Advanced Quantification of Weather Impact on Air Traffic Management

ATMAP 2.0 - Intelligent Weather Categorization with Machine Learning

Stefan Reitmann*, Sameer Alam† and Michael Schultz*‡
*Institute of Flight Guidance, German Aerospace Center, Germany, 38108 Braunschweig
†Air Traffic Management Research Institute, Nanyang Technological University, Singapore, 639798
‡Institute of Logistics and Aviation, Dresden University of Technology, 01069 Dresden

***Abstract*—The classification of weather impacts on airport operations will allow efficient consideration of expected local weather events and in an analysis of air traffic network behaviors. We use machine learning approaches to correlate weather data from meteorological reports and airport performance data contains of flight plan data with scheduled and actual movements as well as delays. In particular, we used unsupervised learning to cluster performance impacts at the airport and classify the respective weather data with recurrent and convolutional neural networks. It is shown that a classification is possible and allows estimates of delay including weather and flight plan data at an airport. This paper serves to illustrate a possible classification with machine learning methods and is the basis for further investigations on this topic. Our machine learning approach allows for an efficient matching of the decreased airport performance and the occurrence of local weather events. Thus, we provide an update of current weather classifications, which will be a basis for a better understanding of interdependencies between local and network-wide effects in the air transportation system.**

***Keywords-component*—airport performance; weather impact; METAR data; machine learning;**

I. Introduction

Weather has a significant impact on airport operations and the performance of the whole aviation network. Delayed operations may caused by airport capacity constraints due to severe weather conditions. The prediction of aircraft processes along their whole trajectories is required to achieve punctual operations. Uncertainties during the airborne phase of flights represent only a minor impact on the overall punctuality. In the current operational environment, ground tasks gain more relevance. The focus on ground operations will allow the different stakeholders to define and maintain a comprehensive 4D aircraft trajectory over the day of operations. Using a reliable and predictable departure time is one of the main tasks of the ground activities. Mutual interdependencies between airports, as departing delays propagate thought the network, result in system-wide far reaching effects. In 2016, reactionary delays continued to be the main delay cause, followed by turn around delays, accounting for 46% of departure delays.

Flight deviations are important for the air traffic management and induced by weather and traffic situations as well as controller actions (e.g. directs [1]). Typical standard deviations for airborne flights are 30 s at 20 min before arrival [2], but could increase to 15 min when the aircraft is still on the ground [3]. As shown in fig. 1, the average time variability (measured as standard deviation) during the flight phase (5.3 min) is higher than in the taxi-out (3.8 min) and in the taxi-in (2.0 min) phases, but it is still significantly lower than the variability of both the departure (16.6 min) and arrival (18.6 min) phases [4]. The changes experienced during the gate-to-gate phase are comparatively small, leading to a translation of departure variability into arrival one [5]. Thus, the arrival punctuality is driven by the departure punctuality and all stakeholders (airlines, airport, network manager, air navigation service providers) play a significant role on the system-wide punctuality performance.

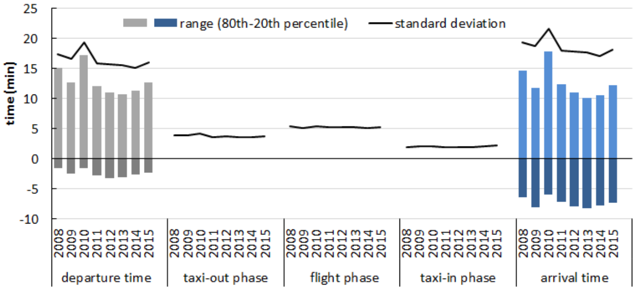


Fig. 1: Analysis of European flights from 2008–2015 regarding variability of flight phases, not considering flights departing to or arrival from outside Europe (for data, see [4], [6]).

For example, airlines strategically implement buffers to absorb a part of the delay generated by tactically reducing its propagation and achieving a desired target of punctuality [7], [8]. In 2016, only 81% of the flights were punctual with a decreasing trend starting from 84% punctuality in 2013 [4]. Weather related delays are reported by the flow management positions as the second most common cause of en-route air traffic flow management (ATFM) delays (18%) [4]. For airports, the closer they operate to their maximum capacity, the more severe is the impact of a capacity loss due to external events such as weather.

A. *Status quo*

Current research in the field of flight and airport operations addresses economic, operational and ecological efficiency [9]–[18]. The propagation of delay in the network is paramount when assessing the impact of congestion [19], [20]. This is particularly critical when estimating the resilience of the Air Traffic Management (ATM) system and the impact of different mechanisms on the expected performances' variations [21]–[23]. Dynamic traffic situations emerge from traffic flow patterns across Europe and to/from intercontinental flows, military operations [24], volcanic ash eruptions [25], zones of convective weather [26], prevention of contrails [27], consideration of commercial space operations [28] and integration of new entrants [29]. Current research also considers passengers metrics as trade-offs between optimisation of flight performances not possibly being aligned with passengers experience [30]. This can be particularly relevant when optimising arrival flows at airports under uncertainty [31], [32]. Thus, delay generation due to weather impacts including location and time of the primary delay generation and its evolution are relevant to capture the complexity of the system dynamics.

With a focus on airport operations, the weather phenomena could be categorized by the ATM Airport Performance (ATMAP) weather algorithm [33] provided by the Eurocontrol's Performance Review Unit (PRU), which aims to quantify the weather conditions at European airports (measure of the intensity and duration of weather phenomena). Thus, a group of experts identifies relevant aviation weather factors and considers that these factors are additionally coupled with the availability of local airport technologies (such as precision approaches in poor visibility conditions) and aircraft characteristics (such as defined tolerances for crosswind and tailwind). Furthermore, the ATMAP algorithm weight the different weather factors, that similar ATMAP scores will result in comparable impacts on airport operations, although they are based on different weather events (such as high wind speeds or low visibility conditions). The ATMAP algorithm considers five weather classes (ceiling and visibility, wind, precipitations, freezing conditions, dangerous phenomena) and also considers different degrees of severity per weather class. In fig. 2 the daily ATMAP weather score (diamond) is displayed against the airport performance at Frankfurt airport using on-time performance (delay $<$ 15 min) and cancellations.

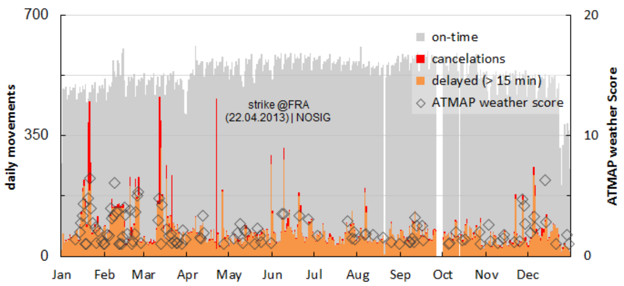
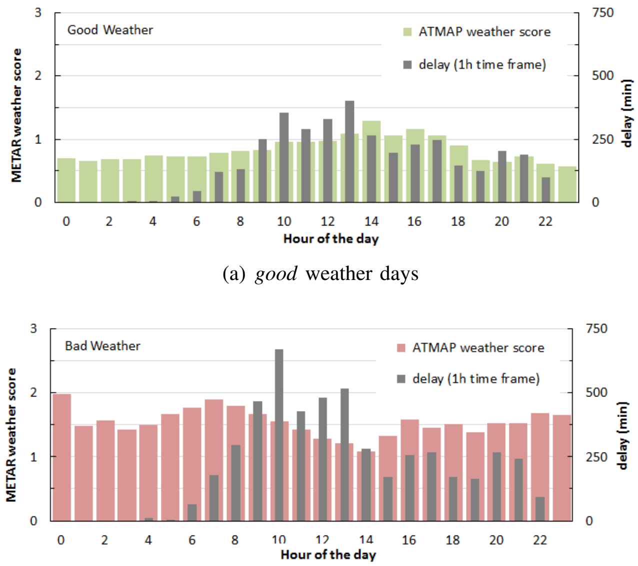


Fig. 2: Airport performance data and ATMAP weather score at Frankfurt Airport (2013).

The following definitions are used in the ATMAP algorithm: *weather phenomenon* is a single meteorological element which impacts the safety of aircraft during air and ground operations; *weather class* is a group of one or more weather phenomena affecting the airport performance; *severity code* is a ranking number of the weather class status (from best to worst); *coefficient* represents the assignment of a score to a given severity code in order to describe the nonlinear behavior of various weather phenomena. The PRU proposes a multi-step procedure to determine the ATMAP weather score: in a first step, a given METAR (Meteorological Aviation Routine Weather Report) observation at the airport will be assessed by specifying the severity code and its associated coefficient for each weather class. This METAR message is parsed, filtered, and transformed to a quantified measures (coefficients). In a second step, these weather class coefficients are summed up to the corresponding ATMAP score (per METAR message). Finally, for a given time interval (hours of operations), the sum of all ATMAP scores are divided by the number of METAR observations to calculate an average ATMAP score per time interval (e.g., per hour, per day). In this context, the ATMAP algorithm separates days of operations into *good* and *bad* weather days, using an average and airport-independent ATMAP value of 1.5 (default European score for bad weather days [33]). In fig. 3 the impact of these different weather classifications are shown by increased number of accumulated delay minutes per hour (gray bars) caused by higher numbers at the ATMAP weather score (green/red bars).



(a) *good* weather days

(b) *bad* weather days

Fig. 3: Different hourly delay characteristics at Frankfurt Airport (2013) considering two weather classifications: (a) *good* and (b) *bad* weather [34].

B. Scope and structure of the document

In this contribution, we provide for the first time a machine learning approach to quantify the impact of local (severe) weather conditions to the airport performance. The ATMAP algorithm by the PRU [33] based on expert judgment and was initially established in 2011. This quantifying method will updated based on actual datasets and appropriate latest methodologies. We propose a machine learning approach, which considers local weather data (METAR) and airport performance data (scheduled and actual flight plan) to correlate the complex dependencies in the airport systems. Herein, we are not focusing on the causes (input) but on the consequences (output) to the airport performance. For this purpose, we categorize the airport performance and backtrack/ evaluate possible causes from the observed weather phenomenon. Finally, we will provide a methodology to map individual weather phenomenon to airport performance impacts, which will be a basis for a new approach to overcome the limitations of the current judgment-based ATMAP algorithm.

The document is structured as follows. Section I provides an introduction of the topic and a status quo of related research activities. In section II, the datasets for weather and airport performance are described, including a brief description of the ATMAP algorithm [33]. The general concept of machine learning, in particular classification and clustering, is addressed in section III. In section IV, several neural networks are applied to the datasets. Finally, the document closes with a conclusion and outlook (section V).

II. Weather and airport performance

The dataset we used for the analysis consists of flight plans and weather data of major European Airports (more than 60 million flights, year 2014-2015). The flight data sets include scheduled and actual time stamps of specific aircraft movements, and air traffic relevant weather data are derived from the airport specific METAR data. From this dataset we used a subset with a focus on London Gatwick airport (EGKK).

A. Weather data

Current weather conditions are usually recorded at each airport in the form of METARs (Meteorological Aviation Routine Weather Report [35]). METARs are reported in combination with a Terminal Area/Aerodrome Forecast (TAF). While TAF provides forecast values, METAR data are measured values. The unscheduled special weather report (SPECI) is another format representing significant changes in airport weather conditions. The time of update and the update interval of a METAR weather report are not harmonized and implemented differently worldwide. For example, at larger airports in Germany, a METAR is released twice an hour (20 min past and 10 to the full hour) while, at small sized airports like Moenchengladbach (EDLN), a new METAR is available once an hour only during the operating times of the airports. Current and historical METAR and also TAF data are accessible at different public available websites (such as https://www.ogimet.com). In addition to information about the location, the day of the month and the UTC-time ("EDDF 190850Z"), the METAR contains information about wind, visibility, precipitation, clouding, temperature, and pressure that are relevant for the air traffic, especially for the airport operations (see tab. I).

TABLE I: Main components of Meteorological Aviation Routine Weather Report (METAR) message.

Parameter	Measurement	METAR Code (Example)
wind	direction azimuth in degrees/speed [kn]	06010KT
visibility	horizontal visibility [m]	7000
precipitation	significant weather phenomenon	−SN
cloud	cover/height*100 [ft] above aerodrome level	BKN019
temperature	air/dew point [°C]	M03/M06
pressure	Sea-level pressure (QNH) [mbar]	Q0998
(trend)	(reported conditions within the next 2 hours)	(NOSIG)

Fig. 4 exhibits exemplary weather information derived from the METAR dataset (average per day): temperature, dew point, wind direction and speed, humidity, and pressure.

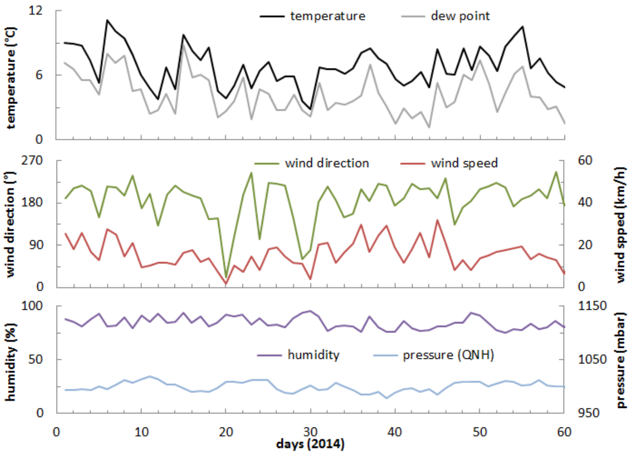


Fig. 4: Weather data from the first 60 days in 2014 at Gatwick airport.

Besides this general weather information, some additional measurements were available related to adverse weather situations, such as information about wind gusts, runway conditions (e.g., ice layer) and thunderstorm related clouds, as well as calculated values of the Runway Visual Range (RVR). The use of METAR weather records for data analysis demands for a detailed analysis, since specific characteristics exist and the data integrity is not assured by the data provider. Typically, data lacks (partial) loss of significant information, such as wind data, dew-point data, or runway condition information (e.g., depth of deposit), variable units of measure, or incomplete information about airport runway conditions. To allow for an appropriate analysis of the weather phenomena, the METAR is decoded stepwise. The information has to be parsed, filtered and transformed to a usable measure in the context of the comparison to the airport performance.

B. Standard ATMAP approach

The current ATMAP algorithm quantifies and aggregate major weather conditions at airports, which have significant impact on the airport operations. Five different weather classes with a significant influence on aircraft and airport operations are included: (1) ceiling and visibility; (2) wind; (3) precipitation; (4) freezing conditions; and (5) dangerous phenomena. In tab. II, these five different weather classes are shown, described with meteorological conditions, and linked to the associated maximum coefficient defined by the ATMAP algorithm. Compared to the other weather classes, dangerous phenomenon have a high particular impact on airport operations which results in the highest coefficients. For both cumulonimbus (CB) and towering cumulus clouds (TCU), the ATMAP coefficients are ranging from 3 to 10 (TCU) or from 4 to 12 (CB) depending on the cloud coverage (FEW, SCT, BKN, OVC). Showery precipitation and intensive precipitation can lead to a further increase of the coefficient values up to 18 or 24 for TCU as well as CB. Other dangerous phenomena with impact on the safety of aircraft operations can be divided into three groups: 30 points (heavy thunderstorm), 24 points (e.g., sandstorm, volcanic ash), and 18 points (small hail and/or snow pellets).

TABLE II: Weather classes defined in the ATM Airport Performance (ATMAP) algorithm.

Weather Class	**Description**	**Meteorological Conditions**	**Coefficient**
(1) ceiling and visibility	deterioration of visibility	precision approach runways (CAT I-III)	max. 5
(2) wind	strong head-/cross-wind	Wind speed $>$ 16 knots (+gusts)	max. 4 (+1)
(3) precipitations	runway friction influencing rwy occupancy time	e.g., rain, (+/$-$) snow, frozen rain	max. 3
(4) freezing conditions	reduced runway friction, de-icing	T $\leq$ 3°C, visible moisture, any precipitation	max. 4
(5) dangerous phenomena	unsafe ops, unpredictable impact	TCU/CB, cloud cover, (+/$-$) shower, storm	max. 30

In tab. III, two examples of METARs from Frankfurt Airport (EDDF) and Munich Airport (EDDM) are given to show the transformation from the METAR message to the ATMAP score.The given METAR observation will be assessed by specifying the severity code and its associated coefficient for each weather class. These weather class coefficients are summed up to the corresponding ATMAP score. In the actual example, both airports are under severe weather conditions (ATMAP score $>$ 1.5) and Munich exhibits ’dangerous weather phenomena’ in particular (scattered sky at 1800ft with cumulonimbus (CB) in combination with showers). Both cases are expected to have significant impact on airport operations.

TABLE III: ATMAP weather score based on local airport METAR messages.

			Weather Class			**ATMAP**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**score**
METAR	EDDF 241320Z **03007KT 9999** $-$**SN** FEW012 SCT018 BKN025 **01**/M02 Q1013 R07L/295 R07C/295 R07R/295 R18/5/295 NOSIG					
values	**9999**	**03007KT**	$-$**SN**	**01**, $-$**SN**	-	
coef.	0	0	2	3	0	**5** (sum)
METAR	EDDM 082120Z **25006KT 3200 SHSN** FEW005 **SCT018CB** BKN025 **M00**/M03 Q1015 TEMPO ...					
values	**3200**	**25006KT**	**SHSN**	**M00, SHSN**	**SCT018CB, SH**	
coef.	0	0	3	4	15	**22** (sum)

According to the time period used at fig. 4 (first 60 days of 2014, daily average), the ATMAP algorithm could quantify these measurements into a ATMAP weather score. The aggregated scores for the particularly observed weather classes indicate the severity of a weather phenomenon with an increasing value (see fig. 5).

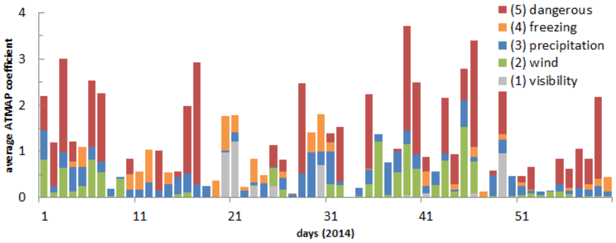


Fig. 5: Corresponding ATMAP score for weather data from the first 60 days in 2014 at Gatwick airport.

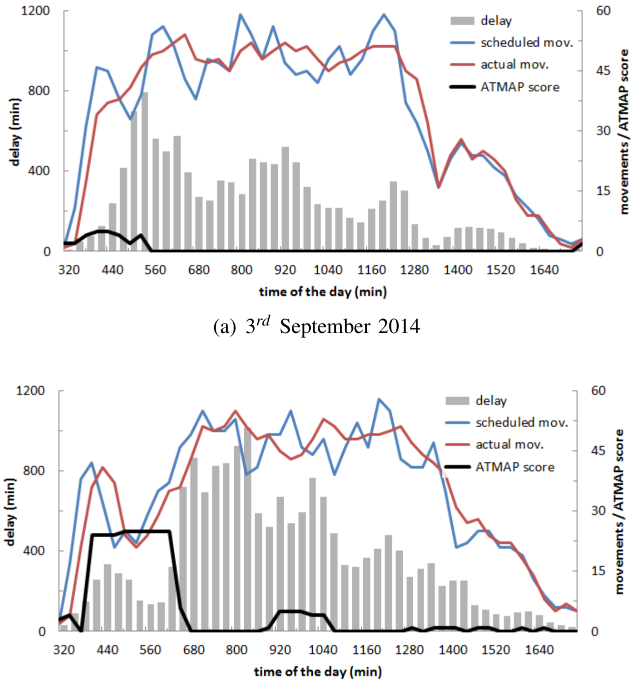
C. Airport performance

The performance of an airport is mainly related to the number of aircraft movements handled (airport capacity). In this case, the term capacity generally refers to the ability of a given transportation facility to accommodate a traffic volume (e.g., movements) in a given time period (e.g., on hourly, daily, or yearly basis). If the air traffic demand approaches or exceeds the given airport capacity, the congestion of provided infrastructure increases which results in delays and cancellations. This demand–capacity imbalance is a key cause of unpunctual operations and affects different components of the whole airport system on airside (e.g., runways, taxiways, aprons) and landside (e.g., passenger handling [36], [37]). Results of a data analysis from Frankfurt airport show that more than 45% of the variability in daily punctuality are related to local weather impacts [38].

Flight delays expressed in minutes are defined as the difference between the scheduled and actual times of arrivals and departures. Reference points for flights are usually their on- and off-block times. Punctuality is determined as the proportion of flights delayed less than 15 min, an internationally accepted performance indicator in air traffic. To anticipate the delay in phases of high traffic demand (peak times), airlines apply buffer strategies, to improve punctuality and mitigate tactical delay costs [4], [21]. The definition of delay can vary according to the stakeholder so that a lot of terms and definitions have been established, such as acceptable delay, network delay, on-time performance, reactionary delays, delays per flight-gate to gate, arrival delays, departure delays, surface taxiing delays, and passenger delay minutes (cf. [36]). In the current contribution, cancellations will not be considered.

D. Flight plan and weather data

If the airport performance and flight plan data are combined with the weather data a more complete picture about airport operations and their weather dependencies will be arise. Fig. 6(a) exhibits how the delay at the airport increases rapidly to 795 minutes (accumulated delay minutes from all flights in a 1 hour period) at the beginning of the day of operations due to a 2 hour period of fog (ATMAP score 5).



(a) 3^{rd} September 2014

(b) 28^{th} July 2014

Fig. 6: Airport performance data and ATMAP weather score at EGKK.

The data from EGKK also confirm that the impact of weather events at the airport performance is even higher with an increasing severity level. Fig. 6(b) shows the consequences of 4 hours (06:50 - 10:20 hours) of thunderstorm and rain in the vicinity of the airport. Since the traffic demand is increasing due to this time as well, the accumulated delay could only reduced slowly over the whole day and effects the operations of nearly all airplanes at this day.

III. Machine Learning Approach

Our approach differs significantly from the basic considerations made at the ATMAP algorithm [33]. ATMAP is based on expert knowledge, which makes a categorized evaluation of the weather phenomena. Certain effects of airport performance are linked to five weather classes (see sc. II-B). One example is wind categorization: coefficient 1 stands for wind speed small than 15 kt, coefficient 4 for greater than 30 kt. These categories are taken as generally valid for European airports and are not linked to specific airports or regions. We take a critical view of this, since the location of an airport and the meteorological conditions there have a significant influence on its performance. Thus, we want to address these points with our machine learning approach and focus on two core features in the model development:

- the model must be impact-based
  (i.e. we link effects to their causes),
- the model must be adaptive
  (i.e. enable an airport-specific assessment).

In order to evaluate the performance of an airport, it is essential to use the performance (delay in scheduled operations) itself as a benchmark. Therefore, in our model we convert a weather categorization into an *inverse problem* - we deduce from an output (airport performance) its causes (weather). The basic idea behind this approach is that weather phenomena are categorized based on an assignment to effect categories and not vice versa. Therefore the method includes the following steps.

1) *data preparation* of flight schedules/ weather data
2) *clustering*, class creation of impact data
3) *model creation*, parametrization and setup
4) *model training*, application of model to data
5) *evaluation*, error measurement

Due to a large amount of data, non-linear time series and interdependencies, self-learning algorithms are used as a model. These algorithms offer possibilities for independent, complex solutions to similar problems.

A. Classification with Neural Networks

At first, it is essential to distinguish between classification and regression. Classification is about predicting a label and regression is about predicting a quantity. Classification predictive modeling is the task of approximating a mapping function f from input variables X to discrete output variables Y. The output variables are often called *labels* or *categories*. The mapping function predicts the class or category for a given observation.

Classification problems can be solved by a variety of methods within and outside machine learning. They all have advantages and disadvantages and their applicability depends on the particular application. Examples are Logistic Regression, k-Nearest Neighbors and Support Vector Machines (SVM). However, neural networks also offer the possibility to classify non-linear, multivariate data. Thus it is possible to build an adaptive decision support that delivers complex but fast outputs to specific input sets. This is the reason why we focus on neural networks in our application.

There are two main approaches to neural networks that are suitable for time series classification and that have proven successful several applications. These are *Convolutional Neural Network Models (CNN)* and *Recurrent Neural Network Models (RNN)*, especially their sub-type *Long Short-Term Memory (LSTM)*.

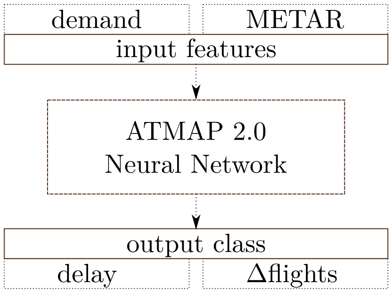


Fig. 7: Neural Network Classification.

RNNs and LSTMs are recommended to detect short-term correlations that have a natural order, while CNN is better able at deriving long-term repeated interdependencies. The reason for this is that RNN could take advantage of the time correlation between measurements, and CNN is better at learning deep traits contained in recursive patterns [39]. The benefit of using LSTMs for sequence classification is that they can learn from the raw time series data directly (see [40]).

B. Impact Clustering

In order to enable the process of classification, the target data streams must be labeled (see fig. 8). This label creation can be done algorithmically or added by expert knowledge. In contrast to ATMAP, however, the effects, not the causes, are labeled here. The labels of the target variables should represent effect categories, which severity the respective effect has on the performance of the airport. Using the general example of the delay, it can be discussed, similar to [41], that a deviation in time of -15 min to 5 min (on time) has no significant influence on the performance of an airport. Based on this, further categories can be found that capture the severity of a growing deviations (delay). This, in turn, is an individual value that refers to the specific airport and demand/capacity conditions there. This form of label creation requires local expertise. It should also be noted that the exclusive consideration of delay is a 1D target label. In this case creating intervals is recommended. The use cases consider these intervals under the abbreviation ”1D” (see sec. IV).

For the algorithmic, multidimensional label creation different methods can be used, which allow a categorization without expert knowledge. In the field of machine learning, this is referred to *unsupervised learning*. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known (or labeled) outputs. A basic method from this area is *k-Mean clustering*. This method searches for a defined number of k groups in a dataset which are similar to each other and takes into account the underlying patterns. To process the learning data, the k-Mean algorithm in data mining begins with a first set of randomly selected centroids used as a starting point for each cluster, and then performs iterative calculations to optimize the centroid positions [42].

IV. Application

The following experiments take up the basic idea of fig. 7 and adapt it to the existing data foundations (see fig. 8). The neural network serves as an adaptive intermediate model and processes weather and airport performance data. The selection of these data can be done algorithmically as well as by expert knowledge. The same applies to the classification of the impact (performance impact at the airport). The neural network itself is determined by its parameters, its structure and the range of data available.

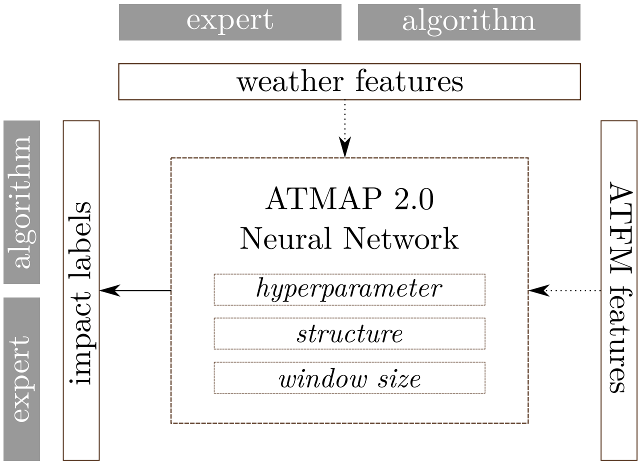


Fig. 8: Classification of time-discrete data streams.

We implemented the given neural networks in *Python 3.6.5* using the open-source deep learning library *Keras 2.2.4* (frontend) with open-source framework *TensorFlow 1.12.0* (backend) and *Scipy 1.0.0* (routines for numerical integration and optimization). Training and testing were performed on GPU (NVIDIA Geforce 980 TI) using CUDA as parallel computing platform and application programming interface. Similar experiments have also been carried out for regression, in which LSTM models were used to predict delays (cf [43]).

A. Data preparation

In the first step of the application, the raw data must be prepared for the use of machine learning. This includes the selection of features for input and output and the classification of output. Similar to clustering, the choice of features also offers the possibility of solving this algorithmically or of drawing on expert knowledge. The feature selection for the following applications is made with expert knowledge. Features are all numerically accessible factors of METAR dataset (see Input A in tab. IV). The airport performance is decisively determined by the relationship between demand and capacity, where capacity significantly influenced by weather events. An imbalance between the demand (scheduled movements) and capacity (actual movements) results in delays, which are added as a supplementary airport performance indicator.

TABLE IV: Feature selection of weather and airport performance data.

	Features
Input A	wind speed, visibility, temperature, humidity, pressure, wind direction, heat index, wind gust speed; actual aircraft movements (arrivals and departures)
Input B	Input A + snow, rain, fog
Input C	Input B + thunder, hail, precipitation; scheduled aircraft movements (arrivals and departures)

The indicators snow, rain, fog, thunder, hail and precipitation have not turned out to be statistically significant (only few events are observed in 2014 and 2015) and are therefore added separately. In the latter case, $n_{arrivals\_scheduled}$ and $n_{departures\_scheduled}$ are added to increase detail.

Both delay and the deviation of $n_{flights\_scheduled}$ from $n_{flights\_actual}$ can be used sensibly as outputs. It should be noted that the METAR data is provided at 30 min intervals and the output values need to refer to these slots. A uniform database is indispensable for the learning process. Either the METAR data is mapped to the single flights or the flight data is aggregated to the 30 min slots of the weather data. Since we consider the constant 30 min time slots as an advantage in the learning process, we have decided for the second variant. As aggregated value, the average, absolute delay as the deviation from the scheduled time to the actual time at the gate is used. The deviation $\Delta\text{flights} = n_{flights\_scheduled} - n_{flights\_actual}$ is calculated absolutely per 30 min slot.

The outputs are either mapped to intervals for the 1D case (delay) or clustered the 2D case (delay and Δflights) (see tab. V). For both cases we create $k = 5$ impact labels. The result of the k-Mean clustering is shown in fig. 9.

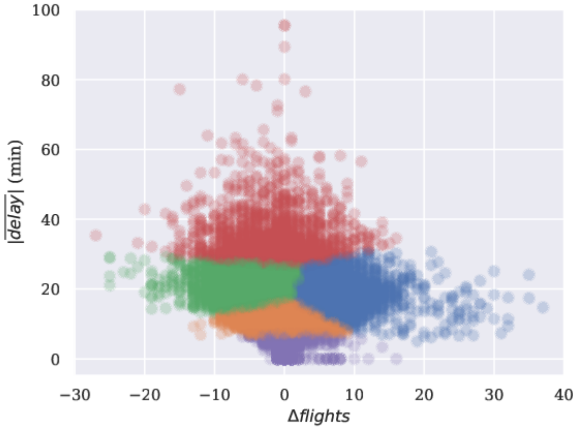


Fig. 9: k-Mean-clustered 2D-data set with $k = 5$ clusters.

The combination of two indicators is intended to reflect the effect of meteorological conditions on the dynamics of an airport in greater detail. The limits of the intervals for 1D are inspired by [41]. The properties of the intervals or clusters are shown in tab. V.

TABLE V: Label creation with clustering and intervalization.

	borders / centroids	sizes
1D output	$\begin{bmatrix} 5 \\ 15 \\ 25 \\ 50 \end{bmatrix}$	$\begin{pmatrix} 0 & 2,740 \\ 1 & 10,340 \\ 2 & 9,066 \\ 3 & 2,515 \\ 4 & 72 \end{pmatrix}$
2D output	$\begin{bmatrix} -0.20 & 11.38 \\ -1.55 & 34.20 \\ 0.15 & 2.42 \\ 6.50 & 18.40 \\ -3.51 & 20.00 \end{bmatrix}$	$\begin{pmatrix} 0 & 1,757 \\ 1 & 9,097 \\ 2 & 4,080 \\ 3 & 3,854 \\ 4 & 5,945 \end{pmatrix}$

For the faultless use of k-Mean, delay outliers with a value of ≥ 100 min were removed. However, only 2 sequences were deleted. The lines of the centroid matrices represent a combination of flights and delay. The first line corresponds for example to -0.20 Δflights and 11.38 min delay, whereby it should be noted that Δflights assume integer values, but this is not relevant for the centroids.

B. Neural Network Setup

The result of a learning process of neural networks depends essentially on their structure and parametrization. A certain set of parameters determines this process. For the following experiments, parameters should remain uniform across applications, others should be changed.

TABLE VI: Hyperparameter setup of the networks.

	individual parameter	constant parameter
Config I	n_{layer} = 100, window size = 50	optimizer = Adam,
Config II	n_{layer} = 100, window size = 6	n_{epochs} = 20,
Config III	n_{layer} = 10, window size = 6	batch size = 64

The optimizer determines the learning rate ($\eta = 10^{-3}$) of the network, the number of epochs the repetitions. This increases to a total of 300, since it is a stochastic process that is performed 15 times each for evaluation. The number of layers determined the complexity of the network, the window size includes the inclusion of past time steps in the calculation. A higher value would have no relation to reality with 48 half-hour sequences per day. A total of 24,733 sequences are available, of which 66% (16,324) are used for training and 34% (8,409) for testing.

In addition to the parametrization, the structure is essential. This differs depending on the paradigm. Tab. VII exhibits the chronological order of the used layers.

Of the core layers mentioned, LSTM, Conv1D and ConvLSTM2D are essential components of the defined paradigms. In addition to these layers, dropout is particularly noteworthy. It helps to avoid overfitting by setting a fraction rate of 0.5.

TABLE VII: Structure of the network paradigms.

	LSTM	Conv1D	ConvLSTM2D	Dropout	MaxPooling1D	Flatten	Dense
LSTM	x			x			x
CNN		x		x	x	x	x
CNN-LSTM	x	x		x	x	x	x
ConvLSTM			x	x		x	x

C. Model fitting and evaluation

As already described, each application is executed 15 times. The reason for this is that neural networks are stochastic, which means that a different specific model is created when training the same model configuration with the same data. This is an advantage of the network, because it gives the model its adaptability, but requires a more complex assessment of it. Tab. VIII shows the final accuracies after 15 applications of the respective model - for a better overview the corresponding σ have been omitted. These values include the correctly mapped input-output ([weather,demand]-delay) combinations and thus the quality of the trained net.

TABLE VIII: Accuracy of LSTM, CNN, CNN-LSTM and ConvLSTM with different data sets and hyperparameters.

		Config I		**Config II**		**Config III**	
		1D	*2D*	*1D*	*2D*	*1D*	*2D*
LSTM	Input A	82.1%	89,6%	72.5%	66.0%	70.5%	72.0%
	Input B	76.8%	73.6%	75.1%	67.0%	71.8%	70.1%
	Input C	64.3%	55.7%	63.8%	52.6%	42.8%	47.9%
CNN	Input A	90.1%	96.9%	74.1%	68.8%	72.1%	62.8%
	Input B	36.4%	23.8%	42.8%	41.0%	42.8%	37.7%
	Input C	36.3%	61.3%	42.8%	37.7%	42.8%	43.0%
CNN-LSTM	Input A	72.6%	74.3%	70.6%	61.9%	66.6%	53.6%
	Input B	42.4%	36.6%	41.6%	53.6%	42.7%	39.3%
	Input C	41.5%	55.2%	41.9%	61.4%	42.2%	55.2%
ConvLSTM	Input A	87.4%	88.2%	74.2%	77.3%	74.2%	77.2%
	Input B	36.3%	61.3%	42.8%	50.3%	42.8%	37.7%
	Input C	42.6%	37.6%	42.8%	34.0%	42.8%	37.8%

The best results in tab. VIII are achieved uniformly with a 2D target label using the most elaborate learning configuration. In all cases, this refers to Input A, i.e. the consideration of statistically relevant weather phenomena and demand. The ConvLSTM achieves similarly strong results with 1D outputs.

Input B has the lowest accuracy of all paradigms (except LSTM). This is due to the addition of weather indicators that have already proved irrelevant in exploratory analyzes. They falsify the result. The results are improved by adding further performance indicators to input C. It shows that the LSTM can handle irrelevant inputs better than the CNN.

A merger of LSTM and CNN is intended to ensure that even irrelevant inputs do not lead to distorted results. The results of input B only partially confirm this for both CNN-LSTM and ConvLSTM. The consideration Config I - 2D - Input B is the most inexact of all paradigms at CNN.

The computing times dropped considerably from configuration I to III. It should also be noted that the use of hybrid paradigms (CNN-LSTM, ConvLSTM) led to a sharp increase in computing time.

D. Model summary

From the results in tab. VIII it can be concluded that there are several satisfactory model solutions for different prerequisites, data and configurations. A decision depends largely on the available data and the level of detail of the investigation.

The results show that a transfer of weather events and performance indicators to a classified delay is possible. It should be mentioned here that extracting knowledge from the neural network can provide added value. Possible approaches exist for all paradigms. However, it is difficult to express the influence of individual meteorological components on the overall delay, like in [33]. A combinatorial conclusion must always be assumed which can describe the delay as the end product of several dependent inputs. The results show that this combinatorial mapping by machine learning is possible.

The trained network can therefore be used as adaptive decision support, taking into account the local conditions of the airport 1:1. A prediction of decision recommendations is determined by the inputs. These are weather data and demand in our experiment. While the demand is anchored in the flight plan by STA and STD and offers a wide forecast horizon, a weather forecast is limited by the size of the TAF. This amounts to a period of 6 hrs and means a forecast horizon of 12 time slots with width 30 min.

Note that this is a multi-step prediction and no one-step prediction. This means that the value predicted by the model is not updated by the real value over the entire forecast horizon. Labels that are predicted incorrectly are carried accordingly. Input values that flow into the calculation of a delay label refer to the window size.

A short, representative example (tab. IX, fig. 10 and 11) refers to flight plan data at EGKK on 3^{rd} September 2014.

TABLE IX: Label prediction for 3^{rd} September 2014.

	timeslot					
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$
EGKK	4	4	1	2	2	2
LSTM	4	4	1	***1***	2	2
CNN	4	4	1	2	2	2
CNN-LSTM	4	***2***	***2***	2	2	2
ConvLSTM	4	4	1	2	2	***1***
	label					

All networks were used in their best configuration (for input A, configuration I, 2D). The start pulse thus comprises a sequence of 50 time steps of preceding time slots. A 6-step prediction is performed. The values of tab. IX and fig. 11 represent the labels (cluster numbers) of the k-Mean clustering. The data refer to the presented example day from fig. 6(a), but are already labeled and assigned to the slots of the weather data. Fig. 10 depicts the underlying dataset.

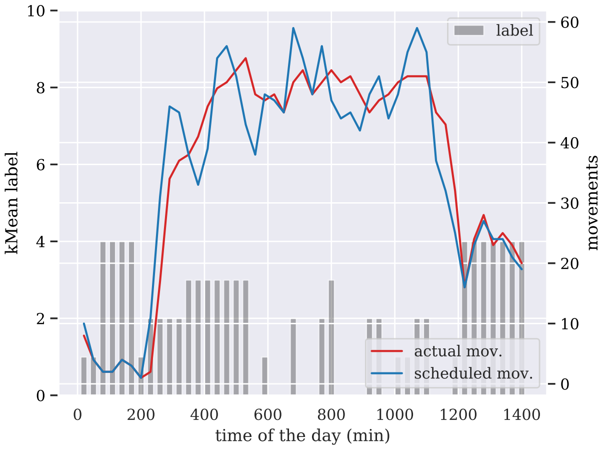


Fig. 10: 3rd September 2014, labeled and slotted.

The advantage of neural networks, especially recurrent paradigms, is the integration of parallel or past knowledge. Thus, errors can occur in predictions which do not lead to a continuation of erroneous predictions. Tab. IX shows that LSTM, CNN-LSTM and ConvLSTM each have one or two prediction errors. The comparison to the actual labels of EGKK is represented in fig. 11.

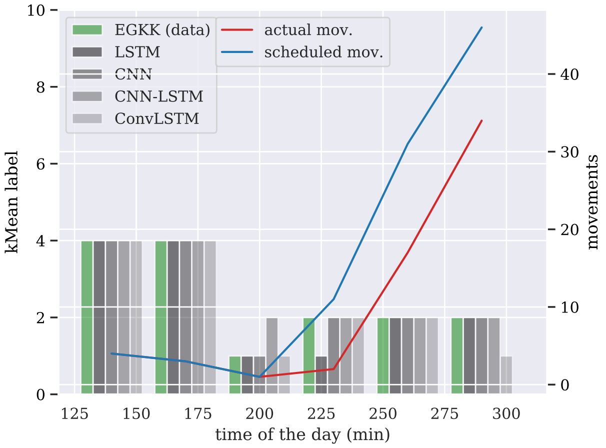


Fig. 11: 3rd September 2014, 6 slot prediction of tab. IX.

The application shown is only intended as an example of a possible use of the model. In further investigation we want to find out, how a decision support for the user can be derived from the trained knowledge. An example would be an optimal adaptation of the demand to meteorological conditions to minimize the overall delay. Furthermore, a complex quantification of the individual components of METAR makes sense, which, however, should be addressed separately due to complex internal interactions of the weather components.

V. Conclusion

In this paper we investigated a quantification of the influence of meteorological conditions on the individual airport performance at London Gatwick airport (EGKK) using machine learning. Different paradigms of neural networks were used and combined in order to process the corresponding data foundation in a target-oriented way. The accuracy of the trained networks was compared and finally exemplarily applied.

The developments in this paper contrast with the mechanism used by the ATMAP algorithm [33], since our approach follows a effect-to-cause relation. From the grouped performance effects at the airport (e.g. delayed operations), machine learning methods were used to draw conclusions about the underlying weather data. The grouping took place by means of unsupervised learning, the classification by supervised learning. Especially pure paradigms from LSTM and CNN show satisfactory results and allow a weather-related decision support for future airport operations.

A more detailed splitting of the data set makes sense in the future in order to gain model complexity. In particular, a separation of arrivals and departures and the associated delays makes sense in the future, as arrival delays also represent reactionary delays of the feeder airports. In this respect, a more specific splitting of the data set would also be useful. The weather will only have a significant influence on the capacity when the airport is working at its capacity limits as far as possible and can no longer serve the demand. We aim to continue our experiments in this direction. Therefore, we want to investigate further airports and airport clusters [34] with regard to their specific weather impacts.

References

[1] C. Bongiorno, G. Gurtner, F. Lillo, R. N. Mantegna, and S. Miccichè, “Statistical characterization of deviations from planned flight trajectories in air traffic management,” *Journal of Air Transport Management*, vol. 58, p. 152–163, jan 2017.

[2] J. Bronsvoort, G. McDonald, R. Porteous, and E. Gutt, “Study of aircraft derived temporal prediction accuracy using FANS,” in *Proceedings of the 13th ATRS World Conference*, 2009.

[3] E. Mueller and G. Chatterji, “Analysis of aircraft arrival and departure delay,” in *Proceedings of the AIAA ATIO Conference*, 2002.

[4] EUROCONTROL, “Performance Review Report – An assessment of air traffic management in Europe during the calendar year 2014, 2015, 2016,” Performance Review Commission, Tech. Rep., 2017.

[5] M. Tielrooij, M. C. Borst, M. van Paassen, and M. Mulder, “Predicting arrival time uncertainty from actual flight information,” in 11th *USA/Europe Air Traffic Management Research and Development Seminar (ATM2015)*, 2015.

[6] EUROCONTROL, “CODA Digest all-causes delay and cancellations to air transport in europe – 2016,” CODA, Tech. Rep., 2016.

[7] A. Cook, G. Tanner, and P. Enaud, “Quantifying airline delay costs - the balance between strategic and tactical costs,” in 14th *Air Transport Research Society (ATRS) World Conference*, 2010.

[8] M. G. Sohoni and S. Erat, “Can time buffers lead to delays? the role of operational flexibility,” *SSRN*, April 2015.

[9] J. Rosenow, M. Lindner, and H. Fricke, "Impact of climate costs on airline network and trajectory optimization: a parametric study," *CEAS Aeronautical Journal*, vol. 8 (2), p. 371–384, 2017.

[10] M. Niklaß, B. Lührs, V. Grewe, K. Dahlmann, T. Luchkova, F. Linke, and V. Gollnick, "Potential to reduce the climate impact of aviation by climate restricted airspaces," *Transport Policy*, p. In Press, 2017.

[11] J. Rosenow, H. Fricke, T. Luchkova, and M. Schultz, "Minimizing contrail formation by rerouting around dynamic ice-supersaturated regions," *Aeronautics and Aerospace Open Access Journal*, vol. 2 (3), p. 105–111, 2018.

[12] I. Gerdes, A. Temme, and M. Schultz, "Dynamic airspace sectorisation for flight-centric operations," *Transportation Resarch Part C: Emerging Technologies*, vol. 95, p. 460–480, 2018.

[13] I. Gerdes, A. Temme, and M. Schultz, "Dynamic airspace sectorization using controller task load," in 6th *SESAR Innovation Days (SIDs)*, 2016.

[14] T. Standfuss, I. Gerdes, A. Temme, and M. Schultz, "Dynamic airspace optimisation," *CEAS Aeronatuical*, vol. 9 (3), p. 517–531, 2018.

[15] B. F. Santos, M. M. E. C. Wormer, T. A. O. Achola, and R. Curran, "Airline delay management problem with airport capacity constraints and priority decisions," *Journal of Air Transport Management*, vol. 63, p. 34–44, aug 2017.

[16] S. Carlier, I. de Lépinay, J. Hustache, and F. Jelinek, "Environmental impact of air traffic flow management delays," in 7th *USA/Europe Air Traffic Management Research and Development Seminar (ATM2007)*, 2007.

[17] J. Rosenow, H. Fricke, T. Luchkova, and M. Schultz, "Impact of optimised trajectories on air traffic flow management," *The Aeronautical Journal*, vol. 123, no. 1260, p. 157–173, 2019.

[18] S. Reitmann, A. Gillissen, and M. Schultz, "Performance benchmarking in interdependent atm systems," in 7th *International Conference for Research in Air Transportation (ICRAT)*, 2016.

[19] B. Campanelli, P. Fleurquin, A. Arranz, I. Etxebarria, C. Ciruelos, V. M. Eguíluz, and J. J. Ramasco, "Comparing the modeling of delay propagation in the US and European air traffic networks," *Journal of Air Transport Management*, vol. 56, p. 12–18, sep 2016.

[20] N. Ivanov, F. Netjasov, R. Jovanović, S. Starita, and A. Strauss, "Air traffic flow management slot allocation to minimize propagated delay and improve airport slot adherence," *Transportation Research Part A: Policy and Practice*, vol. 95, p. 183–197, 2017.

[21] A. Cook, L. Delgado, G. Tanner, and S. Cristóbal, "Measuring the cost of resilience," *Journal of Air Transport Management*, vol. 56, p. 38–47, sep 2016.

[22] S.-L. Proag and V. Proag, "The Cost Benefit Analysis of Providing Resilience," *Procedia Economics and Finance*, vol. 18, p. 361–368, jan 2014.

[23] A. Cook, G. Tanner, V. Williams, and G. Meise, "Dynamic cost indexing – Managing airline delay costs," *Journal of Air Transport Management*, vol. 15, no. 1, p. 26–35, jan 2009.

[24] A. Islami, M. Sun, S. Chaimatanan, and D. Delahaye, "Optimization of military missions impact on civilian 4D trajectories," in *ENRI International Workshop on ATM/CNS (EIWAC 2017)*, 2017.

[25] T. Luchkova, R. Vujasinovic, A. Lau, and M. Schultz, "Analysis of impacts an eruption of volcano Stromboli could have on european air traffic," in *USA/Europe ATM R&D Seminar (11th ATM Seminar)*, 2015.

[26] M. Kreuz, T. Luchkova, and M. Schultz, "Effect of restricted airspace on the ATM system," in *WCTR Conference*, 2016.

[27] J. Rosenow, H. Fricke, and M. Schultz, "Air traffic simulation with 4D multi-criteria optimized trajectories," in *Winter Simulation Conference*, 2017, p. 2589–2600.

[28] S. Kaltenhaeuser, F. Morlang, T. Luchkova, J. Hampe, and M. Sippel, "Facilitating sustainable commercial space transportation through an efficient integration into air traffic management," *New Space*, vol. 5 (4), p. 244–256, 2017.

[29] E. Sunil, J. Hoekstra, J. Ellerbroek, F. Bussink, D. Nieuwenhuisen, A. Vidosavljevic, and S. Kern, "Metropolis: Relating airspace structure and capacity for extreme traffic densities," in *USA/Europe ATM R&D Seminar (11th ATM Seminar)*, 2015.

[30] A. Montlaur and L. Delgado, "Flight and passenger delay assignment optimization strategies," *Transportation Research Part C: Emerging Technologies*, vol. 81, p. 99–117, 2017.

[31] L. Delgado and X. Prats, "Operating cost based cruise speed reduction for ground delay programs: Effect of scope length," *Transportation Research Part C: Emerging Technologies*, vol. 48, p. 437–452, 2014.

[32] G. Buxi and M. Hansen, "Generating day-of-operation probabilistic capacity scenarios from weather forecasts," *Transportation Research Part C: Emerging Technologies*, vol. 33, p. 153–166, 2013.

[33] EUROCONTROL, "Algorithm to describe weather conditions at european airports," Performance Review Unit, Tech. Rep., 2011.

[34] M. Schultz, S. Lorenz, R. Schmitz, and L. Delgado, "Weather impact on airport performance," *Aerospace*, vol. 5, 2018.

[35] F. A. Administration, "Advisory Circular 00-45 – Aviation Weather Services," Federal Aviation Administration, Tech. Rep., 2016.

[36] S. O'Flynn, "Airport capacity assessment methodology - ACAM manual," Network Manager, EUROCONTROL, Tech. Rep. 1.1, November 2016.

[37] M. Schultz and H. Fricke, "Managing passenger handling at airport terminal," in 9th *USA/Europe Air Traffic Management Research and Development Seminar (ATM2011)*, 2011.

[38] P. Röhner, "Modelling of punctuality at Frankfurt airport," Ph.D. dissertation, Gottfried Wilhelm. Leibniz Universität Hannover, 2009.

[39] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep Learning for Sensor-based Activity Recognition: A Survey," *Pattern Recognition Letters*, Feb. 2018, arXiv: 1707.03502.

[40] M. Schultz and S. Reitmann, "Machine learning approach to predict aircraft boarding," *Transportation Research Part C: Emerging Technologies*, vol. 98, p. 391–408, jan 2019.

[41] EUROCONTROL, "A Matter of Time: Air Traffic Delay in Europe," Tech. Rep., Sep. 2007.

[42] S. Na, G. Yong, L. Xumin, K. A. A. Nazeer, M. P. Sebastian, M. A. Dalal, N. D. Harale, U. L. Kulkarni, D. T. Pham, S. S. Dimov, and C. D. Nguyen, "An Improved K-means Clustering," p. 5, 2015.

[43] S. Reitmann and M. Schultz, *Computation of Air Traffic Flow Management Performance with Long Short-Term Memories Considering Weather Impact*, ser. Lecture Notes in Computer Science. Springer, 2018, vol. 11140, pp. 532–541.

Stefan Reitmann is PhD student at the German Aerospace Center (DLR). In 2015 he received his diploma in Traffic Engineering at the Dresden University of Technology with focus on traffic flow sciences. In 2016 and 2017 he was visiting scientist at the State University St. Petersburg (SPbU). His research focus is on machine learning and big data analysis, especially for the usage of neural networks in traffic sciences.

Sameer Alam is an Associate Professor and Program Director of Artificial Intelligence at the Air Traffic Management Research Institute, Nanyang Technological University, Singapore. He obtained his PhD in Computer Sc. from University of New South Wales, Australia in 2008. His research interests are in complex system modeling, machine learning and optimization algorithms applied to air traffic management problems.

Michael Schultz is principle investigator and senior researcher at TU Dresden, Institute of Logistics and Aviation. He was Head of the Air Transportation Department at the German Aerospace Center (DLR, 2014-2019). He holds a PhD in Aviation Engineering (2010) and obtains his Habilitation for Aviation/Aerospace (2019). His research focus on data-driven, machine learning and model-based approaches to improve air traffic management and airport operations. In particular, he researches on dynamic, flow-centric management of airspaces, inter-airport coordination, performance-based airport operations, and advanced concepts for the future air (urban) traffic management.