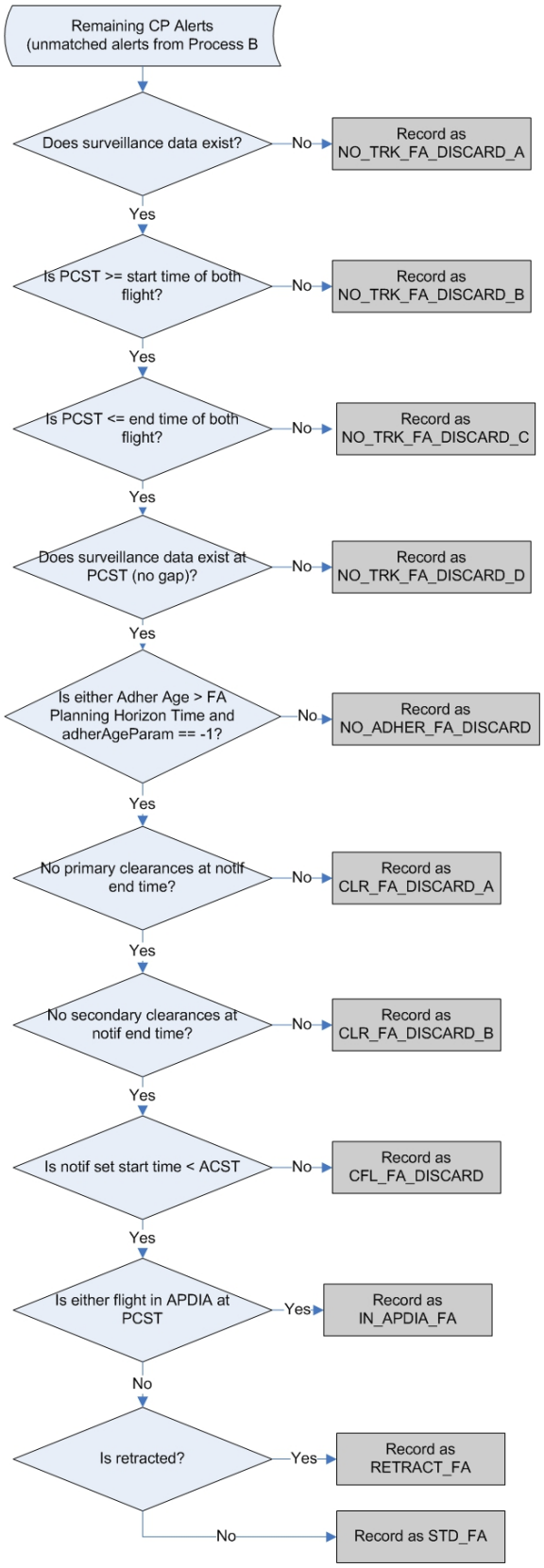


Figure 57. StrategicAlertEvaluator Process C – Evaluating False Notification Set