The Impact of Trajectory Prediction Uncertainty on Air Traffic Controller Performance and Acceptability

Joey Mercer[1], Nancy Bienert[2], Ashley Gomez[3], Sarah Hunt[4],
Joshua Kraut[5], Lynne Martin[6], and Susan Morey[7]
San Jose State University / NASA Ames Research Center, Moffett Field, California, 94035, USA

Steven M. Green[8] and Thomas Prevôt[9]
NASA Ames Research Center, Moffett Field, California, 94035, USA

Minghong G. Wu[10]
University of California, Santa Cruz / NASA Ames Research Center, Moffett Field, California, 94035, USA

A Human-In-The-Loop air traffic control simulation investigated the impact of uncertainties in trajectory predictions on NextGen Trajectory-Based Operations concepts, seeking to understand when the automation would become unacceptable to controllers or when performance targets could no longer be met. Retired air traffic controllers staffed two en route transition sectors, delivering arrival traffic to the northwest corner-post of Atlanta approach control under time-based metering operations. Using trajectory-based decision-support tools, the participants worked the traffic under varying levels of wind forecast error and aircraft performance model error, impacting the ground automation's ability to make accurate predictions. Results suggest that the controllers were able to maintain high levels of performance, despite even the highest levels of trajectory prediction errors.

Nomenclature

ADS-B = Automatic Dependent Surveillance-Broadcast
AOL = Airspace Operations Laboratory
ATA = Actual Time of Arrival
ATC = Air Traffic Control
ATM = Air Traffic Management
ATWIT = Air Traffic Workload Input Technique
DSR = Display System Replacement
DST = Decision Support Tool
D-side = radar-associate controller
ETA = Estimated Time of Arrival
FAA = Federal Aviation Administration
FL = Flight Level
FMS = Flight Management System
MACS = Multi-Aircraft Control System

[1] Research Psychologist, Human-Systems Integration Division, NASA ARC Mail Stop 262-4, AIAA Member.
[2] Research Associate, Human-Systems Integration Division, NASA ARC Mail Stop 262-4.
[3] Research Psychologist, Human-Systems Integration Division, NASA ARC Mail Stop 262-4.
[4] Research Psychologist, Human-Systems Integration Division, NASA ARC Mail Stop 262-4.
[5] Research Psychologist, Human-Systems Integration Division, NASA ARC Mail Stop 262-2, AIAA Member.
[6] Research Psychologist, Human-Systems Integration Division, NASA ARC Mail Stop 262-4.
[7] Senior Research Associate, Human-Systems Integration Division, NASA ARC Mail Stop 262-4, AIAA Member.
[8] Aerospace Engineer, Aviation Systems Division, NASA ARC Mail Stop 210-10, AIAA Associate Fellow.
[9] Research General Engineer, Human-Systems Integration Division, NASA ARC Mail Stop 262-4, AIAA Senior Member.
[10] Research Engineer, Aviation Systems Division, NASA ARC Mail Stop 210-8, AIAA Member.

Mdn = median
NAS = National Airspace System
NASA = National Aeronautics and Space Administration
NextGen = Next Generation Air Transportation System
nmi = nautical mile
RMS = Root-Mean-Square
RUC = Rapid-Update Cycle
STA = Scheduled Time of Arrival
TBO = Trajectory-Based Operations
TOD = Top of Descent
TRACON = Terminal Radar Approach Control
VNAV = Vertical Navigation
WAK = Workload Assessment Keypad

I. Introduction

THE National Airspace System (NAS) forecasts continued growth in traffic demand[1], and under the plans for the Next Generation Air Transportation System (NextGen), the Federal Aviation Administration (FAA) aims to address one of the system's constraining factors, the controller's mental capacity, by increasing the use of automation aids[2, 3]. Typical examples of such Decision Support Tools (DSTs) are Medium-Term Conflict Detection functions, speed advisories, and other arrival and departure management functions. These tools depend on the predicted speed and path of an aircraft: its trajectory. The performance of DSTs, and ultimately their operational acceptance, is likely dependent on the accuracy of the underlying trajectory predictions. NextGen foresees a shift towards Trajectory-Based Operations (TBO)[2, 3], which are expected to be supported by DSTs, further highlighting the importance of trajectory prediction accuracy.

II. Background

Trajectory prediction capabilities are the foundation on which future Air Traffic Management (ATM) systems will likely be built. These predictions will be used by NextGen automation tools to provide advisory aids. The TBO in NextGen rely on an aircraft's four-dimensional trajectory, enabling DSTs for air traffic management tasks, such as conflict detection and resolution, and time-based metering[2, 3]. Trajectory predictions, by their very nature, are not perfect: they are informed guesses.

The general problem with trajectory prediction accuracy has been a topic of study for several years. When predicting the trajectory of an aircraft, several factors must be taken into account. Some factors in the equation are known, but many are not, and become represented by estimates. The collection of factors that must be estimated and considered in order to generate a trajectory prediction is well documented[4, 5]. Often these estimates depend on other estimates, any of which can be modeled to varying degrees of fidelity[5], adding to difficulties of accurately predicting a trajectory. The quality of the input data being used to formulate the estimates also has a clear impact on the resulting trajectory prediction[6].

Advanced DSTs included in future ATM systems will take the predicted trajectory of an aircraft and determine whether an alert, either cautionary (such as highlighting a conflict) or more informational (such as suggesting a speed to meet a schedule), is warranted. Uncertainties in trajectory predictions then, directly impact the DST's ability to help the controller perform their task. While it is possible that the controller could compensate for the errors in 'bad automation,' they can only do so to a limited extent. Figure 1 illustrates the theory that the controller's efforts to compensate for the automation may reach a point at which the controller reaches a workload ceiling (represented in the figure as a vertical dashed line) and performance worsens. The study described here is the first

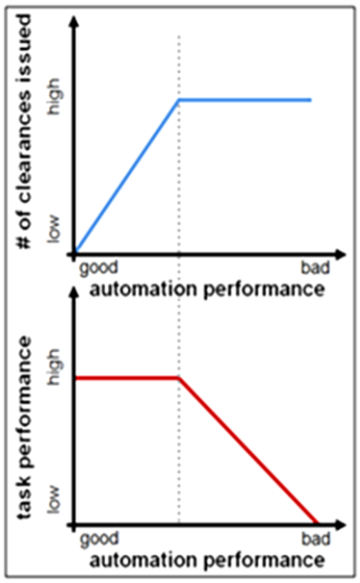

Figure 1. A theoretical effect of automation prediction accuracy on controller workload and performance.

significant attempt to answer an important question regarding trajectory prediction performance: "How accurate does a trajectory prediction need to be to support successful NextGen TBO concepts?"

III. Methods

A Human-In-The-Loop simulation was conducted in January of 2013 in the Airspace Operations Laboratory (AOL)[7] at NASA's Ames Research Center. Shown in Fig. 2, the simulation airspace included two en route sectors (one high-altitude and one low-altitude) feeding the northwest meter-fix of Atlanta's Terminal Radar Approach Control (TRACON).

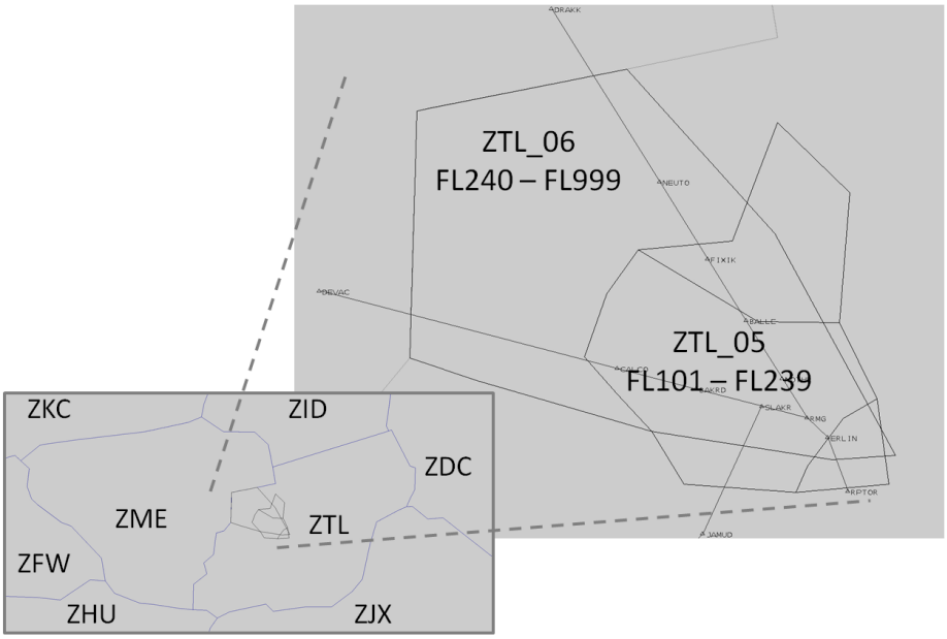

Figure 2. The two test sectors in the northwest portion of the Atlanta en route airspace.

The participants staffing the test sectors were retired air traffic controllers, none of whom were familiar with the test airspace, and had an average of 23.75 years of experience and had been retired for an average of 5.5 years[*]. Confederate controllers, also retired, staffed four additional positions: a 'ghost' sector consisting of the surrounding en route airspace; a TRACON 'ghost' position; and two Radar-Associate (D-side) positions, one for each test sector. Lastly, six student/general aviation pilots staffed four confederate pseudo-pilot positions. During a one-week study, two separate simulations were conducted simultaneously and in parallel, creating two 'worlds'. Although the approach of two worlds required twice the number of positions described above, it doubled the amount of data collected in the same amount of time. Figure 3 depicts the lab layout used during the study.

A. Envisioned Operational Concept

The working environment being studied focused on time-based metering operations working an arrival push, requiring the controllers to deliver arrival traffic in accordance with scheduled times over the meter-fix. Arrival traffic entered the high-altitude sector mainly from the northwest, descending along standard arrival routes into the low-altitude sector. The low altitude-sector was responsible for managing the merging traffic flows at the meter-fix to deliver the aircraft to the TRACON. The high-altitude sector worked to pre-condition the traffic in order to help the low-altitude controller deliver the aircraft to the meter-fix on time. The test participants were instructed that they were responsible for delivering aircraft to the meter-fix within 20 seconds of the scheduled time (information shown to them on their display in a meter list and in aircraft data blocks), as well as providing standard separation services for all aircraft. Over-flight and departure traffic added to the environment, thereby increasing the complexity of the

[*] Participants had 25, 23, 25, and 22 years of experience, and retired in 2007, 2008, 2007, and 2008.

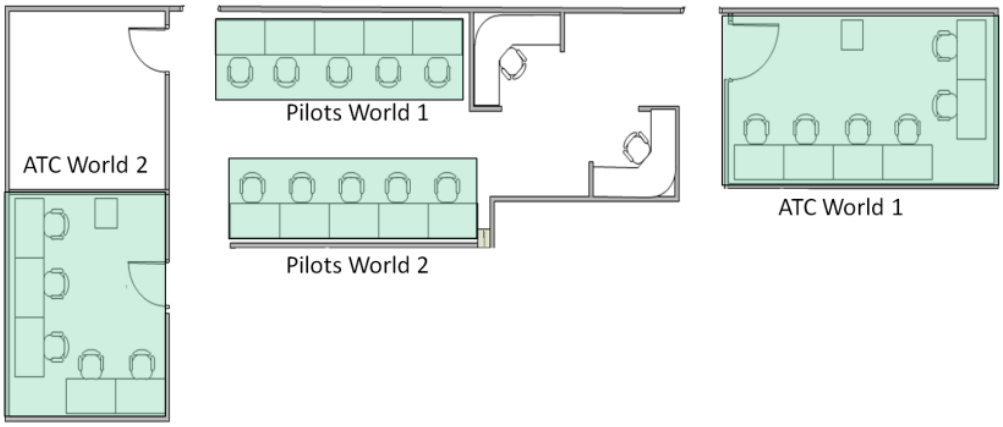

Figure 3. Simulation lab layout

task of providing separation. All simulated aircraft were equipped with a Flight Management System (FMS) and Automatic Dependent Surveillance-Broadcast (ADS-B) -out capabilities. The simulation's primary test conditions did not include any Data Comm-equipped aircraft, requiring controllers to issue all instructions via voice.

All workstations used the Multi-Aircraft Control System (MACS)[7]. The controllers used MACS' Display System Replacement (DSR) emulation hosted on large-format monitors similar to those used in current air traffic control facilities. Specialized keyboards and trackballs like those used in the field helped to further replicate the look and feel typical of these facilities.

In performing their duties, the controllers had several automation aids available to them, as shown in Fig. 4. The time-frame of the operations simulated in this study was anchored in a notional 'mid-term' NextGen environment, where better surveillance data and advanced DSTs could be expected to be available. Meter lists contained arrival sequence and schedule information, while delay indications were also shown in the data block. The precision of the delay values was to the tens of seconds, meaning delay values shown were always in increments of 10 seconds. A conflict probe displayed information regarding detected conflicts in a conflict list as well as in the data block.

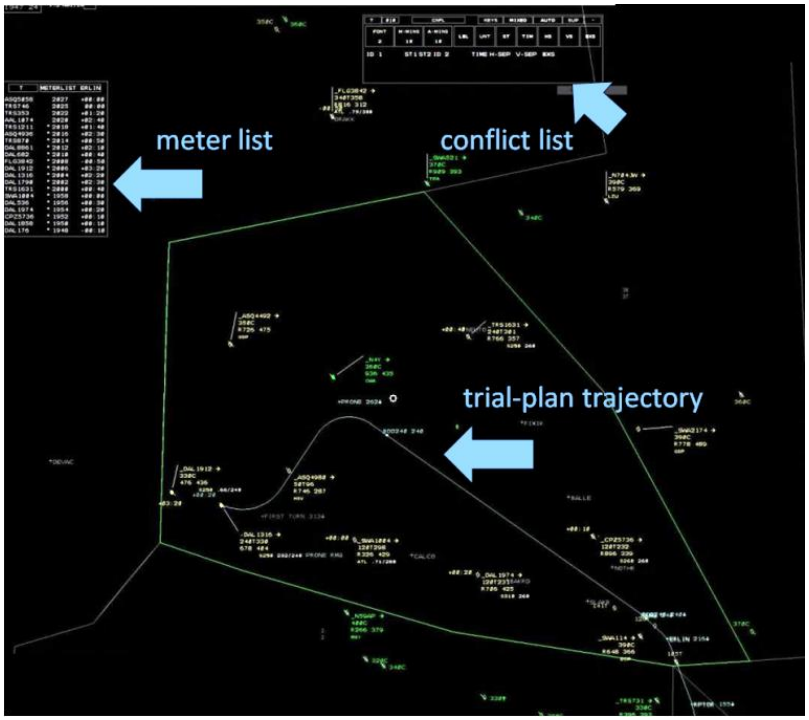

Figure 4. Prototype meter list, conflict list, and trial-planning tools.

Trial-planning tools were available, allowing controllers to manually craft trajectory changes to solve local problems. The provisional trajectories associated with these trial-plans were checked in real-time for potential conflicts and delay impact, the results of which were displayed to the controller, serving as integrated 'what-if' feedback. Controllers could then issue desired route amendments to the aircraft via voice, a process facilitated by the fact that the lateral route changes made with the trial-planning tools automatically snapped to named waypoints. This method avoided any difficulties expected with communicating latitude/longitude coordinates.

Additionally, functions were available to the controllers that allowed them to request a solution from the automation for detected problems. When requested, the automation (using software originally developed as part of the Advance Airspace Concept[8]) would attempt to find a new trajectory for the aircraft, by using lateral, vertical, and speed maneuvers to resolve detected traffic conflicts or address metering problems. The result was then displayed to the controller in the form of a 'trial-plan', at which point controllers could proceed as if they had crafted the trial-plan themselves: either issue the new trajectory clearance to the aircraft, or modify the trial-plan further.

It is important to note that the delay information displayed in the meter list and data block was configured with a unique behavior that could be considered a tool in its own right. Any amendments made by the controllers to an aircraft's trajectory, either manually entered, or trial-plan-assisted, caused the ground system's automation to immediately compute a new trajectory incorporating the newly-available information. For example, if the controller issued a speed clearance to an aircraft, when inputting the new speed as a system entry, the automation would then compute a new trajectory for the aircraft based on the new speed. This had an immediate effect on the delay information displayed in the data block and meter list, which would immediately update to reflect the new trajectory prediction. Making controller actions known to the ground system's automation enabled the trajectory predictions to better represent the controller's current plan. In contrast, a trajectory that is based only on initial 'nominal' speeds would become 'outdated' once the controller issued a new speed instruction, at which point it could only rely on the real-time progress of an aircraft for updated information to be assimilated into new predictions, eventually estimating something more accurate. Thus, it was important that the participant controller not just issue their instructions verbally, but enter them into the system also.

Consider an aircraft with a meter-fix Estimated Time of Arrival (ETA) of 18:28:00 that receives an instruction from Air Traffic Control (ATC) to reduce to 240 knots. If the trajectory predictions remain unaware that this clearance has been issued, it is the position data (representing the aircraft's movement), gathered over time, that will contribute to a new trajectory prediction resulting in a later ETA. The result is that as the aircraft gets closer to the meter-fix, the ETA becomes more accurate. In contrast, and as simulated in this study, when the controller makes a system entry to reflect the new speed, the automation can use this information to assume the aircraft will fly differently, and thus compute a new trajectory. As a result, the automation's new ETA is used to immediately display new delay information in the meter list and data block, expecting the aircraft to arrive later.

Because all these tools were trajectory-based, they were subject to the various simulated errors inherent in those trajectories, meaning that due to uncertainties in the trajectories provided, the tool information displayed to the controllers was imperfect. This highlights one of the simulation's primary objectives: to examine at which point the automation tools would become unacceptable to the controllers and no longer support adequate system performance in terms of separation services or metering conformance.

B. Sources of Simulated Trajectory Prediction Errors

In order to understand when trajectory automation predictions would become unacceptable to controllers and performance targets could no longer be met, the simulation manipulated uncertainties inherent in these trajectory predictions. Serving as the study's primary independent variable, uncertainties were introduced in the form of wind forecast errors and errors in aircraft performance assumptions (e.g., climb/descent rates). A selection of different Rapid-Update Cycle (RUC) wind files created mismatches between environment and forecast wind fields. Wind forecast errors either over-predicted or under-predicted a predominant tail-wind by varying amounts. A baseline condition with no wind errors was included, as well as 'Realistic' Root-Mean-Square (RMS) wind errors of 10 knots, meant to represent typical 'real-world' forecast errors. Other levels of wind error included 'Moderate' RMS errors of 20 knots, and 'Large' RMS errors of 30 knots.

The simulation also investigated errors in the underlying aircraft performance models, which were implemented such that while the ground system's assumptions about aircraft performance remained constant, the actual descent and climb performance of individual aircraft was flown according to modified 'scale factors.' The scale factors were designed to impact the distance normally needed by an aircraft to descend from one altitude constraint to the next, or to climb from one altitude constraint to the next. For example, as determined by its FMS, an aircraft cruising at Flight Level (FL) 350 might normally need 100 nmi to make an idle descent down to the meter-fix's crossing restriction of 12,000 feet. A scale factor of 1.0 would indicate no change to the aircraft's typical

performance, creating the expectation that this aircraft would descend at or near 100 nmi away from the meter-fix. By comparison, a scale factor of 1.1 for the same aircraft would cause its FMS to calculate a Top of Descent (TOD) 110 nmi prior to the meter-fix. Meanwhile, the ground system's automation would still generate a trajectory prediction assuming a TOD 100 nmi out, creating room for errors in ETA predictions, conflict probing, etc. Conversely, a scale factor of 0.9 would produce an FMS-computed TOD 90 nmi prior to the meter-fix, again creating a discrepancy between the expectation of the ground system's automation and the aircraft's actual performance.

The simulation examined a baseline condition with no aircraft performance errors, as well as two additional target levels of aircraft performance model errors. First, all aircraft were split into two populations: arrivals and non-arrivals. Aircraft in each population were then assigned a scale factor across a normal distribution. More specifically, a 'Realistic' error condition targeted aircraft performance errors of 5%, accomplished by assigning one standard deviation (68%) of the population a scale factor between 0.95 and 1.05. Two standard deviations (27%) of the population had a scale factor of either 0.9 or 1.1, and the remaining 5% of the population had a scale factor of either 0.85 or 1.15. A 'Large' error condition, meant to investigate aircraft performance errors of 12.5%, assigned one standard deviation (68%) of the population a scale factor between 0.875 and 1.125, two standard deviations (27%) of the population a scale factor of either 0.75 or 1.25, and the remaining 5% of the population a scale factor of either 0.6875 or 1.3125. Figure 5 illustrates an example of this distribution.

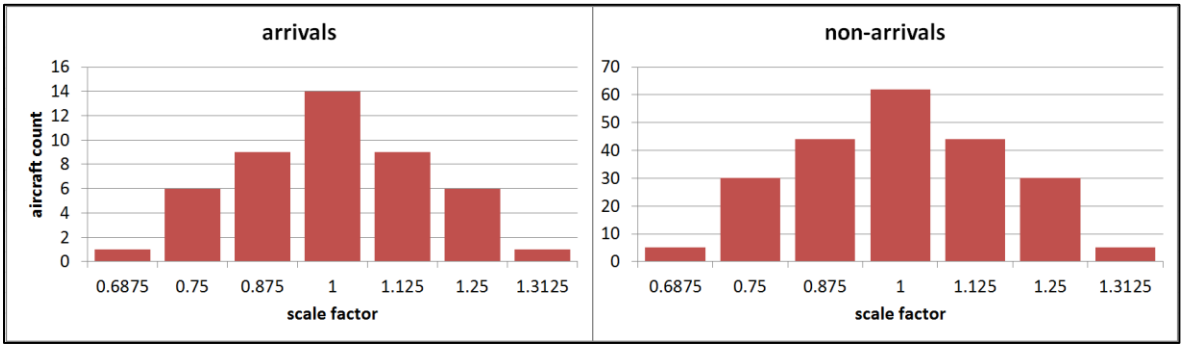

Figure 5. Performance error distributions for arrivals and non-arrivals.

One other source of prediction uncertainty was included in the simulation, meant to mimic typical real-world operations. Flight technical errors were incorporated into the study, simulated as variability in the speed maintained by an aircraft. For the entirety of the simulation, all aircraft were configured with an over-speed tolerance of 10 knots. Only if an aircraft in idle descent was more than 10 knots off its target speed would the FMS' Vertical Navigation (VNAV) guidance then try to correct the error. The potential impact of this is that while the ground system's automation builds a trajectory prediction based on assuming a particular speed, in actuality, the aircraft may be flying a slightly different speed.

C. Traffic Scenario Characteristics

The simulation employed two primary scenarios, intended to minimize potential learning effects. Based on historical real-world traffic, wind, and route data, the scenarios represented traffic consistent with Atlanta operations, but were modified to meet requirements of the study's design. The traffic scenarios consisted of arrivals as well as crossing, over-flight traffic. A light number of departures were also included. The scenarios were designed such that the delays associated with the arrival aircraft would gradually build up to their peak approximately 40 minutes into the scenario, and then decrease for the remaining 15 minutes. Smaller delays ranged from 0 – 2 minutes, while larger delays reached as high as five minutes. The smaller delays were intended to be worked by mostly speed control, whereas the larger delays were expected to require the controllers to vector aircraft.

Coupling the two scenarios with different combinations of forecast and environment ('truth') winds allowed the simulation to not only examine different magnitudes of wind error, but also both directions of error bias. A positive bias (an over-prediction error) was achieved when forecast winds were stronger than the environment winds. Forecast winds that were weaker than the environment winds resulted in a negative bias (an under-prediction error). Environment and forecast winds were paired in these two different ways for each of the four wind error conditions. As an example, 80-knot environment winds paired with 60-knot forecast winds produced a negative bias, whereas

the same environment winds paired with 100-knot forecast winds produced a positive bias. A total of eight pairings were distributed across the two traffic scenarios (scenarios A and B), as shown in Fig. 6.

	N	R	M	L
wind bias	0	10	-20	30
(env/forec)	70/70	70/80	80/60	70/100
wind bias	0	-10	20	-30
(env/forec)	90/90	90/80	80/100	90/60

Figure 6. Wind field combinations used with scenario A (white) and scenario B (green). 'No wind error', and 'Realistic', 'Moderate', and 'Large' wind errors are denoted N, R, M, and L, respectively.

D. Study Design and Conduct

The simulation followed a repeated-measures design with two trials per condition. Six different combinations of wind forecast error and aircraft performance error served as the study's conditions, displayed in Fig. 7. Error conditions are named according to the combination of aircraft performance error (N, R, or L) and wind forecast error (N, R, M, or L), with the first letter representing aircraft performance error. Prior to data collection, the participants received two days of training, during which time they worked variations of the primary traffic scenarios to help gain familiarity with the tools, their roles and responsibilities, as well as their sector's general operations. The run schedule's arrangement presented the different conditions sequentially, gradually increasing the errors over the first 12 55-minute runs, then gradually decreasing the errors over the second half of the study. This design expected to capture noticeable detriments to performance and workload at some point during the 'ramp-up' of the errors, after which the 'ramp-down' period would provide additional opportunities to capture the point at which the error conditions caused problems for the controllers.

wind error / aircraft performance	no error	realistic error	moderate error	large error
no error	NN			
realistic error		RR	RM	RL
large error		LR		LL

Figure 7. Initial combinations of wind forecast error and aircraft performance error.

As an alternative to this 'ladder-up-ladder-down' approach, researchers designed a second run schedule. If, after having experienced all the error levels, no detriments to performance and workload were observed, simply repeating the same conditions in reverse order would unlikely produce useful data. Instead, the alternate run schedule design used the latter part of the study to explore other factors that might impact controller workload and performance in the presence of trajectory prediction errors.

Ultimately, the conduct of the simulation followed the alternate run schedule design, creating opportunities to investigate several additional factors. First, tighter scheduling constraints were explored. Rather than being scheduled at a constant two minutes apart, during one run, aircraft were scheduled 90 seconds apart, with every fourth aircraft scheduled 120 seconds from that in front. This run sought to address concerns that the spacing between consecutive on-time aircraft was large enough that small errors in meter-fix delivery could occur while not jeopardizing separation. A tighter schedule would force the controllers, while coping with the trajectory prediction errors, to be more precise in managing the arrivals. Second, a new scenario was added (scenario C), designed to be somewhat more difficult, with delays that ramped up more quickly, producing a longer 'peak delay' period, where controllers would be working to reduce the high delays. Two runs were used to test scenario C and it's interaction with the RR and LL error conditions, ran in conjunction with environment/forecast wind pairings reflecting negative error biases. Third, as shown in Fig. 8, the study's design was expanded to include a fifth level of wind error: one with 'Extra-Large' 40-knot forecast errors. Data collected during operational field trials verify that this larger wind

wind error / aircraft performance	no error	realistic error	moderate error	large error	extra-large error
no error	NN				
realistic error		RR	RM	RL	
large error		LR		LL	LXL

Figure 8. Tested combinations of wind forecast error and aircraft performance error.

	N	R	M	L	XL
wind bias	0	10	-20	30	40
(env/forec)	70/70	70/80	80/60	70/100	60/100
wind bias	0	-10	20	-30	-40
(env/forec)	90/90	90/80	80/100	90/60	100/60
wind bias		-10		-30	
(env/forec)		90/80		90/60	

Figure 9. Wind field combinations used with scenario A (white), scenario B (green), and scenario C (yellow). 'No wind error', and 'Realistic', 'Moderate', 'Large', and 'Extra-Large' wind errors are denoted N, R, M, L, and XL, respectively.

error is still within the limits of observed, real-world errors[9]. Figure 9 displays the final distribution of wind pairings across the three scenarios.

Additionally, the final six runs investigated the impact of the availability of particular tools. The first of three exploratory toolsets, a no-tools 'Baseline' (BL) set of tools served to represent operations similar to current-day. Another toolset added only one tool to the BL set, investigating the hypothesis that by itself, the tool added significant value. Specifically, having the controller's speed intent immediately available to the automation was predicted to help the controllers better manage the arrival flows in the presence of the simulated trajectory prediction uncertainties. This toolset is referred to as 'Speed intent from the data block's 4th line', (S4). Lastly, the 'Data Comm' (DC) toolset added Data Comm technology to the 'full' toolset, offering the controllers the ability to electronically send any trial-plan directly to the flight deck, rather than issue clearances via voice.

Constraints in the alternate run schedule meant that each toolset could only be run twice. In an effort to capture their possible interactions with the error conditions, the three toolsets were examined in both the 'Realistic-Realistic' (RR) and 'Large-Large' (LL) error conditions. Furthermore, the desire was to test the toolsets in the more difficult scenarios available, in an effort to increase the likelihood of capturing negative impacts on controller workload and performance. In addition to scenario C, scenario A was used in examining the toolsets because observations throughout the first part of the study indicated that scenario A seemed more challenging than scenario B.

Figure 10 presents the final run schedule: runs 1 – 12 comprised the primary investigation, and runs 13 – 23 comprised the exploratory runs. Runs using the same traffic scenario alternated between a base list of aircraft call-signs and a randomized set of call-signs.

System data was collected from each workstation, including aircraft flight states, operator task data, automation states, voice communications, etc. Screen recordings captured as movie files were also saved. Workload Assessment Keypads (WAKs) probed controller workload at three-minute intervals during simulation trials using Air Traffic Workload Input Technique (ATWIT) [9] ratings on a modified six-point scale (1 as low workload, 6 as high workload). The controllers completed questionnaires at the end of each run, as well as a post-simulation questionnaire. Debrief discussions provided an additional opportunity for controllers to offer feedback.

Monday	Tuesday	Wednesday	Thursday	Friday
	5 RL	10 RM	15 LXL	20 RRS4
1 NN	6 RL	11 LL	16 RR	21 LLS4
2 NN	7 LR	12 LL	17 LXL	22 RRDC
3 RR	8 LR	13 LL*	18 RRBL	23 LLDC
4 RR	9 RM	14 LL	19 LLBL	

Figure 10. The simulation's run schedule, indicating traffic scenarios A, B, and C (white, green, and yellow, respectively). Run 13 investigated the tighter scheduling criteria (denoted with an *), runs 15 and 17 investigated the Extra-Large wind errors, and runs 18 – 23 investigated the 'Baseline', 'Speed intent from the data block's 4th line', and 'Data Comm' toolsets (denoted BL, S4, and DC, respectively).

IV. Results

The trajectory prediction uncertainty simulation studied how errors in trajectory predictions might impact the controller's ability to perform their tasks. A rich set of data resulted, allowing researchers to assess the sensitivity of trajectory-based operations to flaws in their underlying trajectories. The results in this paper serve as an overview of the study's outcomes. Results in this paper first present the primary investigation's findings, followed by the findings from the exploratory runs. Follow-on publications discuss detailed analyses that examine controller strategies and tool usage[11, 12].

A. Results of the Primary Investigation

1. Safety (runs 1 – 12)

How close airplanes actually came together offered a measurement of safety for the simulated operations. Losses of separation events can be categorized as Proximity Events or Operational Errors. Proximity events, the less severe of the two, refer to events where any two aircraft were between 4.5 and 5.0 nmi apart laterally and closer than 800 feet vertically. Operational errors involve aircraft pairs with separation less than 4.5 nmi laterally and 800 feet vertically. Either type of event was only considered if its duration was longer than 12 consecutive seconds.

Overall safety was high. One operational error occurred in run 11, which simulated an LL error condition using scenario A. The loss of separation was between an east-bound over-flight and an arrival aircraft on the more northerly flow, taking place in the high-altitude sector.

2. Metering Accuracy (runs 1 – 12)

One of the simulation's primary objectives was to investigate the impact of trajectory prediction errors on the controllers' ability to perform their task. Since delivering aircraft to the TRACON according to the meter-fix schedule was a key aspect of the ATC task, schedule conformance serves as an important performance metric. For this analysis, when aircraft crossed the meter-fix, aircraft were classified into one of three groups: on time, early, and late. An aircraft was considered to arrive at the meter-fix on time if it arrived within 25 seconds of its STA. Aircraft arriving more than 25 seconds early relative to their STA were considered 'early', and those arriving more than 25 seconds late relative to their STA were considered 'late'. The formula used in this analysis was simply STA – Actual Time of Arrival (ATA), with positive values indicating early arrivals, and negative values indicating late arrivals. Overall performance was high; 578 of 598 aircraft (97%) were delivered on time. In comparisons by error condition and traffic scenario, schedule conformance at the meter-fix was always at a 94% success rate or higher (see Fig. 11).

Additionally, the raw STA-ATA values were tested for statistical significance across error condition and traffic scenario. Significant differences in data were not found in either comparison, as seen in Table 1, indicating that the controllers achieved similarly high performance regardless of error condition or traffic scenario.

3. Flight Path Efficiency (runs 1 – 12)

While the controllers appeared to successfully deliver aircraft to the meter-fix on schedule, the manner in which they accomplished this, and its effect on the aircraft merited investigation. An examination of estimated fuel burn is

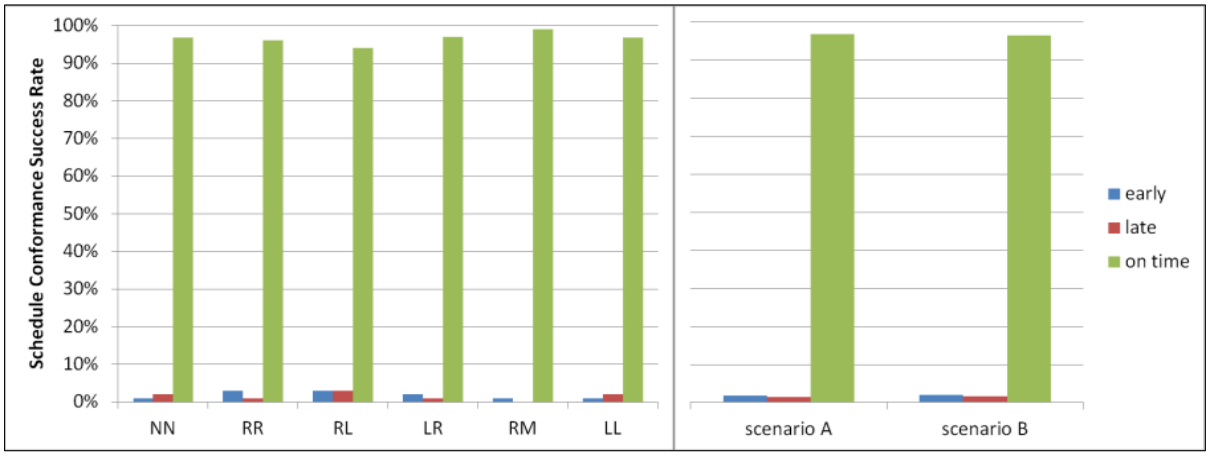

Figure 11. Percentage of aircraft delivered to the meter-fix either early, late, or on time, when compared across error conditions and traffic scenarios.

Table 1. Repeated measure non-parametric comparisons for primary analyses.

Workload	Test	*n*	*df*	χ^2	z	*p*
Error Condition (NN, RR, RL, LR, RM, LL)	Friedman Test (two-tailed)	146	5	19.110	--	†.002
Scenario (A,B)	Wilcoxon Signed-Rank Test (two-tailed)	450	--	--	-8.071	*.000
Meter-fix Schedule Conformance	Test	*n*	*df*	χ^2	z	*p*
Error Condition (NN, RR, RL, LR, RM, LL)	Friedman Test (two-tailed)	96	5	7.039	--	.218
Scenario (A,B)	Wilcoxon Signed-Rank Test (two-tailed)	289	--	--	1.704	.082
Flight Path Distance	Test	*n*	*df*	χ^2	z	*p*
Error Condition (NN, RR, RL, LR, RM, LL)	Friedman Test (two-tailed)	66	5	6.823	--	.234
Scenario (A,B)	Wilcoxon Signed-Rank Test (two-tailed)	154	--	--	-6.053	*.000

Note: Consistent violations of normality (Shapiro-Wilk, $p<.05$) led to non-parametric testing across all analyses. All Friedman tests for significance use post-hoc pair-wise comparisons with Bonferroni corrections.
*Note: * significant at alpha=.05 two-tailed.*
Note: †Error Condition was found to be significant $\chi2(5) = 19.1110$, $p = .002$, but no post hoc pairwise comparisons found significance when applying a Bonferroni correction. Due to the likelihood of type-one error, this analysis is not reporting this result as significant.

currently underway, and will be available in a future publication. As a proxy, path distance flown was used for an initial analysis. This analysis focused exclusively on arrival aircraft, and measured the path distance flown for each flight between a 200 nmi radius arc centered on the meter-fix, and the meter-fix itself. The arc's size matched the scheduler's freeze horizon, and completely contained the test sectors. This approach defined a common 'analysis window' within which to measure a flight's lateral progress. In the ideal case, an aircraft would cross the arc, and fly uninterrupted directly (in a straight line) to the meter-fix. Such a case would result in a recorded distance flown of exactly 200 nmi. Values larger than 200 nmi then, are attributed to indirect routings or controller-issued vectors. Aircraft that start the scenario already inside the arc could produce values smaller than 200 nmi.

The various error conditions exhibit average distances ranging from 203.9 nmi in the RM and LR runs, to 204.4 nmi in the LL runs. Scenarios appear to have a larger impact, with averages of 201.5 nmi for scenario B, as opposed to 206.8 nmi for scenario A (see Fig. 12). The data was further tested to verify these differences, confirming that path distances measured in scenario A (*Mdn*= 207.8 nmi) were significantly longer than in Scenario B (*Mdn*= 201.7 nmi), seen in Table 1. Flight path distance, as a function of error condition was not significantly different. Similar to the findings in the schedule conformance data, this supports the notion that the error conditions were not the most influential factor affecting performance, whereas traffic scenario did appear to affect performance.

4. Controller Workload (runs 1 – 12)

During the simulation, controller responses to real-time workload queries exercised the scale's entire range. Illustrated in Fig. 13, mean workload ratings across the error conditions vary from 2.2 to 2.37, and range from 2.5 to

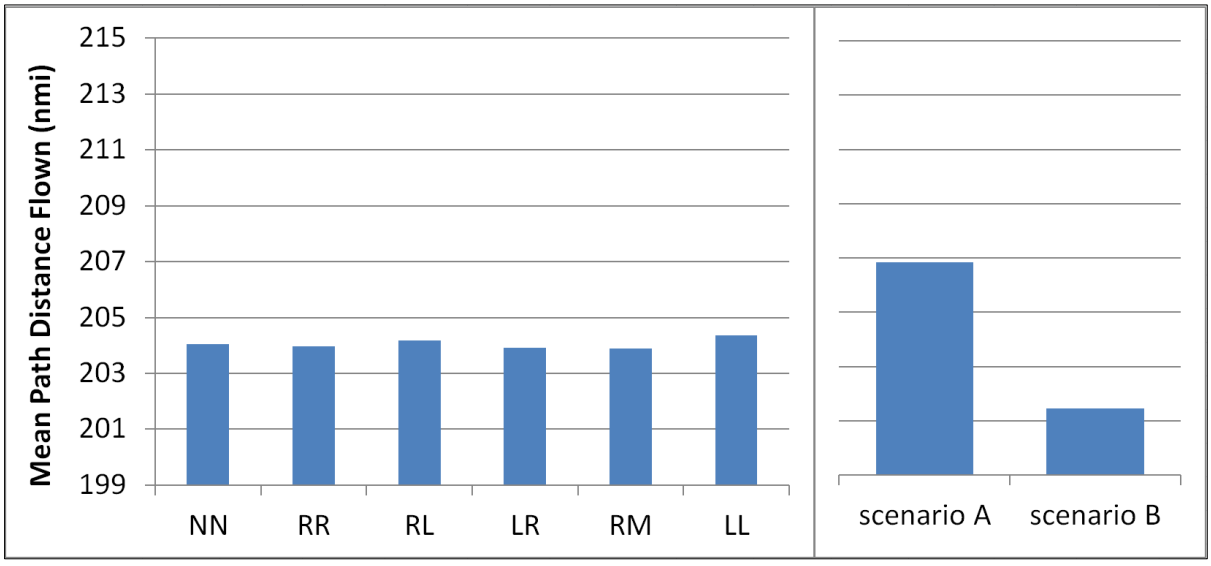

Figure 12. Average path distance flown, according to error condition and traffic scenario.

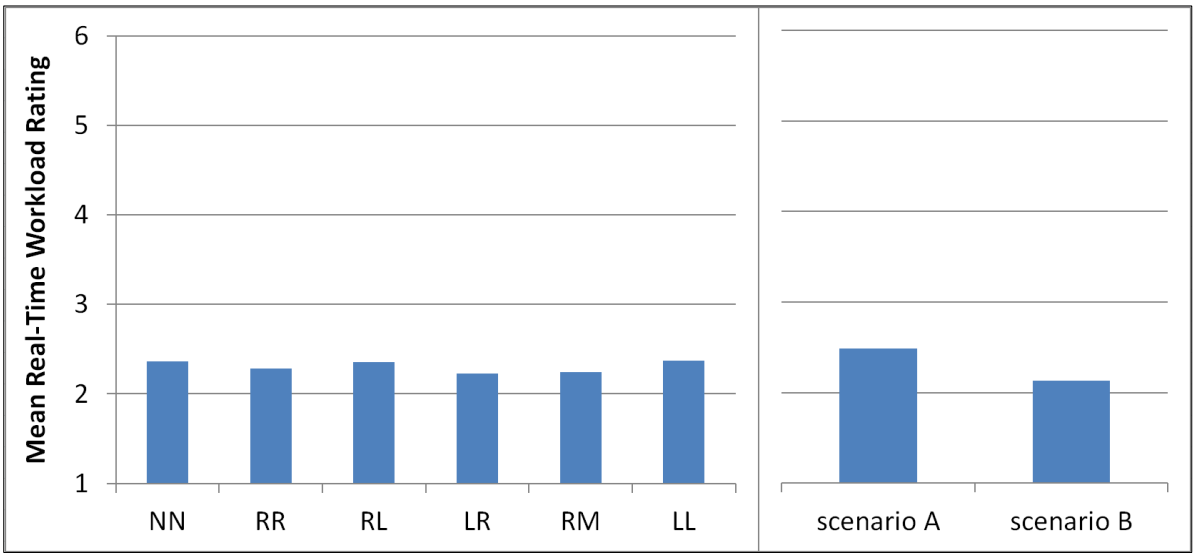

Figure 13. Controllers' average real-time workload rating, when compared across error condition and traffic scenario. A rating of 1 represents low workload, while a rating of 6 signifies high workload.

2.1 for scenarios A and B, respectively. As seen in Table 1, statistical comparisons were statistically different only for traffic scenario. When analyzing the data according to scenario, significant differences were found, showing lower controller workload ratings in scenario B than A.

B. Results of the Exploratory Investigations

Because of constraints in the run schedule, the exploratory runs could not be tested in a balanced matrix, but did afford the opportunity to explore four additional factors that, in the presence of trajectory prediction errors, could potentially impact controller performance and workload. The following results analyze the safety, schedule conformance, flight path efficiency, and real-time workload measures as a function of four different factors: scheduling criteria, traffic scenario, wind error, and toolset availability. To examine the effect of the scheduling criteria, analyses compare runs 13 (tightened scheduler) and 11 (standard scheduler). Two runs for each of the simulated traffic scenarios serve as the means to analyze the effect of traffic scenario. More specifically, comparisons are made between run 14 with 16 (scenario C), run 3 with 11 (scenario A), and run 4 with 12 (scenario B). An analysis of runs 15 and 17 (XL wind error), as compared to runs 11 and 12 (L wind error), examines the effect of increased wind forecast errors. Results regarding the effects of the different toolsets analyze runs 3 and 14 to represent the simulation's typical 'full' toolset, in comparison with the exploratory toolsets simulated in runs 18-23. It is important to understand that co-varying elements are present in the data. Comparative analyses will serve only as indications to possible outcomes; outcomes that will require controlled testing to truly validate. Schedule conformance success rate, and mean values from the flight path distance, and real-time workload ratings described in this section are shown in Figs. 14, 15, and 16, respectively, while details of the statistical analyses can be seen in Table 2.

1. Safety (runs 13 – 23)

A total of three operational errors and one proximity event were recorded during the exploratory runs. They all occurred in the LL error condition, either with scenario A or scenario C. These separation violations are further discussed in the following sections.

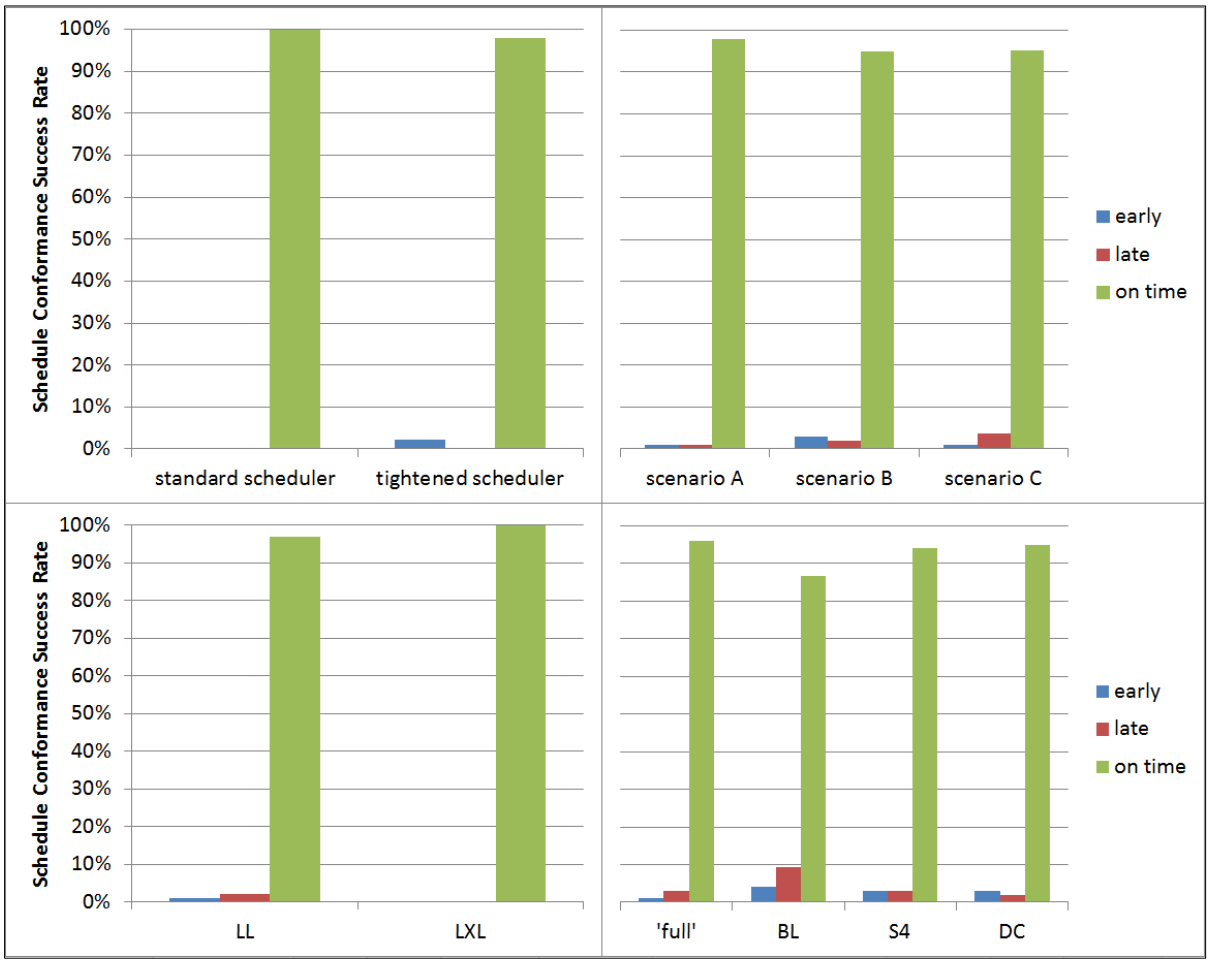

Figure 14. Percentage of aircraft delivered to the meter-fix either early, late, or on time, as reported in the exploratory analyses.

2. Impact of Scheduling Criteria (runs 13 and 11)

Tighter scheduling criteria gave the controllers less room for error when delivering aircraft to the meter-fix. Run 13 tested this new configuration, which is compared against run 11; chosen because, like run 13, it simulated the LL error condition with scenario A. The intention of this manipulation was to increase the difficulty of the metering task by requiring more precision from the controllers, and consequently increasing the risk of separation violations.

During run 13, a proximity event occurred in the high-altitude sector between an east-bound over-flight and an arrival aircraft on the more northerly flow. This event showed striking similarities with the operational error in run 11: both events involved the same high-altitude controller and even the same two aircraft (but with randomized call-signs). An analysis of the meter-fix accuracy data between these two runs shows success rates of 98% (run 11) and 100% (run 13); statistical testing of the data reveals no significant differences. Upon inspecting the flight path distance data, the two runs exhibit identical mean values, and when statistically tested, the medians also show no significant difference. Workload data however, produce mean ratings of 2.3 in run 11, and 2.6 in run 13. Tests for significance confirm that the lower workload reported by the controllers was significantly different with the tightened scheduling criteria (*Mdn*= 2.0) than with the standard scheduling criteria (*Mdn*= 3.0).

These findings suggest that the tightened scheduling criteria was not a problem for the controllers; they were quite capable of delivering aircraft within 25 seconds of the scheduled time. The workload result and the lack of a result in flight path distance allude to the possibility that this condition was potentially easier, which perhaps is a result of the way the simulation was conducted. Since the traffic scenarios were not changed for this manipulation, the tightened scheduling criteria, which scheduled aircraft closer together than in the standard scheduling criteria, had the effect of producing smaller delays. Further investigation is warranted then, to remove this confound.

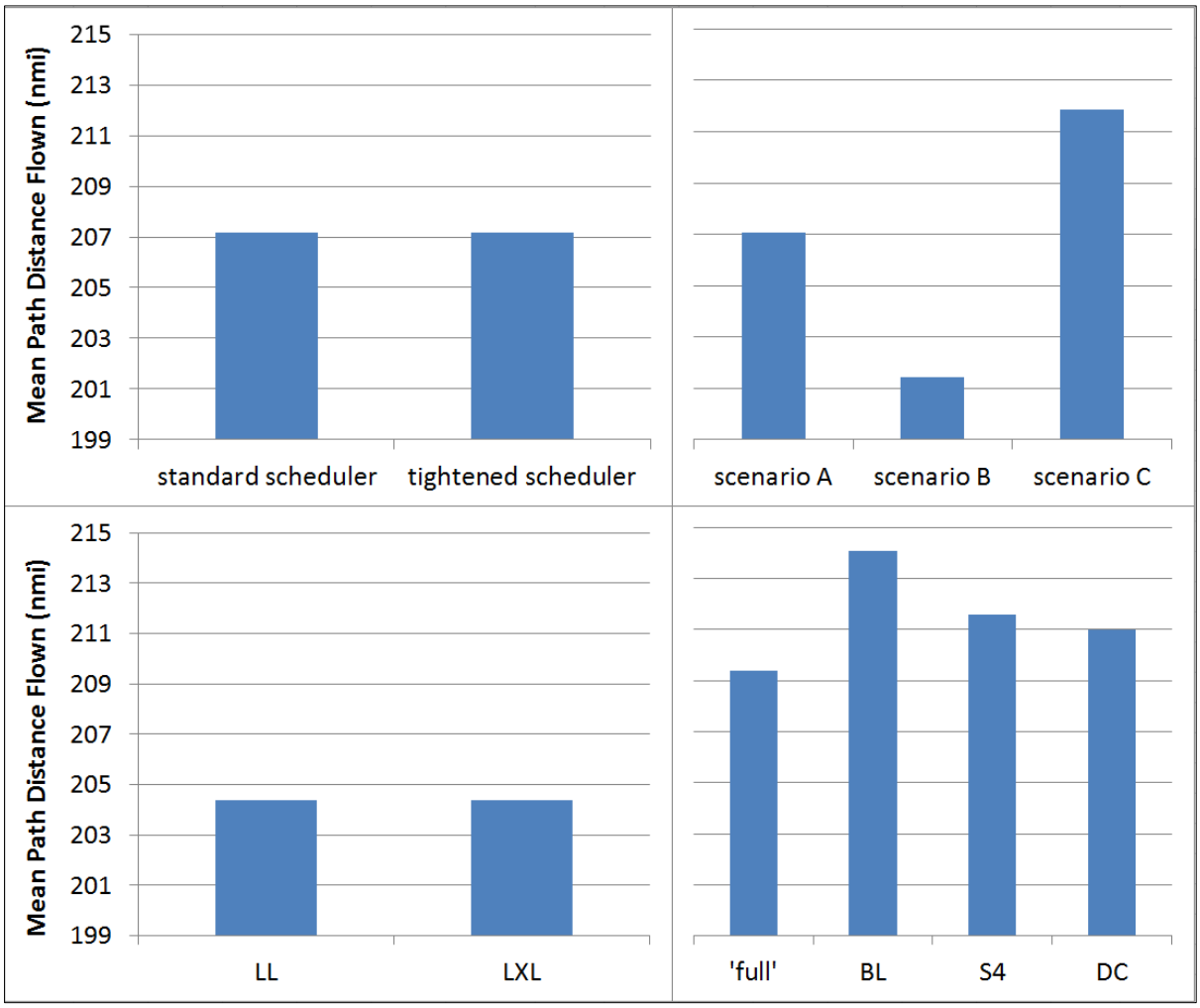

Figure 15. Average path distance flown, as reported in the exploratory analyses.

Scenarios are needed that, when simulated under the different scheduling criteria, produce similar delays, allowing for an even comparison. The recorded separation violation during run 13 is interesting, because it appears unrelated to the scheduling criteria. Since the aircraft involved were in the high-altitude sector, and included an over-flight aircraft (a non-scheduled aircraft), this event may instead be related to either the LL error condition, traffic scenario A, or the combination of the two.

3. Impact of Traffic Scenario (runs 3 and 11, runs 4 and 12, runs 14 and 16)

Three different traffic scenarios allowed researchers to test different levels of difficulty, and investigate the possibility of its interaction with error condition. Task difficulty can be challenging to define: in this context it refers to the heavier push of metered traffic included in scenario C. Over the course of the simulation, each of the three scenarios were run in the RR and LL error conditions, a commonality used to identify runs that allowed for proper comparisons. Runs 3 and 11 simulated scenario A, runs 4 and 12 simulated scenario B, and runs 14 and 16 simulated scenario C, in the RR and LL error conditions, respectively. All analyses in this section, with the exception of the safety data, examine these six runs.

The following discussion of safety issues considers all exploratory runs, rather than just the specific runs selected for direct comparisons. Safety events are typically a rare occurrence, and thus do not lend themselves to statistical analysis. Instead, this discussion will simply make observations of how the separation violations were distributed among traffic scenario. Of the four separation violations recorded during the exploratory runs, one (a proximity event) occurred in scenario A, while scenario C saw three operational errors. These four events were evenly distributed among the controllers; each of the four participants experienced their own separation violation. Common for all four events was that they occurred in the LL error condition. Of the three operational errors, one occurred

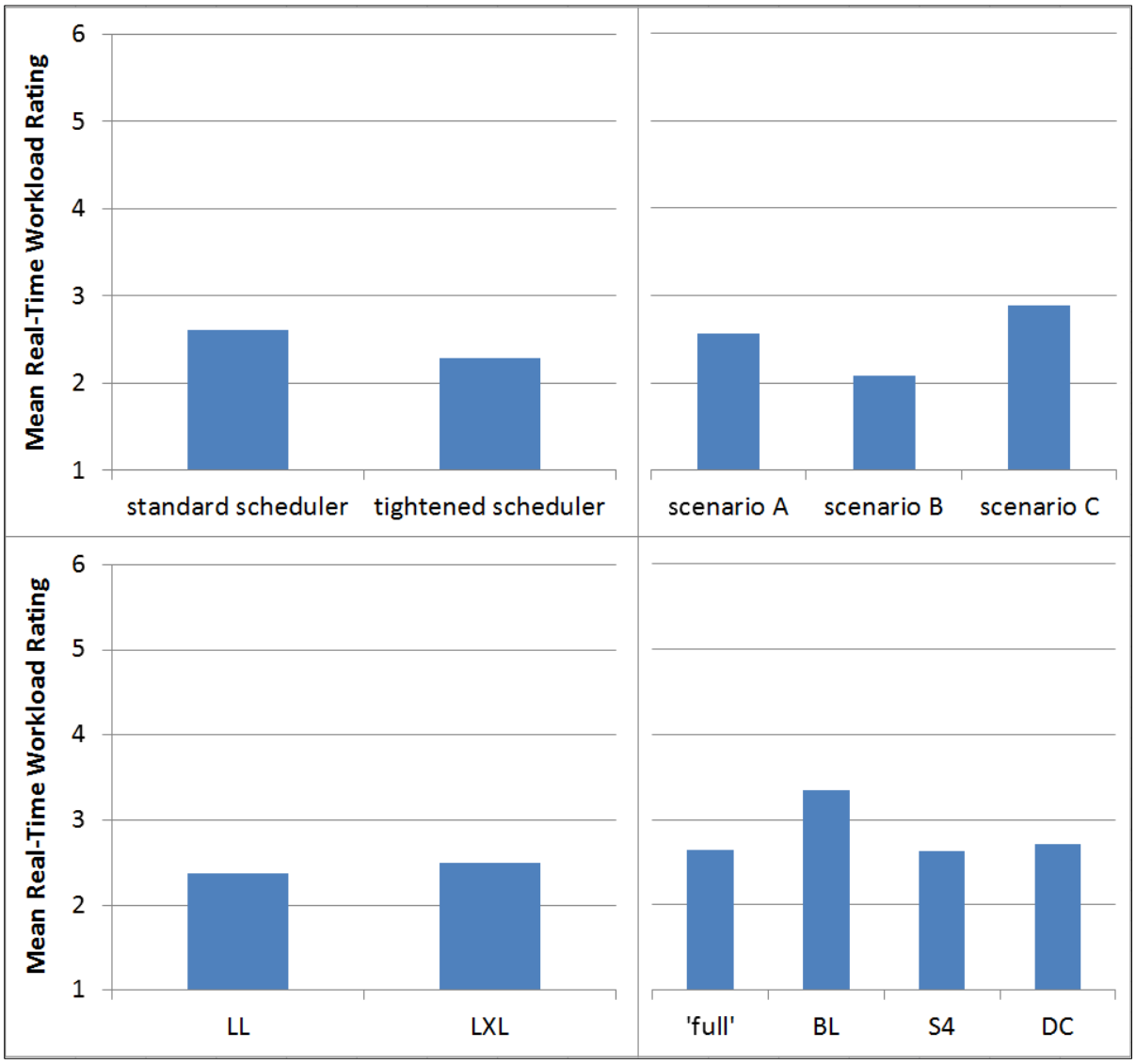

Figure 16. Controllers' average real-time workload rating, as reported in the exploratory analyses. A rating of 1 represents low workload, while a rating of 6 signifies high workload.

between sequential arrival aircraft in the low-altitude sector during run 14; the first time scenario C was used. The other two occurred in run 19, which was a BL toolset condition simulated using scenario C. One of the events in run 19 involved sequential arrival aircraft in low-altitude sector; the other took place in the high-altitude sector, between a north-bound over-flight and an arrival on the more westerly flow. These findings highlight the fact that no separation violations occurred in scenario B, or in any error condition other than the LL condition. It is interesting to note that these traits are also shared by the operational error which occurred in the primary runs (run 11). Additional analyses are needed to better characterize the situations that led to these events, which may help understand the observed relationship to high errors combined with more difficult metering tasks.

An analysis of the meter-fix accuracy data, between the six runs representing the three scenarios, shows success rates of 98% for scenario A, and 95% for scenarios B and C. Statistical testing of the raw STA-ATA data reveals significant differences between all scenarios: scenario B (*Mdn*= 2.0 sec) vs. scenario A (*Mdn*= 2.5 sec), $p<.000$; scenario B vs. scenario C (*Mdn*= 3.0 sec), $p<.000$; and scenario A vs. scenario C, $p=.001$. An inspection of the flight path distance data shows mean values ranging from 201.4 nmi in Scenario B, to 211.9 nmi in scenario C. Tests confirm significance differences of $p<.00$, showing that arrival aircraft flew shorter distances in scenario B (*Mdn*= 201.7 nmi) than in scenario A (*Mdn*= 207.7 nmi) or scenario C (*Mdn*= 209.4 nmi). The workload data, when

Table 2. Repeated measure non-parametric comparisons for exploratory analyses.

Meter-fix Schedule Conformance	Test	*n*	*df*	χ^2	z	*p*
Schedule (Standard, Tightened)	Wilcoxon Signed-Rank Test (two-tailed)	46	--	--	1.268	0.205
Scenario (A,B,C)	Friedman Test (two-tailed)	98	2	16.409	--	*.000
Wind (L, XL)	Wilcoxon Signed-Rank Test (two-tailed)	96	--	--	-.680	.493
Tool Set ('full', BL, S4, DC)	Friedman Test (two-tailed)	96	3	12.086	--	*.007
Flight Path Distance	Test	*n*	*df*	χ^2	z	*p*
Schedule (Standard, Tightened)	Wilcoxon Signed-Rank Test (two-tailed)	31	--	--	-1.156	.248
Scenario (A,B,C)	Friedman Test (two-tailed)	40	2	108.701	--	*.000
Wind (L, XL)	Wilcoxon Signed-Rank Test (two-tailed)	66	--	--	-1.303	.193
Tool Set ('full', BL, S4, DC)	Friedman Test (two-tailed)	60	3	40.872	--	*.000
Workload	Test	*n*	*df*	χ^2	z	*p*
Schedule (Standard, Tightened)	Wilcoxon Signed-Rank Test (two-tailed)	76	--	--	-3.274	*.001
Scenario (A,B,C)	Friedman Test (two-tailed)	152	2	108.701	--	*.000
Wind (L, XL)	Wilcoxon Signed-Rank Test (two-tailed)	152	--	--	-2.253	*.024
Tool Set ('full', BL, S4, DC)	Friedman Test (two-tailed)	152	3	119.927	--	*.000

Note: Consistent violations of normality (Shapiro-Wilk, p<.05) led to non-parametric testing across all analyses. All Friedman tests for significance use post-hoc pair-wise comparisons with Bonferroni corrections.
*Note: * significant at alpha= .05 two-tailed.*

analyzed across these six runs, produces mean workload ratings ranging from 2.1 in scenario B, to 2.9 in scenario C. Further testing of the workload ratings data uncovers significant differences between all scenarios: the controllers reported lower workload in scenario B (*Mdn*= 2) than in scenario A (*Mdn*= 2.5) or scenario C (*Mdn*= 3) $p<.000$, and lower workload in scenario A as compared to scenario C at $p=.001$.

The schedule conformance, flight path distance, and real-time controller workload measures were all sensitive to the different traffic scenarios. Careful consideration must be taken with the meter-fix accuracy results: although statistical significance is shown, differences in STA-ATA data are small and likely do no more than confirm the controllers' ability to successfully deliver aircraft to the meter-fix on time. The apparent trend in the flight path distance and workload data is that scenario B presented less of a challenge to the controllers, suggesting that the traffic scenarios have a bigger impact on performance and workload than error condition. Analyses are currently underway[13] that seek to identify the specific characteristics of these traffic scenarios that possibly created such impact.

One scenario characteristic worth consideration in this regard may be the direction of the wind error bias. Among the runs included in this analysis, traffic scenario A was always paired with positive bias wind errors, whereas traffic scenarios B and C were always paired with negative bias wind errors. Scenarios A and B were comparably designed, yet their data describe scenario B as less challenging. The negative bias wind errors used during trials with scenario B had the effect of presenting the controllers with seemingly smaller initial delay values. In contrast, the positive bias wind errors used during trials with scenario A impacted the trajectory predictions such that the controllers saw seemingly larger initial delay values. For an aircraft left untouched, the delay would gradually correct towards the actual delay (i.e., the delay as originally designed in the traffic scenario) as it came closer to the meter-fix. This is true in either scenario: that is, the seemingly smaller initial delay value in scenario B would gradually increase as the aircraft approached the meter-fix, and conversely, the seemingly larger initial delay value in scenario A would gradually decrease, both converging towards a similar value. In other words, the interaction between wind error bias and inaccurate delay predictions should produce comparable errors, but in opposite directions. However, these findings challenge the theory that the controller responses to inaccurate delays under different directions of wind error bias would be symmetrical. Further investigation is needed to better understand the effect that direction of error has on the manner in which controllers manage traffic.

4. Impact of Wind Errors (runs 15 and 17, runs 11 and 12)

The addition of the 'Extra-Large' wind forecast error attempted to address the concern that the simulated error conditions did not contain large enough errors. The 'Extra-Large' wind errors were run twice, once in run 15 using scenario A, and once in run 17 using scenario B, with both times simulating large aircraft performance errors. These are compared to runs 11 and 12, which only differ in that they used the 'Large' wind errors.

An analysis of the meter-fix accuracy data between these two wind error conditions shows success rates of 97% with 'Large' winds, 100% with 'Extra-Large' winds. Statistical testing of the data shows no significant differences between winds. Upon inspecting the flight path distance data, the two winds exhibit identical mean values, and when statistically tested, reveal no significant differences. The workload data produce mean controller ratings ranging of 2.4 with the selected 'Large' winds, and 2.5 with the 'Extra-Large' winds. Statistical testing of the workload ratings confirms that lower workload during the 'Large' runs was different than in the 'Extra-Large' runs. That the controllers were able to sustain a high level of performance, suggests that the increased workload caused by the larger wind errors is likely not a meaningful change in workload. The lack of separation violations in the LXL error condition also contribute to the notion that error condition by itself may not be enough to cause performance and workload detriments. An important aspect of the simulation's wind errors is that the same error remained constant throughout a given run, giving the controllers the ability to 'learn' the errors and their effect on the automation, and then adjust their clearances accordingly. This process of learning how to compensate for the trajectory prediction errors is likely disrupted were the errors underlying the trajectory predictions changing more frequently than every 55 minutes.

5. Impact of Toolset Availability (runs 18 – 23, runs 3 and 14)

The exploratory runs examined three additional toolsets, enabling the investigation of whether the trajectory prediction errors, when simulated under different working environments, were more or less likely to impact controller performance and workload. Representing current-day operations was a baseline condition (denoted 'BL'). The second toolset was a one-tool condition focusing on the value of trajectory computations reflecting up-to-date amendments made by the controllers. In this condition (denoted 'S4'), speed clearances issued by the controllers, if entered into the data block's 4th line data fields, triggered a new trajectory computation that incorporated the new speeds, reflected by updated delay (i.e., STA-ETA) information. The meter list and conflicts list were still available in this condition, as they are tools currently in use in the field today, but conflict information in the data block, as well as manual- and automation-assisted trial-planning were not included. The purpose of this condition was to assess the value of simply having controller intent information immediately available to the automation. The third examined toolset included all the tools available during the first part of the simulation, and added the ability for the controllers to issue maneuvers via Data Comm (denoted 'DC'). During this condition, all aircraft were equipped for Data Comm, so that the controllers could electronically send any trial-plan directly to the flight deck. The BL, S4, and DC toolsets were each tested twice, once in the RR error condition, and once in the LL error condition. In order to increase the likelihood of capturing any detriments to performance and workload, scenarios A and C, identified as more difficult, were used in conjunction with these toolsets. Runs simulating the RR error condition were paired with scenario A, and runs simulating the LL error condition were paired with scenario C. Runs 3 and 14 employed the 'full' toolset, and were identified as the most suitable for comparison with these exploratory runs.

Meter-fix accuracy data again shows generally high success rates, but with a noticeable impact seen in the BL condition. Success rates were 94% or better, with the exception of the BL condition, which saw a success rate of 87%. Statistical testing of the STA-ATA data shows significant differences between the BL (*Mdn*= -2.5 sec) and S4 (*Mdn*= 1.0 sec) toolsets, at p=0.24. An analysis of the flight path distance data produces mean values ranging from 209.4 nmi when using the 'full' toolset, to 214.1 nmi when using the BL toolset. When statistically tested, the BL condition exhibits significantly longer distances (*Mdn*= 213.5 nmi) than any other toolset: 'full' (*Mdn*= 208.0 nmi, p<.000), S4 (*Mdn*= 211.0 nmi, p=.002), and DC (*Mdn*= 210.0 nmi, p<.000). A comparison of workload data across these four toolsets shows mean workload ratings of 2.6 and 2.7 in the 'full', S4, and DC conditions, and a mean value of 3.3 in the BL condition. Tests for significance reveal that while all toolsets exhibit identical median values (*Mdn*= 3.0), the workload reported by controllers was significantly higher with the BL toolset than any other toolset, at p<.000.

These results serve as another indication that factors other than the simulated error conditions impacted controller performance and workload. The BL toolset, a representation of current-day operations, shows worse performance, as measured by flight path distance flown, and attributes to higher workload. The absence of any significant differences between the 'full', S4, and DC toolsets imply that similar improvements over the BL condition can be achieved, whether adding one or several tools. This may confirm the hypothesis that simply adding a mechanism to make the automation aware of speed intent, for use in trajectory computations, can lead to improved operations.

C. Additional Results

1. Overall Acceptability

During the post-run questionnaires, the participants were asked several questions about the acceptability of the simulated operations. Although the participants filled-out questionnaires after every run, a decision made prior to

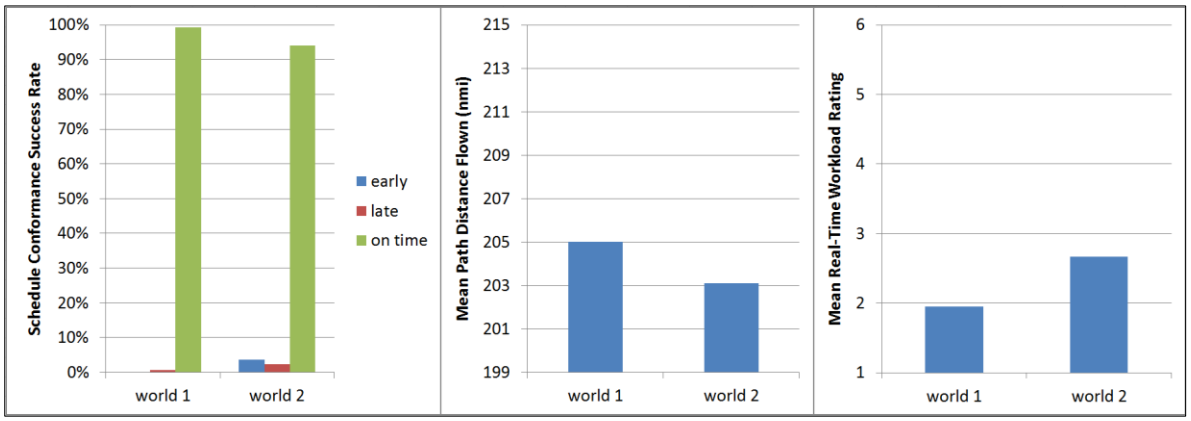

Figure 17. Schedule conformance, path distance, and workload data, according to world, as measured during the primary runs.

the simulation left certain questions asked only every other run, in keeping with the original run schedule's two-run blocks per condition. Acceptability questions were only asked every other run, which unfortunately make rigorous comparisons of the data unfeasible. From the limited set of acceptability data available, noteworthy comments from the controllers indicate that operations were safe and manageable for all runs, with one exception: one of the four controllers rated run 14 (the LL error condition in which scenario C was first used), as unsafe and unmanageable.

Subjective feedback regarding the available tools shows that controllers most used, most preferred, and rated as most stable the meter list, the delay information displayed in the data block, and the speed-related trial-planning functions. While not explicitly mentioned, this feedback data is closely related to the behavior of the delay information displayed in the data block: its ability to update immediately in response to speed amendments.

2. Resulting Operational Concepts

An important result of the simulation was apparent during observations throughout the simulation, and confirmed in the data. Between the two parallel worlds, two different operational concepts surfaced, representing different approaches taken by the high-altitude controllers in using the tools to deliver aircraft to the low-altitude sector. When absorbing delays, both worlds used an initial speed reduction, then added supplemental clearances if necessary, later fine-tuning the situation with further speed changes if necessary. The supplemental clearances issued by World 1 employed the trial-planning tools to create path adjustments, which the controller then issued to an aircraft (e.g., "UAL123, for sequencing, proceed direct BALLE, direct RMG, and the remainder of the RPTOR1 arrival."). World 2's supplemental clearances on the other hand, often were heading vectors. The high-altitude controller in World 2 would first use the trial-planning tools to craft a new route that absorbed the desired amount of delay, similar to the approach taken in World 1. Unlike World 1 however, the high-altitude controller in World 2 would issue heading vectors meant to overlay the trial-planned route (e.g., "UAL123, for sequencing, fly heading 180, expect direct RMG in two minutes."). Following is an analysis of data from the primary runs, offering a closer look at the impact of these two styles of interacting with the automation's trajectory predictions. Schedule conformance success rate, flight path distances flown, and real-time workload ratings described in this section are shown in Fig. 17, while details of the statistical analyses can be seen in Table 3.

The data shows high schedule conformance success rates for both worlds (99% in world 1, 94% in world 2). Tests of the raw STA-ATA data reveal statistical significance between world 1 (*Mdn*= -3.0 sec) and world 2 (*Mdn*= 1.0 sec). An inspection of the flight path distance data shows an average of 205.0 nmi for world 1, and 203.1 nmi for world 2. When tested, the differences between the longer distances flown in world 1 (*Mdn*= 208.4 nmi) than in world 2 (*Mdn*= 206.5 nmi), showed only near-significance. The mean workload rating reported in world 1 was 1.9, and 2.7 in world 2, and tests of the workload data confirm significance between the lower values reported in world 1 (*Mdn*= 2.0) than in world 2 (*Mdn*= 3.0). These findings not only validate the observations made during the simulation that the two worlds worked

Table 3. Man-Whitney U Tests for world-specific analyses.

World (1,2)	*n*	*U*	*z*	*p*
Meter-fix Schedule Conformance	598	49,579	2.316	*.021
Flight Path Distance	396	17,472	-1.870	.061
Workload	906	142,832	10.695	*.000
*Note: * significant at alpha=.05 two-tailed.*				

differently, but also suggest that different strategies associated with different amounts of work can achieve the same goal. More specifically, the workload and path distance data point to the differences between the two approaches, while the schedule conformance data, although different, are well within the goal of crossing the meter-fix within 25 seconds of the STA.

V. Discussion

The initial results presented in this paper suggest that even in the largest trajectory prediction error conditions, controllers were able to learn how to compensate for the errors and adapt their interaction with the tools to deliver arrival aircraft on time and not exceed workload limits. In fact, the error conditions exhibited minimal impact on performance; rather, it was other factors, such as traffic scenario and tool availability, that had measurable impact on system performance. Even still, performance remained fairly high, which raises the question: "What is the cost of the controllers performing at this level under these circumstances?" Five separation events occurred during the simulation, all in cases simulating the LL error condition. This finding perhaps addresses the 'at what cost?' question, by lending to the possible theory that higher levels of error required the controllers to give increased attention to processing the error-prone data provided by the tools, and then filtering / translating that data into the clearances they ultimately issued. It is conceivable that during this process they became somewhat distracted, and missed and/or misjudged the threat posed by a neighboring aircraft.

The idea that the controllers were overly-focused on the arrival task provides an opportunity to more closely examine the details of the operational environment simulated in this study. The traffic scenarios used in the study indicate the main focus was the task of managing arrival aircraft. The scenarios also assumed that the envisioned time frame serving as the backdrop for the simulation was one of current-day traffic levels. The external validity of the results of this study may therefore, be limited. It is not known whether the same results would occur in a similar arrival-management problem simulating higher levels of traffic, or in an airspace with different traffic flow patterns; such as high-altitude en route sectors dealing with higher complexity. Additionally, the scheduling criteria studied in this simulation only showed significance in terms of workload, but it is possible that as simulated, even the 'tightened' scheduling criteria was not tight enough to fully uncover the issues associated with the interaction between inaccurate delivery times and separation violations. A tighter schedule would leave the controllers with even less room for error when delivering aircraft to the meter-fix, requiring them to be much more precise, while simultaneously coping with the trajectory prediction errors.

Another result of the study is the ability of the tools to support different operational concepts. The different operational concepts that emerged during the simulation provide interesting perspectives on the definition of TBO. World 1's approach could be described as trajectory-based, characterized by complete trajectory changes followed by small adjustments. World 2's approach could be described as simply trajectory-informed, characterized by open-ended trajectory changes followed by small adjustments. The controllers' strategy of limiting their exposure to the trajectory prediction errors by taking the advisories merely into consideration- and then formulating their own clearance, is in contrast to the simplistic approach of taking the tools' suggestions at face value, and then issuing corrective clearances as errors become apparent. It is possible that this strategy however, could unintentionally negate certain benefits of the tools, such as the conflict detection performance expected in TBO. The tradeoff between a more conservative use of automation and more overlap between the controller's intentions and the automation's assumptions needs further study.

Were this investigation of trajectory uncertainties applied to far-term ATM environments relying more heavily on automation, the sensitivities of system performance or controller workload to error-prone trajectory predictions might be completely different. A far-term environment in which trajectory changes are sent automatically by the ground system's automation via Data Comm to the flight decks will not have the benefit of a controller recognizing the errors and adjusting the clearances appropriately. An automated system likely would not know that it's wrong, and without an additional error-checking mechanism, would send trajectory clearances without hesitation. Shortly thereafter, as the predictions change, the automation would simply send additional trajectory clearances, effectively correcting itself. Today's and the foreseeable near future's paradigm of having the controller in the loop provides a means for, at a minimum, protection against excessive amounts of corrective clearances. The characteristics of ingenuity, creativity, and flexibility associated with human operators, as well as their ability to learn and adapt on-the-fly, are major contributors to system robustness. However, these same elements also limit a system's scalability. Careful understanding of human-automation cooperation is needed to inform appropriate automation design for both near- and far-term ATM environments.

VI. Conclusion

The fact that even the largest errors did not overwhelm the controllers is a significant finding. Results suggest that the traffic scenarios have a bigger impact on performance and workload than the prediction errors, under the operational environments tested. It is believed that the size of the errors managed by the controllers may prove difficult for a fully-automated system built to issue corrective updates automatically. This suggests the opportunity to enhance research on more far-term concepts and to incorporate corrective learning; much like the controllers did naturally in the simulation. In the presence of these errors, it is also likely that the transition from a controller-in-the-loop paradigm to a fully-automated system may be problematic without significant investigation and improvement to the human-system concept.

Future research on trajectory prediction uncertainty and its impact on controller performance should look to develop methods for the automation to be more aware of what the controller has learned about the errors, and how that information can be used to better support the controller in their task. Other operational environments need investigating, in addition to other means of incorporating errors during a simulation. Hopefully more data can be collected to confirm the belief that automation does not have to be perfect, just predictable.

Acknowledgments

This work was conducted as part of the Functional Allocation and Separation Assurance research focus area, funded by the Concept and Technology Development Project within NASA's Airspace System Program. The authors acknowledge the contributions of many individuals in the Airspace Operations Laboratory and the Human-Systems Integration Division's system support group.

References

[1]Federal Aviation Administration. "FAA Aerospace Forecast, Fiscal Years 2013-2033." FAA, Washington, D.C., 2013.

[2]Federal Aviation Administration. "NextGen Implementation Plan 2013." FAA, Washington, D.C., 2013.

[3]Joint Planning and Development Office. "Concept of Operations for the Next Generation Air Transportation System." Version 3.2, JPDO, Washington, D.C., 2010.

[4]Swiesrstra, S. and Green, S. M. "Common Trajectory Prediction Capability for Decision Support Tools." *Proceedings of the USA/Europe Air Traffic Management Research and Development Seminar*, Budapest, Hungary, 2003.

[5]Mondoloni, S., and Bayraktutar, I. "Impact of Factors, Conditions, and Metrics on Trajectory Prediction Accuracy." *Proceedings of the USA/Europe Air Traffic Management Research and Development Seminar*, Baltimore, MD., 2005.

[6]Mondoloni, S., Swierstra, S., and Paglione, M. "Assessing Trajectory Prediction Performance – Metrics Definition." *Proceedings of the Digital Avionics Systems Conference*, Washington, D.C., 2005.

[7]Prevôt, T., Smith, N., Palmer, E., Callantine, T., Lee, P., Mercer, J., et al. "Integrating Concepts and Technologies in the Airspace Operations Laboratory." *AIAA Modeling and Simulation Technologies Conference*, AIAA, Reston, VA, (submitted for publication) 2013.

[8]Herzberger, H. "Transforming the NAS: The Next Generation Air Traffic Control System." *Proceedings of the International Congress of the Aeronautical Sciences*, Yokohama, Japan, 2005.

[9]Cole, R.E., Green, S. M., Jardin, M., Schwartz, B. E., and Benjamin, S. G. "Wind Prediction Accuracy for Air Traffic Management Decision Support Tools." *Proceedings of the USA/Europe Air Traffic Management Research and Development Seminar*, Naples, Italy, 2000.

[10]Stein, E. "Air Traffic Controller Workload: An Examination of Workload Probe." DOT/FAA/CT-TN84/24, FAA, Washington, D.C., 1985.

[11]Morey, S., Prevôt, T., Mercer, J., Cabrall, C., Martin, L., Bienert, N., et al. "Controller Strategies for Automation Tool Use under Varying Levels of Trajectory Prediction Uncertainty." *AIAA Aviation Conference*, AIAA, Reston, VA, (submitted for publication) 2013.

[12]Kraut, J., Mercer, J., Morey, S., Homola, J., Gomez, A., and Prevôt, T. "How do Air Traffic Controllers Use Automation and Tools Differently during High Demand Situations?" *AIAA Aviation Conference*, AIAA, Reston, VA, (submitted for publication) 2013.

[13]Bienert, N., Prevôt, T., Cabrall, C., Gomez, A., Hunt, S., Martin, L., et al. "Case Study: Analyzing Influences on Traffic Scenario Difficulties for Human-in-the-Loop Simulations." *Proceedings of the Digital Avionics Systems Conference*, Syracuse, NY, (submitted for publication) 2013.