A Remote, Human-in-the-Loop Evaluation of a Multiple-Drone Delivery Operation

Garrett G. Sadler*, Meghan Chandarana†, R. Conrad Rorie*, Terence L. Tyson*,
Jillian N. Keeler‡, Casey L. Smith*, and Megan C. Shyr‡
NASA Ames Research Center, Moffett Field, CA 94035, USA

Dominic G. Wong§
ASRC Federal Data Solutions, LLC, Moffett Field, CA 94035, USA

Scott Scheff¶‖
HF Designworks, Inc., Boulder, CO 80308, USA

Igor Dolgov‖
Joby Aviation, Santa Cruz, CA 95060, USA

Over time, advances in unmanned aircraft systems (UAS) have enabled a shift in the operational paradigm from one operator managing one aircraft to that of multiple operators working together to manage multiple aircraft. This shift has highlighted the need for effective human-autonomy teaming methods to maintain manageable workload levels for operators as well as high standards of system performance and safety. This paper presents a study aimed at evaluating whether automation can help operators manage workload during small UAS (sUAS) package delivery scenarios featuring contingency situations. These contingency situations, resulting from unplanned UAS Volume Reservations (UVRs), required flight path reroutes for multiple aircraft simultaneously. The study manipulated the number of aircraft affected by the UVRs and the level of automation support. The presence of terrain conflicts was also controlled within each scenario. Due to the COVID-19 pandemic, subjects were not able to gain direct access to the Ground Control System (GCS). Therefore, the study was conducted using a subject-surrogate paradigm that required subjects to relay commands through a verbal protocol from remote locations outside of the lab to a researcher surrogate who had direct control of the GCS interfaces at the lab location. Results show that the automated support condition was associated with faster reroute response times, more efficient reroute maneuvers, and significantly lower levels of perceived workload than the manual reroute condition. However, the automation support level did not significantly impact pilots' ability to avoid the UVR successfully; pilots were overwhelmingly capable of avoiding the UVR in all conditions. The presence of terrain conflicts primarily impacted pilot performance by leading to multiple uploads per vehicle, which was not typically required when pilots only needed to maneuver laterally. Although subjects did not have direct control over the GCS, subjective ratings indicate that the displays under test provided them with sufficient information to manage their aircraft and promptly respond to the unplanned UVRs. Overall, the objective and subjective data strongly suggest that the verbal protocol and subject-surrogate paradigm were effective methods for collecting data remotely amid the COVID-19 pandemic.

I. Introduction

For over a century, unmanned aircraft systems (UAS) have increased in application, sophistication, and popularity [1]. Due to the remote nature of unmanned systems, innovations in automation are a key factor in determining the scope

*Research AST Human/Machine Systems, Human Systems Integration Division
†Computer Engineer, Intelligent Systems Division
‡Research Student Trainee, Human Systems Integration Division
§Systems Development Scientist
¶Founder, CEO, and Principal Human Factors Engineer
‖Human Factors Engineer, Air-Taxi Products

and pace of advancements. Human-Autonomy Teaming (HAT) research seeks to study the interaction between human operators and highly automated systems, with the goal of improving operator workload and overall system safety. One recent path of HAT research is termed "*m*:*N*," where multiple operators – *m* – cooperatively manage multiple vehicles – *N*, where *N* is assumed to be greater than *m* [2]. One domain well-suited to *m*:*N* is that of small UAS (sUAS), which relies on high levels of automation and is expected to benefit from a UAS Traffic Management (UTM) system that proposes the use of a federated control architecture to drastically increase the scale of operations [3].

The transition from a 2:1 or 1:1 control paradigm to an *m*:*N* control paradigm could enable significantly more sUAS operations without a commensurate increase in the number of pilots required to control them. Before *m*:*N* can be realized, however, researchers need to better understand the new roles and responsibilities required of operators under this novel operator-to-vehicle configuration. Similarly, the potential limitations or hazards associated with *m*:*N* must also be well understood before moving beyond single vehicle control.

A substantial body of literature already exists around supervisory control, and, more specifically, multi-UAS control [4–11]. A key element of supervisory control is identifying the appropriate level of automation for a given context, which can have direct effects on the human operator's workload, situation awareness, complacency, and skill degradation [5]. Fern and Shively (2009) examined supervisory control in the context of multi-UAS control, tasking participants with managing three UAS at a time while varying the level of automation support between conditions. The authors found faster response times, higher accuracy on questions assessing situation awareness, and lower workload for the automation condition that allowed the operator to simultaneously send out a command (referred to as a "play") to multiple vehicles at once. Conversely, a second automated condition only allowed the operator to affect one vehicle in their fleet at a time, which was found to reduce the effectiveness of the tool relative to the "play" condition [8]. Monk et al. (2019) also examined pilot performance during multi-UAS control and suggested that automation on secondary, lower-level tasks would have been more beneficial than automation that impacted the primary task of maintaining safe separation [7].

The present study is an extension of a cognitive walkthrough detailed in Smith et al. (2021). In that study, Smith and colleagues recruited active manned and unmanned aircraft pilots to review a concept of operations that consisted of supervisory control of multiple sUAS in a food delivery setting in San Diego, CA. The review also included an overview of the prototype displays and automation support tools that could be made available to the operators of the sUAS during nominal and off-nominal situations. Once the summary concluded, participants were shown four different use cases, which gradually increased the number and type of contingency events present in the scenario. The purpose of the exercise was to gather feedback from the pilots regarding the viability of the scenarios and the utility of the various display elements. Pilots rated contingencies regarding sudden airspace restrictions and battery failures as most likely to occur. The participants also noted that the displays would benefit from a greater emphasis on airspace-related information and that the controls should enable more flexibility for directly controlling the aircraft [2].

The study presented in this paper builds upon the work of Smith et al. (2021) to translate its sUAS concept of operations and the associated tools and displays into a high-fidelity human-in-the-loop simulation. The aim of the simulation is to investigate the utility of automation in the context of resolving conflicts with a sudden airspace restriction that affects multiple vehicles in the operator's fleet. This study will continue the trend of examining various levels of automation in the context of supervisory control and its impact on pilot performance. Additional data will include the impact of number of vehicles affected by a contingency on the effectiveness of the automation as well as participant feedback regarding the overall utility of the display and tools presented over the course of the study.

II. Ground Control Station

The study utilized the United States Air Force Research Laboratory's Vigilant Spirit Control Station (VSCS) to provide a Ground Control Station (GCS) graphical user interface (GUI) for the management and monitoring of multiple vehicles. The baseline architecture allows for multiple operators to simultaneously manage and control their own set of aircraft. The core GCS GUI elements include a Tactical Situation Display (TSD), Timeline, and Status & Log (Fig. 1). While the different display elements are typically presented across multiple monitors, they were all presented within a single screen for the present study to provide the participants with the best resolution possible as they viewed the GCS via a Microsoft Teams video meeting.

The TSD acted as the primary display, which showed a map of the operating environment. This included corridors (white outlined rectangles), hubs (white outlined circles), and hives (green circles) (Fig. 2). For the simulation, corridors were bidirectional and included "lanes." Simulated aircraft followed right-of-way rules and always traveled in the lane on the right side of the corridor from the perspective of the aircraft's direction of travel. Hubs acted as pick-up and

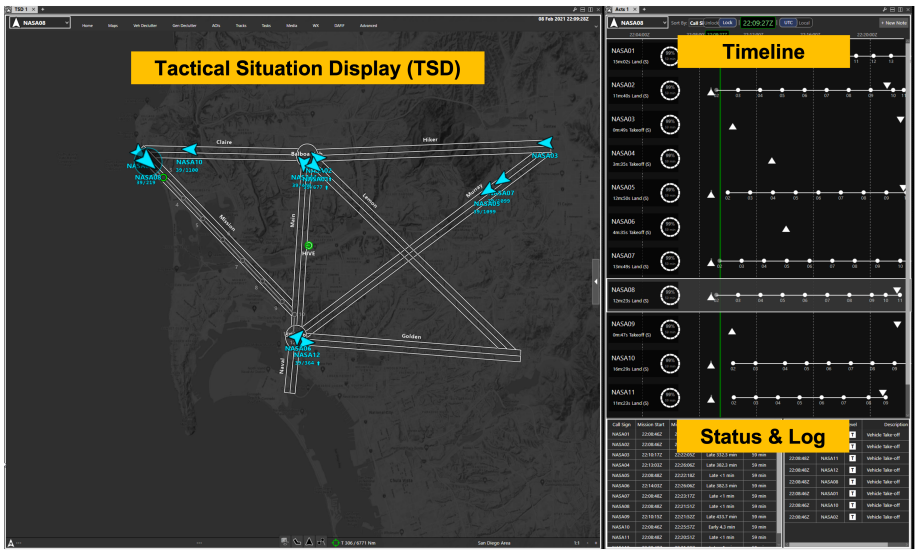


Fig. 1 The Ground Control Station interface includes a Tactical Situation Display (left), the event Timeline for each vehicle (top right), and the Status & Log (bottom right).

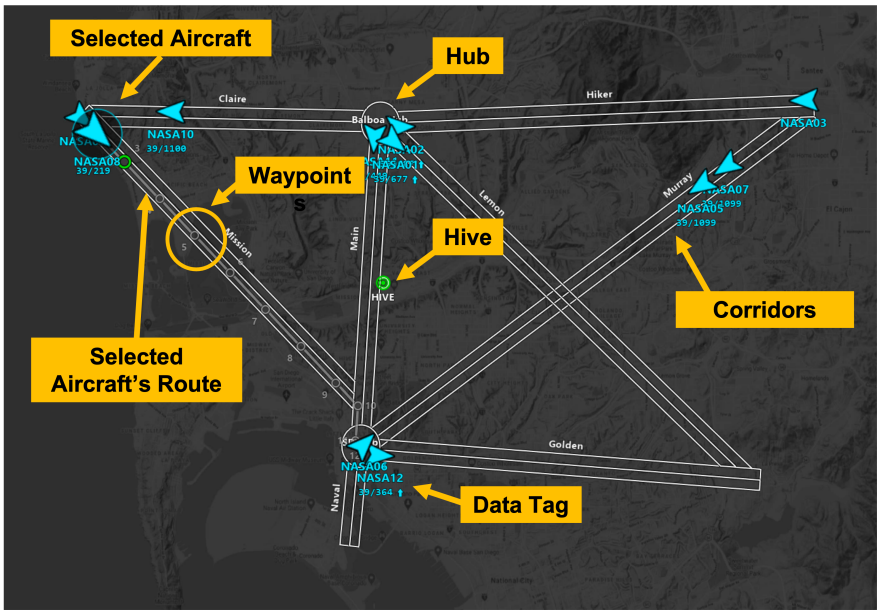


Fig. 2 Tactical Situation Display (TSD) key elements

drop-off locations that connected corridors, while hives were the originating and terminating location for all aircraft operating in the region. The separation line between lanes was shown as a white line in the middle of each corridor. Aircraft being controlled and monitored were displayed as cyan chevrons. Under each aircraft's chevron was a data tag showing the aircraft's callsign, current airspeed, and altitude in feet mean sea level (MSL). The direction of the chevron corresponded to the direction of aircraft travel. When an aircraft was climbing or descending, an up or down arrow was displayed on the bottom right of the aircraft data tag respectively. Selected aircraft were shown with a halo around the chevron icon and were highlighted on the Timeline display. When an aircraft was selected, its flight path was shown as a gray line with its next waypoint shown as a hollow green circle and all ensuing waypoints shown as hollow grey circles. Flight paths allowed for modification by adjusting the position and altitude of the individual path waypoints. When a waypoint was selected for editing, a white crosshair was shown over the selected waypoint with a data tag indicating the waypoint's altitude in feet MSL and above ground level (AGL).

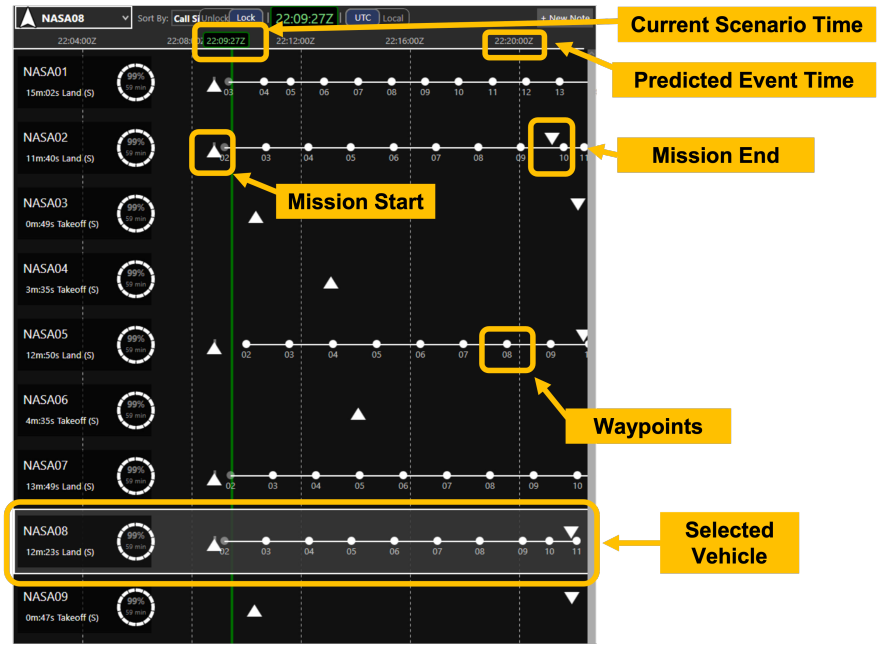
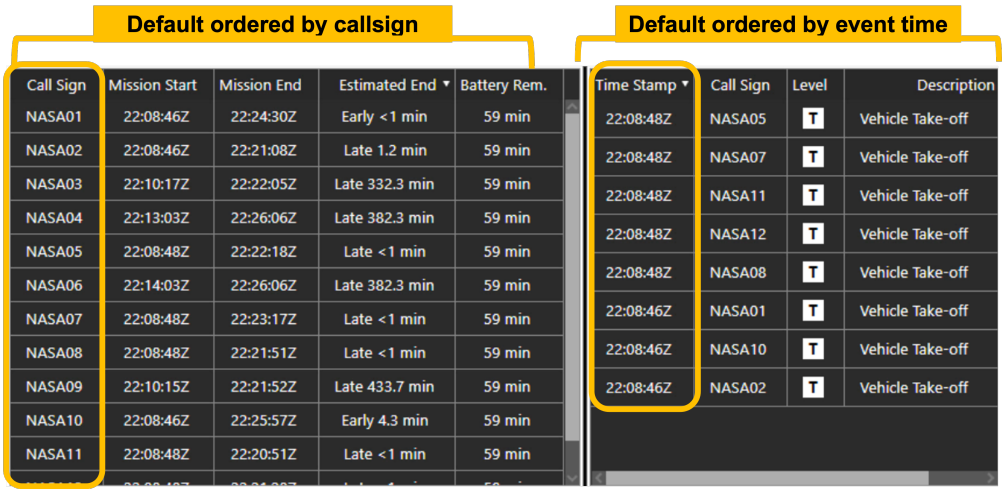
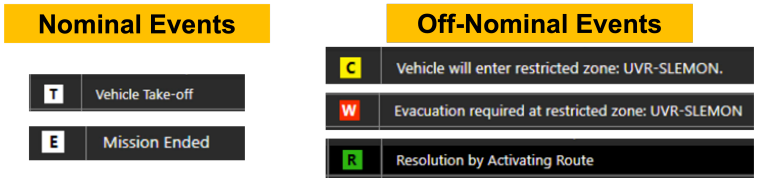


Fig. 3 Screen capture of the Timeline display. All major features are labeled.



(a) Mission status (left) and event log (right)



(b) Nominal and off-nominal logged events

Fig. 4 Status & Log portion of the display.

The right side of the display included the Timeline (Fig. 3) and Status & Log (Fig. 4). The Timeline portion of the GUI provided a way to visualize all major events in the mission for each vehicle in the airspace. Current scenario time

and predicted event time were shown at the top. Vehicle data tags on the left side of the Timeline GUI included the vehicle callsign, battery life and a countdown to the next significant event. Events for each vehicle were shown as dots along the mission timeline. For the study conducted, all events corresponded to waypoints. The current position in the timeline was shown with a vertical green line that extended down the Timeline display for all active vehicles. Start and end times were shown with white triangles that pointed up and down, respectively. Selected vehicles were highlighted with a white box and grey transparent background around the vehicle data tag and its timeline of events. The Status & Log included the current status of all vehicles in the airspace and the Log kept a list of all nominal and off-nominal events for all vehicles.

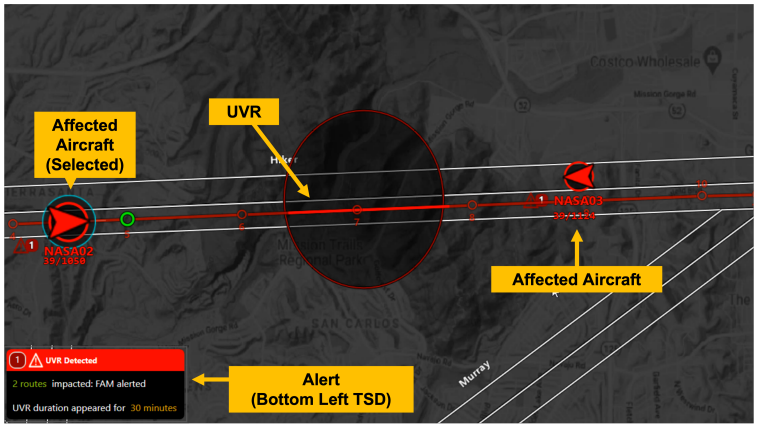


Fig. 5 The visuals shown on the TSD when a UVR is detected.

A. UAS Volume Reservations (UVRs)

Active UVRs were displayed on the TSD in multiple ways (Fig. 5). The first was a dark grey circle outlined in red that appeared on the map portion of the TSD in the location of the UVR. The region encompassed within the circle represented the portion of the airspace (at all altitudes) that was temporarily restricted, requiring all vehicles to be rerouted around it. In some instances, this required that vehicles were rerouted outside the affected corridor. Vehicles impacted by the UVR and associated information (e.g., flight paths) were displayed in red on the TSD, Timeline, and Status displays. The second indication of a UVR was a pop-up message that appeared at the bottom-left of the TSD when the UVR was issued. The pop-up message included how many vehicles were affected and the length of time the UVR was in effect.

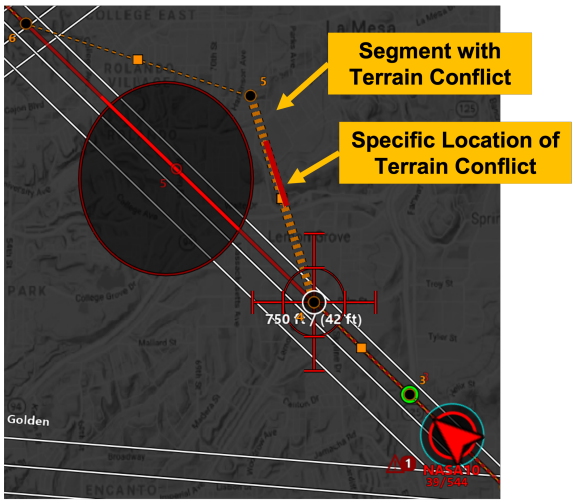


Fig. 6 Flight path segment as displayed when a terrain conflict exists.

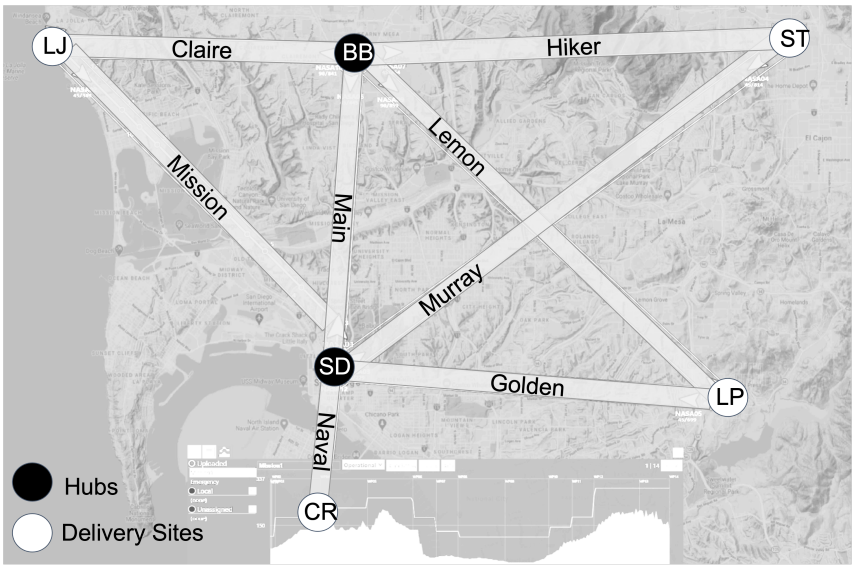


Fig. 7 Airspace and corridors used for the study. Black circles indicate hubs and white circles indicate delivery sites. Corridors are depicted with grey rectangles.

B. Terrain Conflicts

If a portion of a flight path had a terrain conflict, that segment of the flight path was displayed with a thicker, yellow, dashed line (Fig. 6). The specific location of the terrain conflict within that segment was then highlighted in red and flashed until the terrain conflict was resolved. When a waypoint with an associated terrain conflict was selected, the crosshairs that normally appear white were colored red, to emphasize the presence of a terrain conflict. When a waypoint was moved such that a terrain conflict no longer existed, the crosshair returned to the original white color.

III. Experimental Design

The study recruited 12 active pilots for participation, with an average age of 35 years old (SD = 6 years). All pilots had both manned (average of 252 flight hours, SD = 348 flight hours) and unmanned (average of 1100 military combat hours, SD = 488 flight hours) piloting experience. Half of the 12 pilots were IFR rated and five held a Part 107 rating. During the study, subjects played the role of a GCS operator, or Tactical Operator (TO), who supervised highly automated sUAS food delivery missions in the San Diego, CA area. The TOs managed 12 vehicles under the supervision of a Fleet Area Manager (FAM), played by a member of the research team. As part of the operations, two unplanned UVRs resulting from environmental factors (e.g., airspace restrictions from local fire department) occurred, forcing the TO to reroute vehicles around the affected areas. The airspace was modeled using a hub-and-spoke framework, consisting of 8 corridors, 2 hubs, and 4 delivery sites (Fig. 7). The study utilized a 2 X 2 within-subjects design that manipulated the number of vehicles affected by each UVR (2 vs. 4 vehicles) and the automation support level provided, by requiring pilots to either perform reroutes manually or with the assistance of automation. During data collection with the first two subjects, inconsistencies with the trial scenarios were found and, therefore, only data from the last ten subjects were used for the analyses reported in this paper.

In addition to the two manipulations described above, a terrain conflict manipulation was also included and required pilots to make altitude adjustments during the reroute process for a quarter of the vehicles affected by the UVRs. The assisted automation condition provided subjects with lateral reroute suggestions that would avoid the UVR, whereas subjects had to resolve the conflicts on their own in the manual condition. However, in the assisted automation condition, reroute suggestions did not include the altitude corrections necessary to avoid the terrain conflicts, thereby requiring subjects to amend the waypoint altitudes manually in all conditions. The inclusion of the terrain conflict manipulation was designed to increase participant workload in general, and to force participants in the automated reroute condition to actively review suggested reroutes and their proximity to nearby terrain.

Each of the four trials lasted approximately 15 minutes. Trial order was blocked by automation condition and counterbalanced across subjects. The number of vehicles affected by a UVR (2 or 4) within a trial was also

counterbalanced within blocks across subjects. Participants saw two UVRs per trial, which always occurred in two different corridors. The second UVR in a trial only appeared once the first UVR had been completely resolved. Upon the onset of a UVR, subjects had approximately two minutes to reroute the first vehicle before they would violate the UVR. The spacing of the remaining vehicles was staggered by approximately 30-60 seconds, giving subjects slightly more time with the later vehicles to avoid a UVR violation.

All four trials also included five requests from the FAM that asked the subjects to either provide a verbal report on the status of a vehicle and/or UVR, or to change the speed of one of the aircraft in their fleet. The first trial block had two FAM calls before the first UVR, two FAM calls after the first UVR, and one FAM call after the second UVR. The second trial block had one FAM call before the first UVR, three FAM calls after the first UVR, and one FAM call after the second UVR.

At the start of the study, subjects completed an informed consent and a background questionnaire. Afterwards, they were given general training covering the goals of the study, a description of the airspace that would be used in the trials, and an overview of the GCS. The basic controls available to the subject when using the GCS were then explained, including the verbal protocol that was utilized to enable remote participation (explained in more detail in Section IV). Subjects were then able to practice various display interactions indirectly via the verbal protocol to become familiar with both the interfaces and subject-surrogate interaction paradigm. The procedure for modifying flight paths was then explained and they were given the opportunity to practice the verbal protocol with an example scenario. Due to the blocking of trials by automation level, only the flight path modification procedure for the first block was explained initially. The first phase of training ended with an overview of the FAM calls that would occur throughout the trials and a practice session to test the verbal protocol procedures needed to address them. After the trials for the first block were completed, the verbal protocol procedure for the flight path modification during the second block was explained. Practice time was also given for the second block's flight path modification procedure before subjects completed the remaining trials.

After each trial, subjects were asked to complete post-trial subjective questionnaires that included a questions related to their performance with the GCS, and a NASA-TLX form that allowed subjects to rate their workload during the trial. After all trials were completed, a post-simulation questionnaire was given and included general questions about training, the GCS interface, the subject-surrogate interaction paradigm, and the difficulty of the tasks. In addition, subjects participated in an informal verbal post-simulation discussion where they were asked questions about the roles and responsibilities of people who would be involved with similar scenarios in the future (e.g., FAM and TOs), the airspace, and the effectiveness of the automation provided. The objective metrics that were collected during each trial are described in detail in Section V.

IV. Subject-Surrogate Interaction

To adhere to local and federal restrictions resulting from the COVID-19 pandemic, this study utilized a subject-surrogate paradigm for data collection. As part of this interaction between a subject and a surrogate, the subject relayed commands verbally to a surrogate played by a member of the research team that then carried out the actions requested on the GCS. As a result, subjects did not have direct control over the GCS. However, they were able to view a video feed (through Microsoft Teams) of the GCS in real-time as events occurred and as the surrogate carried out the requested actions during the entire length of the trial.

As a way to accommodate this new data collection paradigm, a verbal protocol was developed to provide structure between the interactions of the subject and surrogate. This standardized the set of interactions across subjects and helped to simplify the interactions in order to reduce delays and frustration that would have arisen from a more free-form and unstructured communication style between the subject and surrogate. Although not ideal, the verbal protocol enabled real-time data collection with a subject in the loop while also complying with COVID-19 restrictions.

A. Verbal Protocol

All subjects were trained on the same verbal protocol that was used to communicate commands for specific GCS interactions. This approach paired commanded actions with simulation objects so that the surrogate knew where to focus their interaction. The surrogate performed a readback of all commands.

Each verbal command had an associated action and targetable object(s) for the TSD (Table 1), Timeline panel, and Status panel (Table 2). Display locations that were outside the TSD would have to be referenced in the verbal command (e.g., "Timeline, . . . "). For specific vehicles, changes would be made for their speed, altitude, and lateral movement (Table 3). Subjects practiced all possible verbal commands listed in Tables 1-3 before data collection.

Action	Targetable Objects	Command	Example
Zoom	Asset, UVR, Corridors, Waypoints	"[Object] zoom [in/out AND amount in clicks]."	"NASA03, zoom in 3 clicks."
Center	Asset, UVR, Corridors, Waypoints	"[Object], center."	"NASA03, center."
Click	Asset, Waypoint	"[Callsign of asset]." "[Waypoint Name]."	"NASA03." "Waypoint 5."
Pan	Displays	"Pan [right/left/up/down AND amount]."	"Pan left 3 clicks."

Table 1 Verbal protocol commands for interaction with TSD.

Action	Targetable Objects	Command	Example
Zoom	Timeline	"[Object] zoom [in/out AND amount in clicks]."	"Timeline, NASA03, zoom in 3 clicks."
Center	Current Time	"[Object], center."	"Timeline, NASA03, center."
Click	Asset, Status/log Columns	"[Callsign of asset]." "[Status/log][Column Name]."	"Timeline, NASA03." "Status, Estimated End."
Move	Sliders	"Scroll [right/left/up/down AND amount in clicks."	"Timeline, scroll left 3 clicks."
Sort	Status/log Columns	"[Status/log], sort [ascending/descending AND Column Name]."	"Status, sort descending Estimated End."

Table 2 Verbal protocol commands for interactions with the Timeline and Status & Log.

Action	Targetable Objects	Command	Example
Speed Change	Asset	"[Callsign][increase/decrease] speed by [amount in knots]."	"NASA03, increase speed by 10 kts."
Lateral Movement	Waypoint	"[Callsign], [Waypoint Name], [right/left/up/down AND amount in grid cells]."	"NASA03, waypoint 7, right 3, down 1."
Altitude Change	Waypoint	"[Waypoint Name], [climb/descent AND amount in clicks]."	"Waypoint 5, climb 2 clicks." (1 click = 100 ft)

Table 3 Verbal protocol commands for changing speed and waypoint location for a specific vehicle.

V. Metrics & Analyses

The following metrics were all collected to assess the impact of the automation support level, number of affected vehicles, and terrain conflict variables on pilot performance.

A. Response Time Metrics

- *Initial Response Time (seconds)* – time elapsed from the onset of the first vehicle alert triggered by the UVR to the end of the first verbal command issued by the subject to the surrogate to initiate a vehicle reroute; initially calculated per vehicle and then averaged across the trial
- *First Upload Time (seconds)* – time elapsed from the onset of the first vehicle alert triggered by the UVR to the

first upload (executed by the surrogate) for a corresponding route modification; initially calculated per vehicle and then averaged across the trial
- *Service Time (seconds)* – time elapsed from the start of the surrogate's first interaction with a vehicle affected by a UVR to the final upload made that resolved the UVR and any terrain conflicts; initially calculated per vehicle and then averaged across the trial
- *UVR Resolution Time (seconds)* – time elapsed from the onset of the first vehicle alert triggered by the UVR to the time when all UVR-affected routes have been cleared via route modifications (does not include any additional time spent resolving terrain conflicts); initially calculated per vehicle and then averaged across the trial

B. Reroute Performance Metrics
- *Path Length Difference (nautical miles)* – the absolute difference between the length of the original route and the reroute; initially calculated per vehicle and then averaged across the trial
- *UVR Violations (count)* – instances of a vehicle breaching the UVR boundary
- *UVR Violation Duration (seconds)* – time spent inside the boundaries of an active UVR
- *Flight into Terrain (count)* – instances where a vehicle collided with terrain while rerouting around a UVR
- *Single vs. Multiple Uploads (proportion)* – proportion of vehicles that required one upload or multiple uploads to resolve the conflict with an active UVR and nearby terrain (when applicable); calculated per trial

C. Subjective Metrics
- *NASA-TLX Ratings* – participants' self-ratings of their workload on six different scales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration Level; completed following each trial
- *Display Feedback* – open-ended feedback from participants regarding the utility of the displays they encountered over the course of the experiment; completed following the conclusion of the final trial of the day

D. Analyses

Individual 2 (automation support level: automated vs. manual) x 2 (number of affected vehicles (NAV) condition: 2 vehicles vs. 4 vehicles) repeated-measures ANOVAs investigated if the manipulations used impacted objective performance measurements, including Initial Response Time, First Upload Time, Service Time, UVR Resolution Time, and Path Length Distance, as well as subjective workload captured by NASA-TLX. If a significant interaction was found for any of the above metrics, post-hoc one-tailed paired t-tests were run to examine simple effects of automation support. The ANOVAs assumed sphericity and was verified using Mauchly's Sphericity Test; results are presented with a significance level of $\alpha = .05$.

VI. Results

A. Response Time Metrics

There was no significant main effect of automation support level on Initial Response Times ($F(1, 9) = 1.96, p > .05, \eta_p^2 = .18$, Fig. 8a). Additionally, there was no significant main effect of NAV condition on Initial Response Times ($F(1, 9) = .04, p > .05, \eta_p^2 = .004$). No significant interaction was found between automation support level and NAV condition on Initial Response Times ($F(1, 9) = .46, p > .05, \eta_p^2 = .05$).

There was a significant main effect of automation support level on First Upload Times ($F(1, 9) = 67.94, p < .001, \eta_p^2 = .88$) with First Upload Times taking an average of 20 seconds longer in the manual condition compared to the automated condition (Fig. 8b). No significant main effect of NAV condition ($F(1, 9) = .46, p > .05, \eta_p^2 = .05$) or significant interaction between automation support level and NAV condition for First Upload Times was found ($F(1, 9) = 1.89, p > .05, \eta_p^2 = .17$).

There were significant main effects of both automation support level ($F(1, 9) = 32.89, p < .001, \eta_p^2 = .79$) and NAV condition ($F(1, 9) = 41.33, p < 001, \eta_p^2 = .82$) on Service Times (Fig. 8c), where the absence of automation and the 4 vehicle condition were both associated with significantly slower Service Times. There was also an interaction between automation support level and NAV condition on Service Times ($F(1, 9) = 13.78, p = .005, \eta_p^2 = .61$). The automated condition was found to significantly reduce Service Times in the 4 vehicle condition ($t(9) = 6.69, p < .001$, Cohen's $d = 2.11$), but was not found to have any effect in the 2 vehicle condition ($t(9) = .71, p > .05$, Cohen's $d = .22$).

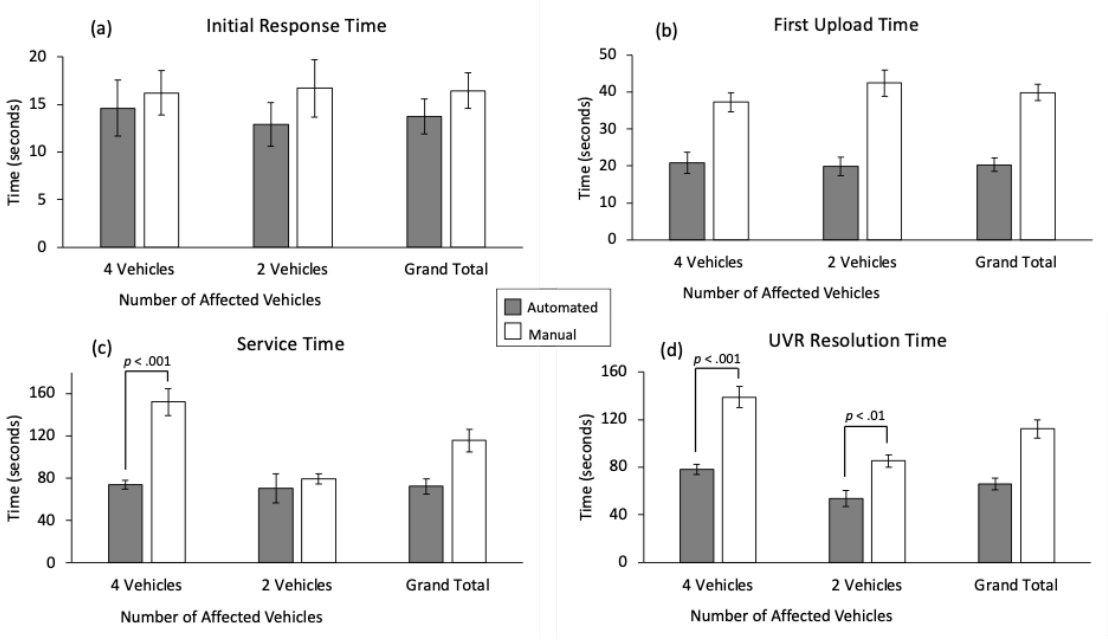


Fig. 8 Response Time metrics (all measured in seconds) by NAV condition and automation support level: (a) Initial Response Time, (b) First Upload Time, (c) Service Time, and (d) UVR Resolution Time. Each bar is depicted as the across-subjects mean and standard error of the mean (SEM).

There were significant main effects of automation support level ($F(1, 9) = 69.03, p < .001, \eta_p^2 = .89$) and NAV condition ($F(1, 9) = 66.494, p < .001, \eta_p^2 = .88$) on UVR Resolution Times (Fig. 8d). Once again, the manual and 4 vehicle conditions resulted in significantly longer response times. There was also a significant interaction between automation support level and NAV condition on UVR Resolution Time ($F(1, 9) = 7.65, p = .02, \eta_p^2 = .46$). While automation support level had a significant effect on UVR Resolution Times in both the 2 vehicle ($t(9) = 3.77, p = .002$, Cohen's $d = 1.19$) and 4 vehicle ($t(9) = 8.88, p < .001$, Cohen's $d = 2.81$) conditions, its effect was more pronounced in the 4 vehicle condition. The automated condition reduced UVR Resolution Times by an average of 61 seconds in the 4 vehicle condition, compared to an average reduction of 32 seconds in the 2 vehicle condition.

B. Reroute Performance Metrics

There was a significant main effect of automation support level on average Path Length Difference ($F(1, 9) = 27.74, p = 0.001, \eta_p^2 = 0.76$, Fig. 9). Flight paths for rerouted aircraft were 26% longer, on average, in the manual condition compared to the automated condition.

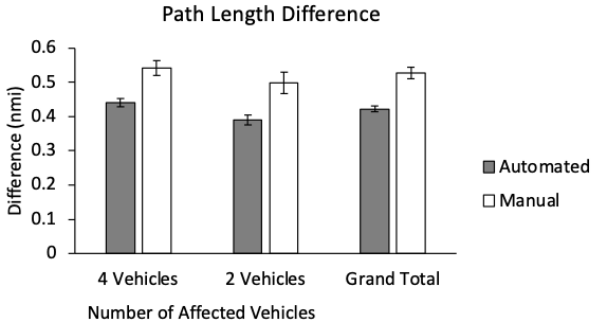


Fig. 9 Path Length Difference (measured in nautical miles) by NAV condition and automation support level, depicted as the across-subjects mean and standard error of the mean (SEM).

There were only two experiment-wide occurrences of UVR Violations (0.83% of cases), and both occurred with the same pilot participant. Regardless, both UVR Violations were recorded in the 4 vehicle condition. In these instances, the average UVR Violation Duration was 21 seconds.

There were five total Flights into Terrain during the study (7.9% of cases), four of which occurred after planned terrain encounters and one after an encounter without a planned terrain conflict. There was no clear trend regarding the effect of automation support level and NAV condition. Three Flights into Terrain occurred in the 2 vehicle condition, compared to two in the 4 vehicle condition, while three occurred in the automated support condition and two occurred in the manual condition.

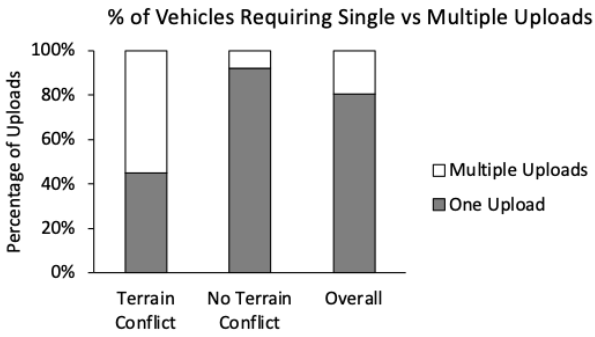


Fig. 10 Percentage of vehicles requiring one vs. multiple uploads to resolve the UVR conflict, as a function of the presence or absence of a nearby terrain conflict during the reroute.

The presence of a terrain conflict had a substantial impact on the Number of Uploads required to successfully resolve a UVR conflict. When there was no nearby terrain, 92% of vehicles required a single upload, with the remaining 8% requiring multiple uploads. However, when there was a terrain conflict nearby, only 45% of vehicles could be rerouted with a single upload; the remaining 55% required multiple uploads (Fig. 10).

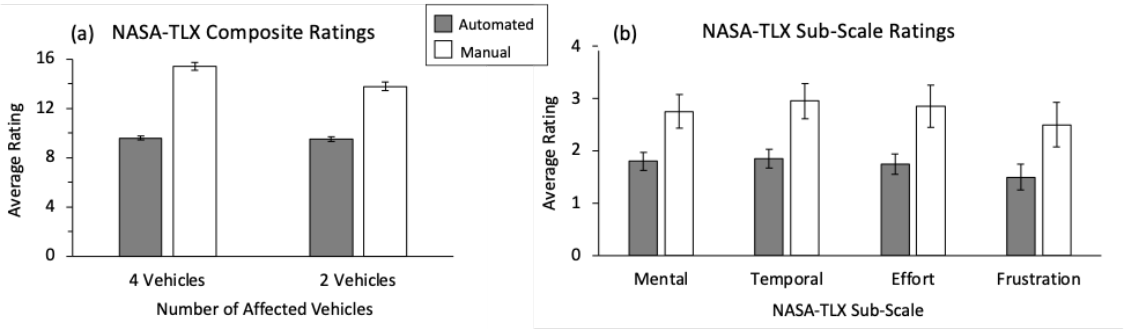


Fig. 11 (a) Composite scores of the NASA-Task Load Index (TLX) and (b) scores on the TLX sub-scales with significant differences between the automation support levels. Each vertical bar and error bar represents the mean and standard error of the mean, respectively.

C. Subjective Metrics

As shown in Figure 11a, pilots' workload ratings on the NASA-TLX were generally low, averaging between 10 and 15, on a scale with minimum and maximum composite ratings of 6 and 42, respectively. Nevertheless, the automated condition was found to significantly decrease workload ratings on the composite score ($F(1,9) = 12.74, p = .006, \eta_p^2 = .59$). The NAV condition variable, however, was not found to have a significant main effect on participants' TLX composite scores ($F(1,9) = 1.18, p > .05, \eta_p^2 = .12$), nor was a significant interaction found between automation support level and NAV condition on composite scores ($F(1,9) = .57, p > .05, \eta_p^2 = .06$). Significant decreases were observed for the automated support condition on the Mental Demand ($F(1,9) = 14.19, p = .004, \eta_p^2 = .61$), Temporal Demand

($F(1,9) = 27.92, p = .001, \eta_p^2 = .76$), Effort Level ($F(1,9) = 14.72, p = .004, \eta_p^2 = .62$), and Frustration Level ($F(1,9) = 7.83, p = .02, \eta_p^2 = .47$) sub-scales of the NASA-TLX, which have a minimum workload rating of 1 and a maximum rating of 7 (Fig. 11b). The reduction in workload ratings from the manual to the automated condition ranged from 35-40%. As seen with the composite scores, the NAV condition was not found to have a significant main effect on sub-scale ratings: Mental Demand ($F(1,9) = .13, p > .05, \eta_p^2 = .01$), Temporal Demand ($F(1,9) = 1, p > .05, \eta_p^2 < .001$), Effort Level ($F(1,9) < .001, p > .05, \eta_p^2 < .001$), and Frustration Level ($F(1,9) = 1.46, p > .05, \eta_p^2 = .14$).

Display feedback indicated that the majority of subjects considered the information shown on the TSD "very useful" ($n = 7$). Subjects did note that the TSD should have included terrain information with the minimum safe altitudes visible on the map. Subjects also requested weather overlays and projected vectors off the nose of their aircraft. Nonetheless, pilots indicated that the TSD information presented to them was ultimately sufficient for resolving the UVR scenario ($n = 10$). The majority of pilots also believed that the Timeline display was "very useful" ($n = 8$), but should have included time to next waypoint, priority of UVR conflicts (e.g., distance to UVR), altitude information, and UVR details (e.g., hover to show point-of-contact/call number) to be considered comprehensive.

All subjects thought they received sufficient training ($n = 10$) but experienced some degree of disruption in performance due to the remote nature of the study. Most subjects believed liaison interaction was sufficient and found the task to be completely manageable ($n = 7$). A large proportion of measured display latency (70%) was in the range of 1-2 seconds, with most of the participant cohort reporting good streaming quality ($n = 8$).

VII. Discussion

Results from the present study suggest that the two primary variables of interest - automation support level, which controlled whether or not subjects received a pre-built reroute suggestion from the GCS, and NAV condition, which varied the number of vehicles (2 or 4) affected by the UVR – had pronounced effects across a large subset of objective measures of performance as well as subjective reports of workload. The only response time metric that did not show a significant main effect of either variable was Initial Response Time. This finding suggests that subjects responded to the UVR with a consistent urgency and did not systematically vary their initial reaction to the emergency across test conditions. The remaining response time metrics, however, did show a significant main effect of automation support level. Unsurprisingly, the automated condition significantly reduced pilots' First Upload Times, Service Times, and UVR Resolution Times. The NAV manipulation, however, only had significant main effects on the Service Time and UVR Resolution Time metrics. This result reflects that the First Upload Time only captured subjects' responses to the first vehicle affected by the UVR; the extra time they spent on the remaining vehicles was only captured in the Service Time and UVR Resolution Time measures. In those conditions, the 4 vehicle condition resulted in significantly longer response times, which was expected.

The Service Time and UVR Resolution Time metrics also produced significant interactions between the automation support level and NAV condition variables. In particular, the automated condition was found to have an outsized effect in reducing response times during the 4 vehicle condition, compared to the 2 vehicle condition. The benefit of the automation was likely less apparent in the 2 vehicle condition since fewer vehicles received the benefit of a suggested reroute in those scenarios. In the case of the Service Time metric, the benefit of the automation condition was completely offset by the extra time required to resolve nearby terrain conflicts, something the automation was not designed to take into account when generating reroutes. For the UVR Resolution Time metric, the presence of the automation did lead to a significant reduction in response times in the 2 vehicle condition, but the effect was slightly smaller, reducing times by 44% in the 4 vehicle condition, compared to 37% in the 2 vehicle condition.

The automation support level continued to have a significant effect on the average Path Length Difference between vehicles' original routes and their amended reroutes around the UVR. Specifically, the automated reroutes reduced the path length by 26% compared to when subjects had to reroute manually. While the automation was not specifically designed to optimize flight path length, it was found to perform better than subjects in the manual condition, who had to verbally coordinate the reroute with a surrogate researcher. Despite a slight loss in path efficiency, subjects in all conditions were highly effective at avoiding the scripted UVRs. Across all subjects, only two UVR Violations were recorded. While both violations occurred in the 4 vehicle condition, there were not enough instances to establish a statistical pattern.

While the majority of the results focused on the impact of the automation support level and NAV condition variables, the presence of terrain near the UVR also resulted in several important findings. The first was the occurrence of five total Flights into Terrain across the study. As with the UVR Violations metric, there were too few instances to establish a pattern, with multiple instances occurring in each condition. Furthermore, the presence of a terrain conflict increased

the number of uploads required to fully resolve a conflict. Whereas subjects only required a single upload for 90% of their vehicles when no nearby terrain was present, that number decreased to 45% when terrain was a factor. This was typically a result of subjects first uploading a lateral maneuver that rerouted the aircraft around the UVR, and subsequently sending one or more additional uploads to raise the altitude for the newly-created waypoints above the local terrain.

Subjective results also revealed an effect of the automation support level on subjects' perceived workload. The automated condition was associated with significantly lower workload ratings, both on the NASA-TLX composite score, as well as on four of the TLX sub-scales: Mental Demand, Temporal Demand, Effort Level, and Frustration Level. One particularly surprising finding was the lack of a significant effect in the TLX composite scores with respect to the NAV condition. The NAV manipulation was designed to increase the pressure on subjects by doubling the number of vehicles they had to reroute around the UVR. Based upon subjects' TLX ratings, however, this increased task load did not correspond to higher workload ratings. One possible explanation is the fact that the vehicles' position relative to the UVR was staggered, such that the fourth vehicle in the fleet in the high workload condition was approximately 2-3 minutes behind the lead vehicle at the onset of the UVR. Thus, while the 4 vehicle condition required a greater total number of reroutes, that same scenario also provided them with more time to respond, potentially mitigating its effect on perceived workload. An additional factor may be that the NAV variable did not change the way in which the subjects interacted with the surrogate researcher during a trial, it only increased the number of their interactions. By contrast, the automation support level variable drastically changed the nature of the subject-surrogate interaction. In the automated condition, subjects no longer had to re-position waypoints individually; instead, they could pull up the recommended reroute for a vehicle and ask the surrogate to execute it immediately, assuming they were comfortable with the suggested reroute provided by the automation. Despite the disparate effects of the experimental variables, subjects' self-ratings indicated relatively low workload levels overall, suggesting that the task at hand was well within their workload limits.

At the end of the experiment, subjects were given an opportunity to rate the overall effectiveness of the displays and tools under test, as well as comment on the effect the remote nature of the study had on their performance. While subjects indicated that they had adequate training on the overall system, and found the tactical situation and timeline displays useful, they consistently mentioned a desire for terrain information – namely, a visualization of the digital terrain elevation database (DTED) - and an ability to perform "bulk action" reroutes (i.e., reroute multiple vehicles with a single upload/action), similar to the multi-vehicle "play" concept described in Fern and shively (2009). Furthermore, subjects reported that the subject-surrogate paradigm that structured their interactions with the researcher was completely manageable, in large part due to their ability to view the GCS remotely without disruptive delays.

VIII. Conclusion

This paper presented a GCS interface that enabled remote TOs to manage 12, highly automated small UAS under a remote subject-surrogate data collection paradigm. Results demonstrated clear effects of each of the experimental variables. Generally, response times increased in the conditions that required subjects to manually reroute vehicles and when there were more vehicles impacted by the UVRs. Automation was particularly effective at reducing subject response times in the 4 vehicle condition and at minimizing the size of the lateral deviation around the UVRs.

Ultimately, subjects performed the primary task of avoiding the UVR at an extremely high rate (99% of the vehicles were successfully rerouted around the UVR). This can be explained by the generally low workload ratings submitted by the subjects and by the largely positive feedback regarding the effectiveness of the displays and tools provided to them. Taken together, these findings suggest that the automation improved the efficiency of subjects' performance, but it did not lead to a significant difference in the number of UVR violations that occurred, which was the primary task subjects were trained on. This finding is even more noteworthy given the experimental setup, which had the subjects participating from a location outside the lab, viewing the experimental interfaces and communicating with researchers remotely, and experiencing the delays and reductions in video quality inherent in such an approach.

Future papers from the authors will elaborate on the subjective feedback submitted by the participants to more thoroughly capture the lessons learned and summarize their recommendations to inform future ground control station interface designs. Additional data on the impact of the subject-surrogate protocol will also be reported to help illuminate the benefits, and drawbacks, of this uncommon methodology. The authors will also explore methods that can be developed to predict the workload of operators based on their real-time interactions with the GCS.

Acknowledgments

Dr. Joel Lachter and Jay Shively of the Human Systems Integration Division at NASA Ames Research Center, and Vernol Battiste of the San Jose State University Research Foundation for their help and support in developing the concepts described in this paper and for their input on the experimental design.

References

[1] Keane, J. F., and Carr, S. S., "A brief history of early unmanned aircraft," *Johns Hopkins APL Technical Digest*, Vol. 32, No. 3, 2013, pp. 558–571.

[2] Smith, C. L., Sadler, G., Tyson, T., Brandt, S., Rorie, R. C., Keeler, J., Monk, K., Viramontes, J., and Dolgov, I., "A Cognitive Walkthrough of Multiple Drone Delivery Operations," *AIAA Aviation 2021 Forum*, 2021, p. 2330.

[3] Federal Aviation Administration, "Concept of Operations v2.0, Unmanned Aircraft Systems (UAS) Traffic Management (UTM)," *URL https://www.faa.gov/uas/research_development/traffic_management/media/UTM_ConOps_v2.pdf*, 2020.

[4] Sheridan, T., "Humans and Automation," *Santa Monica, CA: Human Factors and Ergonomics Society and New York: Wiley*, 2002.

[5] Parasuraman, R., Sheridan, T. B., and Wickens, C. D., "A model for types and levels of human interaction with automation," *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, Vol. 30, No. 3, 2000, pp. 286–297.

[6] Chen, J. Y., Barnes, M. J., and Harper-Sciarini, M., "Supervisory control of multiple robots: Human-performance issues and user-interface design," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 41, No. 4, 2010, pp. 435–454.

[7] Monk, K. J., Rorie, R. C., Sadler, G. G., Brandt, S., and Roberts, Z. S., "A Detect and Avoid System in the Context of Multiple-Unmanned Aircraft Systems Operations," *AIAA Aviation 2019 Forum*, 2019, p. 3315.

[8] Fern, L., and Shively, R. J., "A comparison of varying levels of automation on the supervisory control of multiple UASs," *Proceedings of AUVSI's Unmanned Systems North America 2009*, 2009, pp. 10–13.

[9] Cummings, M. L., Nehme, C. E., Crandall, J., and Mitchell, P., "Predicting operator capacity for supervisory control of multiple UAVs," *Innovations in Intelligent Machines-1*, Springer, 2007, pp. 11–37.

[10] Miller, C. A., and Parasuraman, R., "Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control," *Human Factors*, Vol. 49, No. 1, 2007, pp. 57–75.

[11] Nehme, C. E., Scott, S. D., Cummings, M., and Furusho, C. Y., "Generating requirements for futuristic hetrogenous unmanned systems," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50, SAGE Publications Sage CA: Los Angeles, CA, 2006, pp. 235–239.