Methods and Measurements for the Evaluation of ATM Tools in Real-Time Simulations and Field Tests

Fred Schick
German Aerospace Center
Institute of Flight Guidance
Lilienthalplatz 7
D-38108 Braunschweig, Germany
Phone (49) 531 295 2532
Fax (49) 531 295 2550
email fred.schick@dlr.de

Abstract

The process of developing an ATM tool requires evaluation at different stages. Experiments in the laboratory and in the control centre play an important role as they provide empirical data for the quantitative assessment of traffic-related as well as human-related issues.

Using as an example the guidance of inbound traffic in an Extended Terminal Area (ETMA) the evaluation of ATM tools in real-time simulations and field tests is discussed, with the aim to establish a close methodological link between both types of exercises in terms of common measurements and metrics.

Some implications of a commonality of methods and some benefit from the immediate comparison of results based on common metrics are illustrated. Reference is made to work done at DLR, including earlier studies of the Computer Oriented Metering Planning and Advisory System (COMPAS), more recent contributions to the Programme for Harmonised Air Traffic Management Research in EUROCONTROL (PHARE), and work planned for the near future.

1. Introduction

The development towards improvement of ATM is increasingly making use of automated assistance tools. Assessment of those tools is a challenging task which is permanent throughout the entire development process.

An assessment methodology has to assimilate in the first place the driving forces behind ATM development, among which the most prominent one is the expectation to increase capacity and safety. Traffic-related measurements take the role of evaluating to what extent a certain development meets this expectation.

But an assessment methodology has to assimilate as well the human factors issues of automation assistance. Among the consequences of advanced automation support, the APAS VAPORETO Study of the Commission of the European Community (1996) e.g. listed ”new operating modes, procedures, and methods that change human operator roles and tasks”. Therefore, human-related measurements that address the aspects of operator workload and acceptance provide the second column of methods for the evaluation of ATM tools (Figure 1).

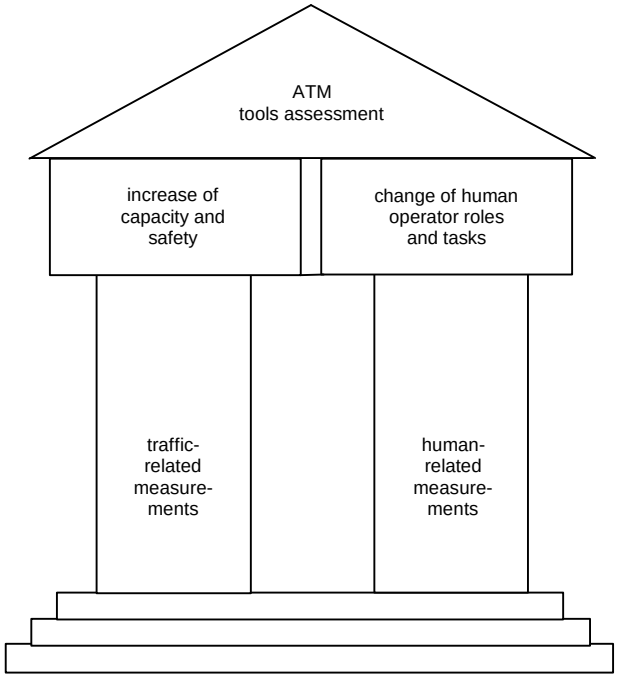

Figure 1 The two columns of ATM tools assessment methodology.

A particular challenge of ATM tools evaluation is to understand the mechanisms of how traffic issues and human issues are related to or dependent from each other. Using appropriate measurements and metrics can help to give insight, and thereby to provide input for the optimisation of a system's design.

Another fact is that during the developmental process of ATM tools there is a permanent need for evaluation, starting from early concept evaluation and ending with in-service evaluation. This may include a variety of different types of exercises, for instance

- modelling,
- prototyping,
- fast-time simulation,
- real-time simulation,
- and finally field tests.

The present paper's focus is on real-time simulation and field tests only. In-between lies the transition from the laboratory to the "real world" which is perhaps the biggest and most challenging step in the life-cycle of a tool or system.

The paper aims at promoting the idea of closer methodological links between laboratory and field tests in terms of common measurements and metrics. Although a system's implementation into operational service has many practical implications for the assessment methods, it is deemed necessary to establish a continuous thread of measurements which allow, at least for some core instances, valid comparisons between simulation and field test results. The following is an attempt to point out the benefit of having some metrics of performance, workload, and acceptance commonly applied in real-time simulations and subsequent field tests as well.

Of course, it would be presumptuous to believe that there existed some universal "best practice" set of measurements. The measurements exemplified below are indeed selective. Using as an example the guidance of inbound traffic in the Extended Terminal Area (ETMA) of Frankfurt/Germany, they have proven their suitability for various assessments of guidance tools. Reference will be made to early simulations and field tests of the Computer Oriented Metering Planning and Advisory System (COMPAS) in the nineteen-eighties/early nineties, to more recent contributions to the Programme for Harmonised Air Traffic Management Research in EUROCONTROL (PHARE), particularly the PHARE Demonstration 2 simulations (1997), and to work planned for COMPAS's successor system, the 4D-Planner, which will cumulate in a six-months field test period at Frankfurt in 1999.

2. The Frankfurt Extended Terminal Manoeuvering Area (ETMA)

Figure 2 gives a schematic picture of the airspace considered in these studies.

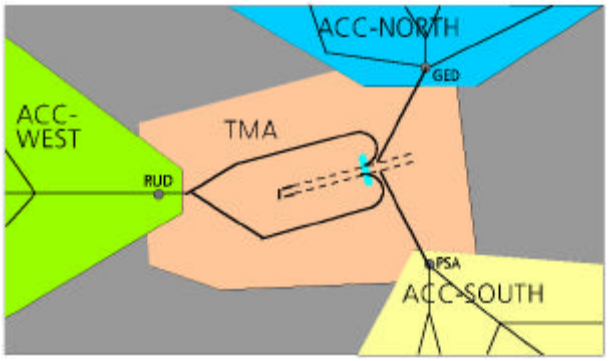

Figure 2 Extended TMA Frankfurt.

The core TMA covers an area of about 30 NM in diameter, adjoining to the Metering Fixes Gedern (GED), Spessart (PSA), and Rudesheim (RUD). The airport's runway system, as indicated in the centre of the figure, consists of two parallel runways (25/07) used for arrivals and departures, and a third runway (18) for departures only. The

TMA is surrounded by three adjacent en-route sectors which are under control of an Area Control Centre (ACC West, ACC South, ACC North). The sectors extend to ranges up to 100 NM from the airport. Corresponding inbound flight times are about 20 to 30 minutes. The standard arrival routes for runway direction 25 are shown as solid lines. Traffic from the north is guided to the extended centreline (dotted) of the northern runway (25 R), traffic from the south to the southern runway (25 L). Arrivals from the west can be guided either on a southbound or on a northbound route to merge 10 NM ahead of the threshold (at the Approach Gate, shortly named "Gate" in the following) with the other arrival routes.

3. Evaluation of the COMPAS System

In order to assist the controllers in a smooth and efficient handling of the arrivals COMPAS (the Computer Oriented Metering Planning and Advisory System) was developed in the early 1980's in co-operation with the German Air Navigation Services (DFS). It was designed to estimate arrival times (based on flight plan data, radar data, aircraft performance data, and wind data) to determine the required separation between aircraft of the same or different weight classes, to plan an optimal sequence of aircraft with regard to a smooth integration of traffic from the different arrival routes, and thereby to make best use of the existing runway capacity.

3.1 Real-Time Simulation Trials

An experimental prototype of the system was tested in the Air Traffic Management and Operations Simulator (ATMOS) of the DLR Institute of Flight Guidance. In a first series of trials, the system optimisation phase, the planning algorithm and the human/machine interface (COMPAS display and interaction devices) were fine-tuned. Then the system layout was frozen for the system evaluation trials. 28 controllers (seven teams of four controllers) were involved in more than one hundred hours of real-time simulation trials in which their work on typical inbound rushes with COMPAS assistance was compared with identical traffic samples without COMPAS, as a baseline.

The task was to assess traffic handling performance, controller workload, and controller acceptance. Traffic data, operational data of controller activity, and subjective estimates and judgements of the controllers were taken as measurements.

Traffic data collection consisted basically of logging

- observed time of aircraft over Metering Fix and Gate,
- planned time of aircraft over Metering Fix and Gate,
- horizontal radar tracks of the aircraft.

Among the metrics calculated from this data were

- sector flight time of aircraft: Metering Fix time minus sector entry time,
- TMA flight time of aircraft: Gate Time minus Metering Fix time,
- compliance with first-come-first-served rule: aircraft rank displacement in the approach sequence between pre-planning estimate (as if there were no other inbound aircraft) and actual Gate time.

Data of operational controller activity were mainly provided by registration of

- radio communication to aircraft (i.e. to pseudo-pilots),
- telephone communication to other controllers for co-ordination purpose,
- actuation of any keys of the COMPAS input device.

From that data metrics were calculated for

- radio communication load: communication time percentage of total simulation time,
- co-ordination effort: frequency of phone calls per 100 aircraft,
- manual input load: number of input actions per 100 aircraft.

For a part of the trials visual scanning patterns of controllers were recorded from eye-point-of-regard measurement.

Subjective estimates and judgements were taken for

- workload estimates by means of the Subjective Workload Assessment Technique (SWAT); controllers' own ratings of workload were collected during simulation runs,
- acceptance evaluation by means of a 39 items questionnaire; controllers rated their agreement/disagreement with the COMPAS human/machine interface (display and input

device) and the plans generated by the system, using a standard rating scale.

To sum up the results briefly, the simulations indicated the capability of COMPAS to enhance a smooth work flow in the management of arrivals. The planning information was highly accepted. Co-ordination effort between ACC and Approach Control was considerably reduced, and the approach sequences established were ”fairer”, showing less frequent violations of the first-come-first-served rule. In the TMA in particular a more direct vectoring was observed from the radar plots, together with a significantly reduced communication load and a significant decrease of average flight time.

3.2 Transition to Field Tests

About two years later a fully operational prototype of COMPAS was installed at Frankfurt ATC, ready for a six-months field test.

From a methodology point of view, evaluation of the same system in real-time simulation and field tests has to account for the different environments, even when the task to assess the effects on traffic flow, controller workload and acceptance remains unchanged. Figure 3 compares typically different issues.

Real-Time Simulation	**Field Test**
suitability of concept or tool; comparison against baseline	verification of simulation results in real-world operation
limited number of controller working positions	fully staffed control center
selected traffic samples	total real-world traffic
isolated simulation runs	long-term observation
data collection at high rates; experimental staff available	economical, event-triggered data collection; no extra staff

Figure 3 Simulation vs. field test: differences relevant for assessment.

In a simulation environment there is better control of the experimental variables, subjects, and traffic samples. A quantitative assessment, including a comparison with existing procedures as a baseline, is supported e.g. by repeated measurements with varying controller teams working on exactly the same traffic samples.

Field tests do not offer this feature. The opportunity of any comparisons with baseline data is much more limited. The attention shifts more to the verification of prior simulation results. Will they be supported and can they be generalised with respect to

- the larger number of working positions, as opposed to the limited number of manned positions in the simulation environment?
- the totality of all traffic situations that do occur in real operation, as opposed to the limited number of traffic samples used in simulations?
- long-term observation, as opposed to relatively short simulation runs?

This has consequences for the data collection procedures.

Recording data continuously at high rates, e.g. the positions of all aircraft, may be justified for simulation periods counted in hours but would be highly uneconomical over periods of months. Data collection triggered by specified events is much more appropriate there. In the COMPAS field tests, for instance, an aircraft’s Gate overflight caused the log of an aircraft data set which contained, among others, actually observed events such as the times of passing Metering Fix and Gate together with the corresponding times planned by the system for that events.

Also measurements requiring extra staff or dedicated equipment at the controller working positions (e.g. eye tracking) cannot be used likewise as well as any measurements requiring extra activity for the controllers during their work (e.g. estimating own workload) are prohibited for safety reasons.

Despite of all these limitations there must be no doubt about the usefulness of having at least some core measurements used likewise in both environments. Some few examples from the COMPAS field tests may illustrate the feasibility of common traffic-related and human-related (operational and subjective) measurements. Moreover, when that measurements are quantified in the same way (in other words, when equal computation rules specify equal metrics in both instances) this is for the benefit of directly comparing quantitative results from field tests with their simulation counterparts.

As an example from the variety of traffic measurements, compliance with first-come-first-served is shown in Figure 4.

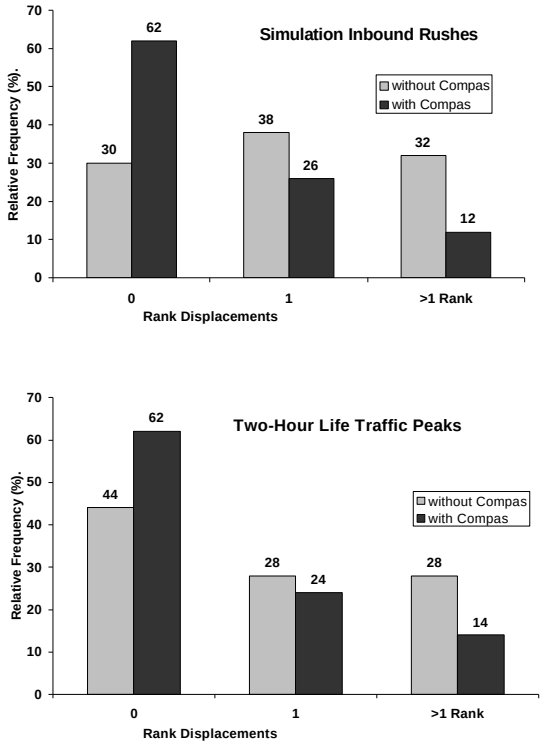

Figure 4 Aircraft rank displacements from first-come-first-served approach sequences in simulated traffic (top) and live traffic (bottom).

The upper part of Figure 4 shows the relative frequency of rank displacements observed from simulations of a traffic sample representing a typical inbound rush on the northern arrival route. Data are pooled from simulation runs of seven controller teams. Particularly the proportion of aircraft with no rank displacement at all against a theoretical first-come sequence was significantly higher with the assistance of the planning system. The bottom graph shows life data for two inbound rushes with approximately the same traffic demand. Data of two two-hour traffic peaks, one observed shortly before the test period, the other one when COMPAS was in use, gave strong support for verifying the above simulation result.

Regarding operational measurements, manual input actions of the controllers may serve as an example. During the simulated inbound rushes on the northern arrival route the ACC controllers in the North Sector used the COMPAS function keyboard for 18.5 input actions per 100 aircraft, on the average over all teams. Figure 5 shows the number of inputs as it developed in the field tests over six months.

Figure 5 Manual inputs using the COMPAS function keyboard during distinct periods of the six-months field tests.

In the first weeks inputs were much more frequent, ranging around 50 per 100 aircraft. At about half time the input frequency fell below the level of the simulations, to stabilise after about five months at or below ten inputs per 100 aircraft. This development was supposed to be mainly due to an increase of trust in the suitability of the plans which was supported by several refinements of the planning software.

Subjective measurements obtained from controller questionnaire responses could be quantified as well by making use of the same questionnaire items and the same standard six-point rating scale. A questionnaire was applied in the simulation trials and was applied again twice in the field tests to a sample of controllers, at the beginning and at the end of the test period. Figure 6 depicts one questionnaire item for illustration.

Controllers in the simulation trials agreed well with a time horizon to display approach planning information about twenty minutes before landing. A contrary picture was observed shortly after the start of the field tests, caused by a few cases in which a technical problem with correlating radar and flight plan data of some aircraft led to some information delay. Improvements during the test period could alleviate this particular problem, so controller responses at the end became more favourable again.

"Plans for the individual aircraft are displayed early enough."

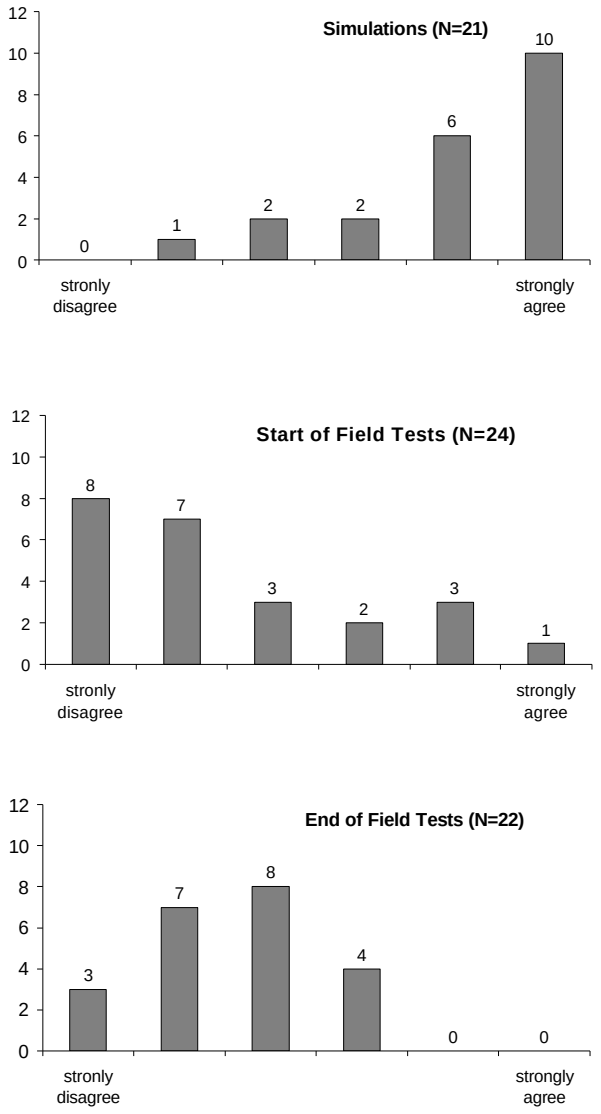

Figure 6 Distribution of controller responses to the same item of a questionnaire in simulations and field tests.

In total, the commonality of methods allowed to draw numerous comparisons of results at the transition from laboratory to operational service.

4. Evaluation of Advanced Concepts

More advanced tools and functions which offer for instance support to 4D-guidance and trajectory negotiation introduce new operating modes and new human/machine interfaces. An assessment of those tools has to account for this development by referring to traffic-related metrics that apply particularly to an analysis of 4D-guidance quality, and by referring to human-related metrics that capture the impact on controller work-style and cognitive functions which add to workload and acceptance.

Also in this context the proven practice of keeping metrics constant over simulations and field trials is consequently pursued further at DLR. The following gives an outline of the metrics used in the PHARE Demonstration 2 simulations. The outline is drawn further up to next year's field tests of the 4D-Planner system at Frankfurt ATC.

4.1 PHARE Demonstration 2 (PD/2)

As part of the Programme for Harmonised Air Traffic Management Research in EUROCONTROL (PHARE), PD/2 formed the second major real-time simulation exercise in a series of three PHARE Demonstrations to support investigations into aspects of the concept of the future European Air Traffic Management System (EATMS). The work programme of PD/2 was to design, implement, and demonstrate the PHARE prototype air and ground computer assistance tools for air traffic management in the ETMA. The participating partners were CENA of France, NATS of UK, NLR of the Netherlands, and the EUROCONTROL Experimental Centre at Brétigny (EEC).

The main objectives of the PHARE Demonstrations were to determine the effect on controller workload and traffic throughput by introduction of computer assistance tools from the PHARE Advanced Tools (PATs) programme, in an environment with an increasing proportion of 4D FMS equipped aircraft with full two-way datalink, and to gain a degree of controller approval for the advanced tools introduced.

The PD/2 system was demonstrated at DLR's real-time simulator ATMOS with 32 controllers from seven European countries participating. The system incorporated advanced controller assistance tools with an associated Ground Human Machine Interface (GHMI) designed in the PHARE GHMI project, as well as, by integration of the DLR Advanced Technologies Testing Aircraft System (ATTAS) Experimental Cockpit, simulated air-ground datalink and 4D experimental flight management system (EFMS).

During the simulation trials a variety of traffic and workload measurements were recorded. Audio and video documentation, observer logs, debriefing sessions, and questionnaires were used to accomplish the PD/2 data collection.

The key traffic metrics produced from the recorded data were

- number of landings per hour:
  simulation runs of 90 minutes were used in PD/2. As they started with an ”empty” ETMA into which aircraft were fed from the adjacent sectors it took about 20 minutes until the first aircraft landed. The time interval between that first landing and the last landing observed before the end of the 90 minutes simulation time served as the basis to calculate the landing rate.
- flight time:
  referring to the aircraft that actually had landed per simulation run, their flight time was calculated as the Time over Threshold minus the Simulation Entry Time.
- inbound delay:
  this metric calculated the difference of Actual Time of Gate Overflight minus Estimated Time of Gate Overflight. The Estimated Time was derived from the aircraft’s ”user preferred trajectory” along the arrival route that could have been flown if there were no other aircraft in the area.
- precision of delivery:
  as a metric of how precisely the plans generated by the PATs tools were actually implemented, Actual Time of Gate Overflight minus Planned Time of Gate Overflight was calculated.
- separation:
  referring to the aircraft’s Actual Time of Gate Overflight, and assuming equal speed at that flight phase, the distances between pairs of aircraft on the same glide-path were calculated.

The key workload metrics were derived from objective and subjective measurements.

Records of controller communication served as an objective workload indicator. Metrics calculated from the records were

- the number of ATC instructions issued per minute,
- the frequency of radiotelephony (R/T) calls per hour,
- the percentage of simulation time spent for R/T calls.

Subjective data were provided by records of

- subjective workload estimates taken during task execution by means of the Subjective Workload Assessment technique (SWAT),
- an overall workload estimate per simulation run was obtained additionally from each controller, using the NASA Task Load Index (NASA-TLX) method.

Questionnaires were used as the main source of information on controller acceptance. A standard six-point rating scale was presented to the controllers to rate their agreement/disagreement with specific properties of the

- human/machine interface (17 items),
- operational procedures (20 items),
- tools and functions (7 items).

The quantitative analysis of traffic data revealed various gains from the introduction of the advanced tools in terms of traffic throughput and quality of service. Overall, benefits were achieved for the number of landings per time unit, average flight time of aircraft, inbound delays, and time precision of delivery particularly under conditions of high traffic load. The separations measured at the Approach Gate showed that these benefits were not achieved at the expense of separation.

Statistical analysis of controller workload revealed some re-distribution of workload between tactical controller position as an effect of the introduction of advanced tools. Furthermore, the introduction of 4D FMS/datalink aircraft in the traffic sample showed a reduction of workload for all tactical controller positions involved.

The PD/2 Ground Human/Machine Interface (GHMI) gained a high degree of controller acceptance. The operational procedures introduced with the PD/2 concept were also approved well in general, although controllers saw some need for further development and improvement of some tools and functions to better support their work under this concept.

Referring to the aim of PHARE, to produce results that help to refine the description of future Air Traffic System concepts, PD/2 was a successful demonstration of the integration of advanced tools, 4D FMS and datalink into an air-ground air traffic management system in an extended terminal area.

The metrics specified and used in the context of PHARE are considered also relevant for the assessment of tools that will be put into service within a short-term time frame.

The 4D-Planner, a system currently developed co-operatively by DLR and the German Air Navigation Services (DFS) to replace the COMPAS System at Frankfurt by 1999 is the candidate to be assessed next. This will include as a major exercise a six-months field test period.

4.2 4D-Planner Assessment

The 4D-Planner is designed to provide time-based arrival planning as well as information to the controller on time-based guidance of inbound aircraft. It goes beyond the functionality of the current COMPAS system by continuous monitoring of aircraft positions, detection of plan deviations, and planning updates if necessary. It generates control advisories which aim to produce arrival sequences with exactly the required separation. This is expected to reduce the variability of the actually observed separation, as a prerequisite for increasing safety and airport capacity.

When the 4D-Planner system will be in operation at the Frankfurt ATC centre, an evaluation of its impact will have to account for the methodological considerations and limitations mentioned earlier. For instance, subjective workload assessment as well as observer logs which require extra staff that can be afforded in simulation trials will not be available then.

Therefore human-related measurements will be concentrating on automatic logging of controller interaction with the system, and on questionnaires:

- Operational data of system use
  The 4D-Planner system has built-in capability to log the time and nature of any controller inputs at any controller working position in the Area Control Centre and in Approach Control.
- Questionnaire data
  Assessment of human/machine interfaces, operational issues, and acceptance in general will be done at least twice, using an underlying standard rating scale which facilitates a quantitative analysis of potential changes of controller responses.

Traffic-related measurements will be available due to the merit of a system called the Flow Monitor.

The Flow Monitor, again jointly developed by DLR and DFS, is in service at the Frankfurt Air Traffic Control Centre since 1993. It has been installed there right for the purpose of on-site assessment of the impact of newly introduced systems and procedures on inbound traffic, particularly regarding their effect on delay and capacity.
The system monitors all arrival traffic and collects data on traffic flow, demand, delay, and aircraft separation.

- Traffic Flow
  is defined as the number of landings actually observed per time unit.
- Demand
  is calculated from data of the aircraft's entry into the ETMA. At that time an estimate of undelayed landing time is logged. The number of arrivals per time unit, based on those estimates that assume conflict-free ideal flight conditions to touchdown, is called Demand.
- Delay
  is the difference between aircraft's actual flight time from ETMA entry to touchdown and the estimated flight time underlying the above Demand calculation.
- Separation
  logging uses the Approach Gate (10 NM from threshold) as the reference point. There the distance to the following aircraft is taken from radar position data.

Flow Monitor data are displayed on-line at the control centre. Flow, Demand, and Delay averages for ten, thirty, and sixty minutes intervals are presented 24 hours a day (Figure 7).

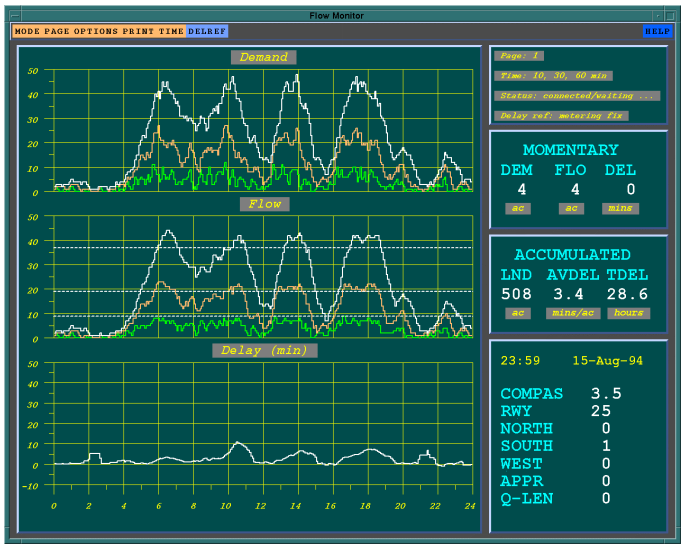

Figure 7 On-line display of Flow Monitor Data.

All data are also stored for statistical analysis over longer periods of time. This is of invaluable importance because it provides baseline data against which the effects of any new tool or procedure can be quantified.

Among the numerous statistics that can be produced from the logged data of all inbound aircraft are for instance daily, weekly, or monthly averages of Flow, Demand, Delay, Flight Time per arrival route, and Separation histograms. An example is shown in Figure 8.

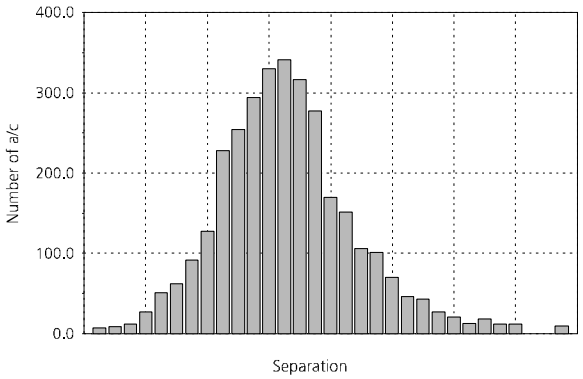

Figure 8 Separation at high traffic intervals (Flow >35 aircraft/hour) over one month.

The versatility of the Flow Monitor statistics can be used to produce exactly the same key traffic metrics that have proven their relevance in the simulation environment.

In conclusion it is expected that the 4D-Planner field tests may serve as another example of the benefit of having a blend of measurements of traffic, controller workload and acceptance, with as many metrics as possible in common with previous exercises.

5. References

1. Commission of the European Community (CEC),
   Validation Process for Overall Requirements in Air Traffic Operations (VAPORETO).
   End Report, Contract No. 8102 – CT94-0002, February 1996.

2. Schick, F.V. and U. Völckers,
   The COMPAS System in the ATC Environment.
   DLR-Mitteilung 91-08, Braunschweig, June 1991.

3. EUROCONTROL,
   Template of Measurements to be used in PHARE Demonstrations.
   PHARE DOC 94-70-07, Brussels, March 1994.

4. EUROCONTROL,
   PD/2 Final Report (Volumes 1 and 2).
   PHARE DOC 97-70-13, Brussels, June 1998.

5. Schenk, H.-D.,
   The Flow Monitor for the Airport at Frankfurt.
   NAV-Canada, Ottawa, March 1997.