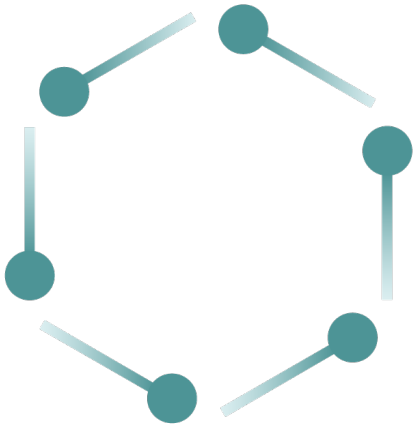


D4.1 Human-machine collaboration in en-route operations

Horizon Europe - SafeTeam

Document number	D4.1
Document title	Human-machine collaboration in en-route operations
Version	0.1
Work package	WP4
Edition date	27.07.2025
Responsible unit	DATABEACON
Dissemination level	PU
Project acronym	SafeTeam
Grant	101069877
Call	Safe, Resilient Transport and Smart Mobility services for passengers and goods (HORIZON-CL5-2021-D6-01)
Topic	HORIZON-CL5-2021-D6-01-13: Safe automation and human factors in aviation – intelligent integration and assistance

This project has been co-funded by the European Union under Grant Agreement 101069877

SafeTeam Consortium

	Innaxis (INX)
	Agencia Estatal de Seguridad Aérea (AESA)
	Technische Universität München (TUM)
	DataBeacon
	ONERA
	Rise Research Institutes of Sweden AB (RISE)
	PEGASUS HAVA TASIMACILIGI ANONIM SIRKETI (PEGASUS)
	CAA INTERNATIONAL LIMITED (CAAi)

Document change record

Version	Date	Status
0.1	27/06/2025	Final document ready for submission

Table of Contents

1 Introduction to complexity management in en-route operations

1.1 En-route complexity management concept

1.1.1 Context - challenges managing workload in enroute ATC

Managing workload and complexity in en-route Air Traffic Control (ATC) poses significant challenges for Air Traffic Control Centers (ACCs). ACCs are responsible for ensuring the safe and efficient flow of air traffic within their designated airspace. However, the increasing number of flights and the complexity of the airspace have made it increasingly difficult to manage the workload effectively. To address this issue, ACCs often limit the amount of traffic per sector.

One of the primary reasons for limiting traffic per sector is to prevent overload and ensure that controllers can effectively handle the workload. Each sector within an ACC has a specific capacity, which is determined by factors such as the number of available controllers, the complexity of the airspace, and the equipment and technology available. By limiting the number of aircraft assigned to each sector, ACCs can ensure that controllers have enough time and resources to handle the traffic safely and efficiently. This approach helps prevent controller fatigue and reduces the risk of errors due to excessive workload.

Additionally, limiting traffic per sector allows ACCs to better manage the complexity of the airspace. En-route ATC involves coordinating the movement of aircraft across vast areas, often involving multiple ACCs and international boundaries. The complexity arises from factors such as varying altitudes, different aircraft speeds, and diverse flight routes. By controlling the amount of traffic in each sector, ACCs can maintain a manageable level of complexity, enabling controllers to effectively monitor and guide aircraft. This approach helps ensure that controllers can maintain situational awareness and make timely decisions, enhancing overall safety in en-route ATC operations.

Understanding how different types of air traffic interact with each other is crucial in effectively managing workload and complexity in en-route Air Traffic Control (ATC). While some traffic may have conflicting paths and require constant monitoring and coordination, others may have non-conflicting routes and pose less of a workload for controllers. Therefore, it is essential for ACCs to have a comprehensive understanding of the traffic patterns and potential interactions before assigning them to specific sectors.

Predicting interactions between aircraft at least 30-60 minutes in advance, or even further, can significantly enhance the management of workload and complexity in en-route ATC. By utilizing advanced technologies and predictive modeling, ACCs can anticipate potential conflicts and plan accordingly. This proactive approach allows controllers to distribute traffic more efficiently across sectors, minimizing the risk of congestion and reducing the workload on individual controllers. Furthermore, by identifying non-conflicting traffic in advance, ACCs can optimize the utilization of airspace and resources, ensuring a smoother flow of air traffic and reducing complexity.

Incorporating predictive interaction analysis into the workload and complexity management strategies of ACCs can also improve safety and efficiency. By identifying potential conflicts early on, controllers can take proactive measures to mitigate risks and maintain separation between aircraft. This not only enhances safety but also reduces the cognitive load on controllers, allowing them to focus on critical tasks and make informed decisions. Additionally, by understanding how traffic will interact in advance, ACCs can provide more accurate and timely information to pilots, enabling them to adjust their flight paths or speeds to avoid conflicts. Overall, integrating predictive interaction analysis into en-route ATC operations can greatly enhance the management of workload and complexity, leading to safer and more efficient air traffic management.

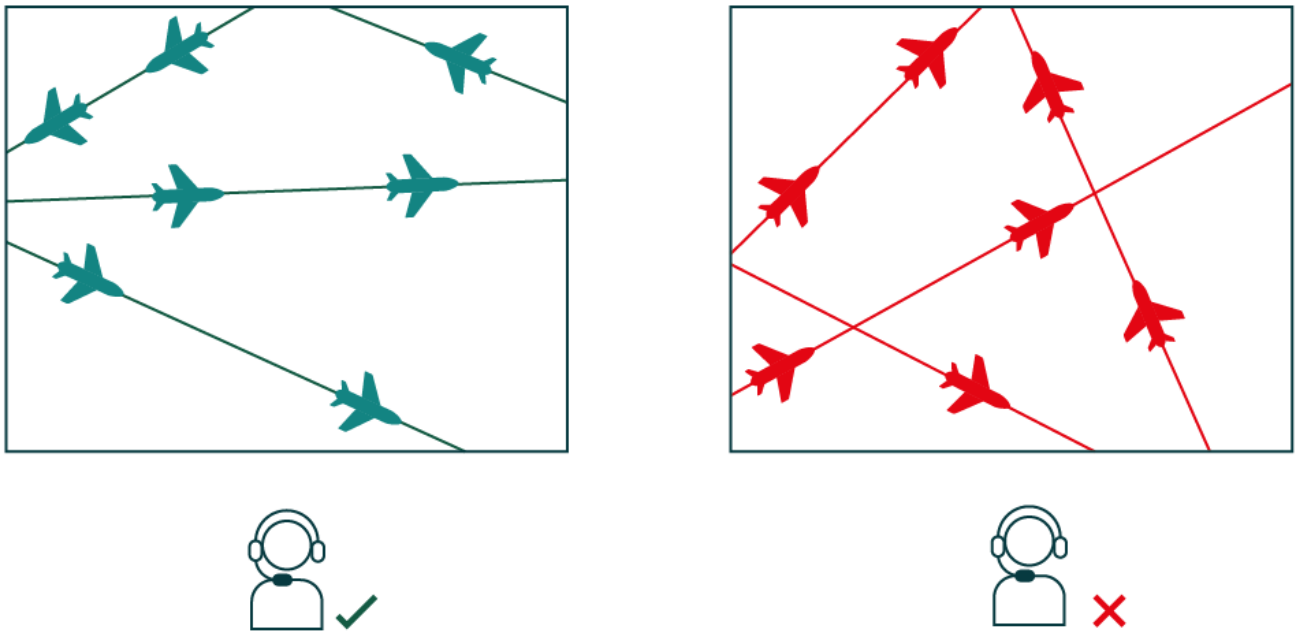
Computing interactions between air traffic in en-route Air Traffic Control is a complex task that requires advanced computation due to several factors. As time progresses, the potential number of interactions increases non-linearly, especially when more traffic flows are involved. This exponential growth in potential interactions poses a significant

challenge for ACCs in managing workload and complexity. Advanced computational algorithms and models are necessary to handle the vast amount of data and calculate potential conflicts accurately. These computations need to consider various factors such as aircraft performance, winds, and potential instructions from adjacent ACCs, which further contribute to the complexity and uncertainty of the task.

The presence of uncertainties in en-route ATC further complicates the computation of interactions. Aircraft performance can vary, and external factors like changing winds can influence the flight paths and speeds of aircraft. Additionally, instructions from adjacent ACCs may impact the flow of traffic and introduce further uncertainties. ACCs must account for these uncertainties and consider multiple scenarios while computing interactions. This requires the development of robust computational models that can handle these uncertainties and generate accurate predictions. By considering various scenarios and accounting for uncertainties, ACCs can better manage workload and complexity, ensuring the safe and efficient flow of air traffic in en-route ATC operations.

Simplifying the problem of workload and complexity management in en-route Air Traffic Control by solely computing the number of expected traffic and identifying evolving traffic (e.g., changing flight levels) is an incomplete approach. While these factors provide some insight into the volume of traffic, they do not fully reflect the expected complexity and workload. This limited perspective can lead to inefficient use of capacity or safety issues within sectors.

Relying solely on traffic volume fails to account for the intricacies of air traffic interactions and the potential impact on workload and complexity. It is possible for a sector to have a high volume of traffic but low complexity if the flights have non-conflicting routes. Conversely, a sector with a lower volume of traffic can experience high complexity due to conflicting flight paths or challenging weather conditions. By overlooking these factors, ACCs may either regulate sectors unnecessarily, resulting in wasted capacity, or underestimate the complexity, leading to potential safety issues.



To address this challenge, ACCs need to adopt more comprehensive approaches that consider not only traffic volume but also the potential interactions, conflicting routes, and other factors that contribute to complexity and workload. Advanced computational models and algorithms can be developed to analyze and predict these complexities, allowing ACCs to make more informed decisions about sector regulations and resource allocation. By incorporating a holistic understanding of complexity and workload, ACCs can optimize sector management, enhance safety, and ensure efficient use of airspace capacity.

1.1.2 The role of the digital assistance in enroute complexity management

Introducing a digital assistant specifically designed to manage complexity in the operational room of Air Traffic Control (ATC) can revolutionize workload management and enhance safety in en-route ATC operations. This digital assistant would serve as a sophisticated system with access to vast computational capabilities, running in the cloud and leveraging a comprehensive set of information to support decision-making processes.

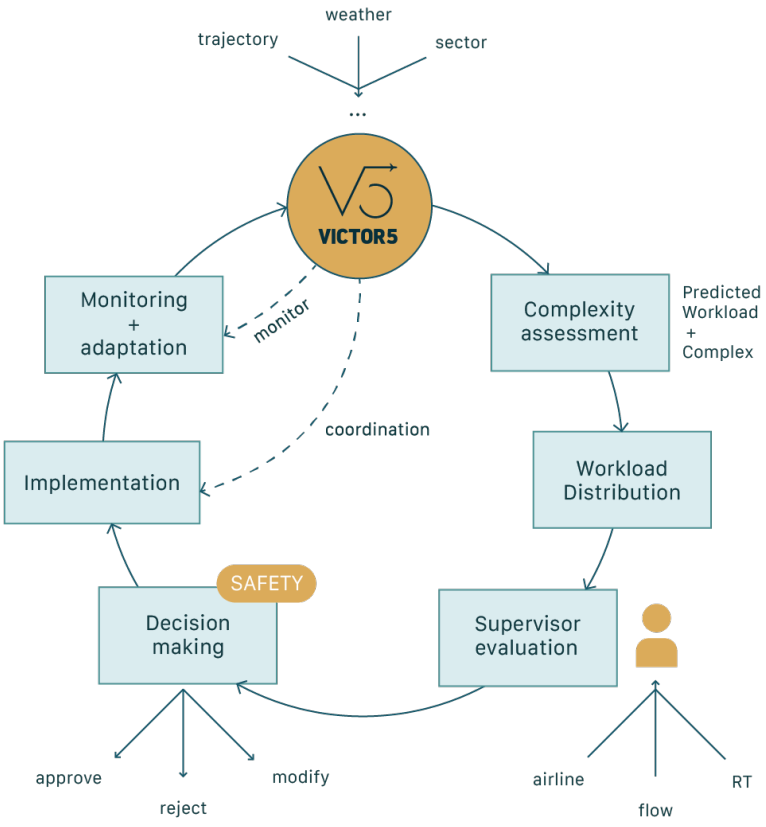
The digital assistant would be equipped with advanced algorithms and models to analyze and predict the complexity of air traffic. It would consider factors such as traffic volume, evolving traffic, potential interactions, conflicting routes, and weather conditions. By incorporating real-time data on winds, aircraft types and performance, complexity of adjacent ACCs, weather forecasts, and other relevant information, the digital assistant would provide a holistic understanding of the operational environment.

With its computational capabilities, the digital assistant would be able to process large volumes of data and generate accurate predictions. It would assist supervisors and operational room heads in making informed decisions about sector regulations, traffic distribution, and resource allocation. By optimizing the workload management process, the digital assistant would help prevent both wasted capacity and safety issues due to underestimated complexity.

Furthermore, the digital assistant would continuously learn and adapt to changing conditions and patterns in air traffic. It would leverage machine learning techniques to improve its predictive capabilities over time, enhancing its ability to anticipate and manage workload and complexity effectively. By constantly analyzing data and patterns, the digital assistant would provide valuable insights and recommendations to support decision-making processes in the operational room.

By harnessing advanced computational capabilities, accessing a wide range of information, and leveraging machine learning techniques, this digital assistant would empower supervisors and operational room heads to make informed decisions, optimize workload management, and ensure the safe and efficient flow of air traffic in en-route ATC operations.

A step-by-step use case illustrating how the concept of **advanced sector and workload management using the digital assistant** could be as follows:



1. **Data Collection and Analysis:** The digital assistant continuously collects and analyzes real-time data, including traffic volume, evolving traffic, weather conditions, airspace restrictions, and other relevant information. It utilizes advanced algorithms and models to predict complexity and workload based on this data.

2. **Complexity Assessment:** The digital assistant assesses the complexity of each sector based on factors such as potential interactions, conflicting routes, and airspace restrictions. It considers the predicted workload and complexity for each sector, taking into account the specific characteristics of the airspace and traffic patterns.

3. **Workload Distribution:** The digital assistant provides recommendations to the supervisor regarding workload distribution. It suggests potential sector closures, openings, mergers, or changes in sector configuration to optimize workload management while ensuring safety and minimizing complexity. These recommendations are based on the predicted complexity and workload, as well as the available resources and expertise of air traffic controllers.

4. **Supervisor Evaluation**: The supervisor evaluates the recommendations provided by the digital assistant. They consider the predicted workload, complexity, and safety implications of the proposed changes. The supervisor may also take into account other factors such as airline preferences, traffic flow management data, and real-time weather updates.

5. **Decision-Making:** Based on the evaluation, the supervisor makes informed decisions regarding sector and workload management. They may approve the proposed changes, modify them based on their expertise and judgment, or reject them if they deem them unsafe or impractical. The supervisor ensures that safety remains the top priority while optimizing workload distribution and complexity management.

6. **Implementatio**n: Once the decisions are made, the necessary changes in sector configuration, airspace allocation, and traffic distribution are implemented. The digital assistant assists in coordinating these changes, providing real-time updates and guidance to air traffic controllers and other relevant stakeholders.

7. **Monitoring and Adaptation:** The digital assistant continuously monitors the implemented changes and their impact on workload and complexity. It collects feedback and data from air traffic controllers and other sources to assess the effectiveness of the decisions made. The digital assistant adapts and learns from this feedback, refining its predictive capabilities and recommendations for future workload management.

Within SafeTeam, and based on the use case definition developed under D3.1, we will limit the role of the digital assistant as a decision support tool, providing enriched information to the supervisor to take informed decisions based on the complexity assessment of the sectors and how the complexity of the air sectors is distributed along the day for the different configurations. The SafeTeam Digital Assistant would therefore provide to the supervisor the different possible configurations with an assigned complexity indicator per sector (or sector partition) for the next blocks of time. This information would help supervisor to balance demand and capacity and to decide on the most optimal configuration based on the complexity assessment. Accordingly, the SafeTeam DA would cover steps 1 to 6. Steps 7 and 8 would be part on a future evolution of the system. By following this advanced sector and workload management concept, utilizing the capabilities of the digital assistant, supervisors can make data-driven decisions, optimize workload distribution, and ensure the safe and efficient management of complexity in en-route Air Traffic Control operations.

The implementation of the SafeTeam digital assistant for sector and workload management targeting TRL6 involves **several specific challenges** that we will address. The DA development focuses on:

1. **Data Integration and Accessibility**: This includes real-time traffic data (ADS-B surveillance data), weather information (wind data and storm data) and airspace restrictions. The Victor5 data platform and interface ensures seamless integration and accessibility of these diverse data sources.

2. **Computational Power and Scalability:** To process large volumes of data, perform complex algorithms and predictive modeling. Victor5 cloud-based infrastructure facilitates scalability. This infrastructure has implemented Uber's H3 (Hexagonal Hierarchical Spatial Index) model. H3 is an open-source geospatial indexing system designed to efficiently partition and index geographic space. Our application of this model is used to divide the airspace volume into hexagonal cells, to paralelize the problem of conflicts detection. It enables advanced geospatial analysis, fast queries, and seamless visualization, even for real time analysis problem (like monitoring the distance among each pair of aircraft).

3. **Accuracy and Reliability:** Although the use case presented is not planned to be used by executive controllers (with high safety critical level), ensuring the accuracy and reliability of the system is key to build trust on the tool. In this line, the current deliverable presents the validation exercises performed with air traffic controllers, its methodology and results. This exercises have been key to validate and calibrate the metric and the tool.

4. **Human-Machine Interaction:** Requires a carefully designed, professional interface. The challenge lies in creating a system that effectively communicates complex information to supervisors and operational room heads, allowing them to understand and evaluate the recommendations provided by the assistant. Balancing the level of automation and human control is also a challenge to ensure the system is trusted and accepted by the specialized audience in ATC.

5. **Regulatory and Safety Compliance:** Implementing a digital assistant in the operational room of ATC requires compliance with strict regulatory and safety standards, even if the tool is not involved in a safety critical application. Ensuring that the system meets these standards and safety-critical considerations, is a significant challenge. Close collaboration with regulatory bodies (as part of work package 5) and adherence to industry best practices is essential to overcome these challenges.

1.2 Sierra5 - Complexity Management tool based on Victor5 platform

The proposed concept would therefore be a digital assistant built on the Victor5 platform, designed to address the challenges of sector and workload management in en-route ATC. It would be specifically developed to overcome the complexities associated with data integration, computational power, accuracy, human-machine interaction, and regulatory compliance. The name given to this tool is Sierra5, using the "S" for "supervisor" according to the aeronautical alphabet and the "5" according to the targeted level of autonomy in the Sheridan scale (see Deliverable 2.1).

To tackle the challenge of data integration and accessibility, the DA shall leverage robust data management systems and interfaces. It shall seamlessly integrate real-time traffic data, weather information and airspace restrictions, ensuring supervisors and operational room heads have comprehensive and up-to-date information at their fingertips.

The proposed Digital Assistant overcomes the challenge of computational power and scalability by harnessing cloud-based solutions and distributed computing architectures. This enables the assistant to process large volumes of data, perform complex algorithms, and scale effectively as air traffic volume increases, ensuring optimal performance and responsiveness.

With a strong focus on accuracy and reliability, the system will utilize validated algorithms and models that are continuously improved and calibrated. The system undergoes extensive testing and validation against real-world scenarios, incorporating feedback from air traffic controllers and supervisors to enhance accuracy and provide reliable predictions and recommendations.

This technology addresses the challenge of human-machine interaction by offering an intuitive and user-friendly interface. It shall effectively communicate complex information to supervisors and operational room heads, enabling them to understand and evaluate the recommendations provided. The system strikes the right balance between automation and human control, ensuring trust and acceptance among the specialized audience in ATC.

In the future, this technology is poised to advance the sector and workload management in en-route ATC. By addressing the key challenges, it will offer a comprehensive and advanced digital assistant that empowers supervisors and operational room heads to make informed decisions, optimize workload distribution, and ensure the safe and efficient flow of air traffic.

2 Sierra5 complexity metric

2.1 Current complexity metrics

The enhanced tactical flow management system (ETFMS) is the Eurocontrol Network Manager system responsible of the coordination and optimize air traffic flow across Europe (Eurocontrol, 2025). To provide this service, it holds two main tasks:

- Calculation of traffic demand in every sector of the NM area of operations, using the flight plan information received from the aircraft operators (AOs).
- Computer-assisted slot allocation (CASA) and distribution of the resulting list of slots to all parties involved: ANSPs, airlines and airports mainly.

It is therefore a core system to balance demand and capacity in the European airspace, including air sectors and airports. As a result of this, ANSPs receive the predicted demand and it will be updated using real-time surveillance data provided by the ANSPs, together with sector configurations, declared capacities and possible regulations or other constraints provided by the ANSPs of the NM area.

To provide this service, this Eurocontrol unit can only monitor the number of flight flights entering a sector once an hour (the entry counts) as well as the number of flights which are actually present in a given sector every minute (the sector occupancy counts).

The ETFM predictions are used by the ATC supervisors to decide sector configurations and personnel rostering during the shift. Accordingly, **this decision is generally made based on these two simple metrics: entry counts and occupancy counts. No information about the complexity of those traffic interactions is available to support the decision-making process of the supervisor.**

On another hand, there is a vast amount of literature about complexity measurement. While there is not a single agreed definition of complexity across the sector (Gianazza, 2017), researchers agree that traffic density is the most important metric to assess traffic complexity. At the same time, it is demonstrated that this parameter alone does not adequately capture or represent the workload of ATCOs and, therefore, has several limitations (Isufaj, 2022). Intense research has been dedicated to the understanding of additional metrics related to complexity and workload (Dmochowski, 2017), even independently of human performance (Perez Moreno, 2022) and a detailed relation is presented in the next section. However, the applicability of those enhanced metrics is limited by the availability of the data required to calculate the relevant parameters, especially of the application requires real time data processing. As an example, a complexity metric was defined to assess the "Applicability of Current Metrics for Benchmarking Purposes" (Standfuß, 2020) using parameters like vertical interactions (for traffics in evolution), horizontal interactions (for headings divergences), speed interactions (for differences in speed), or traffic concentration (flow characteristics) to compare and rank the complexity of the traffic operated by the different ANSPs. This application is related to an historical analysis and, therefore, is not limited by the real-time availability of the data. Additionally, it uses BADA modelled data to calculate the parameters, no real data.

In the current use case application, we are developing a complexity metric to be used by an ATC supervisor at pre-tactical level and, therefore, the defined metric (and the parameters used for it) needs to be available at that level of operations in a control centre. According to Skybrary (2025), "The ATC shift supervisor is a person who is operationally responsible for an ATS unit for the duration of the shift." As the concrete tasks and roles are not prescribe by ICAO, some variations among countries can exist and described in the corresponding Manual of operations. In any case, its routine duties typically include, among others (Skybrary, 2025):

- Choosing the sector configuration - the supervisor usually has final authority on the decision to open new or merge existing sectors, the configuration to be used (e.g. horizontally or vertically split sectors), etc. This includes the decision when (and if) to use single person operations or combine/split working positions (e.g. planner and executive controller). This is the main task SafeTeam DA is assisting.

- Personnel rostering during the shift, determining which controller(s) works in which sector and when. The assessment provided by our digital assistant is intended to be used to support the allocation of resources according to the most accurate estimation possible, on the traffic and its complexity.

Thanks to the adaptation of Victor5 for this purpose, we can enrich the metrics provided by Eurocontrol with parameters based on traffics interactions that can be available real time in an ACC. Other parameters might, in addition, complement our metric and become even more accurate, however, they are not real-time available or impose other legal regulations that limit its usage.

2.2 Parameters to compute traffic complexity

2.2.1 Introduction

Traffic complexity, a multifaceted construct, is influenced by various factors, each playing a critical role in the workload experienced by Air Traffic Controllers (ATCOs). This exploration delves into the traditional key parameters that contribute to the computation of traffic complexity, laying the foundation for an innovative approach capable of comprehensively addressing the challenges in modern en-route ATC.

2.2.2 Factors Influencing Traffic Complexity

1. Number of Planes Overflying an ATC Sector

- **Volume of Air Traffic:** Total aircraft passing through the airspace sector.
- **Traffic Density:** Closeness of aircraft within the sector.
- **Time Distribution:** Distribution of flights over specific time intervals.

2. Airspace Structure

- **Complex Aircraft Routings:** Intricacy and diversity of flight paths.
- **Impact of Restricted Areas and Warning Areas:** Extent of sector impact by restricted zones and associated activities.
- **Size of Sector Airspace:** Physical dimensions of the airspace under sector jurisdiction.
- **Intersecting Flight Paths:** Number and complexity of flight paths intersecting within the sector.
- **Impact of Airline Hubbing:** Extent of sector impact by airline hubbing or major terminal/airport traffic.

3. Radio Congestion

- **Frequency Congestion:** Challenges posed by congestion on radio frequencies.

4. Longitudinal Sequencing and Spacing

- **Sequencing Complexity:** Need for longitudinal sequencing and spacing of aircraft within the sector.

5. Climbing or Descending Traffic

- **Vertical Traffic Movement:** Amount and frequency of aircraft ascending or descending within the sector.

6. Aircraft Mix and Military Flights

- **Types of Aircraft:** Variability in aircraft types, including VFR, IFR, props, turboprops, and jets.
- **Military Aircraft Presence:** Frequency and impact of military flights within the sector.

7. Multiple Functions, Required Procedures, and Coordination

- **Controller Roles:** Variety of tasks a controller must perform, such as approach control, terminal feeder, en route, and in-trail spacing.
- **Procedural Complexity:** Number and complexity of procedures controllers must execute.
- **Interfacing Complexity:** Level of coordination required with adjacent sectors, approach controls, center, and military units.

8. Weather-Related Factors

- **Weather Impact:** How weather conditions affect air traffic control operations.

9. Adequacy and Reliability of Radio and Radar Coverage

- **Communication and Surveillance:** Reliability and coverage of communication and surveillance systems.

10. Use of digital assistants

- To advise of potential conflicts.
- To advise of resolutions to potential conflicts.
- To probe instructions.

2.2.3 Sierra5 Approach

In tackling the complexities of en-route Air Traffic Control (ATC), **Sierra5 pioneers a novel approach that centers on addressing the consequences of the previous factors**. Until now, numerically calculating these consequences has posed a significant challenge, shifting the focus from the causes to their tangible outcomes.

1. Consequences of Traditional Factors and Time Distribution

- **Evolution of Air Traffic Dynamics:** Sierra5 acknowledges the changing landscape of air traffic, where direct routes are increasingly common, disrupting traditional airspace structures. While factors like airspace structure, traffic evolution, and mix of traffic were traditionally influential in workload calculations, as mentioned, Sierra5 emphasizes their consequences rather than their direct impact.
- **Dividing Time Calculations into Smaller Intervals:** Sierra5 enhances accuracy by dividing hours into smaller intervals for more precise time distribution analysis. This approach allows for a finer understanding of traffic patterns and potential workload variations throughout the day.

2. Differentiation by Type of Traffic and Validating Factors by Controllers

While the number of planes remains a powerful variable in workload management, not all flights create the same level of workload for ATCOs. Recognizing this, Sierra5 introduces a groundbreaking approach to differentiate aircraft into three distinct groups:

1. **Aircraft with Potential Conflicts (MTCD)**: This group includes flights that may be involved in potential conflicts or interactions with other aircraft. To identify these flights, Sierra5 employs Medium Term Conflict Detection (MTCD) algorithms, which detect situations where aircraft may come too close in terms of time, even if no separation maneuver is required. This group imposes a high workload on ATCOs.
2. **Non-Conflicting Traffic (NCT)**: NCT refers to aircraft that do not require specific instructions, fly at flight levels with no traffic congestion, and maintain safe distances from destination airports and other aircraft. While experienced controllers naturally recognize this type of traffic, Sierra5's intelligent algorithm can also identify NCT. These flights require less ATCO intervention and contribute to a workload lower than the average.
3. **Other Flights**: This category encompasses the remaining flights that may not be involved in MTCD but do not meet the strict criteria to be classified as NCT. The workload generated by this group is considered medium.

Incorporating MTCD and NCT functionalites into the workload management strategies of Air Traffic Control Centers (ACCs) is essential to effectively navigate the increasing complexity of en-route ATC. By detecting potential conflicts and distinguishing non-conflicting aircraft, ACCs can optimize resource allocation and allow controllers to focus on critical tasks. This integration enhances safety, reduces the cognitive load on controllers, and ultimately enables more efficient workload management in en-route ATC operations.

These factors, crucial in Sierra5's approach, will undergo **validation by controllers as illustrated in D4.1.4**. This validation ensures that the computed traffic complexity aligns with the practical experiences and expertise of controllers, enhancing the reliability and effectiveness of Sierra5's workload management approach.

3. Additional Consideration: Radio Communications

Sierra5 recognizes the Dynamic Nature of Communications. The number of **radio communications is inherently tied to the number of aircraft and the type of traffic** (an aircraft with a potential conflict ahead is more likely to receive more instructions from ATC). Approaches that just rely on communication metrics should be approached with caution, as they may not capture critical situations accurately. For instance, when aircraft are legally separated but close in proximity, the ATCO may refrain from giving additional instructions, instead actively monitoring the situation to ensure the minimum separation is maintained.

2.3 Sierra5 complexity metric: KPIs and formula

2.3.1 Key Performance Indicators (KPIs)

As mentioned before, Sierra5 identifies three fundamental KPIs -number of aircraft in or crossing the sector, potential conflicts assigned to the sector, and Non-Conflicting Traffic (NCTs)- that underpin the calculation of workload and complexity. These KPIs serve as the cornerstone for Sierra5's data-driven approach to complexity assessment.

Sierra5 employs a standardized period of 20 minutes for all its calculations. This duration strikes a balance, allowing for meaningful assessments without making the operational room's sector configuration adjustments impractical in smaller intervals. For each sector, Sierra5 calculates, for the next 20, 40, and 60 minutes, the traffic count of each sector, the number of potential conflicts per sector, and the traffic categorized as NCT. The subsequent development will delve into the specific formula guiding Sierra5's computation of these key metrics.

2.3.2 Enhancing Complexity Calculation with Weighted Factors

In the assessment of air traffic controller workload and sector complexity, it's crucial to recognize that not all components contribute equally to the overall workload. Factors such as the number of aircraft crossing the sector, involved in potential conflicts, and Non-Conflicting Traffic (NCTs) have varying degrees of impact on controller responsibilities.

1. Number of Aircraft in or crossing the Sector: While the number of aircraft provides a foundational understanding of traffic volume, Sierra5's workload calculation formula goes beyond a simple count. It takes into account the specific characteristics of each flight, such as its proximity to other aircraft and the potential for conflicts. Aircraft closer to potential conflicts contribute more significantly to workload.

2. Number of Potential Conflicts: Potential conflicts, as identified by MTCD algorithms, are a significant workload factor. **Sierra5 assigns higher weight to flights involved in potential conflicts**, acknowledging that controllers must allocate more attention and resources to manage these situations effectively.

3. Non-Conflicting Traffic (NCTs): NCTs represent flights that require minimal intervention from controllers due to their adherence to predetermined parameters. Sierra5 considers NCTs with a workload factor lower than the average, recognizing that these flights demand less direct controller involvement.

2.3.3 Variability in Complexity Calculation Formulas

It's important to note that various methodologies exist for calculating air traffic controller workload, and these methods may vary among service providers. However, Sierra5's approach is designed to offer a comprehensive and standardized means of complexity assessment that can be widely adopted across the industry based on real-time accessible sources of data.

Sierra5's key variables—number of aircraft, potential conflicts, and Non-Conflicting Traffic (NCTs)—serve as fundamental Key Performance Indicators (KPIs) that underpin the calculation of workload and complexity. While specific formulas may differ, the industry's move towards a more data-driven and standardized approach necessitates the inclusion of these key variables for accurate and meaningful workload assessment.

2.3.4 Integrating Declared Capacity

To bridge the transition to this innovative tool, Sierra5 can be linked with current practices by incorporating the declared capacity of the sector. This capacity represents the maximum number of aircraft that can be safely managed within the sector, taking into account factors such as airspace configuration and controller capabilities. By aligning Sierra5's workload assessment with declared capacity, the tool becomes a valuable resource for ATC, aiding controllers in effectively managing traffic and maintaining safety.

2.3.5 Formula Development: Mimicking Air Traffic Controller Insights

Sierra5 employs a formula, one of many possible, aimed at mimicking the results derived from complexity polls conducted by controllers. In these polls, a group of Air Traffic Controllers (ATCOs) assesses workload under different scenarios, varying the number of aircraft (ACFT), potential conflicts, and Non-Conflicting Traffic (NCTs), while also incorporating the declared capacity as a reference.

The formula is essentially a curve-fitting endeavor, seeking to replicate the nuanced understanding controllers possess about workload dynamics. By aligning Sierra5's calculations with the insights gathered from these complexity polls, the tool strives to provide a workload assessment that closely mirrors the real-world experiences and expertise of controllers. The methodology and results of these exercises is reported in Annex I.

This approach ensures that Sierra5's formula isn't a rigid, one-size-fits-all model but rather a dynamic and adaptable framework, aligning with the varied perspectives of controllers operating in diverse scenarios. It underlines Sierra5's commitment to offering a tool that resonates with the practical insights of those on the front lines of Air Traffic Control.

2.3.6 Complexity Formula

DATA:

- **CPH** (capacity per hour),
- **T** (total traffic during period of time),
- **P** (number of PCs during period of time),
- **N** (number of NCTs during period of time).

The Sierra5 formula calculates en route air traffic control complexity by summing three weighted KPIs: total traffic (T) at 40%, potential conflicts (P) at 40%, and non-conflicting traffic (N) at 20%. These percentages reflect each KPI's impact, with higher non-conflicting traffic reducing complexity, subject to specific restrictions.

The weights were determined based on input from experienced air traffic controllers, whose insights highlighted that total traffic and potential conflicts significantly drive workload and stress, justifying their higher 40% weights, while non-conflicting traffic, though important for reducing complexity, contributes less directly, thus assigned a 20% weight. It's validation is presented in section 1.4.

Scaled **capacity (C)** to window size (minutes) (i.e. 20 minutes):

$$C = CPH \cdot \frac{window\ size\ (minutes)}{60}$$

T contribution

$$X = \frac{T}{C}\ if\ \frac{T}{C} < 1.5\ else, 2$$

P contribution

$$Y = (\frac{P^2}{C \cdot T})^{1/3}$$

N contribution

$$Z = 1 - \frac{N}{T}$$

Complexity (W)

$$W = MAX(1, MIN(2X + 2Y + Z, 5))$$

Complexity (W) ranges between 1-5

Note: STDEV 15% against Complexity Poll

2.4 Advantages of the Sierra5 approach

Sierra5 brings a paradigm shift to en-route Air Traffic Control, introducing a suite of advantages that redefine the way workload is managed. From standardization and systematic operations to historical complexity insights and formula flexibility, Sierra5 offers an array of innovative features:

1. **Standardization:** Sierra5 introduces a standardized complexity calculation procedure, ensuring consistency across different sectors. This standardization facilitates the establishment of a common method, enabling the comparison of results across various Flight Information Centers (FICs). This uniformity promotes a cohesive and standardized approach to workload assessment in the field of Air Traffic Control (ATC).

2. **Systematic and Automated:** Sierra5 operates systematically and is designed for minimal dependency on human intervention. Its automated capabilities allow for a fully automatic workload calculation process. By reducing reliance on manual inputs, Sierra5 enhances efficiency, accuracy, and operational reliability in en-route ATC.

3. **Historical Complexity Calculation:** Sierra5 goes beyond current workload assessments by providing a historical complexity calculation. By leveraging historical data, the tool can establish meaningful ranges of workload, offering a valuable proxy for sector capacity. This historical perspective enables ATC professionals to analyze trends, anticipate challenges, and optimize resource allocation.

4. **Flexibility with Formula Adoption:** Sierra5 offers flexibility by allowing the adoption of different complexity formulas based on the same foundational Key Performance Indicators (KPIs). This adaptability empowers ATC organizations to tailor the complexity assessment approach to their specific needs, accommodating varying operational contexts and preferences.

5. **Comprehensive Workload Factors:** Sierra5 incorporates factors essential to valuing Air Traffic Controller (ATCO) workload that were previously not fully considered. By accounting for the number of potential conflicts, Non-Conflicting Traffic (NCT), and other critical variables, Sierra5 provides a more comprehensive and accurate representation of the workload dynamics in en-route ATC operations.

6. **Enhanced Decision Support:** Sierra5 serves as a powerful decision support tool, offering valuable insights into workload patterns and trends. ATC supervisors and operational room heads can leverage Sierra5's outputs to make informed decisions, optimize workload distribution, and ensure the safe and efficient flow of air traffic.

7. **Resource Optimization:** The utilization of Sierra5 enables Air Traffic Control Centers (ACCs) to optimize resource allocation based on real-time and historical workload assessments. This optimization contributes to improved operational efficiency, reduced cognitive load on controllers, and enhanced overall airspace management.

3 Sierra5 implementation

3.1 Sierra5 infrastructure and data sources

3.1.1 Infrastructure

The Sierra5 project leverages Automatic Dependent Surveillance-Broadcast (ADS-B) data to compute key performance indicators—total traffic (T), potential conflicts (P), and non-conflicting traffic (N)—for assessing air traffic control complexity. These KPIs are processed and analyzed within a scalable cloud-based infrastructure designed to handle high volumes of real-time aviation data efficiently. The system relies on cloud-based databases to store and manage incoming ADS-B data streams, ensuring low-latency access and robust data handling. Calculations for T, P, and N are performed using optimized algorithms deployed on cloud compute instances, enabling rapid processing of surveillance data, flight plan information, and wind models. This cloud architecture supports scalability to accommodate varying traffic demands across different airspaces. Results are transmitted from the cloud to a web-based user interface, accessible via a standard internet connection without requiring specialized hardware or high-bandwidth networks. The user interface, which visualizes complexity metrics for air traffic controllers, will be detailed in a separate chapter. Security measures, on the data transmission and access controls, ensure the integrity and confidentiality of sensitive aviation data throughout the process.

3.1.2 Data Sources

In order to effectively use the algorithms to detect the number of aircraft, traffic involved in potential conflicts, and non-conflicting traffic, the following data is needed:

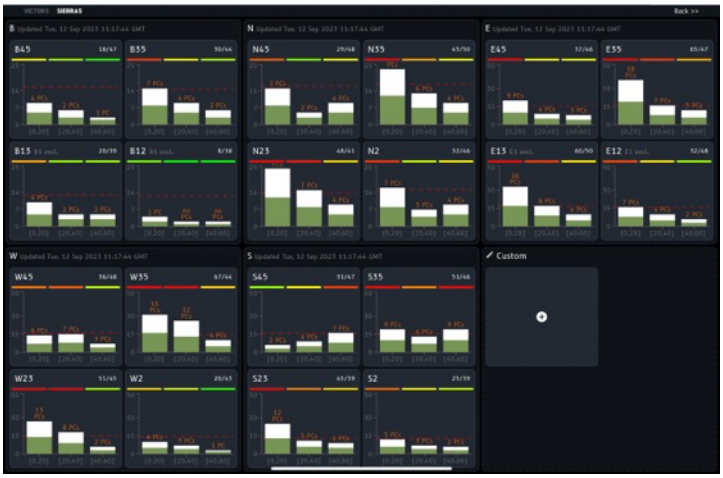
- **Surveillance Data**: Includes position, flight level, selected flight level, ground speed, and additional parameters. The most straightforward way to access this data is through an Automatic Dependent Surveillance-Broadcast (ADS-B) provider equipped with extended squitter capabilities, which provides access to Mode-S data. This ensures comprehensive and precise tracking of aircraft in real time.

- **Flight Plan Information**: Essential for extrapolating aircraft positions to support the calculation of Medium-Term Conflict Detection (MTCD) and Non-Conflicting Traffic (NCT) algorithms. Flight plans provide critical intent data, enabling the system to predict trajectories and assess potential conflicts accurately.

- **Wind Map at All Flight Levels**: Wind conditions have a significant impact on aircraft ground speed, particularly during turns or when encountering varying wind patterns at different altitudes. To account for these effects, access to a comprehensive wind map across all flight levels is essential. While some providers offer rough estimates of wind conditions at various layers of airspace, a more accurate wind map can be generated based on information obtained from ADS-B. Many aircraft transmit real-time wind readings through the extended squitter, allowing for the creation of a detailed and dynamic wind model.

3.2 Sierra5 visual interface and dashboards

3.2.1 Visual interface. Pre-tactical tool

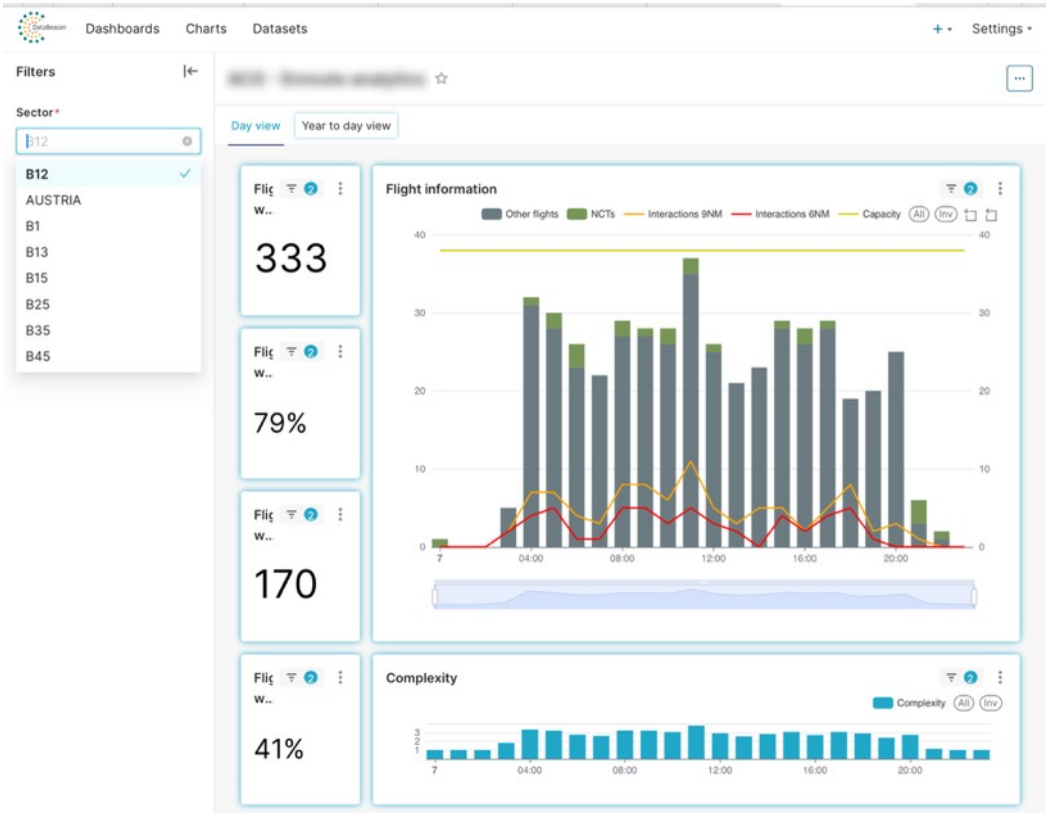


The Sierra5 pre-tactical tool is designed to enhance en route air traffic management by calculating and visualizing airspace complexity up to one hour in advance, enabling supervisors to optimize sector configurations and balance controller workloads. As already mentioned, unlike traditional methods that rely solely on the number of aircraft, Sierra5 incorporates multiple key performance indicators (KPIs)—total traffic (T), potential conflicts (P), and non-conflicting traffic (N)—to compute a comprehensive complexity metric. This metric supports more accurate decision-making for re-sectorization, a daily process

critical for maintaining safety and minimizing regulations that lead to costly airline delays. **By automating the calculation of complexity metrics, Sierra5 reduces reliance on manual assessments**. The tool graphically represents each KPI to maintain transparency and avoid the "black box" effect, allowing supervisors to clearly understand the factors influencing complexity. For example, an en route supervisor can use this information to determine the best sector combinations (e.g., closing one sector to open another) when controller resources are limited, ensuring efficient and safe airspace management.

The pre-tactical tool's user interface is web-based, accessible on any computer following secure authentication, and tailored for supervisors as the primary users. It visually compares pairs of potential sector configurations within the same horizontal constraint, making it intuitive to identify which combination achieves a more balanced workload for air traffic controllers (ATCOs). For instance, volume B can be horizontally divided into two sectors when two controllers are assigned: (B13, B45) or (B12, B35). The interface displays each sector's complexity across 20-minute intervals, with each interval represented by a column. A multicolored segment above each column indicates the overall complexity metric, with colors signaling workload intensity (e.g., red for high complexity in B35's first 20 minutes, suggesting (B13, B45) as a better configuration for workload balance). The height of each column corresponds to the total number of aircraft, with Non-Conflicting Traffic (NCTs) highlighted in green for clarity. The number of potential conflicts (PCs), calculated automatically by Sierra5's Medium-Term Conflict Detection (MTCD) algorithm, is displayed atop each column. A red dotted horizontal line marks the sector's declared capacity (aircraft per hour, divided by three for the 20-minute intervals), providing a reference for assessing traffic loads. This interface, focused on real-time pre-tactical planning, is complemented by a separate post-analysis interface with historical dashboards.

3.2.2 Post-analysis dashboards



The Sierra5 post-analysis dashboards enable managers to conduct in-depth reviews of historical air traffic data, supporting safety analysis, workload assessment, and sectorization optimization. All relevant data—complexity metrics, potential conflicts, interactions (e.g., aircraft proximity within 6 or 9 nautical miles), and controller maneuvers (vectors or flight level changes)—is stored for post-analysis. Users can select any day and any sector or combination of sectors via a typing or scrolling interface, providing flexibility to analyze specific scenarios. The dashboards track maneuvers likely linked to potential conflicts, allowing safety managers to investigate incidents, assess controller workload, and identify preventive measures for future operations. Additionally, the tool supports evaluating whether sectorizations can be improved under similar circumstances, enhancing future airspace management. **The data collected also serves as a valuable resource for training machine learning models to optimize sectorization**, further aligning with Sierra5's focus on automatism through automated data processing and analysis. This web-based interface, accessible on any computer after secure authentication, is designed for managers, including safety managers, to drive data-informed decisions.

The post-analysis dashboards share a similar visual structure to the pre-tactical tool, presenting data in a clear, web-based format. Users can filter by sector and toggle between day or year-to-day views to analyze trends over time.

Key metrics are displayed in dedicated panels, summarizing total flights, flights with interactions at specified distances, and other relevant indicators. A flight information chart visualizes data over a 24-hour period, with bars segmented to show other flights, non-conflicting traffic, and interactions at different proximity thresholds, alongside a capacity reference line. A separate complexity chart below displays complexity metrics across the same timeframe, with bars indicating variations in complexity levels. This intuitive layout ensures managers can efficiently review historical data, assess workload drivers, and identify opportunities for operational improvements.

4 Sierra5 Validation

The Validation of the use case 1 on ATC complexity includes two phases:

1. Validation of the complexity metric defined in Section 2 and the KPIs used.
2. Validation of the Sierra5 pre-tactical tool defined in Section 3 together with its functionalities.

Each of validation exercises, together with the corresponding results are presented in the following sections of this chapter.

4.1 Validation 1: complexity metric and KPIs

Objective

The first experiment was designed to evaluate the accuracy and relevance of Sierra5's proposed complexity metric factoring: total traffic (T), potential conflicts (P), and non-conflicting traffic (N) in assessing air traffic controllers' perception of an environment's complexity. This step does not pertain to the evaluation of the human-machine cooperation between a supervisor and Sierra5. Its objective was to evaluate the relevance of the complexity measure proposed by Sierra5. Particularly, the experiment intended to compare Sierra5's traffic complexity evaluation with those of Air Traffic Controllers for different scenarios (i.e., different traffic). This process goal is to ensure that the tool's metrics align with operational realities by comparing Sierra5's calculated complexity (W) against ATCO's perceptions using scenarios based on real air traffic recordings.

Scenarios & independent variables selection

The first step involved selecting different scenarios for evaluation purposes. Two elements were considered for this selection:

6. Explore as independently as possible the influence of each KPI used by Sierra5 (i.e., total traffic or T, potential conflicts or P, and non-conflicting traffic or N).
7. Don't limit our evaluation to the manipulation of the KPIs.

In that sense, a first approach was to define a set of KPIs value as a control condition (i.e., $T = a$, $P = b$, $N = c$), and derive from it 6 conditions with a higher or lower variation of only one of those KPI's value (e.g., $T = a$, $P = b$, $N = \boldsymbol{d}$; as illustrated in the table below). This set of value was then used to extract real data and select scenarios with visualization of the traffic using real air traffic data from the Austrian airspace.

Variables	Total traffic (T)			Potential conflicts (P)			Non-conflicting traffic (N)		
Conditions	Low	Med	High	Low	Med	High	Low	Med	High
Control		X			X			X	
T	X				X			X	
			X		X			X	
P		X		X				X	
		X				X		X	
N		X			X		X		
		X			X				X

Table 1Approach 1 for Scenario's selection with independent variables and conditions

Despite the selection of various sets of KPIs' value respecting the proposed structure, we did not find in the available dataset corresponding scenarios. This could be due to the relative dependence between the different KPIs or to the limited amount of data available.

We have therefore proposed a second approach based on currently available data. We explored the range of complexity covered by the Sierra5's metric and its associated range of KPI's values. In order to compare ATCO perceived complexity and Sierra5 complexity metrics, we chose 7 complexity values from 1.5 to 4.5 in steps of 0.5, and tried to identify scenarios where the complexity values returned by Sierra5 came as close as possible to these target values. We selected scenarios that best satisfied our first constraint: for each KPI, find different scenarios where the complexity value mainly depends on this KPI, while minimizing the contribution of the two others. In addition and in order to evaluate other factors potentially involved in perceived complexity, we decided to test the impact of a sector's own complexity (which might be affected by its geometry, traffic dynamic, closeness to airports, etc.). In that sense, we selected 3 sectors within the Austrian airspace based on an equal traffic capacity within 15 minutes.

To summarize, we manipulated 2 different factors (i.e., independent variables):

- ***Complexity*** attributed by Sierra5 (7 values from 1.5 to 4.5; ± 0.1)
  For each complexity value, 3 repetitions with a goal to have a local maximum of one of the KPI, while minimizing the contribution of the two others.
- ***Sectors*** considered (3 different ones, E35, S35 & N35)

For each sector, we tried to distribute the local maximum of each KPI (e.g., 2 max T, 3 max P and 2 max N). Only for the S35 sector, we could not find more than a scenario with a local maximum NCT in contrast to the other conditions.

It resulted in 21 trials, with control variables values reported in the table below:

Variables / Conditions	Complexity		T	P	N	Sector
1	1.5	1.56	6	1	5	E35
2		1.5	9	0	8	N35
3		1.46	3	0	0	S35
4	2	2.03	10	1	9	E35
5		1.9	4	1	1	N35
6		2.06	11	0	7	S35
7	2.5	2.54	13	0	6	E35
8		2.5	7	2	2	N35
9		2.44	11	1	7	S35
10	3	2.94	14	1	8	E35
11		2.9	9	3	3	N35
12		3.04	15	0	4	S35
13	3.5	3.55	17	2	10	E35
14		3.55	19	0	7	N35
15		3.59	10	5	1	S35
16	4	3.98	19	1	5	E35
17		3.92	17	4	9	N35
18		3.98	13	5	1	S35
19	4.5	4.52	25	0	12	E35
20		4.56	20	1	15	N35
21		4.45	15	6	0	S35

Table 2. Approach 2 for Scenario's selection with description of independent variables and conditions

Metric / Dependent variables selection

Because it was not possible to replicate a realistic air traffic controlling task which would have allowed for an accurate assessment of ATCO's workload. We therefore considered a simplified approach of ATCO's **perception of the complexity** of air traffic controlling scenarios to eventually leverage on the quantity of data and participants we could gather.

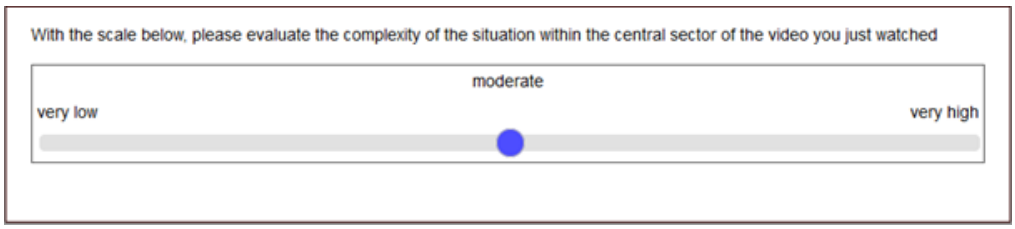


Figure 1. Scale used for the evaluation of the perception of the complexity of each scenario.

Each complexity perception was evaluated on a 1 to 5 continuous scale (*Fig.1*), equivalent to the range of complexity offered by Sierra5's complexity metric. Participants were not restricted on their response time. In addition, each session ended with an equivalent rating (same continuous scale from 1 to 5) of the complexity perception of the sector itself.

Material

Figure 2. Screenshot of a scenario's video, centered on the sector to be considered for evaluation of its' perceived complexity

Each scenario was originally 15-minute scenario video (visual of one of the scenarios in *Fig.2*), accelerated to 5 minutes (thrice the speed) to optimize trial quantity within a minimal time frame while maintaining participant engagement. The videos were uploaded on the Youtube platform, inaccessible with the public search bar but only with unique URLs. As to control the quantity and the quality of information provided to each participant, we chose to rely on video recordings without option to replay, speed up or down, or even pause the content.

This allowed to implement the experimental protocol using the PsyToolkit platform, enabling to program experiments and surveys for academic studies, and make them accessible online to a broader number of targeted participants. Recruiting ATCOs is challenging due to their demanding schedules, so the experiment aimed to be easy to perform on their own time. It could be completed on computers, tablets or smartphones (no headphone needed) without any experimenter present as every instruction required were displayed prior to the evaluation.

The survey was aimed at any Air Traffic Controllers (ATCOs) with various levels of trainings and experience, to bring their operational expertise to the process. To maximize participation, the project used LinkedIn promotions, presented at aviation congresses, and leveraged professional networks.

Participants were informed on the project, the objective of this experiment, their rights regarding their participation and their data, and asked for their consent to participate, their data to be collected and analyzed by the ONERA

research team, and the results to be shared within the European Project. Links to the researchers contact information, a data protection officer of ONERA, and to the European Project official webpage were also made available.

Data collection

Participating ATCOs, having provided informed consent, first completed a survey detailing their expertise and demographics. In this introduction, they were also assigned a unique ID to ensure their anonymity and dataset continuity through the different sessions. They were required to report their ID at every experimental session.

Each participant was presented with the 21 scenarios divided in 3 sessions of 35 mins. To be familiarized with the structure and specificities of the different sectors considered, each session was centered on one sector only (i.e, E35, N35 or S35). To avoid impacts of the order of presentation, sessions were randomized between participants, and conditions within sessions. In addition, participants were asked to perform one session within one sitting, while they could realize the three different sessions on different days. However, they were made aware that the full completions of the 3 sessions was required to analyze their set of data.

Participants were then instructed on the main task as reported below:

> "*You are going to be presented 7 air traffic videos within a sector to provide you with an overview of a scenario and enable you to make an evaluation regarding the complexity of this situation.*
>
> *We ask you to watch entirely each video in full screen without any replay, speeding or pause. Once each video finishes, close the YouTube window by clicking on the X, then press Continue.*
>
> *[…]*
>
> *You should focus on monitoring the traffic affecting the central sector, not the full screen, FL335 and above. The task may be passive but we want you to adopt the controller's point of view. Imagine measuring and solving potential conflicts and the efforts associated with those tasks.*
>
> *You will then be asked to report your perceived complexity of the management of this situation with a scale from "very low" to "very high" complexity."*

Despite the passiveness associated with looking at the video, we wanted them to treat the scenarios as real operations, and try to actively identify potential conflicts in order to elicit realistic perception of potential complex situations requiring management. However, with an online survey, we could not control for their engagement within the task.

Analysis

The main hypothesis of Validation 1 is that Sierra5's complexity metrics accurately assess ATCOs judgment. We will assess this hypothesis by comparisons of the scenario's complexity perception ratings with Sierra5's calculated complexity scores. In order to do so, coefficient of correlation can be calculated between independent and dependent variables. We will also look at the impact of individual KPIs manipulations (T, P or N) on the participants' perception ratings.

Adjacent hypothesis considers that ATCOs judgment will diverge between participants depending on their individual expertise and preferences. The correlations between complexity ratings and metrics will be explored at the participants level, to observe those individual variations depending on the sample of data collected.

Closing hypothesis, assume that the nature of the sector and therefore some additional variables (e.g., geometry, traffic dynamic, closeness to airports, etc.) impacts ATCOs evaluations of the scenarios, not considered by Sierra5. We will therefore consider how sectors complexity perception of the sectors may discriminate variation in complexity perception between them (i.e, E35, N35 or S35), that could be matched to variability in the scenarios evaluation. Therefore, exploration of the coefficient of correlation with or without sector segregation could entail on its impact on complexity evaluation.

Results are documented in report D4.1.4.3

4.2 Validation 2: Sierra5 tool and functionalities

Objective and Test Description

The second validation exercise aimed to focus on Sierra5's cooperativeness. More particularly, it addresses potential human-machine cooperation risks identified through the application of the SafeTeam approach (D2.1.4) to Sierra5's tool. The prior application of the HABA-MABA/LOA assessment (D3.2.2) underlined several potential issues in future development of Sierra5:

- *"Data quality is key to any data-driven tool. The acquisition of data is not a task that the supervisor has to actively monitor but it is possible that some training about the data origin and the overall processing carried out may be necessary for the supervisor to have a better **understanding of the system** and to increase **confidence** and **trust** in it."*
- *"Supervisor does not need to have a deep technical understanding of how the information or predictions have been calculated, although, some high-level training and basic familiarization may be necessary to give them an overview of **what the process does**, in order to better understand any **potential shortcomings** and build **confidence**."*
- *"Having some degree of **understanding of how** the metrics, predictions or solutions have been calculated can be needed."*
- *"**Overreliance** could occur as long as the supervisor continuously uses it to improve the workload balance in the different sectors."*

Those pitfalls are associated with operator's lack in situational awareness and complacency issues. The system development's goal was therefore set to reduce out-of-the-loop-phenomenon, be transparent and predictable, and provide users with situational awareness.

Those problematics directly relate to four of the design principles (D2.1) which can serve as input to the evaluation process (see D2.2.2):

- *Agents should share a common goal*
- *Agents should be able to share their **status** and intentions and observe the intentions of others.*
- *Designers of the system should strive for shared **Situation Awareness***
- *The system should foster mutual **trust***

We therefore, used an experimental approach and sets of metrics described in D2.2 to quantify the pre-tactical tool's ability to support supervisors in optimizing sector configurations by leveraging the automated calculation of complexity, compared to traditional methods based solely on total aircraft count (T); but also, to test the potential of the Sierra5 tool to foster situational awareness, and trust though two levels of explainability of the complexity metric.

Scenarios & independent variables selection

For this exercise, a simple airspace was created, divided into three columns (A, B, C), each sliced into a lower (1) and upper (2) level, resulting in six different sectors: A1, A2, B1, B2, C1, and C2. Sectorization rules permit merging of adjacent columns (e.g., AB or BC) or lower and upper sectors within the same column (e.g., A1 and A2) as seen in *Fig.3*. Note that diagonal sectorization (e.g., A and C) is not allowed, as it does not occur in real-world air traffic control operations. Configurations range from 1 to 6 open sectors, depending on the number of available controllers. For this validation exercise, the airspace was to be allocated to two Air Traffic controllers, as to restrict the number of sectorizations solutions to 3 as seen in *Fig.1* (i.e., (i.e., 1) [AB][C]; 2) [A][BC]; 3) [ABC1][ABC2]).

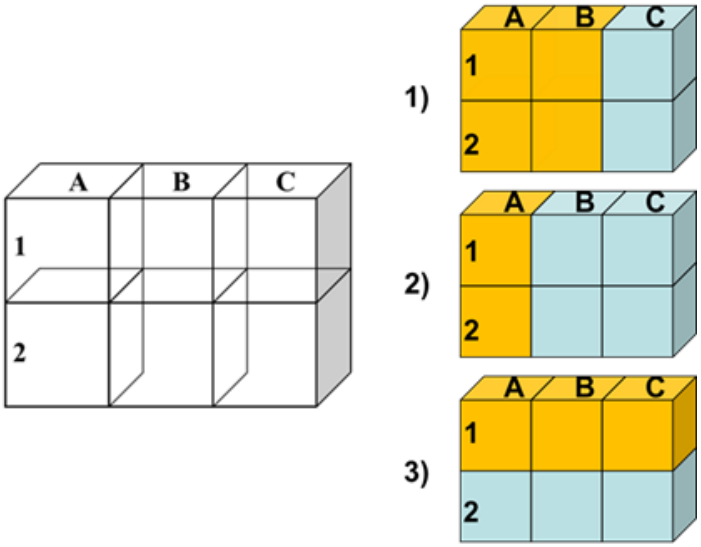


Figure 1. Airspace used in Validation 2 and the corresponding sectorization solutions.

In order to explore Supervisor's ability to optimize sector configurations with or without Sierra5, we selected scenarios for the upcoming hour, derived from real historical data within the Austrian airspace database made available for the SafeTeam project, reflecting various traffic and complexity conditions. We chose complexity metrics from 2 to 4 in steps of 0.5, and tried to identify scenarios where the optimal sectorization solution depended on mainly one of the KPI (i.e., the total traffic (T), potential conflicts (P), and non-conflicting traffic (N)), while minimizing the contribution of the two others.

In addition, in order to evaluate other factors potentially involved in airspace sectorization, we decided to consider the dynamic of the KPI within the predicted hour of each scenario. This manipulation is only applied to the Explained experimental condition (described below), where the complexity metric is accompanied by a graphical representation of the KPIs value and additionally their predicted distribution through 3 periods of 20 mins as reported in *Fig.2*. We defined three manipulations of this dynamic:

- "**Positive dynamic**" as when the dynamic of the KPIs reinforces the optimal solution based on the optimal distribution of the complexity metric
- "**Neutral dynamic**" when no difference in KPI distribution occurred between solutions
- "**Alternative dynamic**" when two solutions (of the three presented) were closely optimal in complexity distribution but the dynamic of the KPIs favored the less optimal one (between the two optimal solutions)

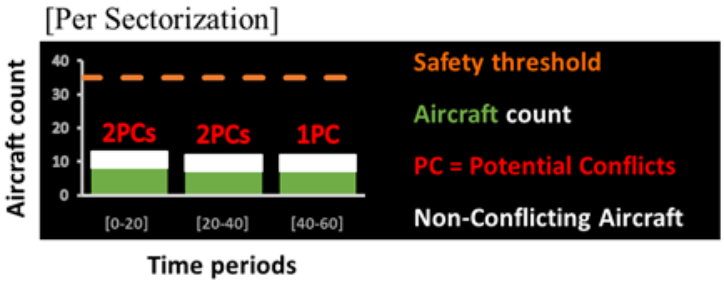


Figure 2. Display of the KPIs values and dynamic through three twenty minutes periods provided for each sectorization solutions considered within the Explained condition

Those manipulations enable to test if participant notify differences in dynamic between solutions and scenarios, and if so, if they accept that supplementary information and modify their response.

To summarize, we manipulated 3 different factors (i.e., independent variables extended in *Table. 3*):

- ***Complexity*** attributed by Sierra5 (5 values from 2 to 4; ± 0.5)
- ***KPI manipulated*** between solutions proposed (T, P and N)
- ***KPIs' dynamic*** between solutions proposed (Positive, Dynamic and Neutral)

Variables / Conditions	Complexity	KPI manipulated	Dynamic of the KPI
1	2	T	Alternative
2		P	Neutral
3		N	Positive
4	2.5	T	Positive
5		P	Alternative
6		N	Neutral
7	3	T	Neutral
8		P	Positive
9		N	Alternative
10	3.5	T	Alternative
11		P	Neutral
12		N	Positive
13	4	T	Positive
14		P	Alternative
15		N	Neutral

Table 3. Description of the independent factors manipulated in Evaluation 2

Conditions selection

The evaluation protocol consisted of 3 different conditions:

- **Control condition**: associated with traditional sectorization methodology where participants accessed only to total aircraft count (T) within each sector of the Airspace as described in *Fig.3*
- Conditions with Sierra5's complexity metric (W)
    - **Unexplained Sierra5 condition:** associating T to Sierra5's assessment of each sectorization solution. This condition displays 2 complexity metric per solution (i.e., one complexity per new sector) described in *Fig.4*
    - **Explained Sierra5 condition:** associating T, Sierra5's complexity metrics and the description of the KPIs (T, potential conflicts P, non-conflicting traffic N) and their dynamic as reported in *Fig.2* and *Fig.5*

The Explained condition allowed to explore the impact of transparency within the calculation of the complexity metric. It also enables to test the impact of their dynamic, to foster trust within the complexity metric, but also avoid overreliance if there are potential issues with the data quality considered or within the calculation of the complexity – avoiding overreliance in the system.

	A	B	C
1	10	8	15
2	9	12	22

1)	A	B	C
1	[AB]		[C]
2			

2)	A	B	C
1	[A]	[BC]	
2			

3)	A	B	C
1	[ABC1]		
2	[ABC2]		

Figure 3. Example of a Control condition

	A	B	C
1	22	12	9
2	15	8	10

1)	A	B	C
1	5.0		2.1
2			

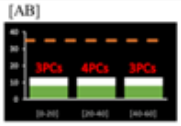
2)	A	B	C
1	3.2	3.7	
2			

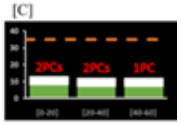
3)	A	B	C
1	3.5		
2	3.5		

Figure 4. Example of an unexplained complexity condition

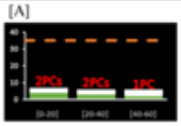
	A	B	C
1	10	8	15
2	9	12	22

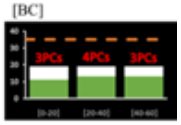
1)	A	B	C
1	3.7		3.2
2			



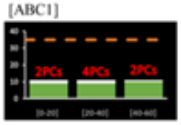


2)	A	B	C
1	2.1	5.0	
2			





3)	A	B	C
1	3.5		
2	3.5		



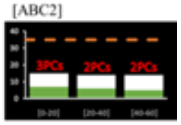


Figure 5. Example of an Explained complexity condition with neutral dynamic

Metrics / Independent variable selection

As described in the introduction to this second Evaluation, the experimental design was developed to quantify Sierra5's complexity metric ability to support supervisors in optimizing sector configurations compared to traditional methods based solely on total aircraft count (T); but also, to test its potential of the Sierra5 to foster situational awareness, and trust. The selection of the relevant metrics was based on the approach and resources listed in D2.2. It was also constrained by the level of development of the system, and the online platform of experiment. Three different dimensions has been explored:

- **Team performance**:

We recorded responses provided by the participants to each scenario presented.

- **Confidence and Trust**

After completion of participants response (for each condition and rounds), they were invited to report their "Confidence in my answer" - defined as the belief that your choice was correct based on the available evidence. They were further asked to report their agreement with "I trust the complexity metric provided" – defined as the belief that the complexity metric give an accurate assessment of the complexity of the situation. Dynamic reporting of Confidence and Trust were both assessed on a 1 to 5 scale from Disagree to Agree.

- **Acceptability and Explainability**

End of the experiment's questionnaires centered on the subjective quality of the Explained display through two important dimensions of the human-machine cooperation. We used the Acceptability scale rating 9 components on a 5 points Likert scale as captured in *Fig.6* . It explores the Usefulness and Satisfaction perceived by participants while interacting with a system.

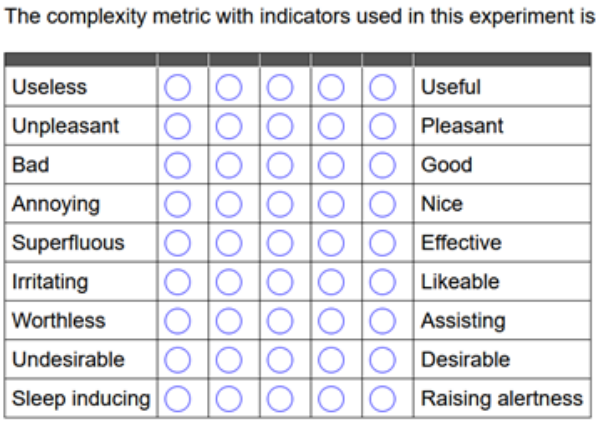


Figure 6. Acceptability scale presented at the end of the experiment regarding the complexity metric within the Explained display

The second questionnaire included in this Evaluation is the Explanation satisfaction scale. Participants have to rate 7 components on a 5 points Likert scale between "*Strongly disagree*" and "*Strongly agree*" as presented in *Fig.7*.

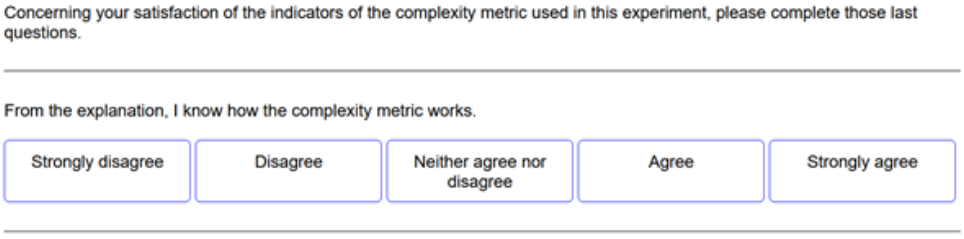


Figure 7. One item of the explanation satisfaction scale presented at the end of the experiment regarding the display proposed in the Explained condition

The main instruction and components are reported below:

> *« Concerning your satisfaction of the indicators of the complexity metric used in this experiment, please complete those last questions."*

"1. From the explanation, I know how the complexity metric works.

2. The explanation of how the complexity metric work is satisfying.

3. This explanation of how the complexity metric work has sufficient detail.

4. This explanation of how the complexity metric work seems complete.

5. This explanation of how the complexity metric work tells me how to use it.

6. This explanation of how the complexity metric works is useful to my goals.

7. This explanation shows me how accurate the complexity metric is."

Material

Each scenario and displays were created based on prior Sierra5's displays propositions. The protocol was implemented with the PsyToolkit platform as a survey, allowing for easy access online to a broad number of targeted participants. Recruiting supervisors is even more challenging than ATCOs as there are even fewer, so the experiment aimed to be easy to perform on their own time. It could be completed on computers, tablets or smartphones (no headphone needed) without any experimenter present as every instruction required were displayed prior to the evaluation.

Participants were informed on the project, the objective of this experiment their rights regarding their participation and their data, and asked for their consent to participate, their data to be collected and analyzed by the ONERA research team, and the results to be shared within the European Project. Links to the researchers contact information, a data protection officer of ONERA, and to the European Project official webpage were also made available.

Data collection

Participating Supervisors, having provided informed consent, first completed a survey detailing their expertise and demographics. Each dataset was assigned to a unique ID to ensure their anonymity.

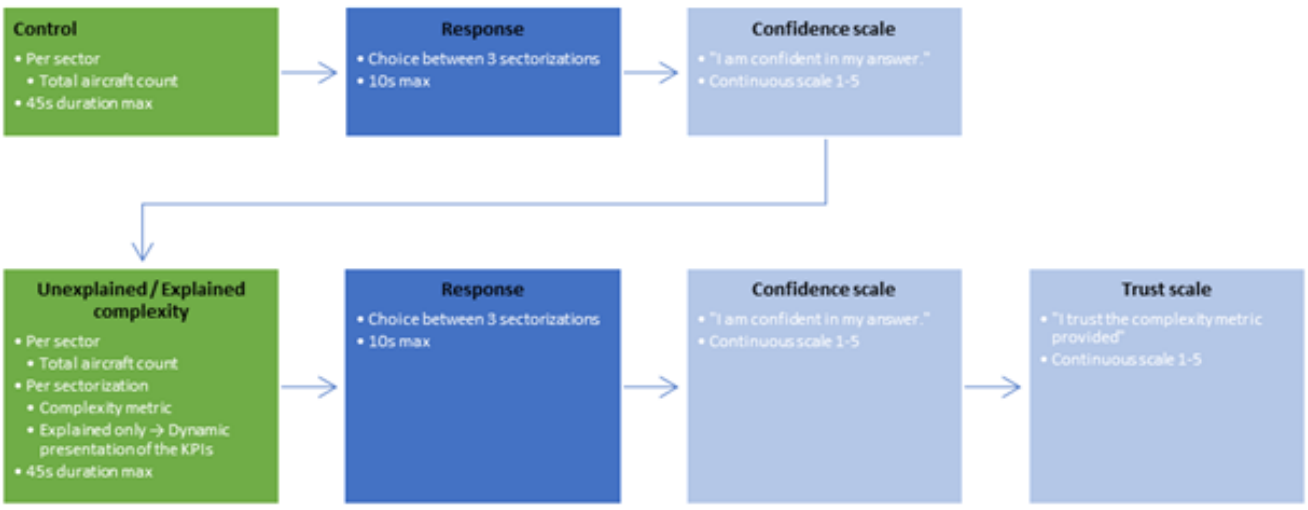


Figure 1 Figure 8. Trial structure

For analysis purposes the Control conditions were merged with Sierra5's complexity metrics condition into 2 total conditions. Each trial shared the same structure illustrated in *Fig.8*, where participants were presented each scenario in two rounds:

- **First round:** Control display
- **Second Round**: Unexplained or Explained display

Within the introduction of the survey, participants were instructed on their sectorization task, the two rounds of evaluation per trial, with the different kind of resources made available (i.e., T for control, W for Unexplained/Explained, and KPI dynamics for Explained only) with appropriate definitions and illustrations to familiarize them with the material. They were also informed on the different questions they would have to complete following each trial (Confidence and Trust ratings), and at the end of the experiment (Acceptability and Explainability ratings). They had the opportunity to complete one training trial before moving on to the experiment.

Each of the 15 conditions were presented twice: 1 for the pair Control – Unexplained and 1 for the pair Control – Explained. In order to avoid familiarization to repeated scenarios, we reversed the values horizontally and vertically between the two conditions so to do not change the complexity metric and dynamic but modify the optimal solution position. We also pseudo-randomized the conditions so as to avoid repetition of the same scenario within one block, by distributing them in two blocks of randomized conditions (e.g., block 1 = even number conditions of Control–Unexplained and odd number conditions of Control–Explained, block 2 = odd number conditions of Control–Unexplained and even number of Control–Explained).

After a maximum presentation time of each display for 45s, as to constrain and balance out the duration of information gathering between participants, they had to select their response between the 3 sectorizations solutions proposed. They then assessed their Confidence in their answer, and for Sierra5's displays their Trust in the complexity metric (W).

They were able to take a pause (without time constrain) at the middle of the experiment. When done with the 30 trials total, they completed both the Acceptability and Explainability ratings concerning the Explained complexity metric display.

Analysis

· Team performance

We first explore how the presence of the Sierra5 complexity metric modulate participant choice. To do this, we can examine the number of changes between the initial choice (Control round) and the choice in the presence of the complexity metric W (complexity metric round). Our hypothesis is that adding Sierra5's complexity metric (Unexplained & Explained condition) should generate reversal between the choice in control and complexity metric rounds, but also that explanation could boost this change rate

The second hypothesis is that giving additional information (KPIs T, P and N and their dynamic) to the display of Sierra5's complexity metric would help supervisors to understand this tool and improve their decision-making process. To do that, we defined the best option amongst the three solutions proposed based on both the complexity metric and the distribution of the three KPIs, then we compare it to supervisors' choice regarding this best option for each condition and round (i.e., Control + Unexplained/Explained complexity metric).

Finally, we specifically explored how the presentation of the dynamic of the KPIs (i.e., evolution through 3 periods of 20 minutes of their predicted distribution) impacts supervisors' accuracy. To do that, we manipulated the adequation between this dynamic and the choice relevant regarding the complexity metric only, as to test if participants would perceive this change in dynamic and modify their choices. If supervisors use this dynamical information, a same level of performance for Positive and Alternative dynamics should be observed (that is, supervisor should change their choice toward the choice supported by this dynamical information). To test that we compare the sum of participants selections of solutions depending on the dynamic manipulation within the Explained condition only.

· Confidence and Trust

The following hypothesis considering Confidence and Trust was raised from the potential risk in human-machine cooperation that the integration of Sierra5's system could lead to. Based on design principles (D2.1), confidence in participant's answer (the belief that your choice was correct based on the available evidence) and trust in the complexity metric (the belief that the complexity metric give an accurate assessment of the complexity of the situation) is expected to be improved through the addition of Sierra5's tool (i.e., Control vs. Unexplained/Explained), even further by the explanation of how does Sierra5's tool work (i.e., Unexplained vs. Explained). We therefore examine Confidence ratings as the average difference between within-trial evaluation (i.e., Confidence in Control answer – Confidence in Unexplained/Explained answer), and the Trust as the average of ratings per displays (i.e., Unexplained vs. Explained trust ratings).

· Acceptability and Explainability

The last part of Evaluation 2 is focused on the subjective quality of the Explained display through two important dimensions of the human-machine cooperation. The first dimension explore the Usefulness and the Satisfaction associated with the tool assessed. We adjust the Acceptability scale scores from (-2) to (+2) and averaged within two subscales: items 1, 3, 5, 7 and 9 for the Usefulness scale; items 2, 4, 6 and 8 for the Satisfying scale. The Explainability scale scores are also extracted and adjusted from (-2) to (2).

Results are documented in report D4.1.4.3.

4.3 Assessment and conclusions

The validation polls and exercises for Sierra5 are essential to ensure that the tool aligns with real-world ATC requirements and effectively assists ATCOs in workload management. The insights and feedback gathered from ATCOs will be instrumental in refining Sierra5, making it a valuable addition to the ATC toolkit. The validation process will ultimately contribute to safer, more efficient, and more reliable air traffic control operations.

4.4 Conclusions Validation 1

The first validation experiment was designed to evaluate the accuracy and relevance of Sierra5's proposed complexity metric factoring: total traffic (T), potential conflicts (P), and non-conflicting traffic (N) in assessing air traffic controllers' perception of an environment's complexity. Participants were asked to report their perceived complexity of some air traffic scenarios of accelerated data recorded from the Austrian airspace. Despite the efforts to ease the access to the experimental protocol (online survey) and to spread its communication to the targeted audience through various general and specific network, securing willing participants was a lengthy and resource-intensive process, highlighting the difficulty of conducting such validations in an operational context.

We collected complete datasets from 8 participants (48.9 year old ± 5.1 years; 7 males), with various level of expertise (2 ATC supervisors, 2 ATC instructors, 4 Radar ATCs) and countries of activity (Czech Republic, Bahrain, Greece, Singapore, South Africa & Spain). Due to the limited amount of participant, the analysis of the data is mainly qualitative and cannot rely on most statistical tools. It also prevents us from considering individual variations in complexity perception ratings.

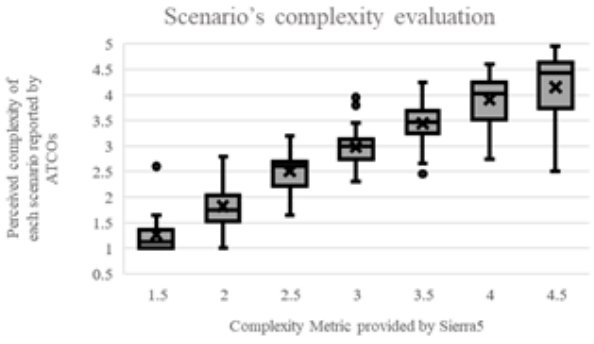


Figure 1. Distribution of dependent ATCOs perceived complexity of the scenarios presented, given the independent complexity metric calculated by Sierra5

Our main hypothesis was that Sierra5's calculated complexity metric based on T, P & N, would be comparable to operator's evaluation of the scenarios. Participants coefficient of correlation between the Sierra5's complexity metric calculated by Sierra5 and the complexity ratings made by the participants lead to an average r = .90 (SD = .07). A one-sample T-test between those coefficients and 0 – the hypothesis that there is no correlation between those two complexities – results in $p = 1.1x10^{-9}$. This score is notable and support the hypothesis of a good fit between Sierra5's calculation and Controllers perceived evaluation of the complexity of air traffic scenarios as observed in *Fig.1*. A replication of this evaluation would be valuable on a larger sample of ATCOs to draw general conclusion.

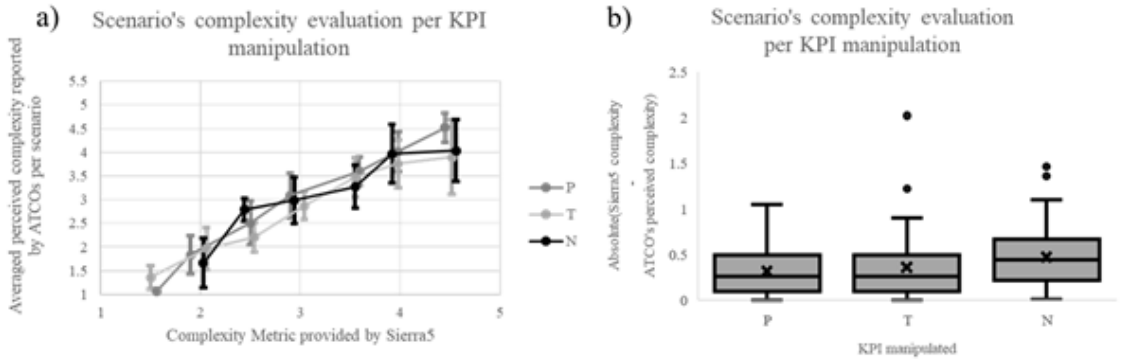


Figure 2. Scenario's complexity evaluation per KPI manipulation

We did test the impact of KPIs manipulations on the observations of perceived complexity as seen in *Fig.2*. The manipulations of T, P and N (see *a)*) do not seem to have an impact on the relationship between complexity metric controlled and measured. Further consideration of the distance between Sierra5's calculated complexity and ATCOs rated complexity against each KPI manipulation in *Fig 2. b)*, may imply larger differences between Sierra5's complexity and participants' perceived complexity when manipulating Non-Conflicting Traffic (N). This observation is supported by a small difference in coefficient of correlation between trials with N manipulations' r = .94 in contrast to T's r = .99 and P manipulations' r = .99. There might be a smaller fit between those metrics but this observation is quite limited per the quantity of participants and the selection of scenarios. As described in D4.1 4.1, we were not successful in selecting scenarios with the most independent manipulations of KPIs contributions so the results observed in *Fig.2* may be more complex and rely on interactions between the three KPIs selected.

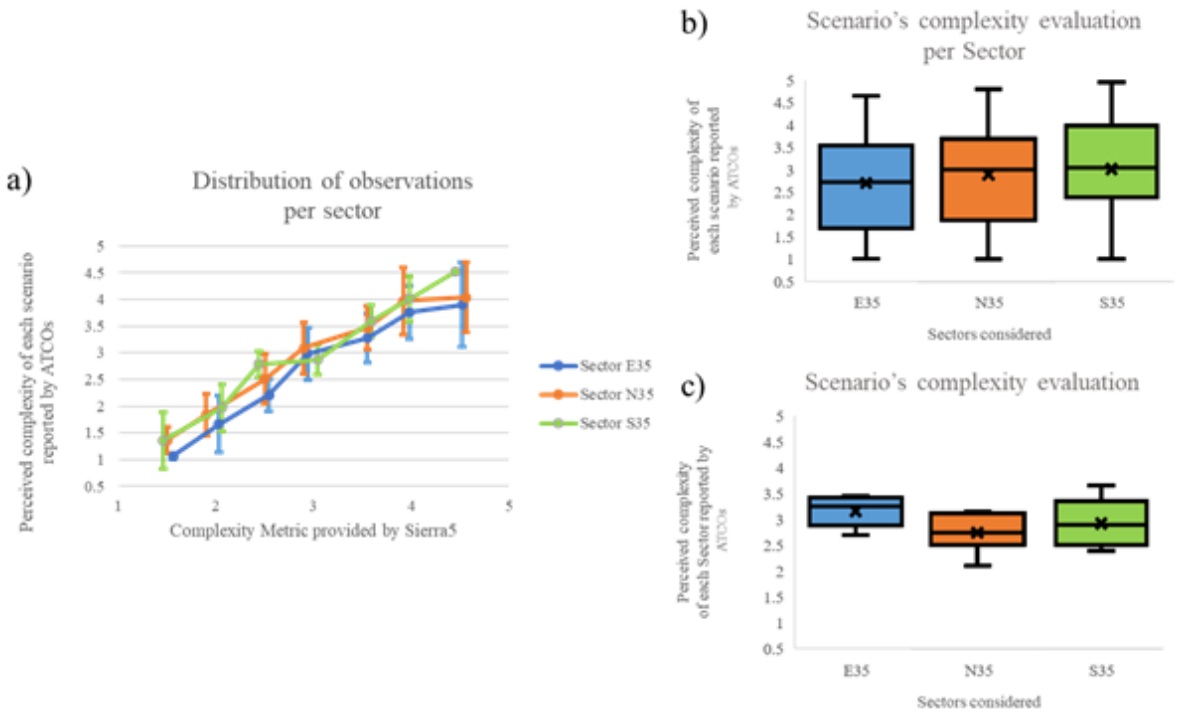


Figure 2 Figure 3. Distribution of dependent ATCOs perceived complexity of a) & b) the scenarios or c) the sector presented given the sectors considered within the experiment

The last hypothesis assumed that further variables impact operator's ratings such as the complexity associated to a sector itself. Despite some slight differences in ratings of sector's complexity at the end of the experiment (as seen in *Fig.3 c)*), with slightly higher average ratings for E35 in contrast to N35 & S35, this independent variable did not seem to impact participant's perception of the scenario's complexity as illustrated in *a)* in between scenarios, or *b)* at the sector level. Coefficient of correlations are high for each sector responses E35 r = .89, N35 r = .89 and S35 r = .94. Differences between S35 and the other sectors correlation might be explained by a more accurate evaluation of the 4.5 complexity metric scenario as seen on *Fig.3 a)*. It would be interesting to explore this local difference in evaluation of the high complexity scenarios.

As for validation of the correlation between complexity metrics, protocol's replications with larger samples and the consideration of further independent variables (e.g., nature of the potential conflicts, meteorological events, …) not associated with the calculation of Sierra5's (i.e., P, T, & N) complexity metric is recommended.

4.5 Conclusions Validation 2

The second validation exercise aimed to focus on Sierra5's cooperativeness. More particularly, we aim to assess:

- Sierra5's ability to support Air Traffic Supervisors in their task to optimize sector configurations through the complexity measurement provided;
- Sierra5's ability to address potential human-machine cooperation risk such as situational awareness and trust through two levels of explainability of the complexity metric.

Participants to this exercise were presented with an airspace and solutions of sectorizations. Their task is to choose the optimal sectorization from 3 possible choices. A first choice (Control round) was performed with only the total aircraft count (T) presented per sector. Following this first choice, participants received the calculated complexity metric as a new information and had to decide again the optimal sectorization (complexity metric round). Two different levels of information were available during this second round: an "Unexplained condition" where complexity metric was presented with only T count per sectors, and an "Explained condition" where the complexity metric was displayed with an overview of the supporting KPIs. The protocol addressed variations in objective performance and subjective evaluations of participants experience of the tool.

Experienced supervisors participated in the validation, bringing expertise in airspace management. We collected complete datasets from 6 participants (50.7 year old ± 8.1 years; 5 males), all supervisors with additional expertise as ATC instructors, ATC examiner and radar ATCO and various countries of activity (3 from Hungary, 2 from Spain, 1 from Italy). Recruiting supervisors was challenging due to their operational commitments, requiring outreach efforts such as LinkedIn promotions, aviation congress presentations, and professional networking. Despite these efforts, securing participants remained a time-intensive process. Due to the limited amount of participant, the analysis of the data is qualitative and cannot rely on most statistical tools.

Team performance

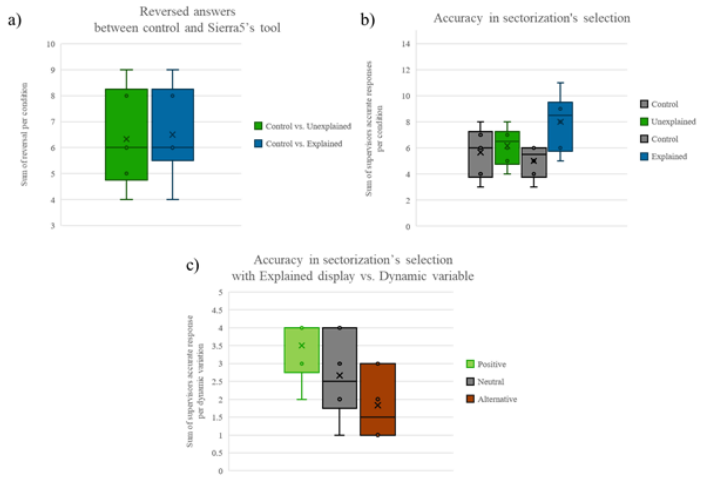


Figure 1. Evaluation of participants performance with traditional approach: T count as control, and with Sierra5's complexity metric Unexplained or Explained

We first explored how the presence of the Sierra5 complexity metric modulated participant choice. To do this, we examined the number of changes between the initial choice (Control round) and the choice in the presence of the complexity metric W (complexity metric round). Our hypothesis is that adding Sierra5's complexity metric (Unexplained & Explained condition) should generate reversal between the choice in Control and complexity metric rounds, but also that explanation could boost this change rate. When observing the average number of reversals in response between the Control round (i.e., T only) and the complexity metric round (i.e., same scenario but addition of the complexity metric, and in the Explained conditions the KPIs) reported in *Fig.1 a)*, supervisors in both Unexplained and Explained condition modified several of their prior selection. Supervisors are consequently modifying their former approach to sectorization by weighing in the complexity metric. Surprisingly, we do not observe any influence of the presence of explanation on this rate of change, indicating that the complexity metric only is enough to modulate supervisor's choice. These findings indicate that supervisors use the complexity measure in their sectorization strategy to optimally divide airspace among available resources, and this whatever the level of explanation.

We then explored the accuracy of this selection. To do that, we defined the best option amongst the three solutions proposed based on both the complexity metric and the distribution of the three KPIs, then we compare it to supervisors' choice regarding this best option. The hypothesis is that giving additional information (i.e., KPIs T, P and N and their dynamic) to the display of Sierra5's complexity metric would help supervisors to understand this tool and improve their decision-making process. We extracted participants answers within each condition and round (i.e., Control + Unexplained/Explained complexity metric) and their accuracy within those scenarios, as reported in *Fig.1 b)*. Results indicate that participants accuracy is higher when presented with complexity metric and detailed values of KPIs and their dynamic through the hour. Supervisors are considering that additional information and making more optimal segregation of airspace judgments.

Finally, we specifically explored how the presentation of the dynamic of the KPIs (i.e., evolution through 3 periods of 20 minutes of their predicted distribution) impacts supervisors' accuracy. To do that, we manipulated the adequation between this dynamic and the relevant choice regarding the complexity metric only. We defined "Positive dynamic" as when the dynamic of the KPIs reinforces the optimal solution based on the complexity metric; "Neutral dynamic" when no difference in KPI distribution occurred between solutions; and "Alternative dynamic" when two solutions (of the three presented) were closely optimal in complexity distribution but the dynamic of the KPIs favored the less optimal one (between the two optimal solutions) as to test if participants would perceive this change in dynamic and modify their choices. If supervisors use this dynamical information, a same level of performance for Positive and Alternative dynamics should be observed (that is, supervisor should change their choice toward the choice supporting by this dynamical information). However, the results show a different pattern. The consequent sum of accurate responses in the Explained condition per the dynamic variable as displayed in *Fig.1 c)*, show higher accuracy with the "Positive dynamic", therefore when this information align with Sierra5 complexity metric distribution between solutions. In contrast, "Alternative dynamic" seem to be associated with lesser accurate responses, which could mean that Supervisors did not take into account this alternative solution and preferred to focus on the optimal complexity metric distribution. It would therefore be important to evaluate how the dynamic of those KPIs may vary in realistic air traffic scenarios, if they play an important role in ATCOs workload, and if so, try to integrate them in the calculation of Sierra5's complexity as participants seem to rely on this element even in presence of the display of KPIs' dynamic.

Confidence and Trust

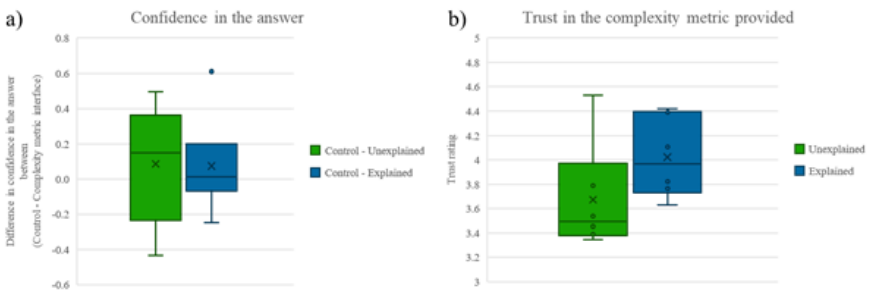


Figure 2. Confidence and Trust ratings between conditions of Evaluation 2

Beyond the impact on performance, an important element concerns the impact of the explanations on the supervisor's attitudes towards the artificial partner. Particularly, we are first interested in change in confidence (the belief that your choice was correct based on the available evidence) and trust (the belief that the complexity metric gives an accurate assessment of the complexity of the situation) regarding the presence or the absence of such explanations. Questions of confidence followed each answer, and Trust each interaction with Sierra5's tool. Based on design principles (D2.1), confidence in participant's answer and trust in the complexity metric is expected to be improved through the addition of Sierra5's tool (i.e., Control vs. Unexplained/Explained), even further by the explanation of how does Sierra5's tool work (i.e., Unexplained vs. Explained).

Confidence is reported as the average difference between within-trial evaluation (i.e., Confidence in Control answer – Confidence in Unexplained/Explained answer). Fig.2 a). In contrast to our hypothesis, results indicate no effect of complexity metric, nor of explanation on the confidence of the supervisors. First, a same level of confidence in supervisors' responses is observed for the control round (without complexity metric) than for the complexity metric round (with the complexity metric). The presentation of the complexity metric does not seem to increase supervisors' confidence in their choice. Similarly, the presence of explanation doesn't appear to improve the average level of confidence. However, we can observe a more distributed range of responses within the Unexplained condition, suggesting a more stable level of confidence in presence of explanation. Taking together, these results indicate that supervisors did not make important changes in their subjective experience of their answer in presence of the complexity metric, nor in presence of explanations. However, the presence of explanations seems to make the level of confidence felt more stable.

Yet, when assessing average evaluation of trust (see *Fig.2 b)*, Trust ratings in the complexity metric seem to be enhanced in the Explained condition. The explanation of the KPIs values used by Sierra5 for the calculation of the complexity metric could play an important and positive role in the interaction of the participants with the system. The explanations would foster the development of trust. This result would benefit from a replication at a larger scale and an examination of the impact of training and expertise with using Sierra5 on the evolution of trust.

Acceptability and Explainability

The last part of Evaluation 2 is focused on the subjective quality of the Explained display through two important dimensions of the human-machine cooperation. The first dimension explore the Usefulness and the Satisfaction associated with the tool assessed. Every 9 items of the Acceptability scale were evaluated positively as reported in *Fig.3*, yet the usefulness of the complexity metric is highly rated (≈ 1.9 on a -2/+2 scale), its satisfaction scale (≈ 1.0 on a -2/+2 scale) seems a bit downgraded. The complexity metric with explanation was therefore well accepted by the 6 supervisors assessed in this Evaluation, and considered as highly useful but some future work may want to consider how to improve the satisfaction experienced by the participants through for example evolution of the display. This observation has to be confirmed at a larger scale with a bigger sample of participants, but more importantly in a realistic environment while performing an airspace management task.

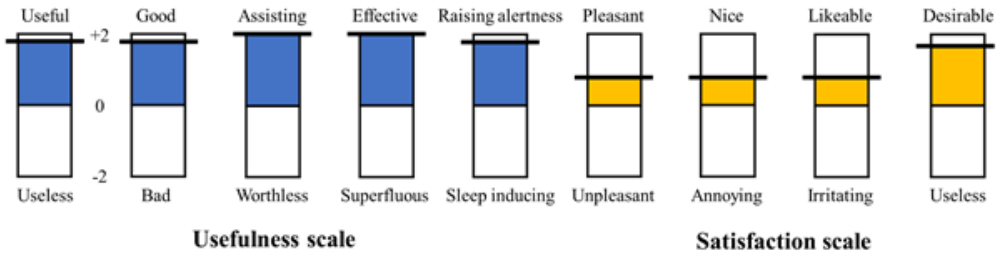


Figure 3 Figure 3. Acceptability scale ratings of the Explained display

The second dimension, through the Explanation Satisfaction Scale explore an a posteriori judgment of an explanation provided. It allows to explore understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy and trustworthiness. The results in *Fig.4* indicate high ratings in each of the 7 items assessed by the participants, translating an overall satisfaction in the explanations provided. A lower score seems to be associated with the last item considering how the system provide information on our accurate the complexity metric is. Future evaluations should entail follow-up questions to understand what information may be lacking or inducing doubt on the accuracy of the complexity metric.

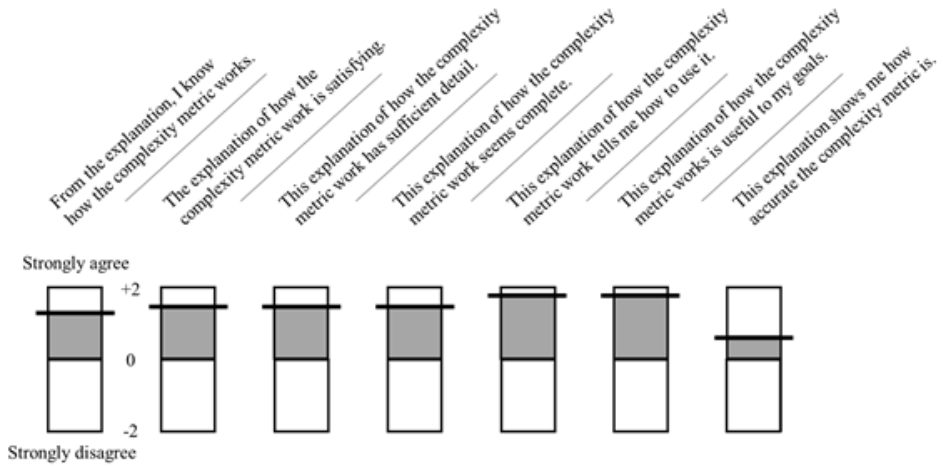


Figure 4. Explanation satisfaction scale ratings of the Explained display

Annex I - Sierra 5 Complexity Metric initial assessment

1 Detailed Explanation of Exercises Conducted with Air Traffic Controllers for Complexity Ranking

Introduction

A series of simulation exercises were conducted as part of a technical evaluation to analyze how ATCOs perceive and rank the complexity of various air traffic scenarios. This study was intended to validate the SafeTeam complexity metric and, ultimately, contribute to the optimization of airspace design and decision-support tools.

Objective of the Exercises

The primary objective of these exercises was to evaluate the subjective (perceived) complexity of real air traffic scenarios as perceived by ATCOs with varying levels of experience. By presenting a range of controlled scenarios with differing traffic compositions, the study aimed to capture insights into the factors that influence perceived workload and situational complexity and, particularly, compare the ATCO's perception on complexity with the SafeTeam calculated metric.

Participants

The exercises involved eight Air Traffic Controllers (ATCO1 to ATCO8) from diverse operational backgrounds, representing a wide range of experience levels (from 8 to 28 years) and with experience in different regions and air navigation service providers (ANSPs), including Spain, Panama, UAE, Bahrain, Qatar, Singapore, and the UK. This diversity ensured a broad spectrum of perspectives influenced by varying operational procedures, airspace characteristics, and cultural factors in ATM practices. The inclusion of controllers with different years of experience allowed for an analysis of how expertise impacts the perception of scenario complexity.

Simulation Setup

The exercises were conducted using a controlled environment reproducing real, historical, air traffic scenarios. Each scenario was characterized by four key parameters that collectively defined the traffic composition and potential conflict load:

1. **Total Number of Flights**: Representing the overall volume of aircraft under control within the simulated sector. This ranged from 20 to 60 flights across the scenarios.
2. **Number of Potential Conflicts Detected (PCDs)**: Indicating pairs of aircraft that were identified as having a risk of loss of separation based on predefined criteria (e.g., proximity in altitude, lateral distance, or converging trajectories). PCDs ranged from 0 to 40 per scenario.
3. **Number of Non-Conflicting Traffics (NCTs)**: Representing aircraft that did not pose any immediate conflict risk with other flights but still required monitoring and coordination. NCTs ranged from 0 to 50.
4. **Rest of Traffics**: Defined as the remaining aircraft that were neither classified as PCDs nor NCTs, ranging from 0 to 30.

A total of 19 distinct scenarios were developed by varying these parameters to create a spectrum of low-to-high complexity situations. Additionally, the scenarios were grouped based on Declared Airspace Capacity, which represents the maximum number of traffics that the simulated sector was designed to handle safely under normal operating conditions. For the first 12 exercises, the declared airspace capacity was set at 50 traffics, while for the following 7 exercises, it was set at 30 traffics. This variation in capacity allowed for an assessment of complexity

perception in relation to sector saturation levels, providing insight into how controllers perceive workload when traffic volumes approach or exceed capacity thresholds.

Methodology

Each ATCO was tasked with evaluating all 19 simulation scenarios independently. After observing each scenario, they were asked to assign a complexity rating on a scale of 0 to 10, where:

- **0** represented an extremely simple scenario with minimal workload or cognitive demand.
- **10** represented an extremely complex scenario with significant workload, high cognitive demand, and challenging conflict resolution requirements.

The ratings provided by each ATCO reflected their subjective assessment of the difficulty in managing the given traffic situation based on factors such as the number of conflicts to resolve, the density of traffic, the need for tactical interventions, and overall situational awareness requirements. **The simulations did not involve real-time control tasks; instead, they focused on observation and post-scenario evaluation to isolate complexity perception from performance metrics.**

Data Collection

For each of the 19 scenarios, data was recorded on:

- The defining parameters (Total Flights, PCDs, NCTs, Rest of Traffics).
- The individual complexity ratings assigned by each of the eight ATCOs.
- The ATCOs profiling characteristics
- The Sierra5 computed metric for each scenario. This resulted in a comprehensive dataset capturing both objective scenario characteristics and subjective controller perceptions. The diversity in ratings across ATCOs also allowed for an exploration of inter-individual variability influenced by experience levels (ranging from 8 to 28 years) and regional operational differences. The inclusion of declared airspace capacity as a parameter further enriched the dataset by providing context on how perceived complexity correlates with sector design limits.

Conclusion

These simulation exercises represent a structured approach to quantifying perceived complexity in air traffic control through direct input from experienced professionals, enabling its comparison with the Sierra5 computed metric. By systematically varying key parameters such as total flights, potential conflicts, non-conflicting traffics, alongside considering declared airspace capacities (50 traffics for the first 12 scenarios and 30 traffics for the last 7 scenarios), a reliable understanding of workload drivers was achieved. Through this exercises, we could confirm the relevance of the KPIs used for the defined Sierra5 metric.

Results

					Workload 1-10							
					ATCO1	ATCO2	ATCO3	ATCO4	ATCO5	ATCO6	ATCO7	ATCO8
				Experience >	25Y	22Y	19Y	14Y	8Y	23Y	14Y	28Y
Capacity	Total	PCD	NCT	Rest	Spain	Spain	Panama	UAE	Bahrain	Qatar	Singapore	UK
50	60	30	10	20	9	9.5	10	10	8	10	10	9
	60	20	30	10	8.5	9	9	9	7	7	9	8
	60	10	30	20	7.8	8	7	8	4	5	8	7
	60	10	50	0	7.9	7	6	7	2	2	8	6
	60	0	40	20	5.5	5	5	1	2	1	7	4
	50	40	5	5	9.5	9	9	8	9	10	10	9
	50	30	10	10	8.8	8.5	7	7	8	8	9	8
	50	10	30	10	6.4	6	6	5	7	4	8	6
	40	30	10	0	8.4	8	8	7	7	7	9	7
	40	20	10	10	8	6.5	6	6	6	6	8	6
	40	10	20	10	7.9	5.5	5	5	5	3	7	5
	30	20	0	10	8	6.5	7	4	5	3	7	6
30	40	0	10	30	4	4	5	1	4	2	7	4
	40	5	20	15	6	4.5	6	1	4	3	8	4
	40	10	20	10	7.8	6	7	3	4	7	9	6
	40	15	0	25	8.3	7	8	4	5	8	10	8
	20	15	5	0	8.1	6.5	7	4	5	5.5	8	5
	20	20	0	0	8.8	7	10	6	6	6	9	6
	20	10	5	5	7	5	8	4	5	4	7	4

S5 Complexity	Poll average	Sigma	Capacity	PCD	NCT	S5 Complexity	Comparison
8.8	9.4	0.7	10.0	0.79	0.83	8.8	93.68%
7.8	8.3	0.9	10.0	0.69	0.50	7.8	93.52%
7.2	6.9	1.5	10.0	0.55	0.50	7.2	105.13%
6.5	5.7	2.4	10.0	0.55	0.17	6.5	113.89%
4.7	3.8	2.2	10.0	0.00	0.33	4.7	122.40%
9.5	9.2	0.7	10.0	0.93	0.90	9.5	103.55%
9.0	8.0	0.7	10.0	0.84	0.80	9.0	111.65%
7.1	6.1	1.2	10.0	0.58	0.40	7.1	118.00%
8.3	7.7	0.8	8.0	0.91	0.75	8.3	108.59%
7.9	6.6	0.9	8.0	0.79	0.75	7.9	120.00%
6.7	5.4	1.5	8.0	0.63	0.50	6.7	123.87%
7.9	5.8	1.7	6.0	0.87	1.00	7.9	135.82%
5.5	3.9	1.8	10.0	0.00	0.75	5.5	141.94%
7.0	4.6	2.1	10.0	0.50	0.50	7.0	153.42%
7.5	6.2	2.0	10.0	0.63	0.50	7.5	120.80%
8.9	7.3	1.9	10.0	0.72	1.00	8.9	121.91%
7.8	6.1	1.5	6.7	0.91	0.75	7.8	127.10%
8.7	7.4	1.7	6.7	1.00	1.00	8.7	117.91%
7.3	5.5	1.6	6.7	0.79	0.75	7.3	133.48%

3 Correlation Analysis and Conclusions from Air Traffic Controller Complexity Rankings

This analysis focuses on the relationships between scenario parameters (Total Flights, Potential Conflicts Detected (PCDs), Non-Conflicting Traffics (NCTs), and Rest of Traffics) and the complexity ratings assigned by the eight ATCOs. Additionally, it address the dispersion of the dataset, the relative ranking of low and high complexity scenarios, and the potential influence of experience (years of service) on the evaluations.

1. Correlation Analysis

The Pearson correlation coefficients between each parameter are calculated together with the average complexity rating across all ATCOs for each of the 19 scenarios. The correlation coefficient ranges from -1 to 1, where a value close to 1 indicates a strong positive relationship, a value close to -1 indicates a strong negative relationship, and a value near 0 suggests little to no linear relationship.

- **Total Flights and Average Complexity Rating**: The correlation coefficient is approximately 0.62, indicating a moderate positive relationship. This suggests that as the total number of flights in a scenario increases, the perceived complexity tends to increase as well. Higher traffic volumes likely contribute to greater monitoring demands and workload.

- **Potential Conflicts Detected (PCDs) and Average Complexity Rating**: The correlation coefficient is approximately 0.85, showing a strong positive relationship. This is one of the most significant findings, as it highlights that the number of potential conflicts is a primary driver of perceived complexity. Scenarios with more PCDs are consistently rated as more complex, reflecting the cognitive demand associated with conflict detection and resolution tasks.

- **Non-Conflicting Traffics (NCTs) and Average Complexity Rating**: The correlation coefficient is approximately -0.15, indicating a weak negative relationship. This suggests that an increase in non-conflicting traffics does not significantly contribute to perceived complexity and may even slightly reduce it.

- **Rest of Traffics and Average Complexity Rating**: The correlation coefficient is approximately -0.25, showing a weak negative relationship. Similar to NCTs, an increase in non-critical traffics does not strongly influence complexity perception, as these aircraft typically do not demand immediate attention or tactical actions.

These correlations indicate that among the parameters tested, the number of Potential Conflicts Detected has the strongest impact on how complex a scenario is perceived by ATCOs, followed by the Total Number of Flights. Parameters like NCTs and Rest of Traffics appear to play a lesser role in driving complexity perception. This result is consistent with the Sierra5 metric, where the PCDs have a squared direct relationship, while the NCTs and the rest of the traffic are inversely proportional.

2. Dispersion of the Dataset

To assess the dispersion or variability in the dataset, it was examined the range and standard deviation of complexity ratings across all ATCOs for each scenario, as well as across scenarios for each ATCO. Dispersion provides insight into the consistency of ratings among controllers and whether there is consensus on perceived complexity.

- **Variability Across ATCOs per Scenario**: For most scenarios, the standard deviation of ratings among the eight ATCOs ranges between 1.0 and 2.5 points on the 0-10 scale. This indicates moderate variability in individual perceptions of complexity. For instance, in high-complexity scenarios (e.g., those with high PCDs), some ATCOs might rate them as 9 or 10, while others might assign a 7 or 8. Similarly, in low-complexity scenarios, ratings might range from 1 to 5. This variability suggests that personal factors such as experience, training background, or regional operational differences may influence subjective assessments.

- **Overall Range Across All Ratings**: The full range of ratings spans from 1 to 10 across all scenarios and ATCOs, showing that the dataset captures perceptions across the entire spectrum of complexity. However, the interquartile range (middle 50% of ratings) is narrower, typically falling between 5 and 8.5, indicating that most ratings cluster around moderate to moderately high complexity levels rather than extreme lows or highs.

The moderate dispersion suggests that while there is general agreement on what constitutes high or low complexity (as evidenced by trends in average ratings), individual differences lead to noticeable variation. This could reflect differing thresholds for workload tolerance or varying interpretations of situational demands.

3. Relative Ranking of Low and High Complexity Scenarios

To understand how low-complexity and high-complexity scenarios are ranked relative to each other, it was computed the average rating for each scenario across all ATCOs and identified the extremes.

- **High-Complexity Scenarios**: Scenarios with higher numbers of PCDs consistently received the highest average complexity ratings. For example, the scenario with 60 total flights and 30 PCDs achieved an average rating of around 9.4, with several ATCOs assigning it a score of 10. Another notable high-complexity scenario was with 50 total flights and 40 PCDs, averaging around 9.2. These results align with the strong positive correlation between PCDs and perceived complexity, confirming that potential conflicts are a dominant factor in workload perception.

- **Low-Complexity Scenarios**: Scenarios with zero or very few PCDs received the lowest average ratings. For instance, the scenario with 60 total flights (the busiest case) but 0 PCDs averaged around 3.9, with some ATCOs rating it as low as 1 or 2. Similarly, a scenario with only 40 total flights, no PCDs, and a high proportion of Rest of Traffics averaged around 3.9. These lower scores suggest that without conflict resolution demands, even relatively high traffic volumes are perceived as manageable.

- **Relative Spread**: High-complexity scenarios generally have average ratings clustering tightly between 8.5 and 9.5, indicating strong consensus on their difficulty. In contrast, low-complexity scenarios show greater spread in averages (ranging from about 3 to 5), suggesting less agreement on exactly how simple these situations are, possibly due to differing interpretations of monitoring workload even in non-conflict situations.

This analysis reveals that while high-complexity rankings are more uniform due to clear drivers like PCDs, low-complexity rankings are more variable, potentially influenced by secondary factors like traffic density or individual controller strategies for handling routine tasks.

4. Influence of Experience on Evaluation

To explore how experience (years of service) might influence complexity evaluations, it was analyzed trends by comparing the ratings of ATCOs grouped by experience levels: less experienced (8-14 years), moderately experienced (19-23 years), and highly experienced (25-28 years). Correlations between years of experience and average ratings per ATCO across all scenarios were also computed.

- **Correlation Between Experience and Ratings**: The correlation coefficient between years of experience and average rating per ATCO is approximately -0.35, indicating a weak negative relationship. This suggests that more experienced controllers tend to assign slightly lower complexity ratings overall compared to less experienced ones. For example, ATCO5 (8 years) often rated scenarios higher (average rating around 5.6) compared to ATCO8 (28 years), who had an average rating closer to 6.2 but was more conservative on extreme highs.

- **High-Complexity Scenarios**: In scenarios with many PCDs, less experienced controllers (e.g., ATCO5 with 8 years) frequently assigned maximum or near-maximum scores (9 or 10), while highly experienced controllers (e.g., ATCO8 with 28 years) were slightly more restrained, often scoring between 8 and 9. This could indicate that experienced controllers have developed better confidence in managing conflicts, thus perceiving them as marginally less complex.

- **Low-Complexity Scenarios**: For simpler scenarios with few or no PCDs, less experienced controllers showed greater variability in their ratings (ranging from 2 to 5), while experienced controllers were more consistent (often rating between 4 and 6). This may suggest that newer controllers are more sensitive to subtle differences in workload even in easier situations, whereas veterans apply a more standardized assessment based on broader benchmarks.

- **Regional/Experience Overlap**: It's worth noting that experience often correlates with operational background (e.g., different regions), which could confound pure experience effects. However, within similar experience brackets (e.g., ATCO4 and ATCO7 both at 14 years), ratings are still somewhat dispersed, hinting that personal factors beyond just years of service play a role.

Overall, while there is evidence that greater experience slightly tempers the perception of complexity the effect is not very strong. Other factors like individual temperament or specific training may also mediate how experience translates into evaluation.

Conclusions

Based on this analysis, several key conclusions can be drawn for application in Air Traffic Management contexts:

- **Primary Driver of Complexity**: The number of Potential Conflicts Detected stands out as the most significant factor influencing perceived complexity, far beyond total traffic volume or other categories like NCTs.

- **Moderate Dispersion**: The dataset shows moderate variability in ratings among controllers, reflecting subjective differences in workload perception. While consensus exists on extreme cases (very high or very low complexity), intermediate scenarios have varied responses.

- **High vs. Low Complexity Perception**: High-complexity scenarios are uniformly recognized as challenging due to clear indicators like numerous conflicts, whereas low-complexity rankings are less consistent, potentially due to differing views on baseline workload elements like monitoring non-critical traffic.

- **Experience Influence**: Experience has a medium-low effect on complexity perception, with veteran controllers tending to rate scenarios slightly lower than less experienced ones; however, since the effect is not dominant, broader human factors should also be considered.

4 Comparison: Sierra5 Complexity Metric

The Sierra5 (S5) complexity metric appears to be an effective measure for evaluating the complexity of air traffic control simulation exercises based on the comparison of the S5 Complexity scores to the average complexity ratings provided by the eight Air Traffic Controllers. One limitation of this study lies in comparing a subjective, experience-based perception with an objective, mathematically derived metric. Accordingly, the comparison needs to be performed more in a qualitative way, comparing the rankings rather than the actual values of the metric. The following points support why the S5 metric aligns well with expert judgment and can be considered a robust indicator of scenario complexity:

- **Close Alignment with Average Expert Ratings**: The data shows a strong correlation between the S5 Complexity scores and the Exercise Average scores (the mean of the ATCOs' individual ratings). Particularly:

    - In scenarios with high complexity, such as the first exercise (S5 = 8.8, Average = 9.4) and the sixth exercise (S5 = 9.5, Average = 9.2), both metrics consistently indicate high difficulty.

    - In lower-complexity scenarios, such as the fifth exercise (S5 = 4.7, Average = 3.8) and the thirteenth exercise (S5 = 5.5, Average = 3.9), both metrics similarly reflect reduced complexity.

This consistent alignment suggests that the S5 metric captures similar aspects of complexity as perceived by experienced ATCOs, validating its relevance.

- **Consistency Across a Range of Scenarios**:

    - Across all 19 exercises, the S5 Complexity scores rarely deviate significantly from the Exercise Average. Even in cases of slight divergence, such as the twelfth exercise (S5 = 7.9, Average = 5.8) or the fourteenth exercise (S5 = 7.0, Average = 4.6), the general trend remains comparable.

This indicates that the S5 metric is stable and reliable across different types of simulation exercises, which likely vary in terms of total flights, Potential Conflicts Detected (PCDs), non-conflicting traffics (NCTs), and other traffic factors.

- **Potential for Capturing Underlying Factors**: The S5 metric may incorporate a balanced consideration of multiple variables or KPIs (e.g., total number of flights, PCDs, NCTs, and other traffic) in a way that mirrors how experienced ATCOs intuitively assess complexity. While individual ATCO ratings might be influenced by subjective factors or specific experiences, the S5 metric provides a standardized and objective benchmark that closely matches their collective judgment (Exercise Average).

- **Practical Implications**: The close correspondence between S5 Complexity and the Exercise Average implies that S5 can serve as a reliable tool for assessing complexity without needing to rely solely on subjective human ratings.