Fifteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2023)

# Contributing Factors to Flight-Centric Complexity in En-Route Air Traffic Control

Gijs de Rooij*, Amber Stienstra*, Clark Borst*, Adam B. Tisza†, Marinus M. van Paassen* and Max Mulder*
*Faculty of Aerospace Engineering, Delft University of Technology, Delft, the Netherlands
Email: {g.derooij, c.borst, m.m.vanpaassen, m.mulder}@tudelft.nl
†Maastricht Upper Area Control Centre, EUROCONTROL, Maastricht, the Netherlands
Email: adam.tisza@eurocontrol.int

***Abstract*—To alleviate the workload of air traffic controllers, part of the air traffic may be handled by a future automated system. When deciding which flights to delegate, a distinction can be made between basic and non-basic flights, with the former being prime candidates for delegation. The human controller can then focus on the non-basic flights, where human competencies are most valuable and more difficult to automate. The classification of flights is preferably based on objective measures relating to the traffic situation. Existing complexity models are, however, often used for capacity predictions or airspace restructuring and primarily to assess the complexity of a sector as a whole. In this paper we use empirically collected flight complexity ratings from 15 professional en-route air traffic controllers. They indicated which other flights contributed to their complexity assessment of a single flight of interest. This exploratory study was able to build a machine-learning model which adequately classifies these flights, based on a qualified majority of controllers. By analyzing the interactions between the included flights, we discuss whether a classification model can differentiate between basic and non-basic flights, and which traffic features play the largest role. Once this can be done reliably and an appropriate complexity threshold has been chosen, a model can be developed as a starting point for an automatic allocation algorithm that distributes flights between a human controller and the computer.**

***Keywords*—air traffic control; complexity; human factors; human-automation teaming**

## I. Introduction

In striving for safe and efficient operation of Air Traffic Control (ATC) in an increasingly capacity-limited Air Traffic Management (ATM) system, Air Traffic Control Officers (ATCO) are progressively supported by automated tools [1]. Decades of human-automation research have shown, however, that humans are bad at supervising automated systems and thus benefit greatly from active involvement [2], [3]. Nevertheless, routine traffic does take away cognitive capacity from ATCOs that could be better used in handling more complex situations.

One way to redistribute workload and cognitive effort is to allocate a subset of the traffic to a computer agent, enabled by the increased use of Controller-Pilot Data Link Communications (CPDLC), freeing up the ATCO's cognitive resources, which are needed for complex problem solving. The human ATCO is then responsible for controlling the remaining traffic with active involvement. Exploratory research showed that such a shared airspace is feasible and accepted by ATCOs under certain conditions [4]. Assigning flights to either a human or a computer agent can be regarded as the next evolutionary phase in Flight-Centric ATC, a concept where specific flights are assigned to different human ATCOs [5].

Eurocontrol's Maastricht Upper Area Control Centre (MUAC), an Air Navigation Service Provider (ANSP) responsible for the upper airspace over the Netherlands, Belgium, Luxembourg and part of Germany, proposes a strategy to initially only allocate *basic* traffic to an automated system, while the ATCOs are kept engaged with the task of handling the more complex *non-basic* traffic [6]. Basic flights are presumably easier to automate and do not evoke the creative problem-solving skills that human ATCOs are known to enjoy, making them a prime candidate for delegation to a computer.

The level of responsibility of the envisioned computer system will be increased in three stages, throughout which the computer will only aid with or control the basic part of the traffic. In the first two stages, the ATCO can still take back manual control over a flight. In Stage 1 all flights are handled with approval of the ATCO. In Stage 2, no ATCO approval is needed for the selection and handling of basic traffic. Any of the ATCOs on the sector may intervene at all times though, as they remain responsible for all traffic in the sector: complex as well as basic. In Stage 3, the computer will autonomously control an entire sector with basic traffic, performing all ATC tasks in that sector. In this stage, the ATCO will no longer be responsible for the traffic and will not be monitoring the sector.

As an enabler for this strategy, it is paramount to understand what differentiates a 'basic' from a 'non-basic' flight. Furthermore, this classification should be automated, based on objective criteria that can be obtained in real-time as a flight approaches a sector. Despite extensive research, the driving factors for air traffic complexity are still not completely understood [7]. Current models predominantly consider the complexity of an entire sector [8] or parts thereof [9], for example to predict sector capacity. This is most commonly done by taking a weighted sum of various contributing factors, such as the rate of flights entering/exiting the sector or the traffic density [10], [11]. The sector-wide approach of the aforementioned methods makes them unsuited for classifying individual flights.

In this paper, we reason from the perspective of a single flight, rather than an entire sector and ask the following questions: What is the relationship between the number of flights ATCOs consider as having impact on a single flight and the *perceived* complexity of that flight? What level of consensus exists among ATCOs on these included flights? What traffic parameters impact the perceived complexity the most? To answer these questions, empirically collected flight complexity ratings and associated flights from 15 professional en-route ATCOs are analyzed using state-of-the-art supervised learning techniques to discover relationships, if they exist, between complexity ratings and traffic factors.

The structure of this article is as follows. First, we distill what lessons can be learned from existing complexity measures that are primarily used to describe entire sectors (Section II). Next, in Section III a human-in-the-loop experiment is described where professional ATCOs had to indicate which other flights they included in their complexity assessment of a single Flight of Interest (FOI) that varied in location and target state over a number of scenarios. Results of the experiment, and subsequent descriptive performance of our machine learning models are given in Section IV. The implications of the findings and an outlook into the future applicability of a flight allocation algorithm are discussed in Section V. Section VI concludes the work.

## II. Background: Modeling Flight Complexity

### A. From Sector-Based Towards Flight-Centric Complexity

Complexity prediction in ATM has predominantly been done in the context of dynamic sectorization to either split or combine sectors based on expected traffic loads and ATCO workload. Over decades, several complexity models have been developed, such as Dynamic Density, Interval Complexity, Fractal Dimension, Input/Output Approach, Lyapunov Exponents and Trajectory-Based Complexity (TBX) [12], [13]. The majority of these complexity models output either a scalar value or a map that represent the sector-based complexity by integrating (e.g., counting and averaging) specific flight characteristics over the entire sector, for example [8]:

- the number of climbing and/or descending flights
- the variance in heading and speed
- the structure of traffic flows (e.g., crossing angle)
- the number of crossing and/or merge points
- distance at, and time to, the closest point of approach (CPA)

It can be argued that sector-based complexity dilutes the complexity contribution of each individual flight. Take for instance the situation illustrated in Fig. 1 where the sector-based complexity map indicates a hotspot in the middle of the sector. This, however, does not mean that all flights passing through the center of the sector are equally complex (or, non-basic). Conversely, flights that do not pass through the center are not all basic flights. Additionally, certain sector disruptions, like local adverse weather or an emergency flight, might not impact all flights equally.

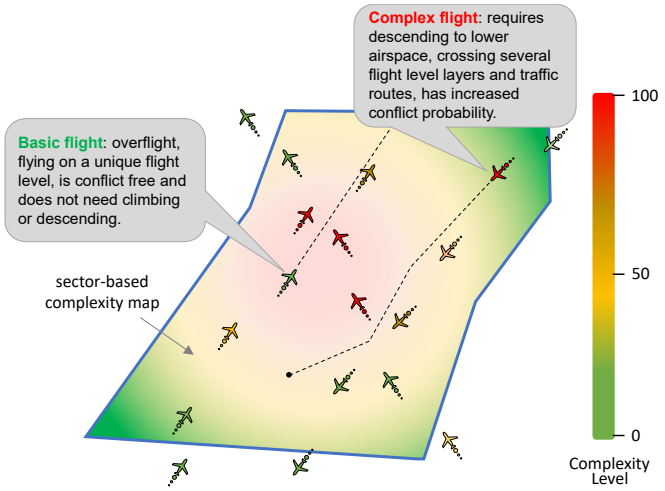



Figure 1: Basic versus non-basic flights.

In an effort to capture the complexity of a single flight, we propose, based on discussions with operational ATCOs and ATM experts at MUAC, that flight complexity centers around *attentional* and *control* demands, such as:

- Attentional demands
  - Trajectory complexity (e.g., winding vs. direct route)
  - Uncertainty (e.g., in climb/descent profiles, arrival time management, pilot delays)
  - Multi-dimensional interaction profile with other flights (e.g., route crossings, altitude overlap, conflict probability)
  - Interaction with environmental disruptions (e.g., restricted airspace, weather cells)
- Control demands
  - Easiness of a conflict resolution (e.g., altitude vs. heading)
  - Conflict geometry (e.g., overtake vs. crossing)
  - Number of required (follow-up) actions (e.g., evade conflict and steer back to target waypoint)
  - Timing of actions (e.g., proactive vs. reactive)
  - Size of the 'solution' space (e.g., sector size for maneuvering flights)

Many of these elements cannot be considered independently in how they impact the complexity of a single FOI and are therefore not easily modeled. For example, given a certain CPA, the convergence angle between flights impacts the time to reach that point. To cope with complexity, ATCOs typically make pair-wise comparisons between flights in a hierarchical manner [14]. For example, to detect conflicts, they first scan the flight labels to detect overlapping altitudes, then narrow down the search to flights with crossing trajectories, followed by anticipating their CPA [15]. As such, the ATCOs' strategies, skills and expertise are expected to play a role in how complexity is perceived.

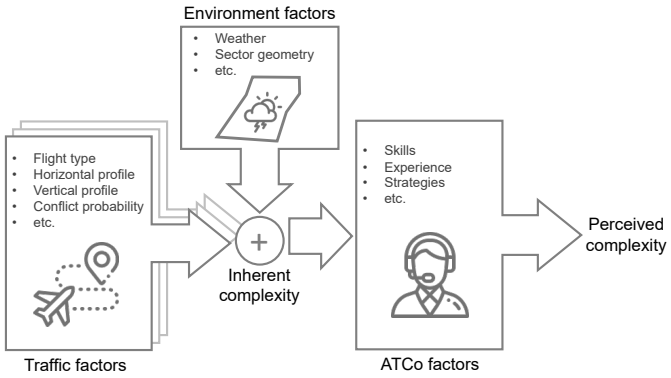



Figure 2: Factors that play a role in ATC complexity.

### B. Inherent versus Perceived Flight Complexity

Similar to the division between taskload and the experienced workload [8], classification of 'basic' and 'non-basic' flights may depend on the preferences, skills and experience of a given ATCO as illustrated in Fig. 2. This notion suggests that the perceived flight complexity may be individual sensitive, similar to the findings of research that studied the impact of personalization on fostering ATCO agreement and acceptance in the context of conflict resolution advisories [16]. Nevertheless, with highly trained professionals, some level of consensus on which flights are more complex than others can still be expected, providing ground for an automated classification algorithm. To be able to discern the contribution of inherent and perceived complexity in determining single flight complexity, labeled data would be needed that allows for relating traffic factors to ATCO complexity ratings.

Currently, such labeled data does not yet exist. Therefore, the study described in this paper aimed to collect labeled data on single flight complexity by designing and conducting a human-in-the-loop experiment.

### C. Supervised Learning

When labeled data is available, *classification* and *prediction* can be done using supervised learning techniques, such as logistic regression, random forests and gradient boosting trees. These have been used, for example, to determine traffic parameters that are most influential to sector complexity [17], [18], but to the best of our knowledge not yet on individual flight complexity.

In a classification problem, unbalanced classes can have a detrimental effect on the model's performance. In our case, the number of flights not contributing to a single flight's complexity vastly exceeds the number of flights that do matter. When mostly trained on non-relevant flights, a model might not be able to predict the important flights. Using ensemble techniques, such as gradient boosting, the impact of included flights on model training can be increased. Similar techniques are done in medical studies, where decease cases are rare, but of paramount importance to discover and predict.

Note that machine learning is mainly used in our preliminary study to examine and describe the complexity factors in a specifically crafted set of scenarios. Creating an operational prediction model for any traffic sample is outside the current scope and would require more extensive data collection.

TABLE I. Participant characteristics.

| | Sector group | | |
|---|---|---|---|
| | Brussels | DECO | Hannover |
| Number of ATCOs | 5 (all male) | 5 (1 female) | 5 (all male) |
| Age, years (std) | 37.0 (4.3) | 40.8 (7.6) | 42.0 (5.4) |
| Experience, years (std) | 13.0 (4.0) | 15.4 (7.2) | 19.4 (6.0) |

## III. Method

### A. Participants and Apparatus

Fifteen professional ATCOs from MUAC voluntarily participated in a simulator experiment. Table I shows their characteristics. All participants provided written consent and the experiment was approved by the Human Research Ethics Committee of TU Delft under number 2206.

During the experiment MUAC's operational interface was mimicked using SectorX, a medium-fidelity Java-based simulator built by TU Delft. Fig. 3 was displayed on a computer monitor and could be controlled with a computer mouse. While only static scenarios were shown, the simulator allowed for some interaction that helped the ATCOs assess the traffic situation. A flight's planned route could be revealed by press-and-hold on the associated label. Furthermore, MUAC's VERification and Advice tool (VERA) was available to see a prediction of the closest horizontal distance between two flights and their corresponding future positions. And finally, the velocity leaders could be extended to show a flight's predicted position one to eight minutes into the future.

### B. Procedure and Participant Tasks

Each participant followed the same procedure, outlined in Fig. 4. At the start they were briefed on their task to assess the complexity of guiding an individual Flight of Interest (FOI) from its current location to the required sector exit point (XCOP) and Transfer Flight Level (TFL). The ATCOs then practiced operating the simulator on a simplified scenario containing only two flights in an artificial sector.

Next, four training scenarios were executed, in which the background traffic, sector and experiment procedure were identical to the measurement scenarios. Each scenario required two consecutive actions from the ATCOs:

1) Indicate which background flights played a role in their complexity assessment (from here on referred to as 'included flights'). If no flights were selected, a confirmation popup was shown before continuing.
2) Register their FOI complexity rating on a 0-100 scale on the screen.

The measurement phase consisted of 36 scenarios and was followed by a review phase where the scenarios that received the highest, lowest and middlemost rating (three each) were revisited. These nine scenarios were presented in the same order as they appeared in the first phase. The participants could see their registered complexity score and which flights they had included, but were not told why these scenarios were selected for review. The ATCOs were asked to fill out a questionnaire about these scenarios to gain more insight into the reasoning behind the reported complexity.

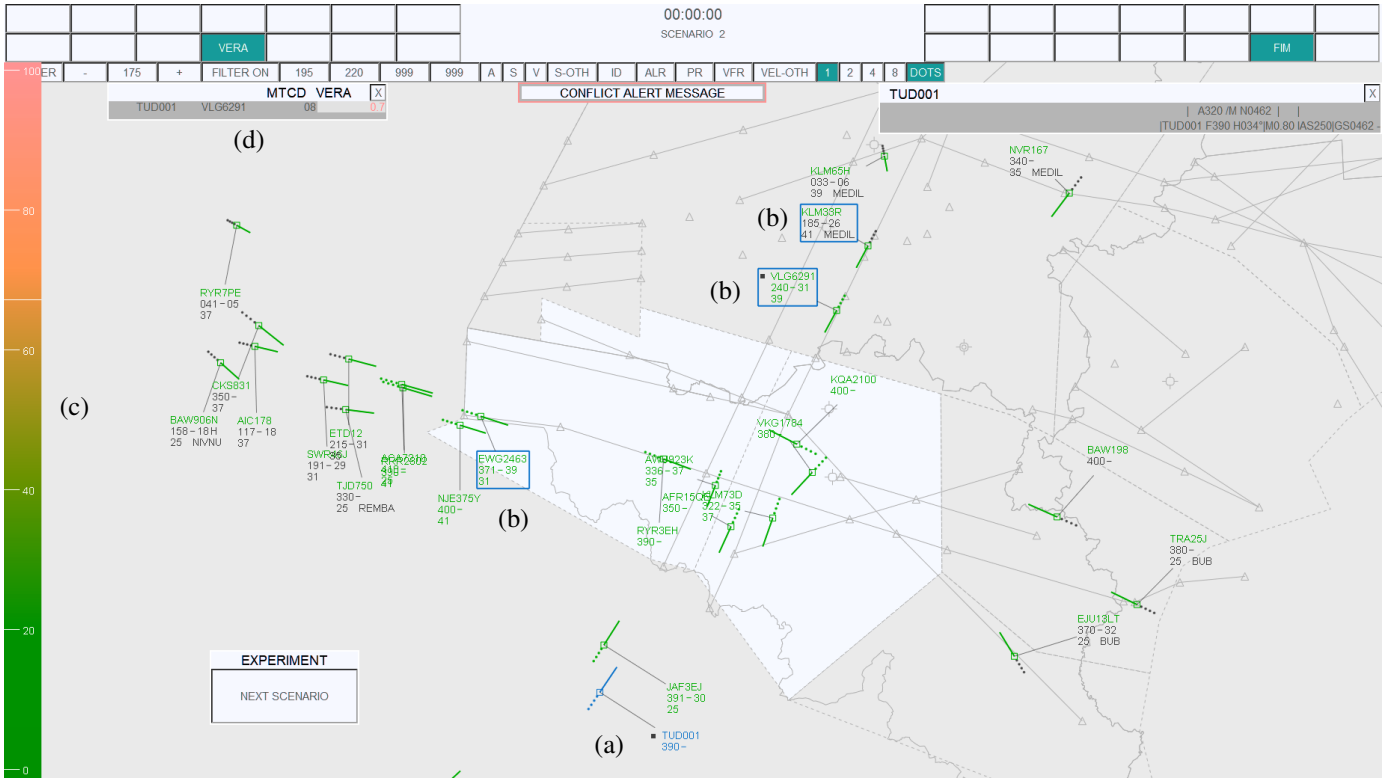



Figure 3: Simulator interface showing a Brussels scenario with flight of interest (a), three included flights (b), complexity rating scale (c) and VERA information (d). Background colors have been inverted for print clarity.

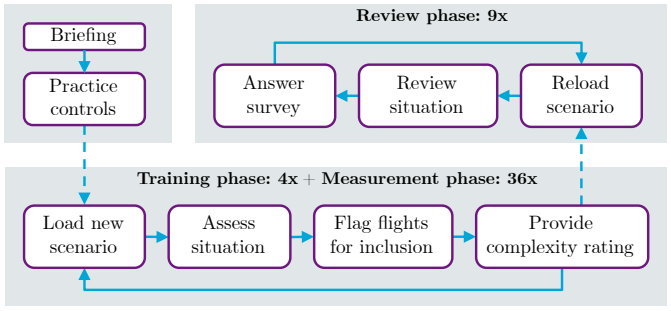



Figure 4: Experiment procedure.

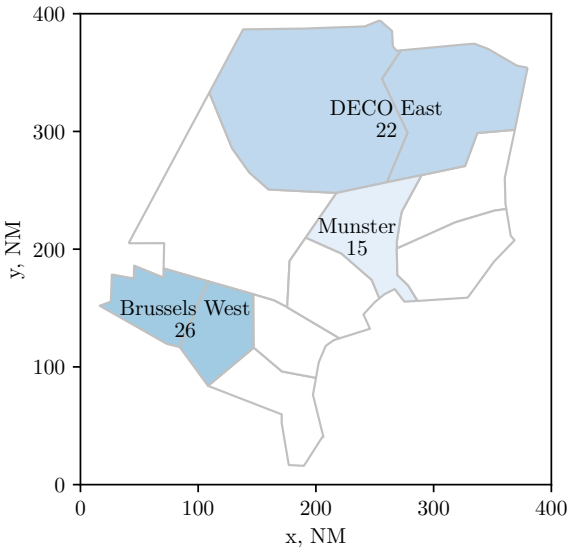



Figure 5: Selected MUAC sectors and their number of flights, excluding the FOI.

### C. Scenario Design

Three distinct MUAC sectors were selected for the experiment, ranging from large and relatively quiet (DECO East) to small and dense (Brussels West). Participants were only presented with the sector they had an endorsement for, ensuring comparable familiarity levels. For each sector, a distinct radar snapshot from 23 March 2022 was taken to serve as background traffic. The snapshots were selected such that it was possible to introduce conflicts with various characteristics. As individual sectors are often combined to balance capacity with demand, we used the same sector configuration as was operational at the time of the corresponding radar snapshot: DECO East contained Jever and Holstein, Brussels West consisted of Koksy and Nicky, and Munster was a sector from the Hannover sector group. Fig. 5 shows these sectors and the number of flights in each of them.

A single FOI was overlaid on the background traffic in a variety of initial positions and exit conditions to create distinct scenarios (see Fig. 3 for an example). It was colored differently to distinguish it from other traffic. Manipulating a single flight, instead of using an entirely different traffic sample for each scenario, eliminated the influence of sector complexity factors external to the FOI (e.g., traffic density, total number of climbing flights) as much as possible.

TABLE II. Scenario design parameters.

| Parameter | Variation | |
|---|---|---|
| Time to CPA (urgency) | Short | (0–300 s) |
| | Medium | (300–600 s) |
| | Long | (600–900 s) |
| Conflict-free direct route (easiness of resolution) | Interactions on current trajectory (I) | |
| | Interactions on current trajectory (II) | |
| | Interactions on current flight level | |
| Flight level change (uncertainty) | Small descent | (0–4,000 ft) |
| | Small climb | (0–4,000 ft) |
| | Large climb or descent | (>4,000 ft) |

Following the flight complexity demands identified in Section II-A, the various scenarios were manipulated to address all of these demands. For example, by using three different sectors with distinct sizes and traffic densities, control demand in terms of available 'solution' space is manipulated. Sector size also impacts attentional demands since the chance of interactions between flights increases. Within each sector, Table II lists the traffic factors that were manipulated to impact the complexity of the FOI. Note that these manipulations might have interactions with one another, meaning that it is not possible to study the impact of each manipulation on complexity separately. The scenarios were presented in a partially randomized order to account for order effects.

### D. Dependent Measures

The experiment resulted in the following output measures:

- Complexity rating for the FOI,
- Flights included by the ATCOs as 'contributing to the complexity rating',
- Usage of VERA, velocity leaders and route preview,
- Current and target (exit) states of all flights, and
- Survey: reasons to include flights and how comfortable the ATCOs would be to delegate the FOI to the computer.

With the (target) states of all flights known, the features listed in Table III have been computed to describe each of the flight pairs including the FOI. The selection of features is based on existing sector complexity research referenced in Section II. Lacking sufficient data to accurately predict climb or descend points, horizontal positions are extrapolated along the current tracks and ground speeds. No advanced trajectory predictions are used yet in this exploratory study. To exclude predicted conflicts beyond a reasonable look-ahead horizon, the calculation of the features was limited to the trajectory before reaching the XCOP for flights descending to a lower airspace within the sector. If a predicted CPA would occur after reaching the XCOP, the CPA was capped to the distance between the flights upon reaching the XCOP. This was only done for these descending flights, as ATCOs do 'look' beyond their sector boundaries to prevent causing any conflicts for their colleagues in adjacent sectors.

To relate a single FOI complexity rating to the characteristics of multiple included flights in a scenario, the aggregated features listed in Table IV have been proposed. Note that these only relate to the FOI itself, or in relation to flights included by the ATCO. Non-included flights may have an impact on the sector-wide complexity, but have been considered irrelevant to the FOI complexity in this study. For all altitude differences absolute values were taken.

TABLE III. Candidate features of included flights relative to the FOI.

| | Feature | Unit | Comment |
|---|---|---|---|
| Attentional demands | Current horiz. separation | NM | |
| | Predicted min. horiz. separation (CPA) | NM | |
| | Time to CPA | s | |
| | Vertical separation | ft | |
| | Exit altitude difference | ft | |
| | Overlapping flight levels | T/F | True if flights *may* be at the same level at some point |
| | Climbing/descending | T/F | |
| Control demands | Convergence angle | deg | |
| | Ground speed difference | kts | |
| | Flight state | - | Assumed or transferred to me |
| | Distance to XCOP | NM | Along a direct path |
| | Required altitude change | ft | From actual flight level to TFL |

* T/F = True/False dichotomous indicator

TABLE IV. Candidate features for the FOI, aggregated over all included flights.

| | Feature |
|---|---|
| FOI | Required altitude change |
| | Distance to XCOP |
| Included flights | Number of flights with altitude overlap |
| | Number of climbing flights |
| | Number of descending flights |
| | Number of flights with CPA under 10 NM |
| | Number of flights with identical TFL |
| | Min./average current separation |
| | Min./average CPA |
| | Min./average distance to XCOP |

## IV. Results

The results are discussed in three steps:

1) The experimentally *collected data* is described in terms of number of included flights with respect to the complexity rating and the level of consensus between different ATCOs, in addition to their use of support tools.
2) A *classification model* is used to determine whether the inclusion of a flight can be linked to objective features and what the relative importance of each feature is.
3) In combination with the FOI's complexity rating, a *regression model* is used to examine the feasibility of predicting a FOI's complexity through its included flights.

Since the number of flights varied over the sectors (see Fig. 5), percentages of the total number of flights shown to participants are used when comparing sectors. Each cell in the tables refers to values belonging to one participant, unless explicitly stated otherwise.

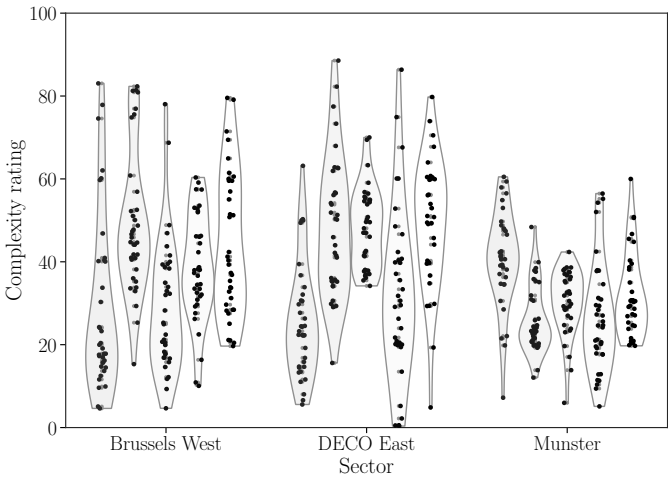



Figure 6: Complexity ratings per ATCO for each scenario.

TABLE V. Total included flights per participant, as share of total flights presented to that participant.

| | Brussels West | | DECO East | | Munster | |
|---|---|---|---|---|---|---|
| | 118 | 12.6% | 60 | 7.6% | 55 | 10.2% |
| | 93 | 9.9% | 70 | 8.8% | 41 | 7.6% |
| | 180 | 19.2% | 97 | 12.2% | 37 | 6.9% |
| | 64 | 6.8% | 56 | 7.1% | 52 | 9.6% |
| | 71 | 7.6% | 77 | 9.7% | 68 | 12.6% |
| Average | 106 | 11.2% | 72 | 9.1% | 51 | 9.4% |
| Std. dev. | 42 | 4.5% | 15 | 1.8% | 11 | 2.0% |

### A. Complexity Rating and Number of Included Flights

Fig. 6 shows the complexity ratings as given by the ATCOs for each of the 36 scenarios of their sector that were presented to them. The large spread in ratings per ATCO shows that the designed FOI manipulations had an effect on the perceived complexity, as the background traffic did not change between scenarios. This effect was less strong in Munster, where the ATCOs gave relatively low ratings compared to their colleagues in the other sectors. To account for between-participant differences, complexity ratings per ATCO are standardized by z-scores in the remaining analyses.

In total, the ATCOs included 1,139 (10.0%) of the 11,340 flights that were presented to them. Although the number of flights was different for each sector, the share of included flights seems to be primarily ATCO-dependent and varies as much as between 6.8-19.2% in a single sector (Table V). One Brussels participant is a noticeable outlier with 180 included flights, significantly skewing the average for that sector.

Similar to the complexity ratings, the number of included flights has been standardized per participant in Fig. 7 to account for individual differences. A Kendall's tau-b correlation test, chosen because of the non-normality of the data, shows a moderate positive correlation ($\tau_b = .547$, $p < .001$) between the complexity scores given by the ATCOs and the number of included flights for all sectors combined. In Table VI the correlations are given per participant and sector. The strongest, yet still moderate, correlation is found for Brussels West ($\tau_b = .611$). The correlations are statistically significant ($p < .001$) for all sectors. In both DECO East and Munster, one participant exhibits noticeably weaker correlations than the other participants.

TABLE VI. Correlation between standardized number of included flights and standardized complexity score (Fig. 7).

| | $\tau_b$, $p$ | | |
|---|---|---|---|
| | Brussels West | DECO East | Munster |
| | .592, $< .001$ | .627, $< .001$ | .381, $= .004$ |
| | .684, $< .001$ | .802, $< .001$ | .561, $< .001$ |
| | .702, $< .001$ | .478, $< .001$ | .563, $< .001$ |
| | .718, $< .001$ | .388, $= .003$ | .553, $< .001$ |
| | .592, $< .001$ | .547, $< .001$ | .773, $< .001$ |
| All participants | .611, $< .001$ | .503, $< .001$ | .527, $< .001$ |

TABLE VII. Number of flight pairs on which VERA was used per participant and the share of those that was included.

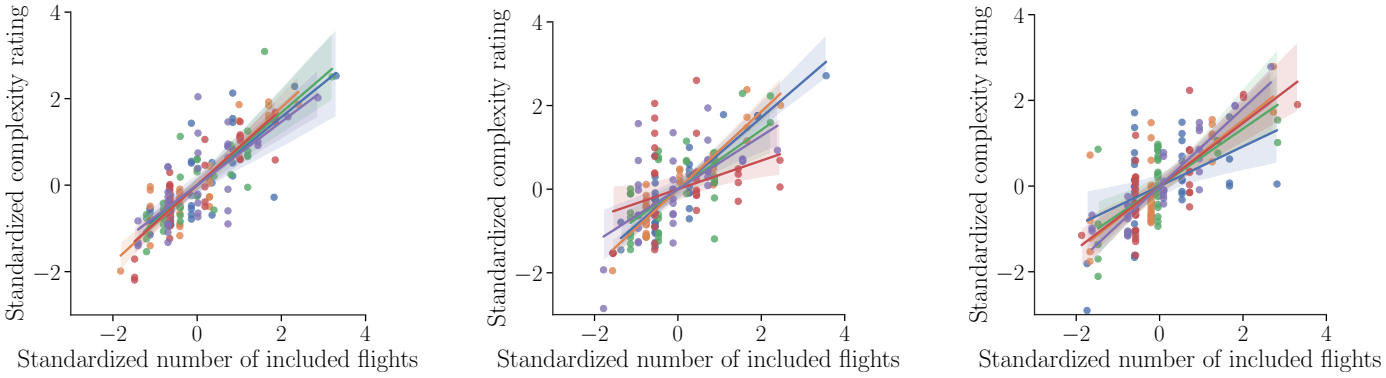
| Checked (included) | | |
|---|---|---|
| Brussels West | DECO East | Munster |
| 38 (37, 97.4%) | 11 (10, 90.9%) | 9 (8, 88.9%) |
| 24 (14, 58.3%) | 35 (29, 82.9%) | 0 (0) |
| 106 (61, 57.5%) | 1 (0) | 33 (25, 75.8%) |
| 63 (41, 65.1%) | 56 (36, 64.3%) | 9 (8, 88.9%) |
| 78 (47, 60.3%) | 49 (36, 73.5%) | 0 (0) |

### B. Usage of Support Tools

To determine (or confirm) whether a flight is in conflict with the FOI, the ATCOs could use VERA to show the predicted minimum separation between two flights. Usage varied greatly over the participants, ranging from not being used at all to checking 106 flight pairs (Table VII). Note that the sectors cannot be readily compared with each other, due to their vastly different number of flights (and thus potential conflicts). The Brussels participant, who included the most flights, was also by far the most active user of VERA. All participants practiced with VERA in the training phase and were thus aware of its availability. In total, 352 (68.8%) of the 512 VERA flights were eventually included (Fig. 8), while only 7.8% of the flights with a CPA below 10 NM was *not* included by the ATCO after confirming this distance through VERA. Presumably, some ATCOs felt comfortable with smaller separation margins and/or would not consider this an immediate problem for far-away flights. Above 20 NM, only some flights were included, mostly by ATCOs who considered any VERA-check an 'include' action.

Besides VERA, extending the velocity leaders beyond the default one minute is another, more crude, technique to check future positions of flights and assess their CPA. As the velocity leaders are adjusted for all flights at once, this cannot be linked to the inclusion of particular flights. Neither did we find indications for velocity leaders being used instead of VERA (i.e., a DECO ATCO who used VERA only once did not extend the velocity leaders at all).

Akin to the usage of VERA, the number of flights for which a visual representation of the planned route on the radar display was requested varied considerably between zero and 54. While the display of routes made flights with




(a) Brussels West (b) DECO East (c) Munster

Figure 7: Standardized number of included flights versus the standardized complexity rating, colored per participant of each sector and shown with 95% confidence intervals. See Table VI for correlations.

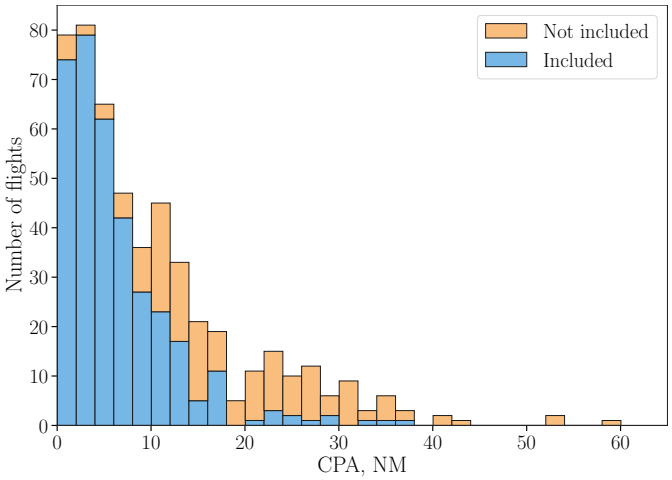



Figure 8: Number of flights on which VERA was used and the share of those flights that was subsequently included.

planned turns more pronounced, the planned turn was already indirectly visible by the waypoint listed in the flight's label. To illustrate the limited predictive value of this measure: three ATCOs requested the route of the same flight that would come into proximity once the FOI commenced a turn, but only one of them decided to include it.

### C. ATCO Consensus

As the number of included flights already shows, there is a level of subjectivity in the data. We therefore introduce three majority levels regarding flight inclusion. Consensus is reached when all five ATCOs of a sector agreed to either include or exclude a flight. For a qualified or simple majority respectively four and three ATCOs were in agreement. The distributions in Fig. 9 show a high level of consensus for all sectors, with the ATCOs unanimously agreeing for 84-88% of the flights, increasing to 94-96% with qualified majorities. Between any two ATCOs in a sector, 88-97% of the flights was identically labeled. The relatively low share of excluded flights in Brussels West, compared to the other sectors, is mostly due to the large number of inclusions by a single participant, as also reflected in Table V.

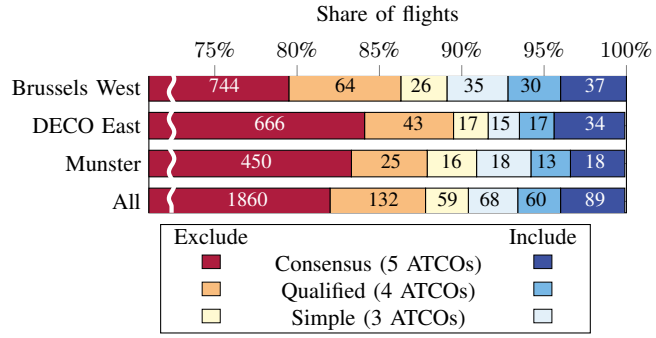



Figure 9: ATCO consensus on flight inclusion per sector and for all sectors combined.

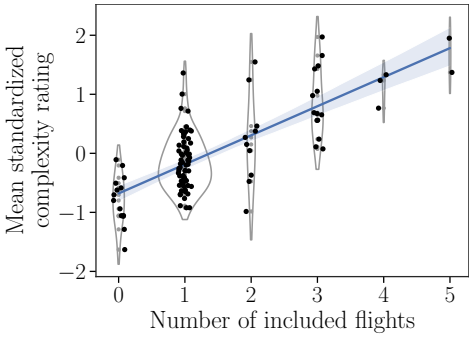



Figure 10: Correlation between complexity rating and number of flights included by a qualified ATCO majority in each scenario. Shown with 95% confidence interval.

Fig. 10 shows the number of flights per scenario that was included by a qualified majority of the ATCOs, versus the average standardized complexity rating for that scenario. Note that the ATCOs agreed on the inclusion of just a single flight in the vast majority of scenarios. Again, a moderate positive correlation is visible ($\tau_b = .553$, $p < .001$), but it is also clear that a higher number of included flights does not necessarily relate to a higher complexity rating. Hence, the features of those particular flights might better explain some of the variability.

### D. Features of Included Flights

To analyze which features play a role in the ATCOs' selection of included flights and see whether this selection can be modeled, we applied a gradient boosting classifier on the features from Table III. Gradient boosting was used in this study, because of its ability to combine weak learners (e.g., due to imbalanced data) into a strong model. The model target was to classify whether a flight was included or excluded by a specified majority of the ATCOs. All flights that did not meet the specified level of consensus were filtered out to ensure that the model was trained and tested on a progressively well-labeled data set. As the label was binary (include or exclude), the simple majority case included all flights.

To avoid under- or overfitting the model, the data was split over four stratified folds, meaning that the share of included flights was equal in all folds. The model was then trained and tested on four splits (each consisting of three training folds and one testing fold) and subsequently tuned through cross-validation and grid search for high F1-scores (a balance between precision and recall).

The resulting confusion matrices, summed over the four splits, are shown in Fig. 11 for each of the majority categories. This clearly reflects the expected increase in performance when filtering on at least a qualified majority that provides more robust labels on the data (Table VIII, averaged over the four splits). 89% of the flights that were included by all ATCOs were correctly classified as 'include' by the consensus model, while only 11% of the included flights were missed.

As a measure for the predictive value of each of the features, their relative importance in the consensus model is given in Fig. 12, as an interval over the four folds. As was expected, the predicted minimum separation (CPA) appears to be the most important feature, followed by the presence of an altitude overlap. Flights where the altitude bands are not overlapping will never be in conflict, unless one of the flights has to deviate to another level. To illustrate, only seventeen (0.4%) out of 4,492 flights without altitude overlap have been included by the ATCOs and never by more than one ATCO at a time. Ten of these were in the Jever scenarios, with a single ATCO including seven. We were unable to identify the reasons for including these particular flights, other than two cases where the included flight would first climb and then descend within the controller sector, whereas our metric purely looked at current, cleared and exit flight levels to assess the overlap. The current horizontal separation and time till CPA are marginally more important than the other features used in this study. Finally, a flight's ATC state and whether it is climbing or descending seem to have negligible impact.

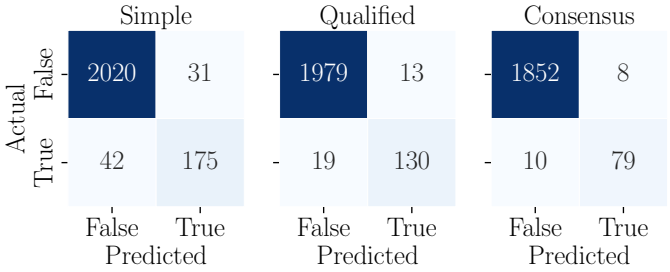



Figure 11: Classifier confusion matrices per majority category.

TABLE VIII. Flight inclusion classifier performance.

| Majority category | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Simple | 0.97 | 0.86 | 0.81 | 0.83 |
| Qualified | 0.99 | 0.91 | 0.87 | 0.89 |
| Consensus | 0.99 | 0.91 | 0.89 | 0.90 |

### E. Predicting Complexity Ratings

Besides identifying the important features of flights that may impact whether a certain flight should be included or not in assessing the FOI complexity, it would also be important to predict the complexity rating associated with the FOI. In that way, the future system envisioned by MUAC would be able to predict the complexity level of a flight entering the sector, classify it as either 'basic' or 'non-basic' and assign the flight to either the computer or the human ATCO, respectively.

Creating such a prediction model first requires that parameters denoting the relationships between the FOI and the included flights are aggregated by descriptive statistics, such as the average, sum, minimum, etc. Table IV lists the relational parameters that we included in this first exploratory study. When a participant included zero flights for a scenario, it was filtered out in this study, as no aggregated features could be computed in that case. The model's goal is mainly to detect the complex cases and scenarios with zero included flights received relatively low complexity ratings anyway. This model is, furthermore, independent of the level of consensus between ATCOs, as they may agree on the inclusion of some flights in a scenario, but may also include flights in their complexity rating for which no consensus was reached. Therefore, we consider their individual combination of included flights and complexity rating.

To test and train the gradient boosting regression model, a fifteen-fold is used with one participant per fold. This ensures that the data belonging to a single participant does not get spread out over the training and test data, such that the model performance is an indication for how well the model generalizes on ratings from other ATCOs for which it was not trained. The model's hyperparameters were tuned through cross-validation and grid search optimizing for high $R^2$-scores.

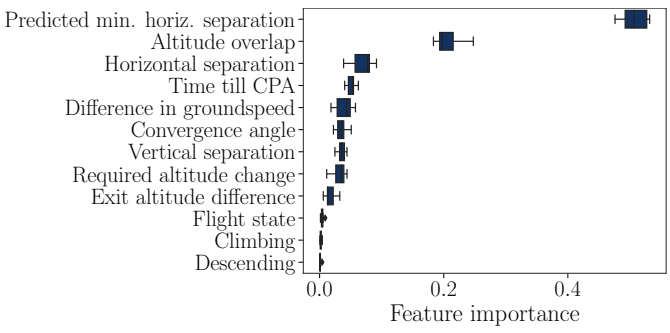



Figure 12: Feature importance of consensus classifier model.

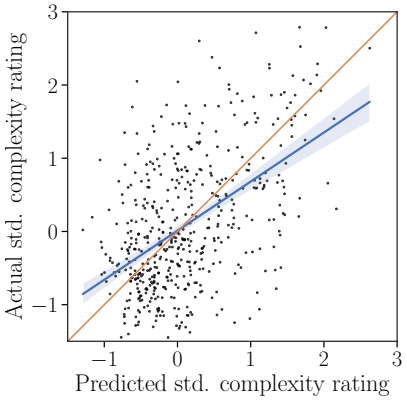



Figure 13: Comparison between the original data and regression model output, shown with 95% confidence interval.

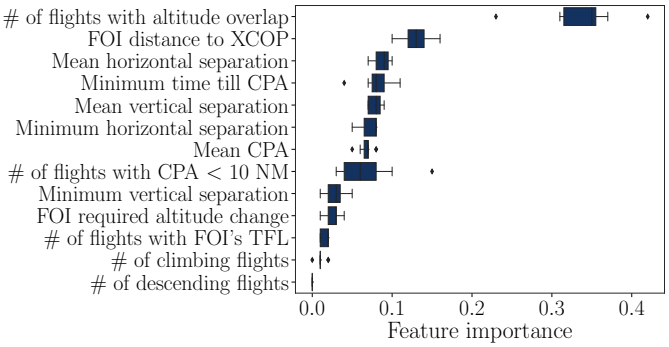



Figure 14: Feature importance of regression model.

Fig. 13 shows an example of the actual complexity ratings by the ATCOs and the corresponding model-predicted rating, for each of the folds combined. With $R^2 = 0.16$, $MSE = 0.68$, $MAE = 0.65$ and $RMSE = 0.34$, the model's performance is relatively weak compared to existing subjective sector-based complexity models, such as those discussed in [19]. The importance of the features, over the fifteen splits, is given in Fig. 14. Despite the weak model performance, some observations stand out. First, the number of flights with an altitude overlap is clearly the most important feature. According to Section IV-D this is closely related to the number of included flights, confirming this metric's moderate correlation with the complexity rating. Furthermore, flights at a closer distance to their exit point (XCOP) are more likely to receive a high complexity rating. Presumably because their solution space is limited. Finally, a group of features is of equal or marginally different importance, confirming that many factors play a role in perceived complexity.

### F. Willingness to Delegate Flights to Automation

With the complexity rating known, a flight can be classified as either basic or non-basic based on a complexity threshold. In the post-measurement reviews, the ATCOs had to indicate how comfortable they were with having the FOI handled by the computer. Unfortunately due to technical issues, only a small part of this data was saved. Based on this limited data and discussions with ATCOs, a higher complexity rating seems to generally match with a lower willingness to delegate the flight, tipping around the zero in their z-scored ratings.

## V. Discussion

### A. Included Flights

Despite a high level of consensus, the ATCOs clearly had different interpretations of which flights to include. This can originate in different working styles, with some ATCOs more pro-actively solving distant conflicts, but may also indicate a lapse in the briefing. Several ATCOs, for example, included all flights that would lead to a loss of separation if no action was undertaken, while other ATCOs did not include such flights if a straightforward solution was available (e.g., descending the flight to its TFL). Furthermore, some ATCOs included flights that did not directly pose a problem for the FOI, but that decreased the solution space for solving conflicts between the FOI and other flights. This was especially evident when the FOI had to fly opposite a stream of flights.

The features that we selected proofed sufficient to correctly classify most of the flights for which there was consensus between the ATCOs though. By focusing on these unanimously labeled flights the results can be considered on the conservative side, but it is inevitable to avoid highly personalized results. The relative importance of the CPA and the presence of an altitude overlap that was found in the included flight analysis, strengthens the hierarchical task analysis presented in [15]. As expected, ATCOs seem to predominantly filter flight pairs based on these two characteristics. Nevertheless, we identified possible improvements for feature calculation during our analysis. Most prominently, we simplified conflict prediction to a mere extrapolation along the current track, ignoring any expected turns or speed changes that the ATCOs may have included in their judgment.

### B. Complexity Ratings

The results show a moderate correlation between the number of included flights and complexity ratings. This confirms the idea behind the Dynamic Density model, where number of flights in a sector is the primary driver for complexity [12]. Brussels West showed the strongest correlation. The sample size is too small to draw definitive conclusions, but it seems probable to attribute it to the relatively large number of flights, and therefore interactions, compared to the other sectors.

Again, a discrepancy in the used definition of 'complexity' cannot be ruled out. Although standardizing the ratings per ATCO is an established method to reduce between-participant differences, it cannot ensure that all ATCOs equally isolated the complexity of the FOI from that of the entire sector. Moreover, individual ATCOs are not always able to provide consistent ratings, even for identical situations [19], explaining some of the variation in the ratings.

The gradient boosting model was able to predict the complexity ratings to some extent, but would have to be improved if it were to be used for flight classification. The input features need to include additional measures for both the FOI and other flights. For example, whether a flight can transit to its TFL unhindered, or whether it has to be put on a heading, requiring prolonged monitoring and rejoining the route. These factors are known to add to the perceived complexity [20].

### C. Experiment Design

The present study only considered a single base traffic sample per sector and is therefore not necessarily applicable to every traffic situation within or outside these sectors. The fact that we observed differences between the three sectors can stem from multiple factors, including the participating ATCOs, the sector geometry or the used traffic sample. The larger number of VERA actions for Brussels West is a testimony of its above average complexity compared to the other sectors. In hindsight, the scenarios for the other two sectors may have been too 'simple', reducing the range of complexity ratings.

While the base traffic was a snapshot from real traffic data, the artificially introduced FOI did not always match ATCO expectations. Some scenarios were rated more complex than initially expected, because they presented an abnormal situation to the ATCOs, and have therefore hindered modeling the complexity ratings. In a small number of scenarios, the FOI was planned to fly a non-straight trajectory. While the retrieval of routes was logged in the experiment, the route points could also be seen in the label without any (logged) action. If ATCOs incorrectly assumed the flight would proceed along its current track, it would most likely have affected their choice of inclusion, as some conflicts only existed along the planned trajectory. Since we primarily focused our analysis on flights for which there was consensus, we expect its impact to be limited, however. Nevertheless, future research should aim to only include realistic FOIs to completely eliminate such inconsistency. For example by taking a large set of radar snapshots and highlighting a single FOI coming towards or just entering the sector.

The relatively small number of participants per sector increased the potential influence of outliers. With a larger sample size, the qualified majority may become more usable. This would increase the certainty about which flights to include beyond just the unanimously included flights.

### D. Operational Relevance

As soon as we can predict an individual flight's complexity based on objective, readily available traffic characteristics, the next step would be to determine the threshold, below which flights are considered basic. The incomplete survey data from the reviewed scenarios does not provide sufficient ground to this cause, other than the observation that the willingness to delegate flights was largest with low complexity. This matches MUAC's proposed strategy about automating basic flights first [6]. The ATCOs indicated that, among others, high trajectory uncertainty of potentially interfering flights was a key reason to be hesitant about delegating a flight. The introduction of computer-directed flights within an airspace may itself have an impact on the perceived complexity of human-directed flights due to the changed teamwork dynamics and tasks and associated uncertainty [12]. This effect is not included in our current analysis and strongly depends on the way the automated system is implemented.

Another intrinsic aspect of ATC is the occurrence of non-nominal situations, such as flight emergencies or adverse weather conditions [11]. When these occur, the complexity of a subset or even all of the flights will inevitably change. This dynamic aspect was excluded from the experiment to first establish a baseline complexity metric before expanding it to a wider range of situations. When part of the flights are computer-directed, failures of the computer and subsequent chains of events should evidently be taken into account as well. Depending on the type of failure, the complexity threshold below which flights are delegated to the computer may need to be lowered.

In an operational context, it would make sense to automatically assign basic flights to the computer, while leaving non-basic and undetermined flights to the human ATCO. Manually handling a basic flight is expected to be a smaller nuisance then prematurely allocating a non-basic flight to a computer. Thus the model should be tuned favoring a high true positive rate (i.e., recall metric of a classifier) over a high precision. Here, expert opinions play an essential role in establishing the threshold in order to increase ATCO acceptance.

Finally, tweaking the model to the individual ATCO might result in a more accurate model and hence increased ATCO acceptance [16]. On the downside, a personalized model might create an unworkable situation where flight allocations change whenever a new ATCO takes over from a colleague. It also means that the computer has to be sufficiently advanced to handle a wider range of complexities than when it is limited to flights about which consensus was reached.

## VI. Conclusion

In the development of a future ATC system where human controllers remain in charge of all non-basic flights while a computer handles all basic flights, this paper demonstrated the feasibility of classifying basic and non-basic flights, based on features extracted from their interaction with surrounding traffic. We showed that, in static scenarios, the perceived complexity of a single flight of interest can be related to the combined sum of interactions that this flight has with other traffic. Professional controllers showed high levels of consensus on which flights to include or not, although personal preference and working styles still play an important role. The aggregated features of these flights resulted in a weak relation with their perceived complexity.

Follow-up research should first aim to improve the complexity model by including additional features. Next, the complexity threshold below which flights can be considered basic should be determined. In addition, dynamic scenarios are needed to validate to what extent the results are generalizable. Subsequently, the operational applicability should be validated by simulating a shared human-automation airspace with flights automatically assigned to either agent based on the presented model. With increasing model accuracy leading to a larger share of confidently classified flights, increasingly more flights can be automatically allocated to the computer.

## Acknowledgment

We express our gratitude to the participating ATCOs from Eurocontrol's MUAC and to MUAC for providing the scenario traffic data and hosting the data collection sessions.

## REFERENCES


[1] SESAR Joint Undertaking, "European ATM Master Plan 2020," 2019.

[2] M. R. Endsley, "From Here to Autonomy: Lessons Learned from Human-Automation Research," *Human Factors*, vol. 59, no. 1, pp. 5–27, 2017.

[3] B. Strauch, "Ironies of Automation: Still Unresolved After All These Years," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 419–433, 2018.

[4] G. de Rooij, C. Borst, M. M. van Paassen, and M. Mulder, "Flight Allocation in Shared Human-Automation En-Route Air Traffic Control," in *International Symposium on Aviation Psychology*, Online, 2021, pp. 172–177.

[5] P. Volf, "Comparison of the Flight Centric and Conventional Air Traffic Control," in *2019 Integrated Communications, Navigation and Surveillance Conference (ICNS)*. IEEE, 4 2019, pp. 1–10.

[6] P. Hendrickx and A. B. Tisza, "EUROCONTROL Maastricht Upper Area Control Centre OPS & Automation Strategy," 2019. [Online]. Available: https://skybrary.aero/bookshelf/books/5341.pdf

[7] B. Antulov-Fantulin, B. Juričić, T. Radišić, and C. Çetek, "Determining Air Traffic Complexity – Challenges and Future Development," *Promet*, vol. 32, no. 4, pp. 475–485, 2020.

[8] B. Hilburn, "Cognitive Complexity in Air Traffic Control - A Literature Review," Eurocontrol, Tech. Rep., 2004.

[9] D. Schaefer, C. Meckiff, A. Magill, B. Pirard, and F. Aligne, "Air Traffic Complexity as a Key Concept for Multi-Sector Planning," in *20th Digital Avionics Systems Conference*. Daytona Beach, FL, USA: IEEE, 2001.

[10] K. Lee, E. Ferons, and A. R. Pritchett, "Describing Airspace Complexity: Airspace Response to Disturbances," *Journal of Guidance, Control, and Dynamics*, vol. 32, no. 1, pp. 210–222, 2009.

[11] R. H. Mogford, J. A. Guttman, S. L. Morrow, and P. Kopardekar, "The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature," FAA, Tech. Rep., 1995.

[12] M. Prandini, L. Piroddi, S. Puechmorel, and S. L. Brázdilová, "Toward Air Traffic Complexity Assessment in New Generation Air Traffic Management Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 3, pp. 809–818, 2011.

[13] T. Prevot and P. U. Lee, "Trajectory-Based Complexity (TBX): A modified aircraft Count to Predict Sector Complexity During Trajectory-Based Operations," in *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*. Seattle, WA, USA: IEEE, 2011.

[14] E. M. Rantanen and A. Nunes, "Hierarchical Conflict Detection in Air Traffic Control," *The International Journal of Aviation Psychology*, vol. 15, no. 4, pp. 339–362, 2005.

[15] G. de Rooij, A. B. Tisza, C. Borst, M. M. van Paassen, and M. Mulder, "Towards Human-Automation Teamwork in Shared En-Route Air Traffic Control: Task Analysis," in *Proceedings of the 2022 IEEE International Conference on Human-Machine Systems, ICHMS 2022*, 2022.

[16] C. Westin, C. Borst, and B. Hilburn, "Strategic Conformance: Overcoming Acceptance Issues of Decision Aiding Automation?" *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 41–52, 2016.

[17] F. Pérez Moreno, V. F. Gómez Comendador, R. Delgado-Aguilera Jurado, M. Zamarreño Suárez, D. Janisch, and R. M. Arnaldo Valdés, "Determination of Air Traffic Complexity Most Influential Parameters Based on Machine Learning Models," *Symmetry*, vol. 14, no. 12, p. 2629, 12 2022.

[18] J. A. Pérez-Castán, L. Pérez-Sanz, J. Bowen-Varela, L. Serrano-Mira, T. Radišić, and T. Feuerle, "Machine Learning Classification Techniques Applied to Static Air Traffic Conflict Detection," in *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1226, 2022.

[19] P. Andraši, T. Radišić, D. Novak, and B. Juričić, "Subjective Air Traffic Complexity Estimation Using Artificial Neural Networks," *Promet*, vol. 31, no. 4, pp. 377–386, 8 2019.

[20] S. Fothergill and A. Neal, "Conflict-Resolution Heuristics for En Route Air Traffic Management," in *Proc. Human Factors and Ergonomics Society 57th Annual Meeting*, San Diego, CA, USA, 2013, pp. 71–75.