A Queuing Theory Approach to Pilot-Controller Coordination for *m*:*N* Operations

Meghan Saephan* and Garrett G. Sadler†
NASA Ames Research Center, Moffett Field, CA, USA

In recent years, attention and interest by industry and researchers has grown in a control paradigm for remotely piloted aircraft termed "*m*:*N* operations." In an *m*:*N* operation, a team of *m* remote pilots in command (RIPCs) collaboratively manage the flights of *N* aircraft. A consequence of an *m*:*N* concept of operations is that the RPICs will have to switch attention from one aircraft to another and from one task to another. Previous research in *m*:*N* operations has focused on the workload experienced by an RPIC and their level of situation awareness on their flights. Researchers have found that RPIC workload and situation awareness are generally sensitive to increasing *N*, although NASA's Multi-Vehicle (*m*:*N*) Working Group has suggested that the driver of workload/situation awareness is the number of exceptions requiring human intervention as opposed to the value of *N* itself. In any case, a natural antecedent of workload is task load. In this paper, queueing theory is applied to a 1:*N* Urban Air Mobility (UAM) air taxi operation in order to estimate pilot task load for managing radio communications with air traffic controllers (ATCs) under increasing *N*. An M/M/1 queueing system is used to model the RIPC's servicing of calls and clearance requests (e.g., departure, arrival, or airspace transition) to ATC for the *N* aircraft. Important parameters for the queueing model are the task arrival rate and the average service time for task completion. Radio communication times from past human-in-the-loop simulation studies are used to measure service times for a 1:4 and 1:12 UAM operation and to interpolate service times for $4 < N < 12$. A Monte Carlo method is then employed, using the measured and interpolated service times, to estimate arrival rate and related queueing statistics. The paper concludes by considering the estimated queuing statistics, particularly the RPIC's utilization (i.e., proportion of time actively servicing tasks), the length of the task queue over time, and the implications for task-balanced system design.

I. Nomenclature

m	=	number of remote pilots in command
N	=	number of aircraft being managed
M/M/1	=	single-server queuing model with exponential arrival and service rates
λ	=	arrival rate
μ	=	service rate
$E[s]$	=	average service time
ρ	=	utilization
L	=	total number of tasks (including those in the queue and those being serviced)
L_q	=	mean queue length
L_s	=	mean number of tasks being serviced
$L_{\max}$	=	maximum observed queue length
p_0	=	probability of zero tasks in queue
$t_{CS,\max}$	=	maximum amount of time spent continuously in service (i.e., without a break)
t_{serv}	=	service time
t_{start}	=	time at which a radio call begins service
t_{stop}	=	time at which a radio call completes service
W	=	mean total time a task takes to complete (including wait time and service time)

* Computer Engineer, Intelligent Systems Division
† Human Factors Researcher, Human/Machine Systems, Human Systems Integration Division.

W_s = mean service time
W_q = mean task wait time

II. Introduction

Future Advanced Air Mobility (AAM) concepts envision a complex aviation ecosystem with diverse sets of aircraft ranging in size at various altitudes. These aircraft will include a mixture of both traditional crewed commercial aircraft along with uncrewed urban air transport aircraft (for cargo and passengers) and small uncrewed aerial systems (sUAS) [1]. In recent years, interest has grown in a concept of operations (ConOps) for uncrewed aircraft (UA) in which fleets of aircraft—whose levels of automation range from highly automated to fully autonomous—are managed by a team of remote pilots in command (RPICs). This new paradigm of many people collaboratively managing multiple aircraft is eponymous with the suggested RPIC-to-Vehicle-Load ratio: *m*:*N*. In *m*:*N, m* is the number of RPICs and *N* is the number of aircraft managed [2] between them. In contrast to a traditional flight deck, where the ratio is 2:1, *m*:*N* operations typically assume $N > m$. As of November 2024, five US operators have received Part 135 certification to perform beyond-visual-line-of-sight, *m*:*N* package delivery operations in limited areas using sUAS [3].

To ensure that *m*:*N* operations guarantee an acceptable level of safety for all flights, it is important that all RPICs maintain effective levels of workload [4] and situation awareness [5] of the UAs they are managing as well as any factors that may affect 1) the planned mission(s) of one or more of their UAs or 2) the level of safety of their operations. And, indeed, research studies of *m*:*N* and multi-aircraft control, more broadly, have examined workload and situation awareness in a wide range of pilot-to-aircraft configurations.

A natural antecedent of workload is task load. In this context, workload is the mental and physical demands associated with performing tasks, whereas the task load is the difficulty and number of tasks [6]. When task load becomes excessive to the point of overwhelm, the one executing the tasks experiences task saturation. This breakdown-condition of task load is analogous to the so-called "red line" of workload excess [4]. Previous research found that RPIC performance in multi-UAS applications degrades when utilization rates (i.e., the proportion of time actively busy) drop below 40% [7] or exceed 70% [8, 9]. One potential way to evaluate task saturation during *m*:*N* operations, is to model task load using queuing theory. Tasks, executed serially, potentially form a queue. The size of that queue and how long tasks "wait" in line for service depend on how frequently tasks arise and how long they take to complete. Applying queueing theory to a given *m*:*N* ConOps is one way to probe the relationship between the *m*:*N* ratio and RPIC task load. By varying common drivers of task load and measuring the corresponding changes to queueing parameters, one can roughly produce "boundaries of feasibility" for an *m*:*N* operation.

Previous applications of queueing theory to aviation scenarios generally pertain to passenger-class aircraft traveling between airports [10, 11]. These applications bring unique challenges for a queueing application, as airports have limited resources that can be used to serve demand, including runways, taxiways, and gates. Compared to these studies of limited physical resources [11], the demands on personnel (e.g., [10]) are not as often modeled explicitly.

In this paper, we aim to evaluate the impact of the arrival rate and service rate of tasks in an *m*:*N* mission on task saturation and the implications for RPICs' ability to manage multiple aircraft simultaneously. This paper will utilize data collected from a previous study focused on comparing the effectiveness of various operator-ATC communications modalities in an Urban Air Mobility (UAM) operation [12]. Monte Carlo simulations were performed using radio communication service time data from [12] to estimate parameters of task queueing models for increasing *N*. The implications of the queueing models and their statistics on operator task load are discussed and may provide insight for balancing task load in system design.

III. UAM Communications Modality Experiment

A major challenge outlined in the AAM concepts is the control of aircraft within urban environments. To mitigate some aspects of this challenging environment, UAM assumes the presence of highly automated services for air traffic management (ATM), such as a Provider of Services for UAM (PSUs) [1, 13]. Such services are intended to provide strategic deconfliction of remotely piloted aircraft at scale; however, for the case of initial UAM operations, the Federal Aviation Administration (FAA) expects that remote pilots of UAM aircraft will coordinate with air traffic controllers (ATC) using traditional voice-over-radio communications [14]. In a previous human-in-the-loop (HITL) simulation study [12], the authors examined the effects of ATC communication for a simulated *m*:*N* UAM operation on pilot workload and performance. Communication modality (voice, datalink, and a hybrid—i.e., a combination of voice and datalink modes) and the *m*:*N* ratio (1:4 and 1:12) were manipulated. Response times, error rates, and

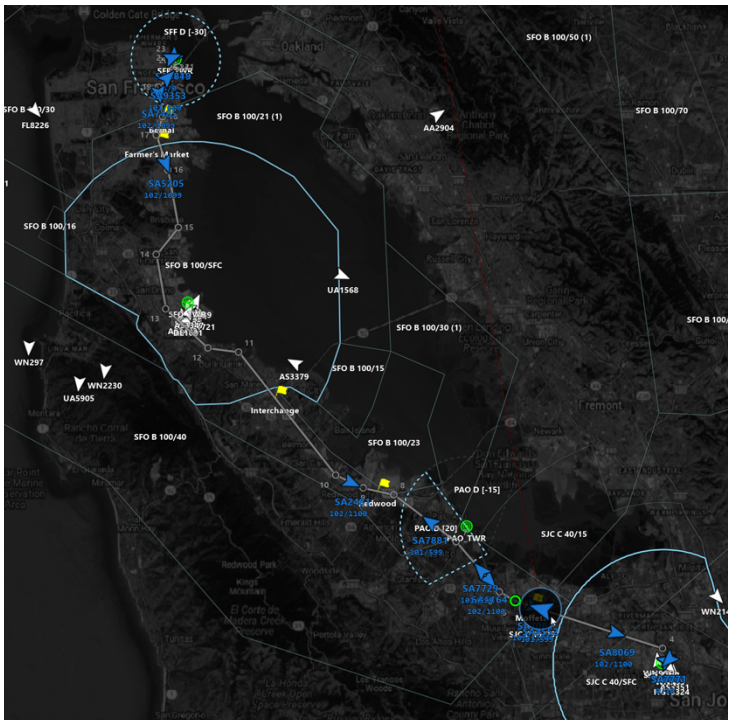

Fig. 1. Example of the Tactical Situation Display used for the UAM Communications Modality Experiment. Cyan airspace boundaries (Class B/C airspace – solid, Class D airspace – dashed) highlight those that are co-altitude with the UAM flights. Checkpoint locations for radio calls are designated with yellow flags [12].

subjective workload ratings (NASA-TLX [6], administered post-trial) were recorded. Twelve participants (M = 38.75 years of age), each in possession of a Private Pilot License, were recruited for the study. Pilots had an average of 5,019 hours flying crewed (i.e., onboard pilot) aircraft. Three of the pilots held Part 107 certificates and had an average of 28.3 hours of remote flying sUAS weighing less than 55 lbs. Data collection occurred between September 7 and September 30, 2022.

Table 1. UAM Communication Modality Experiment Sequence of Calls to ATC for a Northbound Flight

No.	Call Type	Recipient	Example Verbiage
1	Departure Request	SJC	"San Jose Tower, [callsign], request Bayshore Departure."
2	Clear of Zone	SJC	"San Jose Tower, [callsign] at Moffett 600 ft, clear of control zone."
3	Transition Request	PAO	"Palo Alto Tower, [callsign] at Moffett 600 ft, request permission to transit through control zone."
4	Clear of Zone	PAO	"Palo Alto Tower, [callsign] at Redwood 600 ft, clear of control zone."
5	Transition Request	SFO	"San Francisco Tower, [callsign] at Interchange 600 ft, request permission to transit through Class Bravo."
6	Clear of Zone	SFO	"San Francisco Tower, [callsign] at Farmer's Market 600 ft, clear of control zone."
7	Arrival Request	SFF	"Ferry Tower, [callsign] at Bernal 600 ft with information Alpha, request Embarcadero Arrival."

Participants in the m:N UAM Communication Modality study played the role of the RPIC tasked with managing, depending on condition, four or twelve active UAM flights traveling between fictional "vertiports" in the San Francisco Bay Area. Aircraft traveled along a route connecting a vertiport collocated with the San Jose International Airport (SJC) at the southern end and terminating at a vertiport at the San Francisco Ferry Terminal (SFF) at the northern end (Fig. 1). The route followed, roughly, the right-of-way of the 101 Freeway along the peninsula, passing through intermediate controlled airspaces. Participants were responsible for obtaining departure and arrival clearances

at the two vertiports and transition clearances for the intermediate airspaces: a Class D airspace surrounding Palo Alto Airport (PAO) and a Class B airspace around San Francisco International Airport (SFO). Additionally, participants in the experiment were instructed to inform the appropriate ATC authority when they were exiting their airspace. Marked and named checkpoints along the route indicated points at which calls to ATC should be performed. Northbound flights traveled at a cruise altitude of 600 feet mean sea level (MSL) and Southbound aircraft traveled at a cruise altitude of 1100 feet MSL. A list of expected ATC calls for a single UAM flight traveling from SJC to SFF is provided in Table 1. Unsurprisingly, results of the study found that workload was significantly higher, nearly maxing out the NASA-TLX scale, when performing comms for twelve versus four aircraft. On the other hand, and unexpectedly, no significant difference in workload was recorded between the three communication modalities examined. Communication errors, such as using the wrong callsign or speaking on an incorrect frequency, were common in both vehicle load conditions. Participants' response times were substantially faster in the voice condition, though that condition also contained the highest frequency of errors. For more details on the study, see [12].

Whereas participants in the experiment utilized voice, datalink, or a hybrid communication system to coordinate with ATC, depending on experimental condition, the analysis of this paper will focus attention to service time data for departure, arrival, transition, and clear of zone clearance calls from trials using voice-over-radio. For our purposes here, service times for calls to ATC will be measured according to conventional radio communications etiquette consisting of a clearance call/request from the RPIC, a response from ATC, and a readback of ATC's response by the RPIC. This service time corresponds to the interval between when the PIC first "keys" the push-to-talk (PTT) button on the headset to make the clearance call and the release of the PTT button at the conclusion of readback.

IV. Method Overview

A. Queuing Model

Queuing theory lends itself to modeling the performance of *m*:*N* operations. In such systems, the tasks coming in for the *N* aircraft are either completed immediately if the operator is available, or it must wait to be completed until the operator finishes their previous tasks. The tasks that must wait are assumed to be waiting according to some priority scheme within a "queue." The rate at which these tasks arrive is known as the arrival rate. Tasks are completed by one of *m* RIPCs (also referred to as a server in the queuing literature).

This paper applies the M/M/1 queueing model to the voice communications data from the UAM Communications Modality Experiment. This queueing model nomenclature is standardized across queueing literature to allow for easy comparison and identification of relevant models [15]. The first term in a queueing model notation denotes how events or tasks arrive, termed the "arrival rate." The second term denotes how events or tasks are completed, termed the "service rate." The final term (here, the number 1) denotes how many servers are available. In an *m:N* context, this would refer to the number of RPICs. The M denotes an exponential event arrival/service rate, which has the memoryless property that states the probability of a new task arriving in a time period between times t and $t + k$ is independent of whether or not a new task arrived in the time period between $t - k$ and t [15].

The M/M/1 model has the advantage of being computationally tractable, with many important metrics easily derived from system characteristics [9, 11]. However, we do not just apply the M/M/1 model in this paper for ease of calculation; rather, the model is appropriate for the specific scenario we model in this paper. This is not necessarily intuitive. For example, any individual UAM aircraft has a set of tasks that an RPIC must complete, and each task *for each aircraft* is deterministically spaced. However, in the aggregate, the chance of an operator receiving a task to complete in any given interval is independent of whether the operator completed a task in the prior interval because the operator is managing multiple aircraft at different states. Thus, the inter-task arrival rate for the operator (rather than for the individual aircraft) is exponential. Likewise, the inter-task service rate is exponential, as the operator can only service tasks that have arrived in the queue, and each task's service time is a random variable that is independent of the length of prior tasks completed. Finally, for this paper we assume that a single operator is controlling a number of UAM aircraft, making the single-server model appropriate (i.e., 1*:N*). For the more general case of *m*:*N* ($m > 1$), a multi-server model for an M/M/c (with $c = m$) queue may be applied.

Although we believe that the M/M/1 model is appropriately applied in *m*:*N* (where $m = 1$), it is nonetheless important to spell out the key assumptions that we are making in applying this model. They are:

1) Tasks arrive and are completed stochastically;
2) The time between task arrival and completion follows an exponential distribution;
3) There is a single operator completing tasks; and
4) The queue of tasks is unlimited in size [15].

These assumptions are quite restrictive, and it is unlikely that any given aviation task meets each of these assumptions completely; however, other researchers have shown that the results from applying a queueing model that requires these assumptions are as good as those from more sophisticated, non-analytically soluble methods [15]. Further, as described below we have sculpted the particular problem addressed in this paper so that these assumptions are appropriate.

As defined in [15], tasks arrive in a queueing system at rate λ and are completed at rate μ. The average time to complete a task (i.e., the service time) is $E[s]$. On average, there are L tasks. Tasks are either being serviced or are waiting in a queue. Therefore,

$$L = L_q + L_s\,, \tag{1}$$

where L_q is the mean number of tasks waiting in the queue at any given time and L_s is the mean number of tasks being serviced at any given time. Similarly,

$$W = W_q + W_s\,, \tag{2}$$

where W is the total time a task takes to be completed, W_q is the mean waiting time for a task, and W_s is the mean service time for a task. The system has ρ utilization (i.e., proportion of time spent servicing tasks) where,

$$\rho = \frac{\lambda}{\mu}. \tag{3}$$

For an M/M/1 queuing system where tasks arrive at and are completed at an exponential rate, the steady-state values for L, L_q, and L_s are given respectively by:

$$L = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho} \tag{4}$$

$$L_q = \frac{\lambda^2}{\mu - \lambda} = \frac{\rho^2}{1 - \rho} \tag{5}$$

$$L_s = \rho. \tag{6}$$

For queuing systems for which a steady-state exists, Little's queuing formula states that:

$$L = \lambda W, \quad L_q = \lambda W_q, \quad \text{and } L_s = \lambda W_s \tag{7}$$

Applying Little queuing formula to an M/M/1 queuing system where arrival and service rates are exponential yields the steady-state values for W, W_q, and W_s:

$$W = \frac{1}{\mu - \lambda} \tag{8}$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} \tag{9}$$

$$W_s = \frac{1}{\mu}. \tag{10}$$

From the equations above, we can also define the probability of no tasks being in the queue as:

$$p_0 = 1 - \frac{\lambda}{\mu}. \tag{11}$$

The probability of zero tasks being in the queue implies that there is zero waiting time when a new task arrives. Therefore,

$$W_q(0) = p_0 . \tag{12}$$

B. Service Time Measurement and Interpolation

In the section that follows, a Monte Carlo method is described to produce estimates of the M/M/1 queueing statistics described above: λ, μ, ρ, L, L_q, W, W_q, and p_0. In order to perform the Monte Carlo simulation, it will be necessary to have service time measurements for performing radio calls. This section describes how HITL data from the UAM Communication Modality Experiment was used to measure service time averages for the cases of $N = 4$ and $N = 12$. The average service time computed is the value of $E[s]$ for the corresponding value of N. And whereas $E[s]$ is empirically determined for $N = 4$ and $N = 12$, service time averages were interpolated for N, $4 < N < 12$. Both linear and exponential least-squares approximations were tested and, as elaborated on below, an exponential approximation was selected for the analysis herein.

Table 2. Definition of t_{start} and t_{end} by Transmission Category

	Prototypical Example Call Sequence				
Category	#	Speaker	Transmission (*italics to distinguish speakers*)	t_{start}	t_{end}
Departure	1	RPIC	"San Jose Tower, [callsign] requesting Bayshore Departure."	Start #1	End #3
	2	*SJC*	*"[callsign], San Jose Tower, clear for departure and report leaving the control zone."*		
	3	RPIC	"Clear for departure, report leaving, [callsign]."		
Arrival	1	RPIC	"Ferry Tower, [callsign] at Bernal 600 ft with information Alpha, request Embarcadero Arrival."	Start #1	End #3
	2	*SFF*	*"[callsign], Ferry Tower, clear to land, Embarcadero arrival, vertispot 1."*		
	3	RPIC	"Clear to land, spot 1, [callsign]."		
Transition	1	RPIC	"San Francisco Tower, [callsign] at Interchange 600 ft, request permission to transit through Class Bravo."	Start #1	End #3
	2	*SFO*	*"[call sign], San Francisco Tower, clear for the transition and report leaving the control zone."*		
	3	RPIC	"Clear for the transition, reporting leaving, [callsign]."		
Clear of Zone	1	RPIC	"Palo Alto Tower, [callsign] at Redwood 600 ft leaving control zone."	Start #1	End #2
	2	*PAO*	*"[call sign], Palo Alto Tower, roger, good day."*		

To measure service times, audio recordings and corresponding communication log files collected during the UAM Communications Modality Experiment were reviewed and coded to produce records that included, for every transmission, who was speaking (ATC, participant, background transmissions), the call category (see Table 2), frequency of transmission (SJC, PAO, SFO, SFF), and transmission durations (time from when PTT button is depressed to released). As mentioned above, calls/requests between pilots and ATC follow an etiquette consisting of a call-response-readback sequence. The service time, t_{serv}, associated with a given request to ATC, therefore, is the span of time from the start of the initial, t_{start}, call to the time of completion of the readback, t_{end} (i.e.,. $t_{serv} = t_{end} - t_{start}$). Table 2 provides the definitions for t_{start} and t_{end} for each transmission category. Note that, in contrast to the other categories, only two transmissions comprise the "Clear of Zone" category.

A total of 1,117 service time values were obtained from Departure, Arrival, Transition, and Clear of Zone call sequences. Of those 1,117 data points, 42 call sequences (3.8%) were excluded due to qualitatively breaking from the form of the call sequences in Table 2. Of these calls, there were 26 cases where the RPIC failed to perform a readback, 12 cases where the RPIC grouped multiple call categories and/or multiple callsigns together, 1 case of multiple callsigns in a single transmission combined with no readback, 1 case of an aborted call, and 2 cases where the RPIC called Clear of Zone and ATC never responded. The remaining 1,075 service time values were used for the analysis.

Measured service time averages, for each category and overall, are provided in Table 3. In Section V below, formulae for the aforementioned queueing statistics are provided. These formulae assume a single service time, notated $E[s]$, representing the average service time for all tasks. When multiple service time categories are assumed, producing

closed-form solutions for these formulae becomes intractable [15]. For that reason, the overall average radio call service time is used as $E[s]$ for each value of N:

$$E[s] = \begin{cases} 14.551 \text{ sec} & N = 4 \\ 13.603 \text{ sec} & N = 12 \end{cases} \tag{13}$$

Since the Vehicle Load variable of the UAM Communication Modality Experiment was only two levels — 4 and 12 — the two values of $E[s]$ provided by Eq. 13 are the only empirically measured service times used in the Monte Carlo method below. Empirically measured data was not available for $4 < N < 12$; therefore, their values were interpolated.

Table 3. HITL Service Time Averages by Category

	Radio Call Service Time, sec			
	$N = 4$		$N = 12$	
Category	*M*	*SD*	*M*	*SD*
Departure	20.136	7.168	16.813	5.352
Arrival	18.579	4.301	16.059	4.253
Transition	16.989	3.775	15.856	4.037
Clear of Zone	9.589	1.736	9.428	2.396
Overall = $E[s]$	14.551	5.820	13.603	4.875

One challenge of interpolation in this case, which is another byproduct of the binary Vehicle Load variable, is a lack of basis for reckoning the "shape" of the service time curve between the $N = 4$ and $N = 12$ extrema. Should a linear or nonlinear fit be sought? With no intermediate service time data, there is no inherent or obvious answer.

One way to address this question is to reference other existing radio call service time data and trends for similar tasks. For this purpose, the authors turned to data from a multi-UAS study performed by the Monk et al. in 2018 [16]. The focus of the 2018 study was the "detect and avoid" (DAA; a subject of tactical traffic conflict management) problem [17, 18] in a multi-UAS context. Crucially, in that study, participants played the role of the RPIC for 1, 3, or 5 aircraft. Intruder aircraft were injected into the scenario as traffic conflicts leading to DAA alerts if unmitigated by the RPIC. When alerted of a DAA conflict, maneuver guidance was provided to assist in maintaining DAA "Well Clear" (traffic separation) [18]. Of particular importance for our purposes here: RPICs were required to coordinate or inform (depending on DAA threat level) ATC of their maneuver actions.

The standard procedure for responding to a DAA alert is 1) if the alert is a yellow-level, "corrective" alert the RPIC is expected to contact ATC and coordinate their path deviations due to traffic prior to performing the evasive maneuver or 2) if the alert starts at or elevates to a red-level, "warning" alert, then the RPIC is expected to maneuver for traffic immediately and inform ATC after the fact. We were able to pull service time data for the calls made to ATC during the DAA encounters and computed each participant's average service time for calls made in the $N = 1$, $N = 3$, and $N = 5$ conditions. We then averaged across participants to get the service time averages and distribution shown in Fig 2. The obtained service time means (and standard deviations) are $M = 8.59$ sec ($SD = 3.27$ sec), $M = 8.15$ sec ($SD = 2.71$ sec), and $M = 7.94$ sec ($SD = 2.93$ sec), for the $N = 1$, $N = 3$, and $N = 5$ aircraft conditions, respectively.

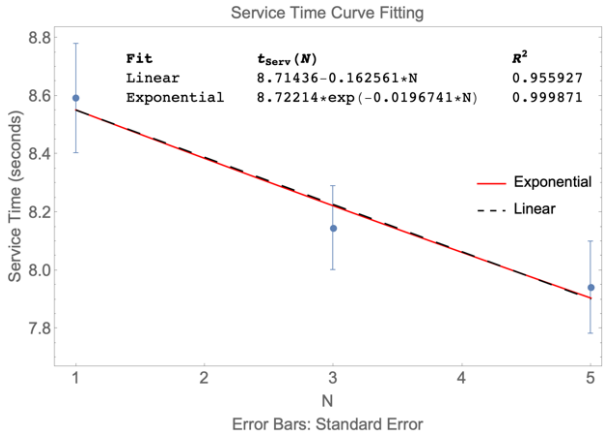

Fig. 2. Multi-UAS DAA encounter ATC call service times for various number of managed aircraft. Linear and exponential regression lines are provided.

As was observed for the radio call service times in Table 3, the 2018 DAA study data reveal a similar trend of decreasing service times for increasing values of N. And since the 2018 study had a three-level manipulation for N, the goodness-of-fit (R^2) of both linear and exponential least squares service time approximations were tested to decide the best approach for interpolating $E[s]$ for $N \in \{5, \ldots, 11\}$ in the UAM Comms context. The calculated R^2 value for the exponential approximation was *slightly* higher than that for the linear approximation (0.999871 vs 0.955927), although it should be noted that the two approximations nearly overlap entirely (Fig. 2). As a result, an exponential least squares approximation was selected as the method to interpolate the intermediate values as $E[s] \approx b_0 \cdot e^{b_1 N}$ with $b_0 = 15.117$ sec and $b_1 = -7.66114 \times 10^{-3}$. The full range of values for $E[s]$ used in the Monte Carlo simulations are shown in Table 4.

Table 4. Values of $E[s]$ for $N \in \{4, ..., 12\}$

N	$E[s]$, sec	Measured or Interpolated Value?
4	14.661	Measured
5	14.549	Interpolated
6	14.438	Interpolated
7	14.328	Interpolated
8	14.218	Interpolated
9	14.110	Interpolated
10	14.002	Interpolated
11	13.895	Interpolated
12	13.789	Measured

C. Monte Carlo Method and Parameter Estimation for Queuing Model

For each $N \in \{4, \ldots, 12\}$, a Monte Carlo experiment consisting of 1,000 simulations was used to estimate the arrival rate (λ) used in the M/M/1 model. This experiment simulated the complete missions (i.e., full vertiport-to-vertiport flights) of aircraft within in the UAM Communications Modality Experiment's assumed airspace and ConOps. For each simulation, random initial locations and directions of travel (northbound or southbound) for the N aircraft are generated. The length of a simulation is 1,800 sec = 30 mins, which was the trial length of scenarios in the HITL. Time is advanced in 1-second steps, and after each time increment a check is performed to determine if any aircraft have passed a checkpoint indicating the "arrival" of a call for which communication with ATC is necessary. If one or more calls have arrived, they are added to the queue on a first-in-first-out (FIFO) basis for servicing. (If the queue length is zero at time of arrival, the first call immediately goes into service and any new arrivals are serviced serially, FIFO.) To ensure a constant value of N active aircraft in the simulation, aircraft immediately add a departure request to the end of the queue upon landing at a vertiport and takeoff in the reverse direction when cleared for departure. When a call is completed, and there is a non-zero queue, the next call goes into service immediately (i.e., there is zero switching time moving between back-to-back calls[‡]). All calls to ATC use the fixed value of $E[s]$ for the corresponding N (see Table 4). The assumption is made that the RPIC can only service one call at a time, and that each call is serviced from beginning to end without interruption.

At the completion of the 1,800 time steps, the task arrival rate, λ, is calculated simply as:

$$\lambda = \frac{(total\ number\ of\ call\ arrivals)}{1800\ \text{sec}}. \tag{14}$$

Similarly, the task completion rate, μ, is calculated as:

$$\mu = \frac{(total\ number\ of\ calls\ completed)}{1800\ \text{sec}}. \tag{15}$$

The values of λ and μ are then used to calculate queueing statistics ρ, L, L_q, W, W_q, and p_0 using the equations presented in Section IV-A. In addition, each Monte Carlo run tracks the evolution of the queue length over the course of the run

[‡] Certainly, this is an unrealistic assumption for a human RPIC. A modification could be made, if desired, to add some small switching time ε between back-to-back calls.

and produces statistics for L_{max}, the maximum length of the queue during the run, and $t_{CS,max}$, the longest continuous stretch of time spent actively in service without a break.

Following the completion of the 1,000 simulations, averages of the queueing statistics across the 1,000 simulations are produced, the value of N is incremented, $E[s]$ is updated to the value for the new N, and the next set of 1,000 simulations are run until queueing averages for all $N \in \{4, \ldots, 12\}$ have been obtained.

Fig. 3 provides a visual representation of the task arrivals and executions for two selected example runs of the Monte Carlo simulation—a 4-vehicle run (Fig. 3a) and a 12-vehicle run (Fig. 3b). Time moves horizontally, increasing from left to right. The labels along the left vertical edge of the diagram represent the N aircraft. Solid, horizontal line segments in the figure represent a period of time in service of a call, where the length of the segments is $E[s]$. Dotted trails leading up to solid lines indicate periods of time in which a call is waiting in the queue. Vertical arrows represent transitioning from a completed call to the next call in the queue.

Uninterrupted "daisy chains" of calls represent an extended period of uninterrupted service, i.e., an expanse of time without a break. In the selected $N = 4$ run (Fig. 3a), the longest stretch without a break is an easily manageable $t_{CS,max} = 59$ sec. This uninterrupted period begins at $t = 13{:}14$, when a call is made to SJC on behalf of ac4. Seven seconds later, at $t = 13{:}21$, ac2 is due to contact SFO and is first in the queue at this point in the run. (Note the short dotted blue trail up until $t = 13{:}29$ when the call goes into service.) At the same time the RPIC begins the call to SFO for ac2 ($t = 13{:}29$), ac1 is due to contact PAO and is added to the queue. Three seconds later, at $t = 13{:}32$, another call to PAO is due, this time on behalf of ac3. This back-to-back chain of calls terminates at the completion of ac3's call to PAO at $t = 14{:}13$.

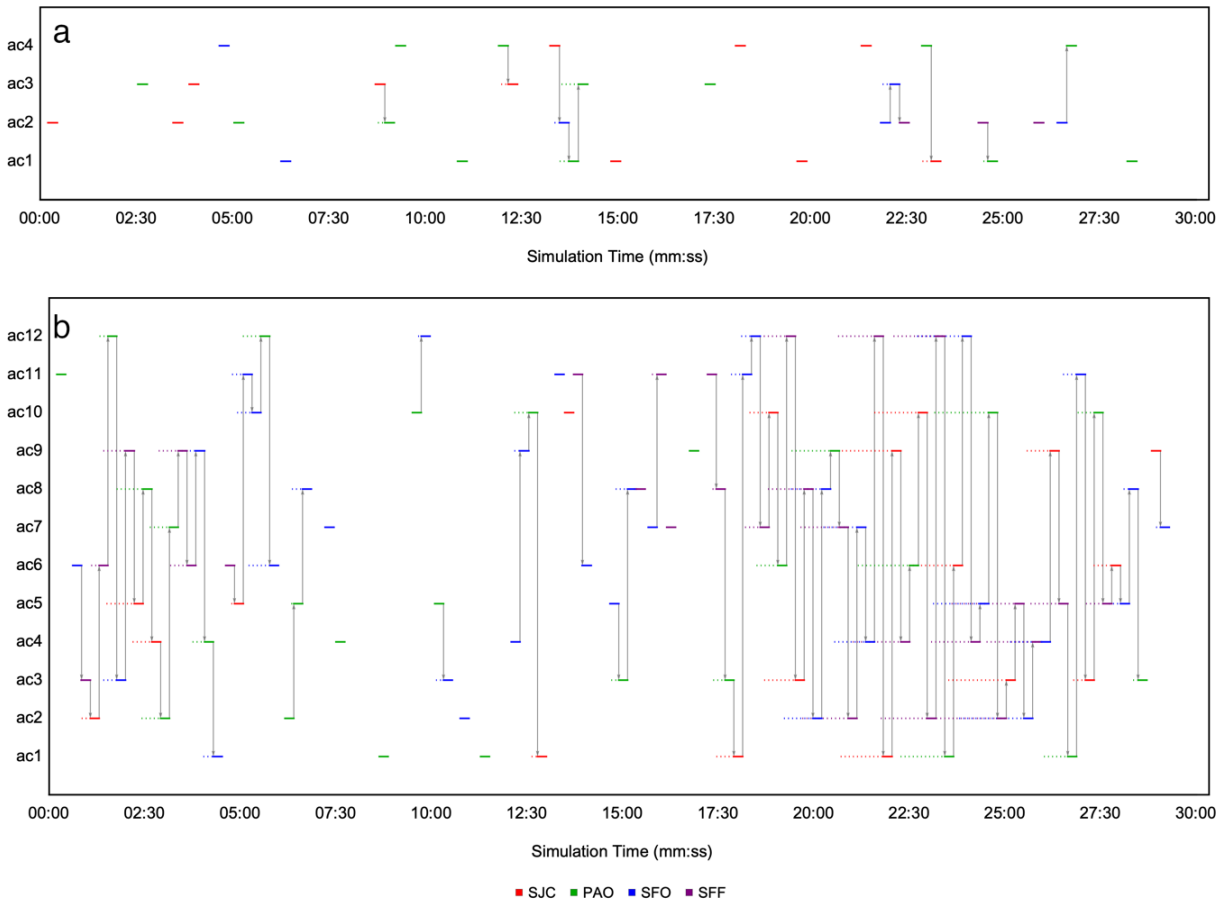

Fig. 3. Visualization of call arrivals and service for example runs of the Monte Carlo simulation: a) N = 4 aircraft and b) N = 12 aircraft. The 30-minute simulation time runs from left to right. Vertical labels represent different aircraft under the RPIC's responsibility. Solid lines within the rows represent time intervals in which calls to the appropriate ATC authority (SJC, Red; PAO, Green; SFO, Blue; SFF, Purple) were actively in service. The length of the solid line is the average call service time, as measured in Section IV-B. A dotted trail indicates a time interval in which a call had to wait in the queue for service. When a queue has formed, vertical arrows indicate transitions between back-to-back calls.

The situation for the RPIC of Fig. 3b is a bit more dire. In that case, the longest uninterrupted stretch $t_{CS,max}$ = 690 sec (11 min 30 sec), beginning at t = 17:13 when ac11 calls SFF and ending at t = 28:43 when ac3 completes its call to PAO. Long waiting times are visible in the many dotted trails during this period. That stretch could perhaps be manageable when rare; but as is evident in Fig. 2b, most of the scenario is spent working through one queue after the next. The RPIC of Fig. 3a has a utilization is ρ = 0.269, compared to ρ = 0.797 for the likely fatigued RPIC of Fig. 3b.

V. Results

A. Queueing Statistics

Mean values of λ, ρ, p_0, W, W_q, L_q, and L_{max} were computed by averaging across the 1,000 Monte Carlo runs performed for each N. Table 5 below shows how these parameters vary, on average, with N. Least-squares linear and exponential regressions were performed to see how these parameters scale for increasing N. The regression tests revealed linear relationships between N and λ (R^2 = 0.99954), ρ (R^2 = 0.999691), and p_0 (R^2 = 0.999691); exponential relationships were observed between N and W (R^2 = 0.995572), W_q (R^2 = 0.998941), L_q (R^2 = 0.999648), and L_{max} (R^2 = 0.999828).

Table 5. Monte Carlo Simulation Results: Averaged Queueing Statistics for $N \in \{4, ..., 12\}$

		λ, sec^{-1}		ρ		p_0		W, sec		W_q, sec		L_q		L_{max}	
N	$E[s]$	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
4	14.661	0.019	0.001	0.284	0.013	0.716	0.013	20.473	0.382	5.812	0.382	0.113	0.013	1.412	0.533
5	14.549	0.024	0.001	0.352	0.015	0.648	0.015	22.462	0.527	7.914	0.527	0.192	0.021	1.795	0.618
6	14.438	0.029	0.001	0.419	0.016	0.581	0.016	24.859	0.676	10.421	0.676	0.303	0.031	2.107	0.626
7	14.328	0.034	0.001	0.484	0.017	0.516	0.017	27.810	0.911	13.482	0.911	0.457	0.047	2.461	0.706
8	14.218	0.039	0.001	0.551	0.019	0.449	0.019	31.707	1.325	17.489	1.325	0.679	0.074	2.868	0.789
9	14.110	0.044	0.001	0.615	0.019	0.385	0.019	36.710	1.844	22.6	1.844	0.987	0.112	3.349	0.910
10	14.002	0.048	0.001	0.677	0.020	0.323	0.02	43.534	2.712	29.531	2.712	1.432	0.174	3.863	1.050
11	13.895	0.053	0.001	0.739	0.021	0.261	0.021	53.557	4.329	39.662	4.329	2.115	0.291	4.505	1.300
12	13.789	0.058	0.002	0.801	0.021	0.199	0.021	70.054	7.766	56.265	7.766	3.279	0.543	5.292	1.596

We were interested to see what relationship exists between utilization, ρ, and maximum queue length, L_{max}. Naively, these parameters together show "how busy" (ρ) and "how backed up" with tasks (L_{max}) the RPIC was during the simulations. Regression tests showed that the mean L_{max} grows exponentially with increasing ρ (Fig. 4).

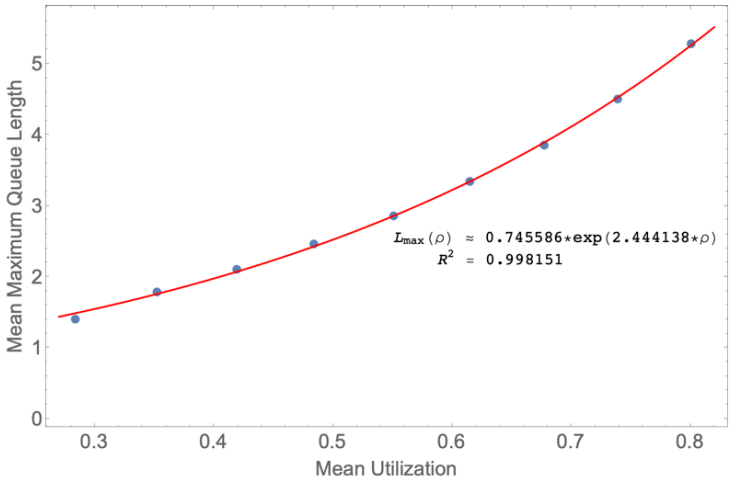

Fig. 4. Exponential relationship between the means of L_{max} and ρ.

B. Queue Length

While Fig. 4 shows an exponential relationship between the *mean* utilization and *mean* maximum queue length, the story of these two parameters is more nuanced than it may appear. To highlight this nuance, Fig. 5 shows plots of the queue length over time for 9 different runs of the Monte Carlo simulation arranged into a 3-by-3 grid. The charts in Fig. 5 show selected runs for $N = 4$, $N = 8$, and $N = 12$, arranged into columns. The runs selected for each N were those corresponding to the minimum, median, and maximum utilization values, which form the rows of Fig 5. In cases where the minimum, median, or maximum utilization was "tied" for multiple runs of a given N, a random selection of tied-utilization runs was made from those with the median value of L_{max}.

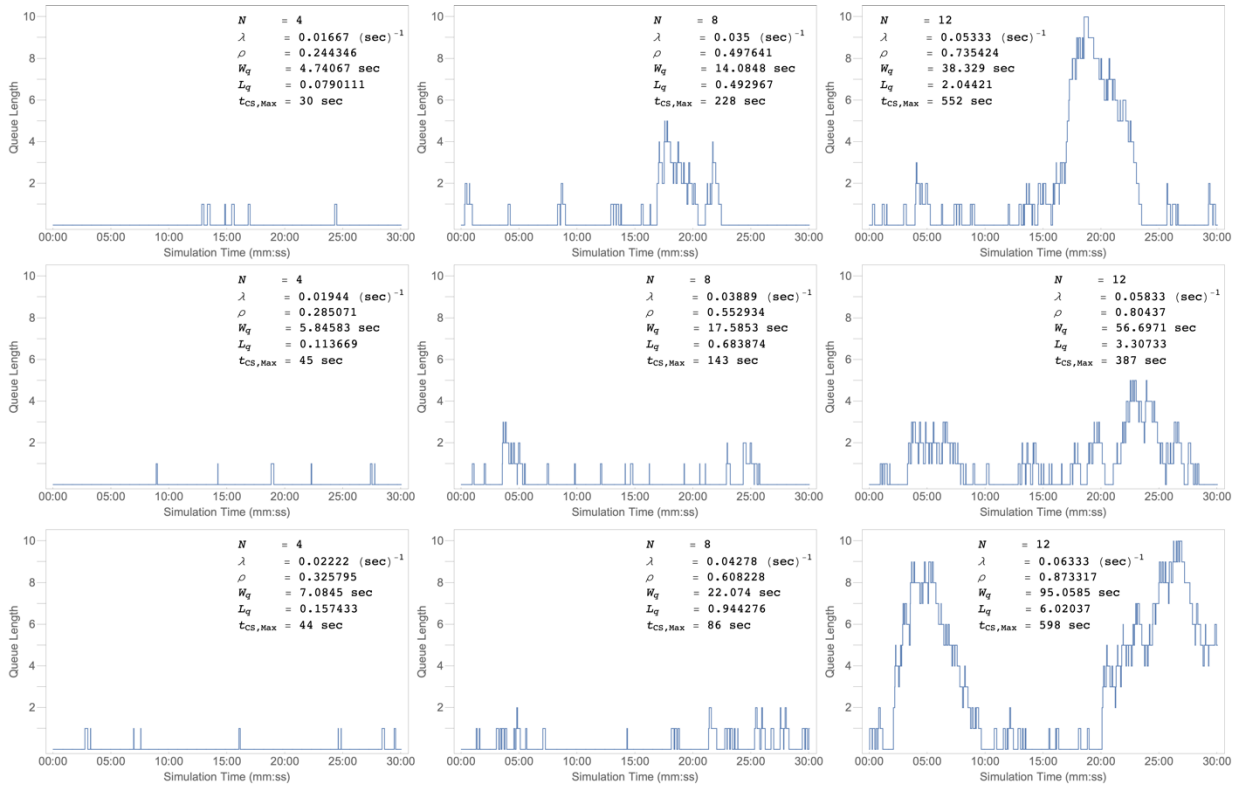

Fig. 5. Three-by-three grid of plots showing the queue length over time for selected runs of the Monte Carlo simulation. Runs are arranged left-to-right by N (4, 8, and 12) and top-to-bottom by values of ρ (minimum, median, and maximum).

When utilization is low (top row), the RPIC has more down time overall. On its own, a low utilization value might suggest that the RPIC was under-tasked. It would be a mistake, however, to conclude that a lower utilization value implies a smaller overall number of tasks. Given the relatively small variance of λ (Table 5) for each given N, lowering or raising the value of ρ changes the *density* of tasks, as opposed to the number of tasks. The top row of Fig. 5 (minimum ρ) show dense periods of activity in which the size of the queue grows relatively large in "bottle-necks" surrounded by relatively long stretches of inactivity. At the other extreme, for maximum ρ (bottom row), the RPIC spends an overall larger amount of time servicing calls and queues grow, though not to the same length as when ρ is minimized. In the cases of the median values of ρ (middle row), tasks are spread out more evenly and queue lengths tend to be shorter than when ρ is minimized or maximized.

VI. Discussion

The results shown in Figs. 4 and 5 indicate that, on average, system utilization increases as the number of aircraft managed increases. In addition, as the utilization increases, tasks become more spread out, the average queue length is longer, tasks wait longer to be completed, and neglect time decreases. As a result, RPICs are more likely to be task

saturated for longer periods of time, with less down time in between task groupings. Table 5 shows that as the number of aircraft managed increases, the probability of a zero-queue length decreases linearly. When $N \geq 8$, tasks are more likely to have to wait to be completed than receive immediate service (i.e., $p_0 < 0.5$). Lastly, when $6 \leq N \leq 10$, the utilization falls within the range for expected effective human performance as outlined in [6-8].

Table 5 and Fig. 5 show that queue length increases as the number of vehicles managed increases. However, while the average queue length, L_q, increases with utilization as the number of vehicles managed increases (Table 5), if the number of vehicles managed remains the same, the maximum queue length is not necessarily proportional to the utilization (Fig. 5). Instead, tasks are more clustered together within lower utilization missions, often resulting in larger maximum queue lengths. Maximum queue lengths dip when the utilization is at its median value compared to when it is at its minimum or maximum.

VII. Conclusion and Future Work

This paper presented an M/M/1 queuing system model for analyzing the performance of RPICs during m:N operations. The impact of aircraft loads between $N = \{4, \ldots, 12\}$ on queue length, task wait times, and utilization were explored. Results show that as the number of aircraft managed increases, the service time decreases and the overall utilization increases. This leads to more spread-out tasks, with less time between groupings of tasks. In addition, the results indicate that queue length is not proportional to utilization.

The results provided in this paper assumed zero time between a task ending a new task starting, which represents the best-case scenario. A more realistic representation would include an additional buffer time between tasks to represent the time taken by an RPIC to cognitively shift between tasks. Data from human-in-the-loops studies should be utilized to determine a realistic average for such a buffer time and additional analysis should be performed to evaluate its impact on utilization, queue lengths, wait times, and overall system performance.

Future studies should also be conducted to further explore the relationship between actual human performance and utilization. While previous studies indicate an acceptable range for utilization, it is not clear the previously determined relationship holds for *m*:*N* applications as compared to traditional multi-vehicle supervisory control scenarios. In reality, it may turn out that the ideal human performance range for utilization is highly application specific. Additional factors such as the maximum continuous service time may contribute to mental workload, which may lead to more errors, ultimately reducing the effectiveness of an RPIC regardless even if the explicit utilization of the RPIC is high.

Acknowledgements

The authors would like to acknowledge Rene Mai of Rensselaer Polytechnic Institute for her contributions to the applicability of queuing systems to *m:N* applications. The authors additionally thank Krish Pradhan of San Jose State University, Melat Dereje of California State University, Long Beach, and Alisa Braun of the Human Systems Integration Division at NASA Ames Research Center for their help in coding data utilized for analysis in this paper.

References

[1] Federal Aviation Administration, "Urban Air Mobility: Concepts of Operations v2.0," 2023. URL https://www.faa.gov/sites/faa.gov/files/Urban%20Air%20Mobility%20%28UAM%29%20Concept%20of%20Operations%202.0_0.pdf.

[2] Sadler, G., Chandarana, M., Rorie, R. C., Tyson, T. L., Keeler, J. N., Smith, C. L., Shyr, M. C., Wong, D., Scheff, S., and Dolgov, I., "A Remote, Human-in-the-Loop Evaluation of a Multiple-Drone Delivery Operation," AIAA Aviation 2022 Forum, 2022, p. 4002.

[3] Federal Aviation Administration, "Package Delivery by Drone (Part 135)," 2023. URL https://www.faa.gov/uas/advanced_operations/package_delivery_drone

[4] Grier, R., Wickens, C., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., and St. John, M., "The Red-Line of Workload: Theory, Research, and Design," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 52, Sage Publications Sage CA: Los Angeles, CA, 2008, pp. 1204–1208.

[5] Endsley, M. R., "Measurement of Situation Awareness in Dynamic Systems," Human Factors, Vol. 37, No. 1, 1995, pp. 65–84.

[6] Hart, S., "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," Human Mental Workload, Elsevier, 1988.

[7] Cummings, M. L., and Guerlain, S., "Developing Operator Capacity Estimates for Supervisory Control of Autonomous Vehicles," Human Factors, Vol. 49, No. 1, 2007, pp. 1–15

[8] Rouse, W. B., Systems Engineering Models of Human-Machine Interaction, Vol. 6, North Holland New York, 1980.

[9] Schmidt, D. K., "Queuing Analysis of the Air Traffic Controller's Workload," Tech. rep., 1978.

[10] Shone, R., Glazebrook, K., and Zografos, K. G., "Applications of Stochastic Modeling in Air Traffic Management: Methods, Challenges and Opportunities for Solving Air Traffic Problems under Uncertainty," European Journal of Operational Research, Vol. 292, No. 1, 2021, pp. 1–26.

[11] Kim, J., Tandale, M., and Menon, P., "Air-Traffic Uncertainty Models for Queuing Analysis," 9th AIAA Aviation Technology, Integration, and Operations Conference (ATIO) and Aircraft Noise and Emissions Reduction Symposium (ANERS), 2009, p. 705.

[12] Sadler, G., Chandarana, M., and Keeler, J., "Effects of Communication Modality on Pilot-Controller Coordination during a Simulated m:N Operation," AIAA Aviation 2023 Forum, 2023, p. 3265.

[13] Price, G., Helton, D., Jenkins, K., Kvicala, M., Parker, S., and Wolfe, R., "Urban Air Mobility Operational Concept (OpsCon) Passenger-Carrying Operations," NASA CR-2020-5001587, 2020. URL https://ntrs.nasa.gov/api/citations/20205001587/downloads/UAM%20Passenger-carrying%20OpsCon%20-%20v14%20GP%20accept.pdf.

[14] Federal Aviation Administration, "Advanced Air Mobility | Air Taxis," URL: https://www.faa.gov/air-taxis

[15] Shortle, J. F., Thompson, J. M., Gross, D., and Harris, C. M., Fundamentals of Queueing Theory, Vol. 399, John Wiley & Sons, 2018.

[16] 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.

[17] Federal Aviation Administration (FAA). Integration of Civil UAS in the NAS Roadmap, First Edition. FAA, Washington, D.C., 2017.

[18] RTCA, "DO-365 – Minimum Operational Performance Standards (MOPS) for Detect and Avoid (DAA) Systems," RTCA Inc., Washington, D.C., 2017.