Enhanced Stochastic Optimization Model (ESOM) for Setting Flow Rates in Collaborative Trajectory Options Programs (CTOP)

Robert Hoffman* and Bert Hackney†
Metron Aviation, Dulles, Va., 20166, USA

Peng Wei‡ and Guodong Zhu§
Iowa State University, Ames, Iowa, 50011, USA

We present a stochastic optimization model for setting flow rates during a new type of traffic management initiative known as a Collaborative Trajectory Options Program (CTOP). CTOP allows traffic managers to restrict traffic flow through a network of flow constrained areas (FCAs). FCAs can be lines or polygonal regions of airspace that traffic managers may have tailored to the traffic flow situation at hand. As of yet, there is no guidance for traffic managers to set the maximum flow rates (flights per unit time period) at each of the FCAs. This is compounded by the stochasticity of the problem: the FCA capacities are usually a function of weather, which is not known in advance. Traffic demand levels at the FCAs are also stochastic, since they depend on the routing and delay tradeoff preferences that airlines submit during the CTOP process. The optimization model we present provides to the traffic managers time-varying flow rates at the FCAs that minimize total expected delay costs, taking into account forecasted traffic demand, airline routing preferences, and forecasted probabilistic weather. Because the model is aggregate, it allows the resource allocation algorithm in CTOP to make final flight-to-route and flight delay assignments. In this respect, the model is highly consistent with CTOP functionality and the collaborative decision making (CDM) paradigm in traffic flow management. In this paper, we explore important characteristics of the model, such as hedging across a range of possible weather outcomes and convergence with airline routing preferences. We demonstrate its use on a realistic air traffic scenario.

I. Nomenclature

$\mathcal{F}$ = set of FCAs (uncapacitated, but flights may be ground-held)
$\mathcal{P}$ = set of PCAs (capacitated)
$\mathcal{R}$ = set of resources = $\mathcal{F} \cup \mathcal{P}$
$\mathcal{C}$ = set of ordered pairs of connected resources: $(r, r') \in \mathcal{C}$ if and only if r is connected to r' in the directed graph
$\mathcal{T}$ = set of time intervals, t=1,...,T
$\Delta^{r,r'}$ = number of time periods to travel from resource r to r', defined for all pairs $(r, r') \in \mathcal{C}$
D_t^r = demand (number of flights predicted to arrive) at resource r in time interval t
$\mathcal{S}$ = set of capacity scenarios
p_s = probability of scenario s occurring
$M_{t,s}^r$ = maximum capacity of resource $r \in \mathcal{P}$ in time interval t under scenario s
c_g = cost per unit time period for holding one aircraft on the ground
c_a = cost per unit time period for holding one aircraft in the air (or by any other tactical method)

* Vice President, Advanced Research and Engineering Services, AIAA Senior Member.
† Senior Analyst, Advanced Research and Engineering Services, AIAA Member.
‡ Assistant Professor, Aerospace Engineering Department, AIAA Senior Member.
§ Graduate Research Assistant, Aerospace Engineering Department, AIAA Student Member.

P_t^r = number of flights planned to arrive at resource $r \in \mathcal{F}$ during time interval t
$L_{t,s}^r$ = number of flights that are accepted to ("land at") resource $r \in \mathcal{P}$ in time interval t under scenario s
$A_{t,s}^r$ = number of flights held over in the air at $r \in \mathcal{P}$ from time interval t to t+1 under scenario s
G_t^r = number of flights whose arrival time at $r \in \mathcal{F}$ is adjusted from time interval t to time interval t+1 (or later) using a ground delay at their point of origin
$f_t^{r,r'}$ = fraction of flights from resource r directed to resource r' in interval t

II. Introduction

The Collaborative Trajectory Options Program (CTOP) is an air traffic management initiative that allows FAA traffic managers to curtail the flow of traffic through airspace or airport resources. A key feature of CTOP is that it admits control of traffic through multiple flow constrained areas (FCAs). This allows traffic to be coordinated over a larger portion of the airspace system, rather than running parallel, independent traffic management initiatives that may conflict with each other.

Under CTOP, the traffic manager determines the location, geometry, temporal scope, and traffic filters to place on each of the FCAs. The geometry of an FCA could be a polygonal region anywhere in the airspace, but it is usually a line segment or piecewise line segment with an altitude range. These are often placed at the boundary between two air route traffic control centers (ARTCCs), as this provides a convenient way for traffic managers to keep flights grouped and managed at the facility level. Flows into or out of airports can be addressed by establishing FCA rings around departure or arrival fixes or around the entire airport. The concept is broad enough that CTOP has the potential to subsume the functionality of its predecessor traffic management initiatives, the ground delay program (GDP) and the airspace flow program (AFP).

When setting up a CTOP, the planning time horizon is divided into discrete time intervals (e.g., 15 minutes each), and the traffic manager must set a planned acceptance rate (PAR) for each future time period for each FCA. This is the maximum number of flights that should be allowed to enter the FCA during that time period. When the CTOP resource allocation algorithm is run, it divides each time period evenly into 'arrival' slots, and flights are assigned to the slots. In time periods where demand exceeds the number of available slots, ground delay is assigned to flights to keep demand below the PAR level. Flight operators can submit rerouting options to the resource allocation algorithm as an alternative to the ground delays.

In this paper, we provide optimization-based decision support models for the FCA rate-setting problem—that is, how to set PARs on the FCAs in a CTOP. Related topics, which will be discussed in future or concurrent work, are where to locate the FCAs, how many FCAs to create, where to locate them, and how to set the traffic filters. For now, we assume that the FCAs are already established and that the traffic manager just needs to know how to set the PARs. In the event that there is only one FCA and that it coincides with a physical resource, then the PAR in each time period should be set the same as the forecasted capacity of the resource. For instance, if the FCA is a ring around an airport with constrained arrival rates, then the PARs should be set to the forecasted arrival rates. In this case, the CTOP is effectively reduced to a GDP. But even this rate-setting problem can be quite challenging. Usually, the FCAs are implemented in anticipation of adverse weather. Since weather is highly stochastic, the FCA rate-setting is stochastic as well. For a single FCA (an airport GDP, to be more precise), prior models have been proposed in the literature, and we will show in this paper how to extend them to the case of multiple FCAs. Our proposed optimization model provides a way to plan FCA rates in the face of airspace or airport capacity uncertainty. Section III describes a geometric, network flow representation of the model, while section IV presents the algebraic formulation.

The other complication in setting FCA rates under a CTOP is that, in addition to capacity, traffic demand can be stochastic as well. Earlier optimization models that were developed for a GDP setting (i.e., controlling flow into a single arrival airport) generally assume that demand can be estimated well in advance. This holds for GDPs, since the vast majority of aircraft bound for the GDP airport are scheduled commercial flights, already committed to their destination airports. But in the CTOP setting, the FCAs could be lines or polygons in space designed to curtail the flow of traffic to downstream constrained resources. Flight operators are interested in passing through them only to the degree that they lie along an expedient path. The routing options that flight operators submit to CTOP (via trajectory options sets) specify their preferences for rerouting options in or around the FCAs. This means that demand to the FCAs is highly volatile, and subject to change in reaction to FCA rates set by traffic managers. The rates are, in turn, dependent upon demand estimates. In section VI, we will show how to break this cycle of dependency by applying our FCA rate-setting model in an iterative loop with flight operator rerouting options.

In section V, we present some behavioral characteristics of the optimization model, while in section VII, we present experimental results for a realistic CTOP application. Section VIII provides high-level conclusions from our work, ramifications for CTOP, and suggested future research.

III. Model Description

Our approach to providing decision support for rate setting under capacity uncertainty is based on linear integer programming combined with ensemble-based weather forecasting. The model we present in this paper is a generalization of a static stochastic model (SSM) developed by Ball et al. [1] for setting PARs at a single constrained resource (a GDP airport) in the face of uncertain airport capacity with known demand. To accommodate CTOP, we adapted SSM to take into account an arbitrary number of resources.

A key aspect of our approach is that we have allowed for the possibility that the FCA, which is a flow control mechanism, may not be co-located with the physically constrained resources of concern. This is entirely consistent with modern traffic management practices. For instance, Fig. 1 shows three line FCAs applied by FAA traffic managers on May 28, 2010. Their objective was to reduce the flow of traffic into the northeastern United States, which was being affected by convective weather. Note that the FCAs fall along the boundaries between ARTCCs. An advantage of placing the FCAs upstream of the constrained resources is that it alleviates the need to accurately predict, hours in advance, exactly where the constraints will occur, which is often difficult. The FCAs can be placed in highly strategic locations that can persist throughout the life of the CTOP.

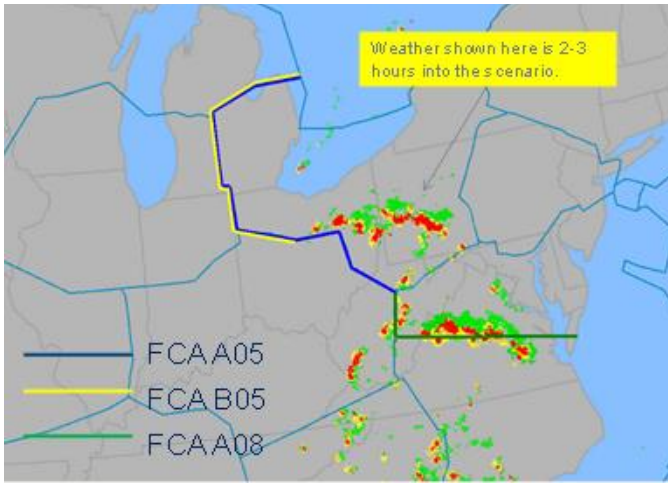

Fig. 1 Three FCAs applied on May 28, 2010, to curtail flow into New York ARTCC

In addition to the FCAs, we propose that traffic managers identify potentially constrained areas (PCAs), which more directly coincide with physically constrained resources, such as sectors of airspace and airport arrival or departure fixes. This allows us to directly model demand and capacity at the resources of true concern. (Under today's CTOP paradigm, the PCAs are left unstated.) The FCAs would generally be placed upstream of the PCAs. We still allow for the possibility that the FCA lies atop an actual constraint; this is just the special case in which the FCA is the same as the PCA.

So, we assume that the traffic manager has identified a set of PCAs of concern, and has created a set of FCAs to control the flow into the PCAs. (We are also researching decision support tools to assist with this.) As part of our decision support modeling process, we construct directed arcs between FCAs and/or PCAs to represent the flow of traffic among these objects. Collectively, this creates a network topology.

The problem we are trying to solve is how to set PARs on the FCAs to efficiently control the flow of traffic through the PCA network. Only the FCAs would have PARs, because these are the flow control mechanisms, and only the PCAs would have realized capacity, because these capture actual airspace system constraints. To capture future capacity uncertainty, our model assumes that for each PCA, alternate future capacity realizations are represented by a finite set of scenarios, each with an associated probability. This technique has been applied by other air traffic management researchers, such as Vranas [2–4], Richetta and Odoni [5], and Ball et al. [1]. More specifically, we divide the future time horizon into a series of time intervals, t = 1, 2, … T. For each PCA, a given capacity scenario provides a capacity value for each of the future time intervals. An example is shown in Fig. 2 for a single PCA. For instance, if the blue scenario occurs, then the capacity of the resource would be 55 flights in the first time interval, 30 flights in the second time interval, and so on. We consider the scenarios mutually exclusive, and their probabilities must sum to 1. The scenarios would apply across all the PCAs. That is, for a second PCA, there would also be a blue scenario, but the PCA would have its own capacity profile. If the blue scenario were to occur, then the corresponding blue capacity profiles would be realized at each of the PCAs.

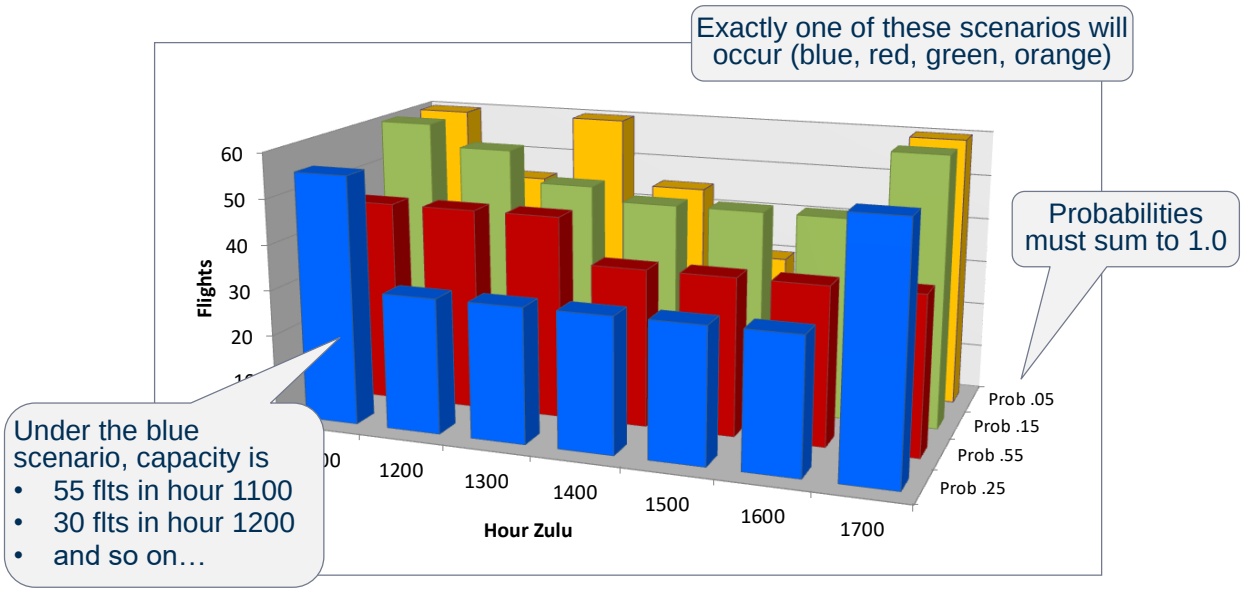

Fig. 2 Example of four capacity scenarios for a PCA

We assume that in each time period, traffic demand to each PCA and FCA can be accurately forecasted. Another key assumption is that by mapping forecasted aircraft trajectories across our FCA-PCA network, we can compute the percent of flow out of a given FCA or PCA into all downstream FCAs or PCAs that it feeds. For instance, a given FCA may have 20% of its outbound traffic flow into PCA1 and 80% flow into PCA2. Our model assumes that these traffic 'splits' must be maintained even after the PARs are set. That is, the outbound flow from the FCA could be cut, say, in half, but the 20/80 split will still hold. (In section VI, we will show how to address demand uncertainty and changes in splits due to rerouting effects of CTOP.)

The optimization model we introduce is the Enhanced Stochastic Optimization Model (ESOM). Like its predecessor SSM, ESOM balances planned ground delay against expected airborne holding costs over an ensemble of possible future capacity profiles. ESOM is an "aggregate" model in the sense that it advises how many flights to allow through each FCA in each time period, regardless of which flights they might be. This makes the approach compatible with the CTOP resource allocation algorithm that actually assigns flights to routes and ground delay amounts. It is also compatible with collaborative decision making (CDM) practices such as airline cancellations and substitutions.

The inputs to ESOM are as follows:

- A set of future time periods (e.g., 15-minute time buckets over a four-hour planning horizon)
- A network topology of FCAs and PCAs (how they are connected)
- Approximate split of demand over resources in the network topology
- For each PCA, the capacity profile (maximum number of flights in each future time period) under each probabilistic forecast

The model outputs are as follows:

- Number of flights that should be allowed to enter each FCA in each time period (the FCA rates)
- Number of flights held in a state of ground holding in each future time period
- For each probabilistic forecast, the number of flights that will be held in a state of airborne holding in each future time period
- Expected combined cost of ground holding and airborne holding

In section IV, we provide the mathematical formulation for ESOM. Since it is motivated by a network flow model, we believe it is best to first present the graphical formulation, as this provides the intuition for the model. For ease of exposition, we start with a special case in which there is only one FCA, which feeds only one PCA, with no travel time between them, and we assume only two capacity scenarios at the PCA, although in practice there could be any number of capacity scenarios. Figure 3 shows a flow diagram that represents the aggregate flow of traffic into the FCA, then into the PCA, within each time period. (This is derived from, but not the same as, the original FCA-PCA network, which in this case would comprise just one FCA node feeding one PCA node.) The demand variables D_t are

exogenous parameters provided to the model as input. For instance, D_1 is the number of flights scheduled or estimated to use the resource in time interval t=1. The horizontal arcs running along the upper (gray) portion of the diagram are ground-holding arcs. For instance, if G_1=2, then two of the flights trying to arrive during time period 1 must be ground-held at their departure airport to postpone their arrival to a later time period. Since this is an aggregate (single-commodity) model, it makes no mention of which flights those should be.

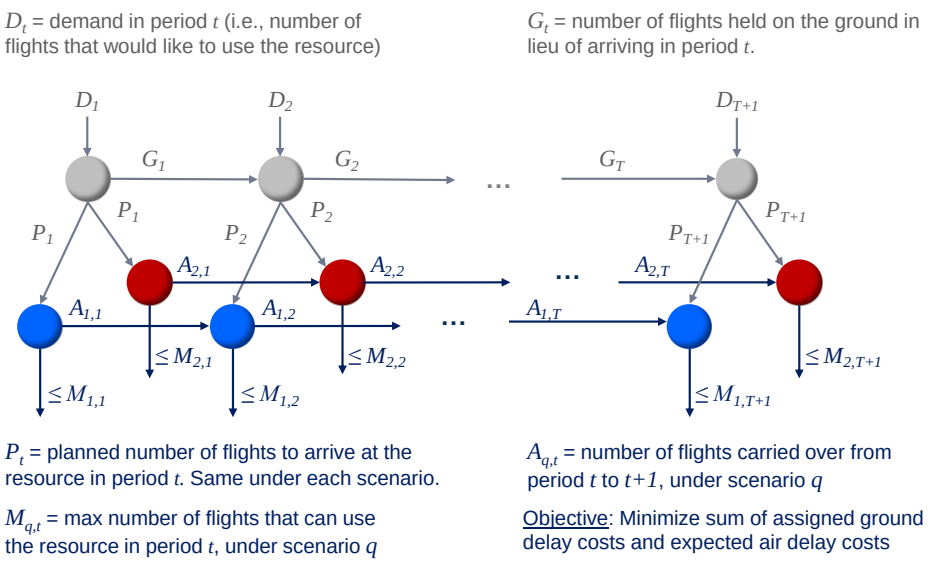

Fig. 3 Network flow design and variable definitions for the SSM

The gray downward arcs indicate the number of flights that we plan to have arrive at the resource in question. For instance, P_1=30 would mean that we plan to have 30 flights arrive; the other flights would have to have been ground-held over the G_1 arc. The actual number of arrivals that can be accommodated is limited by the capacity for the time period (the M variables), which is peculiar to the capacity scenario. In this example there are two such scenarios, one blue and one red. No matter which scenario occurs, we will have sent P_1 flights to the resource in time period 1. Those flights that can be accepted to the resource flow out the bottom of the diagram. But if we send too many flights to the resource under a given scenario, then excess flights will be held in the air from one time period to the next along the A arcs. For modeling purposes, one can think of these excess flights as being held in the air at or near the resource, but in practice, traffic managers can implement this tactical delay anywhere along the flight path using techniques such as vectoring, speed reduction, or miles-in-trail restrictions.

The model records the number of flights that would be airborne-held (tactically) in each time period and under each scenario, then each scenario is weighted by its likelihood of occurrence. Therefore, the total amount of airborne holding is computed in expectation. The expected airborne holding is weighted by a marginal airborne-holding cost (dollars per flight per time period). To this we add the total cost of deterministic ground holding, which is the sum of the G variables, multiplied by the marginal cost (per flight, per time period) of ground holding. The optimization model chooses the values of the P variables that minimize this total expected cost.

For the case in which one FCA feeds one PCA, the resulting algebraic formulation is shown below. This is the static stochastic model presented by Ball et al. in 2003 [1]. Even though the model is not strictly a network flow model (because the P arcs are repeated), the constraint matrix is totally unimodular. Therefore, an integer solution can be obtained from the linear programming (LP) relaxation when the right-hand side of the constraint matrix consists of integer values (which it will, because they are numbers of flights).

$$\text{Minimize } \sum_{t=1}^{T} c_g G_t + \sum_{t=1}^{T} \sum_{q=1}^{Q} c_a p_q A_{t,q}$$

subject to:

$$G_t + P_t - D_t - G_{t-1} = 0 \quad \forall t$$

$$-P_t - A_{(t-1),q} + A_{t,q} + M_{t,q} \geq 0 \quad \forall t, \forall q$$

$$G_t, P_t, A_{t,q} \geq 0 \quad \forall t, \forall q$$

$$G_0 = G_{T+1} = A_{0,q} = A_{(T+1),q} = 0$$

The novelty introduced by this paper is to adapt this base model to accommodate multiple FCAs and PCAs so it can be applied in CTOP. To accomplish this, we associate with each FCA one set of flow control (gray) nodes, and we associate with each PCA one set of capacity realization (blue and red) nodes. Figure 4 shows this association for the case in which two FCAs feed three PCAs.

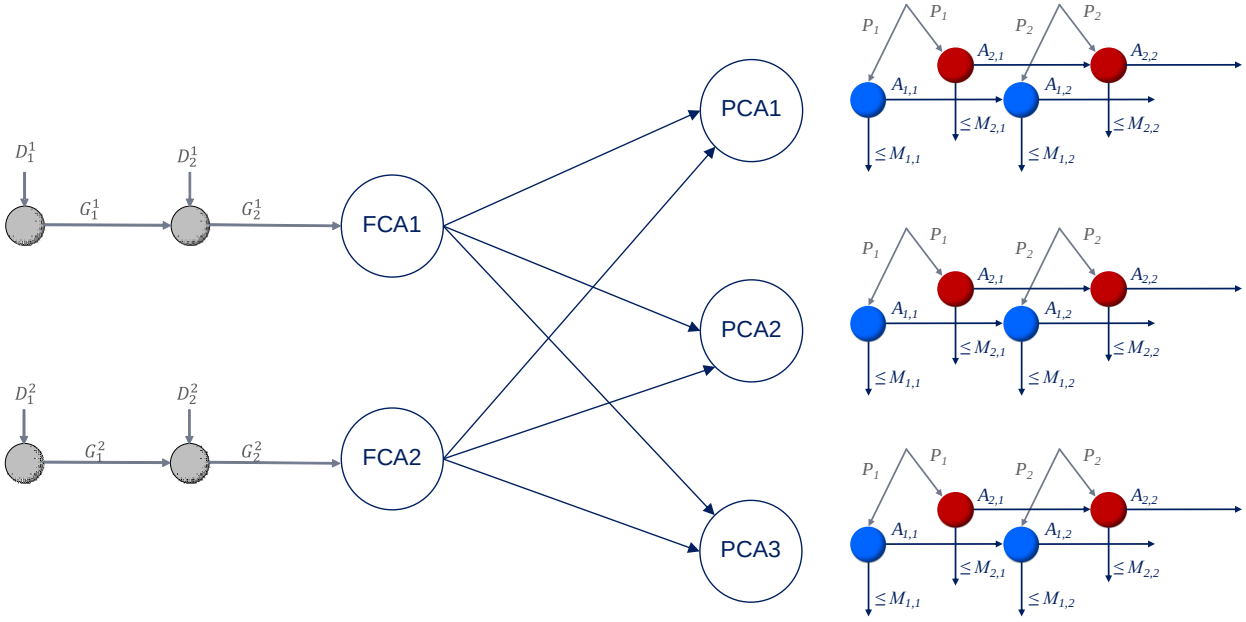

Fig. 4 Ground-holding arcs are associated with FCAs, while air-holding arcs are associated with PCAs

Next, we form arcs within each time period to correctly model the possible flow from FCAs to PCAs, as shown in Fig. 5. For instance, since FCA1 can feed any of the three PCAs, we create arcs in the first time period going from the first FCA to each of the PCAs in the first time period. These are the finely dotted gray lines in the upper left of Fig. 5. When we formulate ESOM, we will insist that the distribution (split) of traffic across the three arcs be fixed. For instance, if the PAR for the first time period at the leftmost FCA is set to 100 flights, and if the traffic splits to the three PCAs are 20%, 30%, and 50%, then 20 flights must be sent in the first time period to the red and to the blue capacity scenarios at PCA1, 30 flights must be sent in the first time period to the red and to the blue capacity scenarios at PCA 2, and 50 flights must be sent in the first time period to the red and to the blue capacity scenarios at PCA 3. We maintain these splits to disallow rerouting flights through the network; the model is only allowed to reduce the overall flow by moving flights to later time periods. Arcs and splits are similarly constructed for the other time periods. The resulting configuration is shown in Fig. 5 for the first three time periods (split values not shown). We model the average transit time from one resource to another by staggering the downward arcs from the FCA (gray) nodes to the PCA (blue/red) nodes. For instance, if it takes two time periods to transit from FCA 1 to PCA 1, then the arcs flowing out of FCA 1 in the first time period would connect to time period 3 in PCA 1.

Using this construction method, we can model an arbitrarily complex FCA-PCA network. A given PCA may feed any number of downstream PCAs; the downward arcs of out of the PCA (blue/red) nodes would feed directly into the downstream PCA nodes. This would represent the case, for instance, in which an airport arrival fix or a sector of airspace feeds an airport. However, we have one restriction on the network design: no resources (PCA or FCA) can be upstream of an FCA. This is not an anomaly of our modeling technique. Rather, it avoids a real-world ambiguity of control. Recall that when flights are delayed by an FCA, the delay occurs at the origin airport (not at the physical location of the FCA). If we were to allow a flight to be delayed by one FCA (at point of origin), then move on to a PCA (and possibly be held there in the air), then move on to another FCA that also dictates ground delay at point of origin, how do we know how much ground delay to apply? The second ground delay could only be applied after we observe the capacity of the interim PCA after the flight departs. But this prevents us from applying the second ground delay at point of origin. And the reason we disallow one FCA from feeding another FCA is that it creates an ambiguity of control: which FCA takes priority? (Note: current CTOP implementation allows traffic managers to have one FCA

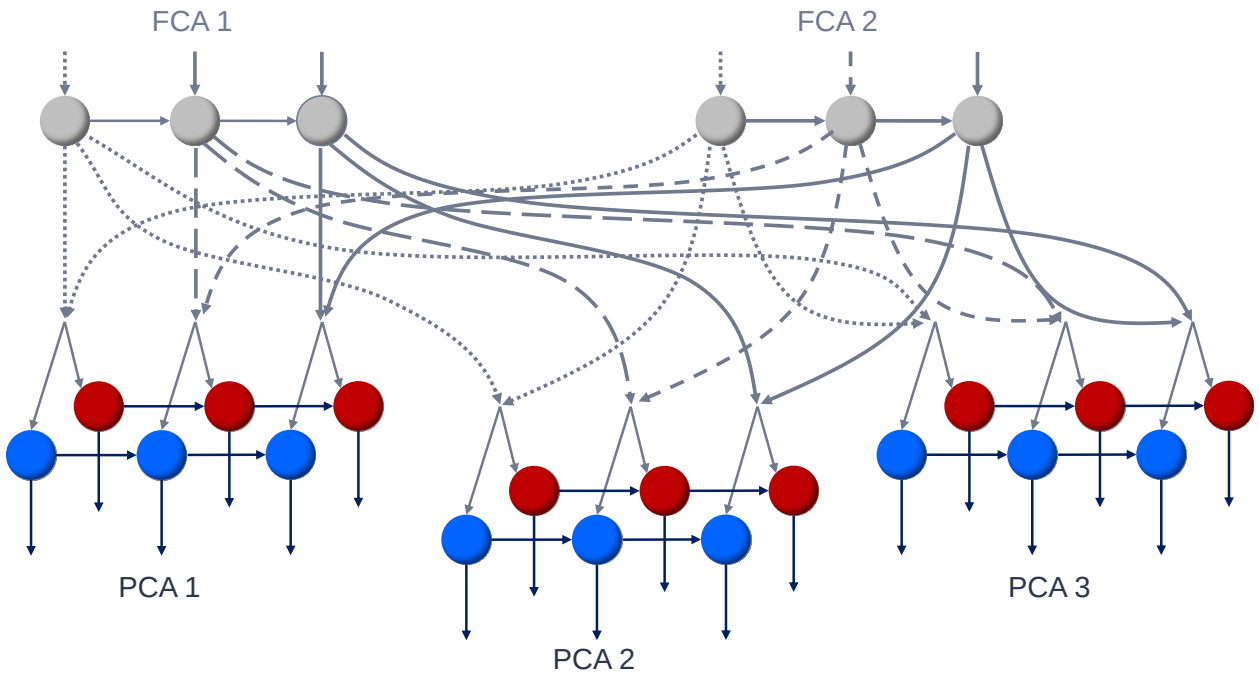

Fig. 5 Connecting FCA nodes to PCA nodes for the case with two FCAs and three PCAs

feed another FCA, but they must specify which one takes priority.) To summarize our network design requirements: the FCAs must be upstream of the PCAs, any PCA can feed another FCA, but no FCA can feed another FCA.

In practice, the traffic managers need not be exposed to the type of network wiring diagram shown in Fig. 5. We have presented this simply to provide the logic and intuition for the algebraic formulation of our optimization model, which we present next.

IV. Algebraic Formulation of ESOM

The mathematical programming formulation of ESOM is derived by enforcing conservation of flow constraints, splitting constraints, and PCA capacity constraints on the type of network flow diagram already presented. We assume we have a pre-defined set of resources, which will remain in effect throughout the life of the CTOP (no new resources are added). The set of resources is the union of the set of FCAs and the set of PCAs. Any two resources may be connected, where "connected" means there is a possibility that traffic will flow from one to the other. (These connections are determined by an a priori analysis of the traffic.) We use the notation shown in section I.

We assume that the planning time horizon has been divided into contiguous time intervals of equal length, and that we know (approximately) how many time intervals it would take for an aircraft to travel from one connected resource to another.

The demand on each resource in each interval has been pre-computed.

For each PCA, a number of future capacity profiles have been forecast, as we described earlier. It is important to keep in mind that when one of the capacity scenarios occurs, the associated capacity profile for every resource will occur (and those capacity profiles can all be different). For instance, if there were 20 time intervals and two resources, r_1 and r_2, and if scenario 5 were to occur, then the capacity profiles for r_1 and r_2 would be $M_{1,5}^{r_1}, M_{2,5}^{r_1}, \cdots, M_{20,5}^{r_1}$ and $M_{1,5}^{r_2}, M_{2,5}^{r_2}, \cdots, M_{20,5}^{r_2}$, respectively.

The primary decision variables are the number of flights planned (by CTOP) to arrive in each time period at each FCA. Those planned FCA arrival rates are exactly what would be fed into CTOP. This is the primary guidance we are giving to the traffic managers as an output of our model.

In principle, the cost parameters could be set by traffic managers in each problem instance of CTOP, but they would likely be set once for all CTOP applications, based on FAA policy. Though their values may be influenced by average industry standards for airline operating costs, we emphasize that these cost parameters reflect overall system

costs from the perspective of traffic managers – specifically, how much they weigh strategic ground holding versus tactical air holding. The ratio of the two cost parameters is a reflection of their attitude toward risk.

The optimization model requires establishment of several auxiliary decision variables:

The landing variables ($L^r_{t,s}$) are not strictly necessary—they can be inferred from the final solution—however, we add them to make the problem formulation and post-analysis more readable. One should not take the notion of landing literally—we just mean they passed into or through the resource without further delay.

We also have the splits of traffic flow proportions coming out of a given resource, which are added as pre-computed parameters.

We already noted that these could be added as free variables for the model to determine, but then the model would be doing its own rerouting, which we reserve for the airlines.

Lastly, we introduce two other pieces of notation, simply to give shorthand to some complex summations (these will make the constraints more intuitive and easier to read):

In the following formulas, to clarify, the sum is over $\{r'|(r',r) \in \mathcal{C}\}$, i.e., the set of r' such that the pair (r',r) is connected.

$$UpFCA^r_t = \sum_{(r',r)\in\mathcal{C}} f^{r',r}_{t-\Delta^{r',r}} \cdot P^{r'}_{t-\Delta^{r',r}}$$

$$UpPCA^r_{t,s} = \sum_{(r',r)\in\mathcal{C}} f^{r',r}_{t-\Delta^{r',r}} \cdot L^{r'}_{(t-\Delta^{r',r}),s}$$

These two summations have a physical interpretation: $UpFCA^r_t$ is the contribution of all upstream FCAs into a given resource *r* during time interval *t*, and $UpPCA^r_{t,s}$ is the contribution of all upstream PCAs into a given resource *r* during time interval *t* under scenario *s*.

The full formulation of ESOM is as follows:

$$\text{Minimize} \quad \sum_{r\in\mathcal{F}}\sum_{t\in\mathcal{T}} c_g G^r_t + \sum_{r\in\mathcal{P}}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}} c_a p_s A^r_{t,s}$$

subject to:

$$G^r_{t-1} + UpFCA^r_t + D^r_t = G^r_t + P^r_t \quad \forall r \in \mathcal{F}, \forall t \tag{1}$$

$$A^r_{(t-1),s} + UpFCA^r_t + UpPCA^r_{t,s} = A^r_{t,s} + L^r_{t,s} \quad \forall r \in \mathcal{P}, \forall s, \forall t \tag{2}$$

$$L^r_{t,s} \le M^r_{t,s} \quad \forall r \in \mathcal{P}, \forall s, \forall t \tag{3}$$

$$\sum_{(r.r')\in\mathcal{C}} f^r_t = 1 \quad \forall r \in \mathcal{F}, \forall t \tag{4}$$

$$G^r_t, P^{r,r'}_t, L^r_{t,s}, A^r_{t,s} \ge 0 \quad \forall r, \forall s, \forall t \tag{5}$$

$$G^r_0 = G^r_{T+1} = A^r_{0,s} = A^r_{(T+1),s} = 0 \quad \forall r, \forall s \tag{6}$$

The objective function is the same as it was for the original model, SSM: the sum of planned ground delay and expected air holding over all the capacitated resources. The first constraint (1) is the flow through the FCA nodes in the network flow diagram. This is depicted in the left-hand side of Fig. 6. Similarly, constraint (2) is the flow at each PCA node, depicted in the right hand side of Fig. 6. Constraint (3) limits the capacity at the PCA, which is the flow coming out the bottom of the node. Constraint (4) says that the split of flows out of each resource must sum to 1. The constraint should be checked when formulating the fractional flows as input parameters; we have added it for sake of completeness in the most general form of the model, which would allow the model to choose the splits of flows. (Again, we are not proposing to do this.) Constraint (5) enforces non-negativity of the decision variables, and constraint (6) sets the boundary conditions and the start and end of the time horizon.

We note that it would be trivial to add capacity limitations on the FCAs (right now, only the PCAs are capacitated). We have avoided this to emphasize that FCAs are flow-control mechanisms, rather than physical aviation resources. If the FCA has a physical capacity, then it should be modeled as an uncapacitated FCA feeding into a capacitated PCA.

As written, ESOM is a linear program (constraints and objective function are linear) and will be solved as such. The number of variables is small by modern optimization standards. A realistic sized problem (at least as we have envisioned it) would be more like 40 time intervals, 20 resources, and 5 scenarios for a total of 4,000 variables for each variable category. There are four such categories (P, G, A, L variables), for a total of 16,000 variables, which can be solved efficiently using the add-in software that comes with a spreadsheet. Even if the size were to climb into

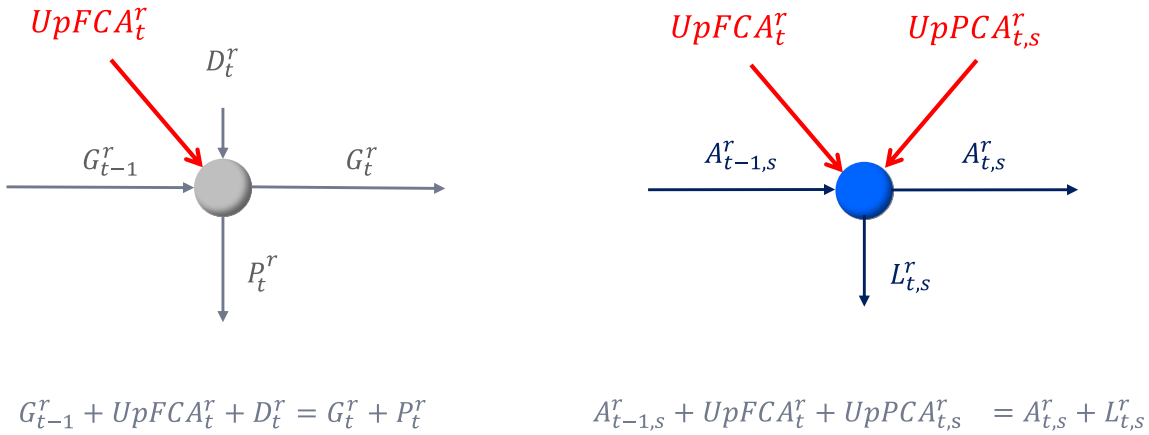

Fig. 6 Constraints corresponding to FCA nodes (left) and to PCA nodes (right)

the hundreds of thousands, it can be solved very quickly (a few minutes) using modern commercial solver software (e.g., CPLEX or Gurobi).

Ideally, we should enforce the decision variables as integer values, because they are numbers of flights. For practical application, it will suffice to solve the linear program and round to the nearest integer values. (In future research, we will confirm whether this works in all cases.) The fractional nature of the flow-splitting variables prevents integrality from being achieved.

V. Behavior of the Model

Before demonstrating practical application of ESOM, we explore its sensitivity with respect to two types of input parameters: holding costs and probabilities of scenarios. This is important to understand, since there is no rigorous method yet established for how to set these values. Surprisingly, although SSM was introduced in 2003, there has been little work to demonstrate its behavior. For ease of exposition, we ran controlled experiments (idealized traffic conditions) for the case of one FCA feeding one PCA, which is exactly SSM. Since ESOM is a generalization of SSM, we fully expect its behavior to be inherited from SSM.

A. Sensitivity to Cost Ratio

Recall that ESOM finds the optimal rate-setting policy across an ensemble of capacity scenarios by minimizing overall expected ground delay and air delay costs. In this sense, the model is designed to hedge across the scenarios. Given the linear nature of the objective function, the degree of hedging is determined by the ratio of the marginal air-holding cost to the marginal ground-holding cost, c_a/c_g. We assume that 10 flights want to arrive in each of 7 time periods at a single resource (FCA). There are two capacity scenarios, each with probability 0.5 of occurring, as shown in Fig. 7. Clearly, not all flights can be accommodated under these scenarios, so some ground holding must take place. The question is how much, given the uncertainty of the scenarios. Since this is a single-PCA scenario, we can assume it is an airport GDP or an AFP.

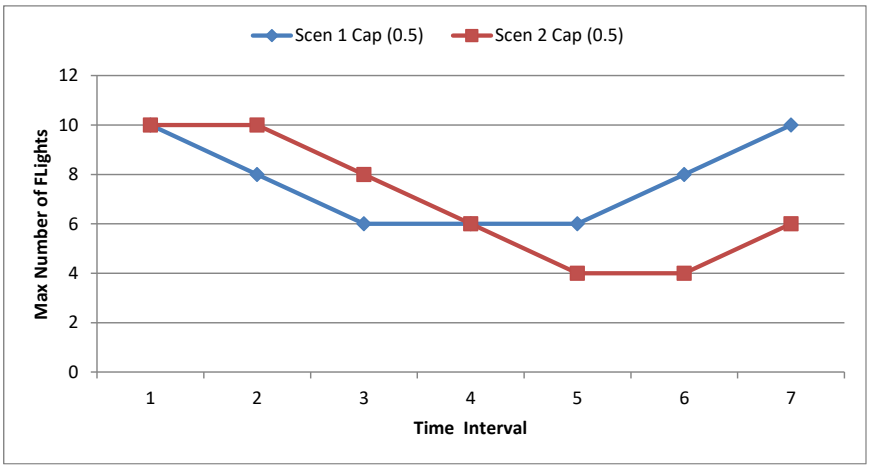

Fig. 7 Capacity scenarios

We ran ESOM to solve for the optimal arrival rates to the FCA; however, we varied the air/ground cost ratio from 0.5 to 12.0, as shown in rows 8–17 of Table 1. (Rows 2–6 show the demand and capacity data.) Each solution row is an optimal FCA rate. For instance, row 8 shows that when the cost ratio is 12:1, the FCA should allow 10 flights in period 1, 8 flights in period 2, and so on.

Key observations about cost ratio sensitivity from Table 1 are:

- Row 8: When the cost ratio is extremely high (12:1), the solution follows a profile equal to the minimum of the two capacity profiles. This makes sense, because the cost of air holding is so high that the model will never send more flights than can be accommodated under any of the scenarios. This reflects risk-averse behavior.
- Row 16: When the cost ratio is extremely low (1.1:1), the solution follows a profile nearly equal to the maximum of the two capacity profiles. This makes sense, because the cost of air holding is so low that the model will send as many flights as can be accommodated under any of the scenarios, accepting the risk that it might have to hold some of them while airborne downstream. This reflects risk-tolerant behavior.
- Row 17: When the cost ratio is less than 1.0, the model allows all the flights to enter, because ground holding is more expensive than air holding. We do not believe that this has any practical application.
- Rows 9–15: These are hybrid solutions, meaning some mix of the two extreme solutions. If you trace down any one column for a fixed time period, you can see the optimal rate increase as the cost ratio drops.

Many of these cost ratio values are unrealistic—they would never be used in practice—but we have included them to deliberately drive out extreme behavior of the model. The most realistic cost ratio is probably in the range of 2:1 or

Table 1 Sensitivity to cost ratio

1	Time Int	1	2	3	4	5	6	7	
2	**Demand**	10	10	10	10	10	10	10	
3	**Scen 1 Cap (0.5)**	10	8	6	6	6	8	10	
4	**Scen 2 Cap (0.5)**	10	10	8	6	4	4	6	
5	**Max Capacity**	10	10	8	6	6	8	10	
6	**Min Capacity**	10	8	6	6	4	4	6	
7	CostRatio								Solution
8	**12**	10	8	6	6	4	4	6	Min capacity
9	**10**	10	8	6	4	4	6	10	Hybrid
10	**8**	10	8	6	6	4	6	10	Hybrid
11	**6**	10	8	6	6	4	6	10	Hybrid
12	**5**	10	8	6	6	4	6	10	Hybrid
13	**4**	10	8	6	6	4	6	10	Hybrid
14	**3**	10	8	8	6	4	6	10	Hybrid
15	**2**	10	10	6	6	4	6	10	Hybrid
16	**1.1**	10	10	8	6	4	6	10	Max capacity
17	**0.5**	10	10	10	10	10	10	10	Demand

3:1. In rows 14 and 15 one can see that the model is "hedging" against extreme solutions. This is a form of risk mitigation, a desirable property of the model.

B. Sensitivity to Capacity Scenario Probabilities

We designed another experiment to understand the model's sensitivity to probabilities of the capacity scenarios. This is important to understand because the science of ensemble weather forecasting is still evolving, and there may be some subjectivity in setting the probabilities. Our experimental scenario (traffic demand, number of scenarios) is the same as for the cost ratio experiment, except that the probabilities are varied from 0.0 to 1.0, while the cost ratio is fixed (at 2:1). The resulting optimal FCA rates are shown rows 6–11 in Table 2. The probabilities shown are the probability of scenario 1; therefore the probability of the other scenario is 1.0 minus that value.

Table 2 Sensitivity to probability of capacity scenarios

1	Time Int	1	2	3	4	5	6	7	
2	Demand	10	10	10	10	10	10	10	
3	Scen 1 Cap	10	8	6	6	6	8	10	
4	Scen 2 Cap	10	10	8	6	4	4	6	
5	Prob Scen 1								Solution
6	1	10	8	6	6	6	8	10	Scenario 1
7	0.8	10	8	8	6	4	8	10	Hybrid
8	0.6	10	10	8	6	4	6	10	Hybrid
9	0.4	10	10	8	6	4	4	6	Hybrid
10	0.2	10	10	8	6	4	4	6	Hybrid
11	0	10	10	8	6	4	4	6	Scenario 2

Key observations about probability sensitivity from Table 2 are:

- Row 6: When the probability of scenario 1 is 1.0, the optimal rates follow the capacity profile in scenario 1. This makes sense, because scenario 1 is certain to happen.
- Row 11: When the probability of scenario 1 is 0.0, the optimal rates follow the capacity profile in scenario 2. This makes sense, because scenario 2 is certain to happen.
- Rows 7–10: For probabilities strictly between 1.0 and 0.0, the optimal rate profile is a hybrid of the two extreme solutions. The model is hedging between extreme solutions.

By having constructed other small examples (not shown in this report), we have seen cases in which the model is overly sensitive to the probabilities, meaning that the solution stays the same for changing probability, then suddenly jumps to another extreme solution at a critical value. Though this behavior is mathematically correct, in practice, this is not a desirable property, because the probabilities will probably not be precisely computed. Traffic managers will be concerned about extreme outcomes as well as expected values, so it will be important to investigate robustness properties of the model, similar to Saraf et al. [6].

VI. Addressing Demand Uncertainty due to Rerouting

In this section, we consider traffic demand uncertainty due to airline voluntary rerouting through or around FCAs. One possible approach would be to reformulate ESOM in a way that includes airline preferences for rerouting. Though we are considering this in future research, such a model may become unwieldy for practical use in a decision support tool. Also, since it would trade off airline-specific cost utilities for rerouting between airlines and with system costs that traffic mangers apply to manage the overall system, this type of comprehensive model may be better suited for descriptive applications (e.g., benchmarking idealized solutions).

The solution we propose is to embed ESOM in an iterative heuristic algorithm that allows ESOM rate suggestions to interact with, and be tempered by, airline preferences for tradeoffs between rerouting and ground delay. One situation we would like to avoid is thrashing that could occur between FCA rates and airline rerouting responses. Recall that FCA rates are a function of forecasted demand and capacity levels at the PCAs. These rates could become suboptimal if demand levels at the PCAs change significantly due to airline reroutes adopted in response to the ground delays imposed by the rates. The new demand forecasts could require traffic mangers to impose a new set of rates, which may induce a new demand pattern, and so on.

Another closely related problem we wish to address is accurately forecasting the traffic demand splits required as input to ESOM. Recall that these splits are the forecasted distribution of traffic across outbound arcs of the FCAs and PCAs in our network. If airlines reroute their aircraft in response to imposed FCA rates, then these demand splits will change also. One may consider this an anomaly of our modeling approach, but really, this is just another way of saying that demand levels at the resources will change.

Our workaround to both these problems is to make an initial demand forecast at all of the PCAs (which also provides us the traffic splits), then apply a model of airline rerouting responses, then rerun ESOM using the new demand forecasts, which again invokes airline rerouting responses, and so on. We repeat this iterative process until the FCA rates converge (that is, ESOM keeps outputting the same FCA rates after each run). We call this the Rate Computation Loop (RCL). The RCL is depicted in flow diagram format in Fig. 8. The written description of the process is below.

A. Rate Computation Loop (RCL) Algorithm (Note: Steps refer to Fig. 8)

Initialization: First, we initialize the resource topology (step A); we input the capacity scenarios (step B); and we input TOS options either provided by the airlines or modeled by our decision support software (step C). Then in step D, we compute the baseline flight trajectories. (A flight is captured in the CTOP only if it has an initial trajectory that goes through one of the declared FCAs.)

The first time the algorithm is run, we skip steps 1 and 2 in the procedure shown in Fig. 8.

Demand Modeling: In Step 3, we project the forecasted flight trajectories onto the resources to obtain the demand variables (D_t^r). We also use this to compute the fractional flows coming out of each resource (step 4).

FCA Rate Optimization: In steps 5 and 6, we formulate and run ESOM, thereby obtaining the recommended FCA rates. Then we ask if those rates have stabilized from prior iterations of the loop. The first time we run the loop, there is nothing to compare it to; therefore, we go to the CTOP resource allocation process.

CTOP Allocation: In these steps (1 and 2), we run the CTOP resource allocation algorithm exactly as would happen in CTOP. This algorithm assigns flights to the FCA slots we have created by setting the FCA planned rates. (E.g., if we call for 30 flights at an FCA in a given hour, then there are 30 slots available during that hour.) Because this algorithm takes carrier TOS options into account, some of the flights will be voluntarily rerouted through our resource network (i.e., take a different sequence of FCAs and PCAs). Dummy FCAs can be added to capture flights routing completely out of the program. This means that the demand on resources has changed, and we probably need to change the FCA rates we already set. Therefore, again we run ESOM in steps 5 and 6.

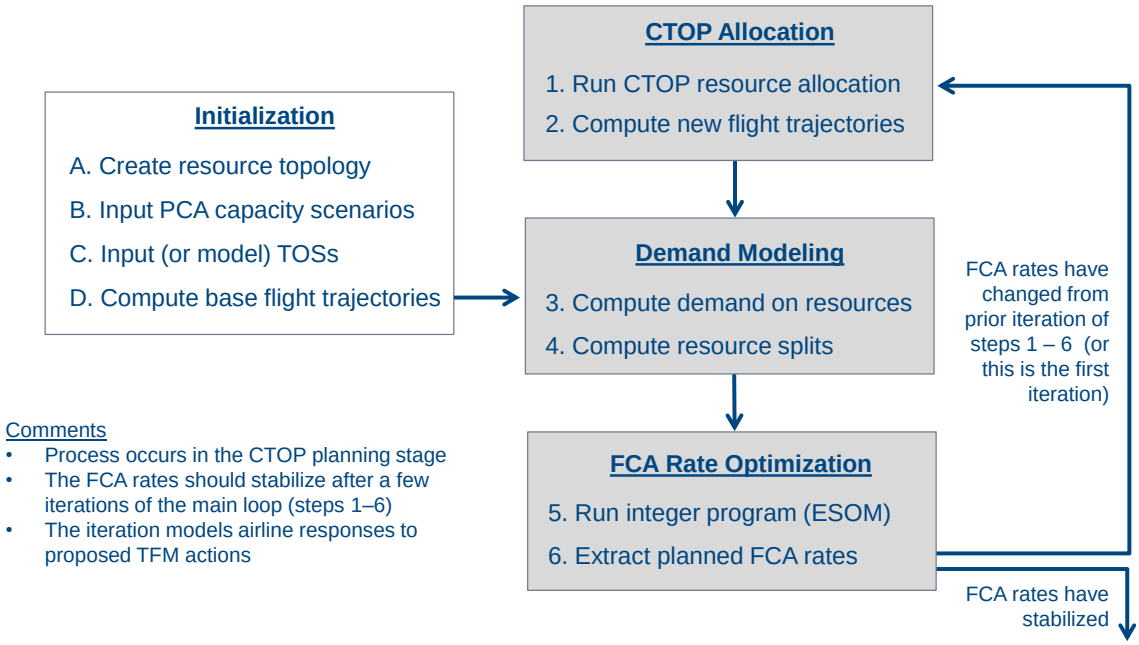

Fig. 8 FCA Rate Computation Loop (RCL), which shows how to apply ESOM to compute FCA rates, taking TOS options into account

The result of RCL should be a set of FCA rates that are in equilibrium with what the airline routing preferences are (or at least what we forecast them to be). In other words, assuming that we have modeled airline rerouting responses well (parallel research we are conducting), we are confident that once the final FCA rates are implemented, they have already taken into account airline responses, and there is no need to reset the rates. The degree to which convergence of this loop will hold is not yet fully understood. Our intuition is that it should produce a fairly stable solution, for the following reason. In practice, the airline reroutes are conducted inside the CTOP resource allocation algorithm at the same time the FCA rates are imposed. Airlines submit their tradeoff preferences a priori. The mechanism for expression of these utilities is designed to be invariant with the FCA rates. Crudely put, the airlines express a number of rerouting options and how much ground delay they would be willing to tolerate before being moved to an alternative route. In contrast, if the airlines resubmitted these utility expressions en masse after having seen the proposed FCA rates, then one could easily construct a counter-example in which the loop permanently thrashes. The convergence issue may well be a moot point: one could always conduct numerous iterations of the RCL and then select the PAR values that provide the least overall cost, when system costs (ground plus air holding) are added to airline rerouting costs.

We should point out that the convergence issue is not an anomaly of our modeling approach to setting FCA rates. No matter what method is used, the traffic management community may have to address the need to reset rates based on airlines responses. Since CTOP has been run only on a few trial instances, there is no experimental data to collect, making solution stability (robustness) with respect to demand changes a matter of outstanding research.

In the next section, we will show in a realistic example how ESOM and the RCL can be used to provide recommended FCA rates.

VII. Experimental Results

We created a realistic example showing how ESOM would be used in planning a CTOP and how ESOM and the RCL would be used to account for rerouting options chosen by flight operators. The example is based on arrival flows into Newark (EWR). Figure 9 shows the location of the PCAs used in this example, located roughly over Washington Center (ZDC) and a fourth PCA over EWR airport. We created one FCA for each of the three large PCAs, modeled as east-west oriented polygons segments to the south of this region. The FCAs capture most of the traffic coming from the south and heading toward EWR and the Northeast, through the PCAs. Based on historical traffic rates for each of the PCAs, we created a baseline capacity scenario and then two other scenarios, with capacities assumed reduced due to convective weather.

In the example, the CTOP planning horizon spans 4 hours, with 1 additional hour to handle possible overflow traffic. In ESOM, we created 20 15-minute time periods. The capacity for the PCAs is shown in Table 3. Only the capacity at PCA1 and PCA2 is affected by the weather scenarios. In scenario 1, the capacity at all PCAs is at its nominal values, 5 flights per quarter-hour at PCA1 and 10 flights per quarter-hour at PCA2. In scenarios 2 and 3, the capacity at PCA1 is reduced to 2, and in scenario 3, the capacity at PCA2 is also reduced to 2. The probability of scenario 1 or scenario 2 occurring is set equal to 0.3, while the probability of scenario 3 is 0.4.

We set the capacity at 10 flights per quarter-hour for PCA3 and for the PCA at EWR airport, under all scenarios, even

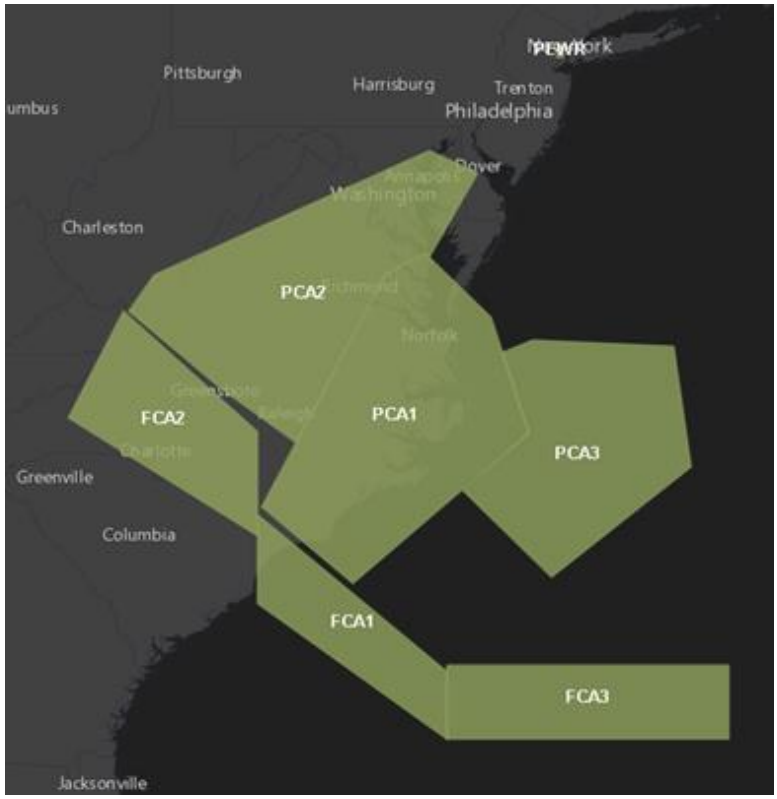

Fig. 9 FCAs and PCAs used for example CTOP

though the demand for those resources is lower, because we are not modeling a demand-capacity imbalance there.

The demand for each FCA in each time period is shown in Table 4. This is derived from actual flights operated on a sampled day during August 2017. Very few flights—only one on this day, from the Caribbean to New York—file through FCA3 initially, but if congestion and delays are restrictive enough over FCA1 or FCA2, flights may choose to reroute through FCA3.

Table 3 Capacity scenarios for PCAs in example CTOP

		time *t* =																			
resource	scenario	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	s1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	25	25	25	25
PCA1	s2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	25	25	25	25
	s3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	25	25	25	25
	s1	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	25	25	25	25
PCA2	s2	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	25	25	25	25
	s3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	25	25	25	25
	s1	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	25	25	25	25
PCA3	s2	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	25	25	25	25
	s3	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	25	25	25	25
	s1	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	25	25	25	25
PEWR	s2	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	25	25	25	25
	s3	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	25	25	25	25

Table 4 Demand at FCAs in example CTOP

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
FCA1	8	4	15	10	9	13	11	7	6	10	5	6	6	8	4	9	2	1	0	0
FCA2	42	54	48	43	48	43	28	43	32	40	55	42	48	25	31	36	12	11	1	0
FCA3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Table 5 shows the traffic "splits," i.e. the portion of traffic passing through each resource that later flows through another resource downstream. We only capture traffic flowing from an FCA to a PCA, or from a PCA to another PCA, because in this use case we are only interested in northbound traffic. Resource pairs not shown (e.g., FCA1 to PEWR) had no traffic flowing between them in our example.

The initial ESOM solution in the RCL (steps 5 and 6 in Fig. 8) recommends ground-holding a portion of the demand at FCA1 and FCA2 and sending the rest to try to arrive on time. There will be a small amount of airborne holding at PCA1 if scenario 2 occurs, or at PCA2 if scenario 3 occurs.

Table 5 Demand splits

from	to	%
FCA1	PCA1	36%
FCA2	PCA1	2%
FCA2	PCA2	30%
FCA3	PCA3	100%
PCA1	PCA2	31%
PCA2	PCA1	9%
PCA2	PCA2	2%
PCA2	PEWR	4%
PCA3	PCA1	20%

We then used the values of the *P* variables (planned acceptance rates) to specify capacity for the FCAs in the CTOP allocation algorithm (steps 1 and 2 in Fig. 8). Because the *P* values are non-integer, we rounded down to the nearest whole number. Even though *P* for FCA3 is 0 (except for the 1 in time period 8), we kept the capacity for FCA3 high, because we don't want to prevent any flights from using routes through FCA3. The zero values in the ESOM solution are a function of the low demand at that resource. In practice, using an artificially high demand as ESOM input should result in rates that are most useful for traffic management. We are exploring this "saturation" technique in related research.

After the CTOP allocation algorithm was run, the demand at the three FCAs met the capacities recommended by ESOM. We computed new split percentages based on the numbers of flights flying from one resource to another downstream (steps 3 and 4 in Fig. 8), and ran ESOM again using this input. The results were similar, with the recommended rates at FCA1 reduced to zero this time. After another iteration, the *P* values did not change, so the loop was complete. This shows that the Rate Computation Loop converges for a realistic example. The *P* values at each iteration are shown in Table 6. (The values in the last hour don't converge, but those only represent spillover traffic.)

Table 6 Optimal rates (values of P) at each iteration of RCL

		time *t* =																			
iteration	resource	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	FCA1	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3	60	18	0	0
1	FCA2	30	30	30	30	30	30	30	30	30	30	30	30	30	30	31	31	57	62	72	0
	FCA3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	FCA1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	44	44	35	0
2	FCA2	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	50	50	56	43
	FCA3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	FCA1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	42	37	0
3	FCA2	29	29	29	29	29	29	29	29	29	29	29	29	29	29	29	29	48	48	51	64
	FCA3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	FCA1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	40	40	3
4	FCA2	29	29	29	29	29	29	29	29	29	29	29	29	29	29	29	29	47	47	47	71
	FCA3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

VIII. Conclusions

This paper introduces an optimization model to minimize traffic flow management delay costs over a network of airspace resources being controlled by a CTOP, when faced with capacity uncertainty. Because the model works with aggregate numbers of aircraft, it is compatible with CDM resource allocation principles and solves quickly. We conducted sensitivity analysis to show properties of the model related to capacity scenario probabilities and to cost structure.

We also showed how to use the model to address traffic demand uncertainty that arises from flight operators exercising the rerouting options that are inherent to CTOP. Because those options are a reaction to flow control rates that would be set by traffic managers (rates that could have been recommended by a model such as ESOM), we devised an iterative loop that would allow them to adapt to these preferences. Ideally, this loop will converge to a stable solution, as we showed with one small example. In future work, we will test a broader range of use cases and examples to determine more general conditions under which the loop will converge.

Although this model does not formally plan for a recourse stage, we envision that a rolling horizon paradigm would be used in CTOP planning. In future work, we will compare ESOM against a truly dynamic model that plans for this, such as the model described in Mukherjee and Hansen [7].

Currently available weather products may not generate capacity profiles with the fidelity needed to capture the uncertainty in a CTOP. However, even for the deterministic case, it is challenging for humans to determine the optimal flow rates for a small network of FCAs, so decision support from a model such as ESOM would still be valuable.

In this work, we assumed a network of airspace resources was already established, but another rich area of future research is determining how to construct the network of FCAs and PCAs.

Acknowledgments

This research was funded under NASA Research Announcement contract #NNA16BD96C. The authors wish to thank NASA technical monitor Deepak Kulkarni, and NASA researchers Heather Arneson, Paul Lee, Nancy Smith, and Antony Evans for their insightful comments and support throughout the research.

References

[1] Ball, M. O., Hoffman, R., Odoni, A., and Rifkin, R., "A Stochastic Integer Program with Dual Network Structure and its Application to the Ground-Holding Problem," *Operations Research*, Vol. 51, No. 1, 2003, pp. 167–171. doi: 10.1287/opre.51.1.167.12795

[2] Vranas, P., "The Multi-Airport Ground-Holding Problem in Air Traffic Control," Sc.D. Dissertation, Ocean Engineering Dept., Massachusetts Inst. of Technology, Cambridge, MA, 1992.

[3] Vranas, P., Bertsimas, D., and Odoni, A. R., "The Multi-Airport Ground-Holding Problem in Air Traffic Control," *Operations Research*, Vol. 42, No. 2, 1994, pp. 249–261. doi: 10.1287/opre.42.2.249

[4] Vranas, P., Bertsimas, D., and Odoni, A. R., "Dynamic Ground-Holding Policies for a Network of Airports," *Transportation Science*, Vol. 28, No. 4, 1994, pp. 275–294.

doi: 10.1287/trsc.28.4.275
[5] Richetta, O., and Odoni, A. R., "Solving Optimally the Static Ground-Holding Policy Problem in Air Traffic Control," *Transportation Science*, Vol. 27, No. 3, 1993, pp. 228–238.
doi: 10.1287/trsc.27.3.228
[6] Saraf, A., Ramamoorthy, K., Stroiney, S., Sawhill, B., and Herriot, J., "Robust, integrated arrival-departure-surface scheduling based on Bayesian networks," *IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC)*, Colorado Springs, Colo., 2014, pp. 2B4-1–2B4-14.
doi: 10.1109/DASC.2014.6979424
[7] Mukherjee, A., and Hansen, M., "A dynamic stochastic model for the single airport ground holding problem," *Transportation Science*, Vol. 41, No. 4, 2007, pp. 444–456.
doi: 10.1287/trsc.1070.0210