Applied Data Science Profile

Assessment of Cirrus Cloud Properties under Anthropogenic Forcing through High-Density Air Traffic over Europe

Research Project Report

Utrecht, April 26, 2021

Edited by

Timothy van der Duim
Faculty of Natural Sciences

2021
University of Utrecht

Summary

This report is the final deliverable of a research project which is part of the Applied Data Science profile at Utrecht University (MSc phase). A brief description is given of the project, including the final results. The purpose of this research is to provide more insight into the role of anthropogenic cirrus formation through air traffic, by investigating the high-density European airspace over a period spanning several years (2015-2020). This is attempted by combining several data sources into a ”Big Data” project, exploiting the strengths of each product to create a data-based model that works on high spatial and temporal resolution.

Data from Meteosat SEVIRI have been combined and validated with CALIPSO's CALIOP data to deduce temporal cirrus cover variability over a rectangular region bound by ($10°$W-$35°$N) and ($40°$E-$60°$N). Cirrus clouds, predominantly occurring between 7 and 14 km altitude, are correlated with air traffic flying on corresponding altitudes. ERA5 Reanalysis data, provided by the EMCWF, delivered data on air temperature and relative humidity in the upper troposphere and lower stratosphere. Meteorology was incorporated into the model as it is of major influence on the formation, lifetime and optical properties of cirrus clouds. More specifically, the impact of aviation on cirrus cover in supersaturated air (relative humidity w.r.t. ice 100% or larger) an sub-saturated air (relative humidity w.r.t. ice smaller than 100%) have been evaluated in parallel. In the former case a statistically significant correlation was found between air traffic density and cirrus cover. In the latter case there was no statistically significant relation found.

In conjunction with aforementioned a Logistic Regression model and a Random Forest model were trained and tested on meteorological data, with as aim a more data-centralized handling of the data. In the end the model performances were not sufficiently adequate to apply them on the data. The long-term time series analysis has resulted in a signal strongly dictated by meteorological conditions. No long-term (6 year) trend was found in mean cirrus cover following the annual increase in air traffic.

Contents

1

Project Outline

Cirrus clouds are ubiquitous clouds composed of ice crystals that are long-lived under the right conditions, and that reside predominantly in the upper troposphere and lower stratosphere. Their major importance relates to their effect on Earth's radiative balance. High, optically thin clouds like cirrus have the tendency to be nearly transparent to incoming shortwave radiation, while they are opaque to planetary longwave radiation [8]. Hence they tend to trap and reflect down more radiation than is reflected or re-emitted into space, leading to a net warming effect on the lower atmosphere.

Cirrus clouds can be formed either naturally or due to anthropogenic activity, in the latter case due to air traffic (aircraft-induced cirrus or *AIC*). Out of aviation-derived radiative forcing components, the AIC component is estimated to be potentially larger than the direct component from greenhouse gases (particularly CO_2 and NO_x) [10]. These observations, in combination with low confidence estimates of cloud radiative forcing feedbacks, exemplify the need of a thorough assessment of cirrus cloud properties (*CCP*) and the way its properties are affected by anthropogenic activity.

The meteorological state of the atmosphere plays a key role in the formation and on the lifetime of cirrus clouds. Cirrus clouds exist of ice particles which can only be formed and prevail if the air is supersaturated with respect to ice [10]. Ice-supersaturation can occur under sufficiently cold and moist conditions. For condensation it is required that $RH \geq 100\%$ where RH is the relative humidity of ambient air. Water droplets consequently freeze under sufficiently cold conditions. The Schmidt-Appleman criterion, which formulates the conditions under which supersaturation takes place, can be used to assess whether atmospheric conditions are cold enough for cirrus clouds to form, see Figure 1.1.

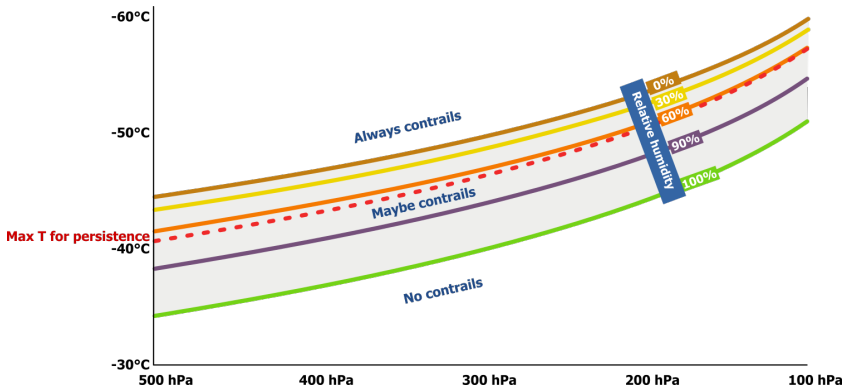


Figure 1.1. Appleman chart. Retrieved from [1].

The main question this research wants to answer is whether the data at hand provides evidence for a spatial-temporal correlation of air traffic density with *cirrus cloud cover*. This is done using state-of-the-art satellite products and a big data approach on high resolution data. The research focuses on European airspace due to its high air traffic density and data coverage.

The persistent rise in global European air traffic up till the COVID pandemic [12] (Figure 1.2) raises the hypothesis that cirrus cover might be rising along, assuming a positive relation between cirrus formation and air traffic density. The COVID pandemic furthermore offers the possibility to compare cirrus cloud cover during the pandemic, with exceptional low air traffic density, with previous years.

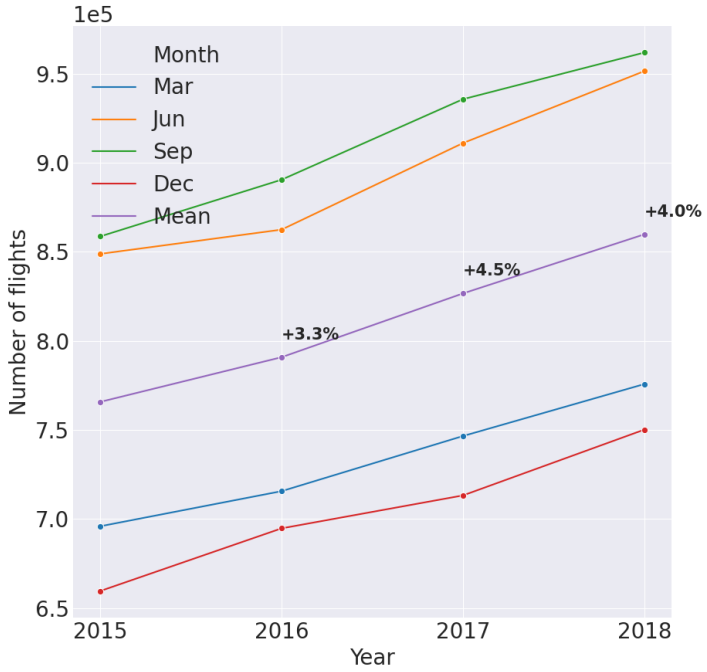


Figure 1.2. Total number of flights per month detected by European radar, for four months a year over the period 2015-2018. Data derived and analyzed from EUROCONTROL OneSky database [retrieved on Jan 12 2021].

Besides meteorology and air traffic density, engine combustion type also relates to cirrus formation [15]. The engine type determines the amount and composition of engine combustion products which could potentially lead to cirrus formation. In addition, [15] argues that high by-pass engines lower the exhaust temperature leading to sooner (i.e. lower temperature threshold) contrail formation. By looking into engine types, the impact of combustion type on cirrus occurrence is investigated.

In summary, two research questions arise:

1. How does contrail formation evolve over time between 2015 and 2020 under 1) increasing air traffic and 2) a global pandemic?
2. How large is the impact of air traffic on cloud cover and does combustion type significantly alter the results?

A major challenge involved in this research is the existence of confounding variables. Cirrus formation is influenced by many factors amongst which atmospheric conditions, and its formation might be naturally or aircraft-induced. In addition, when considering a closed volume of air, any appearance of cirrus at a given time might be either formed locally or advected from adjacent regions. In short, meteorology cannot be disregarded.

2

Data Description & Data Wrangling

Cirrus clouds are due to their optically thin nature not easily detectable by satellites. Over the past decades more satellites have been equipped with sensors that are able to detect cirrus, using the Shortwave-Infrared (*SWIR*) portion of the electromagnetic spectrum at 1.38μm. At this wavelength a large portion of the radiation is reflected by ice crystals contained in cirrus clouds, while nearly all radiation beneath the cloud is absorbed by water vapor. The inability for passive remote sensing satellites like MODIS, Landsat and Sentinel-2 to detect optically thin ($\sigma < 0.2$) cirrus remains. Active satellites, like the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (*CALIOP*) carrying on board the Cloud-Aerosol Lidar with Orthogonal Polarization (*CALIOP*), have the advantage over passive satellites to detect a broader optical range of clouds while not being restricted to daytime sampling. This section provides an overview of the data products involved in this research, including the data relating to air traffic, and the wrangling techniques used.

2.1. Air Traffic

The EUROCONTROL OneSky network database suits the purpose of this project to analyze high-quality air traffic data over Europe on a fine spatial and temporal resolution. The R&D data archive [4] provides data on all commercial flights operating in and over Europe, supplemented with data originating from air navigation service providers' radar and datalink communication.

2.1.1. Air Traffic Data Description

The temporal domain of the EUROCONTROL data spans the years 2015-2018, and the months March, June, September and December. Those datasets contain all flights that pass through European airspace, i.e. they are not restricted to flights that either depart or arrive at a European airport. In fact, those "pass-over" flights account for approximately 1.5% of all registered flights included in the data sets. Each monthly data bundle inter alia consists of a *Flights* data set, including all flights that were registered during that month, and a *Flight Points* file including all filed or actual flight paths. All data sets are in CSV format.

Within the *Flights* data set each row is a unique flight, hence the total number of rows is the total number of registered flights. The data sets include a unique identifier for each flight *ECTL_ID*, the ICAO departure airport code *ADEP*, the latitude of the departure airport *ADEP Latitude*, the longitude of the departure airport *ADEP Longitude*, the ICAO destination airport code *ADES*, the latitude of the destination airport *ADES Latitude*, the longitude of the destination airport *ADES Longitude*, the planned arrival time *Filed Arrival Time*, the time an aircraft departs from its parking spot *Actual Off-Block Time*, the *Actual Arrival Time*, the aircraft type *AC Type*, the aircraft operator *AC Operator*, the unique aircraft identifier *AC Registration*, the *ICAO Flight Type* (either S - scheduled or N - Non-scheduled), the market segment of the operation *STATFOR Market Segment*, the requested cruising flight level *Requested FL* and the distance flown in nautical miles *Actual Distance Flown (nm)* (see Figure 2.1a).

Within the *Flight Points Actual* data sets each row is a reported aircraft location at a given time based on

radar tracking. The data sets include a unique identifier for each flight *ECTL_ID*, a numeric sequence number of the points crossed by the flight *Sequence Number*, the time at which the point was passed *Time Over*, the flight level (altitude) at that point *Flight Level*, the latitude at that moment *Latitude* and the longitude at that moment *Longitude* (see Figure 2.1b). For March 2015 the *Flights* data set consists of 698,715 rows (unfiltered) and the *Flights Points Actual* data set of 17,549,122 rows (unfiltered), meaning flights were tracked about 25 times on average during their operation.

flights_0315 - DataFrame

Index	ID	ICAO_dep	CAO_dest	AC_type	C_operato	AC_regis	light_type	market	dist	dep_airport	dest_airport
0	184408024	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
1	184411600	KORD	EHAM	B77L	QTR	A7BFB	S	Traditional Scheduled	3685	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
2	184412879	KORD	EHAM	B748	ABW	VQBLQ	S	All-Cargo	3977	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
3	184414552	KORD	EHAM	B744	CLX	LXRCV	S	All-Cargo	3880	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
4	184429962	KORD	EHAM	B744	KLM	PHBFO	S	Traditional Scheduled	3756	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
5	184430455	KORD	EHAM	B763	UAL	N662UA	S	Traditional Scheduled	3699	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
6	184454965	KORD	EHAM	B744	KLM	PHBFV	S	Traditional Scheduled	3668	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
7	184461620	KORD	EHAM	B744	ABW	VQBHE	S	All-Cargo	3675	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
8	184467593	KORD	EHAM	MD11	MPH	PHMCP	S	Traditional Scheduled	3609	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
9	184479879	KORD	EHAM	B763	UAL	N660UA	S	Traditional Scheduled	3689	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
10	184504092	KORD	EHAM	B744	KLM	PHBFU	S	Traditional Scheduled	3674	Chicago O'Hare International Airport	Amsterdam Airport Schiphol

(a) Flights March 15

flight_mar15 - DataFrame

Index	ECTRL ID	uence Nur	Time Over	light Leve	Latitude	Longitude	Index	ECTRL ID	uence Nur	Time Over	light Leve	Latitude	Longitude
26	184408024	26	2015-03-01 07:17:44	350	50.3667	-8	97	184408027	25	2015-03-01 06:00:56	360	51.9947	-5.04028
27	184408024	27	2015-03-01 07:32:02	350	51.0695	-4.50389	98	184408027	26	2015-03-01 06:08:30	295	51.61	-3.28361
28	184408024	28	2015-03-01 07:46:30	350	51.65	-1.00806	110	184408028	5	2015-03-01 00:23:47	200	48.7022	2.24306
29	184408024	29	2015-03-01 08:00:54	230	52.1147	2.48806	111	184408028	6	2015-03-01 00:33:24	330	47.6392	3.75139
60	184408025	26	2015-03-01 06:57:32	350	50.3667	-8	112	184408028	7	2015-03-01 00:43:16	330	46.6078	5.33167
61	184408025	27	2015-03-01 07:15:54	320	50.5742	-3.46389	113	184408028	8	2015-03-01 00:46:22	330	46.3025	5.87972
62	184408025	28	2015-03-01 07:19:02	277	50.7075	-2.75	114	184408028	9	2015-03-01 00:54:00	287	45.5278	7.08889
63	184408025	29	2015-03-01 07:20:13	263	50.7497	-2.51861	134	184408029	15	2015-03-01 01:43:16	360	36.087	28.4203
64	184408025	30	2015-03-01 07:22:12	229	50.6753	-2.12583	135	184408029	16	2015-03-01 01:46:39	360	36.3397	28.0822
95	184408027	23	2015-03-01 05:44:57	360	52.9089	-8.45361	136	184408029	17	2015-03-01 01:59:31	360	37.3989	26.8383
96	184408027	24	2015-03-01 05:57:38	360	52.2417	-5.68028	137	184408029	18	2015-03-01 02:02:49	360	37.685	26.5286

(b) Flighttracks March 15

flights0319 - DataFrame

Index	callsign	number	icao24	registration	typecode	origin	destination	firstseen	lastseen	day	latitude_1	longitude_1	altitude_1	latitude_2	longitude_2	altitude_2
0	DAL28	nan	abca08	N859NW	A332	nan	EGLL	2019-02-28 00:14:29+00:00	2019-03-01 07:55:29+00:00	2019-03-01 00:00:00+00:00	13.7516	-13.6857	9753.6	51.465	-0.421044	30.48
1	AEA016	nan	34444f	EC-LXR	A333	LEMD	LEMD	2019-02-28 00:46:59+00:00	2019-03-01 02:55:52+00:00	2019-03-01 00:00:00+00:00	40.4923	-3.57474	304.8	40.4877	-3.56786	495.3
2	CSN327	nan	780737	B-6137	A388	YSSY	KHHR	2019-02-28 00:48:19+00:00	2019-03-01 03:19:24+00:00	2019-03-01 00:00:00+00:00	-33.9235	151.17	0	33.9568	-118.356	213.36
3	CCA839	nan	780783	B-6533	A332	nan	LEBL	2019-02-28 00:52:20+00:00	2019-03-01 06:03:21+00:00	2019-03-01 00:00:00+00:00	-37.4077	144.647	2743.2	41.3806	2.37891	800.1
4	ETH506	nan	04004c	ET-AOV	B788	OLBA	SBGR	2019-02-28 01:20:57+00:00	2019-03-01 18:14:55+00:00	2019-03-01 00:00:00+00:00	33.7825	35.4583	609.6	-23.4267	-46.4424	807.72
5	CES771	MU771	780de5	B-5973	A332	YSSY	EHAM	2019-02-28 01:24:15+00:00	2019-03-01 04:22:59+00:00	2019-03-01 00:00:00+00:00	-33.922	151.169	0	52.314	4.74488	15.24
6	CSN461	nan	780aec	B-2042	B77L	KORD	EDDF	2019-02-28 01:25:50+00:00	2019-03-01 10:39:57+00:00	2019-03-01 00:00:00+00:00	41.9725	-87.9977	914.4	50.0368	8.57292	83.82
7	LAN805	nan	e80214	CC-BGA	B789	KMIA	YMML	2019-02-28 02:36:23+00:00	2019-03-01 05:53:16+00:00	2019-03-01 00:00:00+00:00	25.7854	-80.3252	0	-37.6928	144.842	121.92
8	KQA100	nan	04c118	5Y-KZB	B788	nan	EGLL	2019-02-28 02:43:18+00:00	2019-03-01 15:49:44+00:00	2019-03-01 00:00:00+00:00	18.1612	71.2549	10972.8	51.4777	-0.428096	7.62
9	SIA285	SQ285	76cd68	9V-SKH	A388	NZAA	NZAA	2019-02-28 02:50:11+00:00	2019-03-01 00:19:58+00:00	2019-03-01 00:00:00+00:00	-37.0137	174.78	-304.8	-37.0076	174.802	-99.06
10	ACI140S	nan	3a1e43	F-OHSD	A332	RJBB	YSSY	2019-02-28 02:57:43+00:00	2019-03-01 00:07:51+00:00	2019-03-01 00:00:00+00:00	34.4355	135.256	0	-34.0152	151.206	160.02

(c) Flighttracks March 19

Figure 2.1. Example of flight datasets for March 2015 and 2019 (**(a)** and **(b)** from EUROCONTROL database, **(c)** crowdsourced database).

In order to extend the temporal coverage of air traffic data to COVID-19 lockdown period (2020), another data source is consulted. Worldwide crowdsourced air traffic data from the same OpenSky network has been made available for 2019 and 2020 on Zenodo [16]. From this source only one CSV file is provided

for each month, with features highly similar to aforementioned *Flights Points Actual* data sets, the main difference being that those crowdsourced data sets merely contain two reported locations per flight. The first rows for March 2019 are shown in Figure 2.1c.

2.1.2. Air Traffic Data Wrangling

Figure 2.1a shows that the columns '*dep_airport*' and '*dest_airport*' are added. Each unique ICAO airport identifier is coupled to the full airport name by use of an airport data set retrieved from OpenFlights [13].

Conditional filters are applied concerning invalid (NaN) entries. Omitting all flights containing invalid arguments would lead to an undesirable data reduction, particularly for the 2019 and 2020 data sets. E.g. any missing values within the *market* or *AC_operator* column are acceptable, while in case of any missing value in the *Latitude* or *Longitude* column the row would be omitted. The latter applies to 0.02% of all data rows in the *Flight Points* data set.

The temporal resolution at which flights are geolocated is variable and ranges from minutes to more than an hour in some cases. This is shown in the barchart of Figure 2.2, which shows the frequency of time gaps between each pair of consecutive tracked locations on individual flights, for a random collection of 10,000 flights. In order to divide the spatial domains into small cells (see Section 3.1) it is desirable to have a more continuous representation of each flight path, as otherwise data might be "missed" by the algorithm. Take e.g. a grid cell of 0.25° in width (roughly 27 km) which is crossed by an aircraft flying to the East (or West). If the aircraft flies with 1,000 km/h, this means the aircraft crosses the region within ≈ 100 seconds. The time gaps between consecutive flight points should be at most ≈ 50 seconds in this case to ascertain the cell contains two positions of the aircraft: a requirement for the algorithm to compute a travelled distance and assign the correct distance flown to the particular cell. In case two consecutive flight points are located in different cells, the distance flown between the points is assigned to the cell where the aircraft was flying at first - this results spatially in some inaccuracies, but is of no influence to the results when averaging over the entire region. Clearly there are many different ways in which a rectangular cell can be crossed - some paths even shorter than the one described above. Small time intervals in combination with the grid cell size should prevent exclusion of a significant proportion of the data.

Another important notion relates to the time gaps of flight points in conjunction with the time discretization of the temporal domain, which is another parameter of choice. Say the temporal domain is discretized into 15-minute intervals. Within each of those intervals, flights are being filtered out based on the timestamps at each reported location. The algorithm identifies the flight points belonging to the same flight, but only finds the flight points that were within that specific 15-minute time window. Hence the distance-based metric is applied to each 15-minute chunk of each flight path separately, thereby overlooking each portion of the total flight path between two flight points at the adjacent edges of two different time windows. This inaccuracy can be mitigated to a large extent by taking larger time windows (which decreases the temporal resolution of the analysis), and/or by decreasing the time gaps between flight points. Issues related to the spatial-temporal discretization of flight paths result in a "patchy" appearance when mapping air traffic density, which in summary can be mitigated be decreasing the time interval between flight points.

Using a one-minute interval would mean that the time gap between roughly 75% of all consecutive flight geolocations would decrease, as 7% of the time gaps are less than half a minute and 18% around one minute (the rest larger than one minute), see Figure 2.2. For this analysis the time gaps are set to half a minute, as this yielded sufficient model robustness when changing other hyperparameters like the time window and grid cells. An even higher sampling frequency was not attainable due to computational limits. The data is resampled using a linear interpolation method between reported flight points. Performing the interpolation on data of March 2015 with an original file size of around 18 million flight points results in a file containing roughly 83 million flight points.

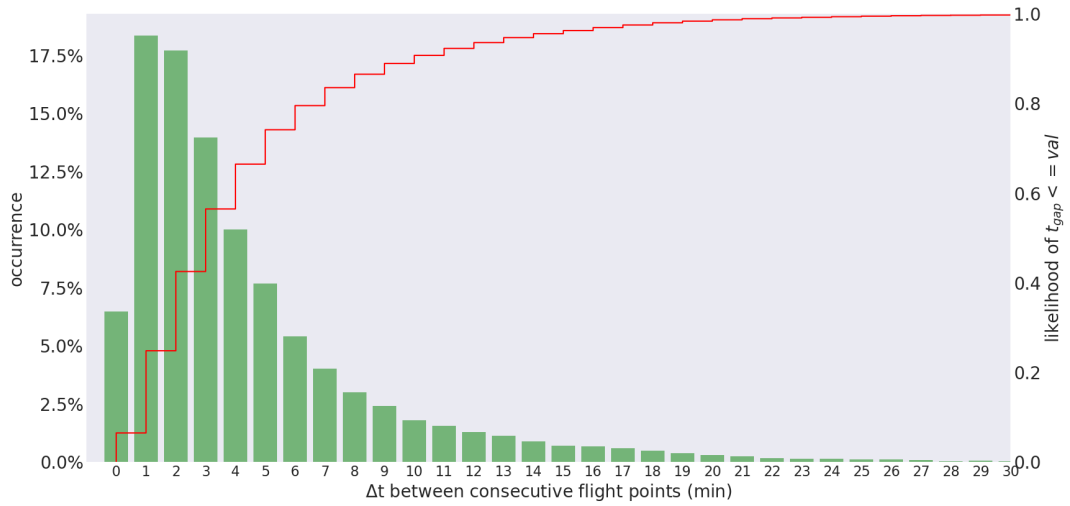


Figure 2.2. Distribution of time intervals between two consecutive flight path locations for 10,000 randomly selected flights.

Regarding the data sets retrieved for 2019 and 2020 from [16], which do not contain more detailed along-flight location tracking and include non-European flights, the data sets from 2015 to 2018 are used in the data wrangling process. A dichotomous approach is used to extract the subset of flights that flew through the region of interest (*ROI*). The unique set of departure-arrival airport pairs over 2018 are used to filter out the flights that flew over the ROI in 2019 and 2020 with high probability. In addition, all flights with a departure or arrival location within the ROI are included. Upon removing all duplicates (single flights might be detected by both methods) an estimated flight data set for the considered month is obtained. This approach fails to detect flights that neither have a reported geolocation within the ROI nor a pre-occurring departure-arrival combination from 2018, which might be the case for new flight routes and new airports. Any concern should be raised regarding the unofficial source of the data, some unrealistic entries related to flight levels and the many missing values.

A final data filtering approach relates to flight levels. As cirrus clouds do not generally form below 6 km, all data points of aircraft flying below this level could safely be filtered out without corrupting the analysis. Six km corresponds to a flight level of 197 (hundreds of feet). Compared to the original data set, around half of the data is removed after performing all those data wrangling steps, the majority taken up by the flight level filter.

Valuable features for the sake of the research are contained in both the *Flights* data set (A/C type, departure and destination airport) and in the *Flight Points* data set (discretized flight path with reported geolocations). The two data sets have one column in common, being '*ID*' in Figure 2.1a and '*ECTL_ID*' in Figure 2.3). Using this column the cleaned and processed data sets can be merged into one, see Figure 2.3[1]. These 'merged' data sets are the ones that are used for the analysis.

ECTRL ID	Time Over	light Leve	Longitude	Latitude	ICAO_dep	CAO_dest	AC_type	C_operate	AC_regis	flight_type	market	dist	dep_airport	dest_airport
184408024	2015-03-01 07:17:30	350	-8	50.3667	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
184408024	2015-03-01 07:18:00	350	-7.87944	50.3909	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
184408024	2015-03-01 07:18:30	350	-7.75889	50.4151	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
184408024	2015-03-01 07:19:00	350	-7.63833	50.4394	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
184408024	2015-03-01 07:19:30	350	-7.51778	50.4636	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
184408024	2015-03-01 07:20:00	350	-7.39722	50.4878	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
184408024	2015-03-01 07:20:30	350	-7.27667	50.5121	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
184408024	2015-03-01 07:21:00	350	-7.15611	50.5363	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol
184408024	2015-03-01 07:21:30	350	-7.03556	50.5605	KORD	EHAM	B763	UAL	N661UA	S	Traditional Scheduled	3953	Chicago O'Hare International Airport	Amsterdam Airport Schiphol

Figure 2.3. Screenshot of merged dataframe Flights 03/15 and Flighttracks 03/15. Example flight shown went from Chicago O'Hare Airport to Amsterdam Schiphol.

[1]The data wrangling steps discussed in this section have been already implemented before taking a screenshot.

2.2. CALIPSO Lidar

The CALIPSO satellite is a nadir-viewing active satellite, equipped with the CALIOP lidar that is polarization-sensitive and emits dual-frequency pulses with a repetition rate of 20 Hz. From the backscatter signal CALIPSO is able to construct high-resolution vertical cloud profiles, including optically thin cirrus. Hence, for an optimal representation of cirrus properties one could argue CALIPSO is the most accurate product. A major drawback of CALIPSO for the purpose of this research is its polar orbit and limited swath width, which make its temporal resolution low considering a fixed location on Earth. For this research the *CAL_LID_L2_01kmCLay-ValStage1-V3-41* [3] (and its precursor, *CAL_LID_L2_01kmCLay-ValStage1-V3-40*) is used, as those products provide a sufficiently high spatial resolution of 0.1×0.1 km while being the most mature product of all since they passed the validation stage.

The CALIPSO Lidar products used are level 2 products, meaning they come in non-uniform temporal-spatial grid scales. The products are formatted in Hierachical Data Format (HDF4). For the ROI under consideration the average overpass time for CALIPSO is 7 minutes. In order to simplify the analysis without comprising its accuracy too much, the mean timestamp during the overpass is taken as fixed timestamp for that overpass, hence leading to a maximum time offset of 3.5 minutes of each individual observation. Moreover, the data is resampled over the spatial domain into a fixed, uniform lon-lat grid of $0.25° \times 0.25°$.

Also, the CALIPSO data files provide a binary encoded single 16-bit integer which contains classification flags, i.e. for each detected layer by the backscatter signal it contains 1) feature type (e.g. cloud), 2) feature sub-type (e.g. cirrus) and 3) quality assessment. The full table that is used to decode can be found in Appendix A. Upon decoding, only the detected cirrus clouds which have been provided a medium or high confidence level (both feature type '*cloud*' and feature subtype "*cirrus*") is kept. Additionally, cloud optical thickness (COT), cloud mid-layer pressure and cloud mid-layer temperature are extracted. Those are all derived properties from the lidar backscatter signal. The cloud mid-layer pressure is used as a cloud locator in the vertical to bin the detected cirrus into predefined layers. This will be elaborated on further in Section 3.

2.3. Meteosat SEVIRI

EUMETSAT's Meteosat Second Generation (MSG) satellite carries on board the "Spinning Enhanced Visible and InfraRed Imager" (SEVIRI), which is an imaging instrument in geostationary orbit that continuously observes the ROI. Eight of the twelve spectral channels operate in the thermal infrared spectrum and generate data on e.g. (cirrus) clouds. The strength of this product lies within its very high temporal and spatial resolution (15 mins and 5×5 km, respectively). However, as it is a passive sensor it misses very thin clouds, and the product runs up to 2017. Nonetheless, the fact that the relevant channels for cirrus detection operate in the thermal infrared spectrum, make the use of this product possible during the night. The product contains both cirrus cover and (cirrus) optical depth.

The CLAAS-2.1 record [6] provides data derived from SEVIRI on cloud properties. The record covers the period 2014-2017 and does hence not extend till the COVID pandemic. The CLAAS product is validated and intercalibrated with data from MODIS Aqua. The spatial and temporal resolution of the product corresponds to the native Meteosat SEVIRI resolution and features cloud type as well as cloud microphysical properties such as COT. Each 15-minute mapping is stored in a separate NETCDF file, and an accompanying auxiliary file allows transformation of this L2 product into a geospatial lon-lat grid mapping. This gridding is done on the same resolution as the aforementioned CALIPSO product, being $0.25° \times 0.25°$.

The high temporal resolution of the product over the entire ROI makes this product an attractive complement to this research. More specifically, this product will be used to assess the second research question upon product validation with CALIPSO. For the time series analysis this product does not suffice due to its temporal coverage.

2.4. Meteorology

Due to the existence of confounding meteorological variables that restrict and affect cirrus formation, meteorological variables cannot be disregarded. ERA5 hourly reanalysis data on pressure levels [7]

from the ECMWF is available for on a horizontal resolution of $0.25° \times 0.25°$ and at various pressure levels. A selection of pressure levels is made, based on the pressure levels at which cirrus clouds generally may form: 100 hPa, 125 hPa, 150 hPa, 175 hPa, 200 hPa, 225 hPa, 250 hPa, 300 hPa, 350 hPa, 400 hPa and 450 hPa, which is roughly in between 7 and 16 km geopotential height, i.e. including the upper troposphere and lower stratosphere. From this data source the ambient air temperature and relative humidity at each pressure level is extracted. The data is formatted into NETCDF files.

2.5. Data Modelling Approach

2.5.1. High-level Flowchart

Figure 2.4 shows a high-level flowchart of the data analysis process, with the red and green boxes indicating the start and end, respectively. The blue boxes are the envisioned end products of the analysis. The main pre-processing steps have been described in Section 2. Intermediate steps will be elaborated on in Section 3.

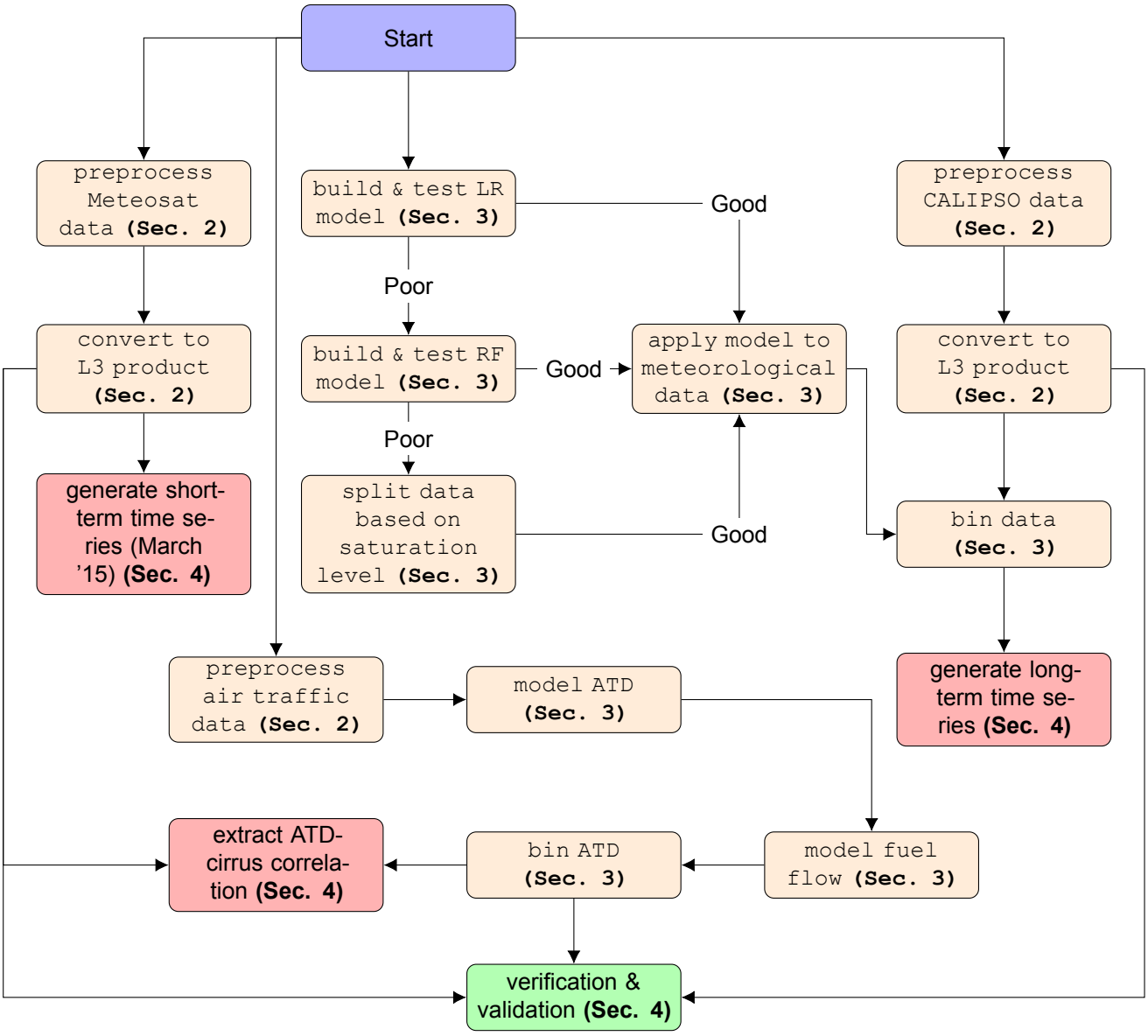


Figure 2.4. Project high-level flowchart.

2.5.2. Processor Allocation & Version Control

This project involves importing, processing and assimilating large chunks of data (only the selected air traffic data comprises roughly 27 GB of data). Data is stored on an external hard drive and chunks of the data are backed up on my own PC and on the server of the University Utrecht. Regarding data processing and analysis, I use parallel with my own PC the Gemini cluster GRID engine from the University of Utrecht. Furthermore the version control of my coding is taken care of by a Git repository, which is also publicly accessible.

3

Methodology

This section focuses on the main methodologies adhered to in order to answer the research questions. This includes the parametrization of air traffic into a quantitative parameter (Section 3.1, the determination of air traffic for missing years (2019 and 2020) in order to include the effects of COVID-19 on cirrus cover (Section 3.2), the handling of meteorological variables that might obscure the results (Section 3.3), and the used modelling approach applied to long-term trend modelling (Section 3.4) and short-term modelling (Section 3.5).

3.1. Parametrization of Air Traffic

After performing all data wrangling steps described in Section 2.1.2, the resulting data sets are parametrized into an interpretative quantity that can be projected onto the fixed spatial grid of $0.25° \times 0.25°$. The spatial domain is chosen as a rectangle bound by coordinates ($10°$W-$35°$N) and ($40°$E-$60°$N), which includes over 80% of European air traffic data obtained. The parameter will be referred to as Air Traffic Density or *ATD* and carries the unit of distance per km^2 per hour, in line with [11]. Essentially all flight paths are integrated and subsequently aggregated within a defined atmospheric box. This is shown in Equation 3.1, which shows that the ATD in each box i is computed by the summation of the (assumed) linear flight paths between consecutive flight points j for each flight k within the total set of flights flying within a time window, i.e. using the Euclidean distance metric. The traversed distance in longitude and latitude coordinates is converted to kilometers.

$$ATD_i = \sum_{k \in \{K\}} \sum_{j \in \{k_j\}} \sqrt{\left|\text{lon}_j - \text{lon}_{j-1}\right|^2 + \left|\text{lat}_j - \text{lat}_{j-1}\right|^2 + \left|h_j - h_{j-1}\right|^2} \tag{3.1}$$

Each atmospheric box is bounded by two pressure levels (which is a modelling parameter). Another control parameter is the time window over which ATD is computed. Increasing this window masks some spatial-temporal variability in ATD, albeit its implications depend on the purpose of the analysis. In any case the time window should be larger than the discretization in time of the flight paths, as flights that pass through the box might be completely missed otherwise. This readily exemplifies the importance of interpolating the flight paths as described in Section 2.1.2. Figure 3.1 shows a contourmap of ATD as retrieved by the algorithm on the 1st of March 2015 between 12 PM and 12:15 PM, for all flight points between 6 and 14 km altitude. The data has been interpolated in time with $\Delta t = 30$ seconds, where some major European flight hubs (Paris, Brussels) are clearly detectable. Note that as this data only include flight levels of 197 of higher, not all airport hubs stand out (see Appendix B). The linear pathways emerge due to the existence of flight corridors, but may also be related partially to the linear interpolation.

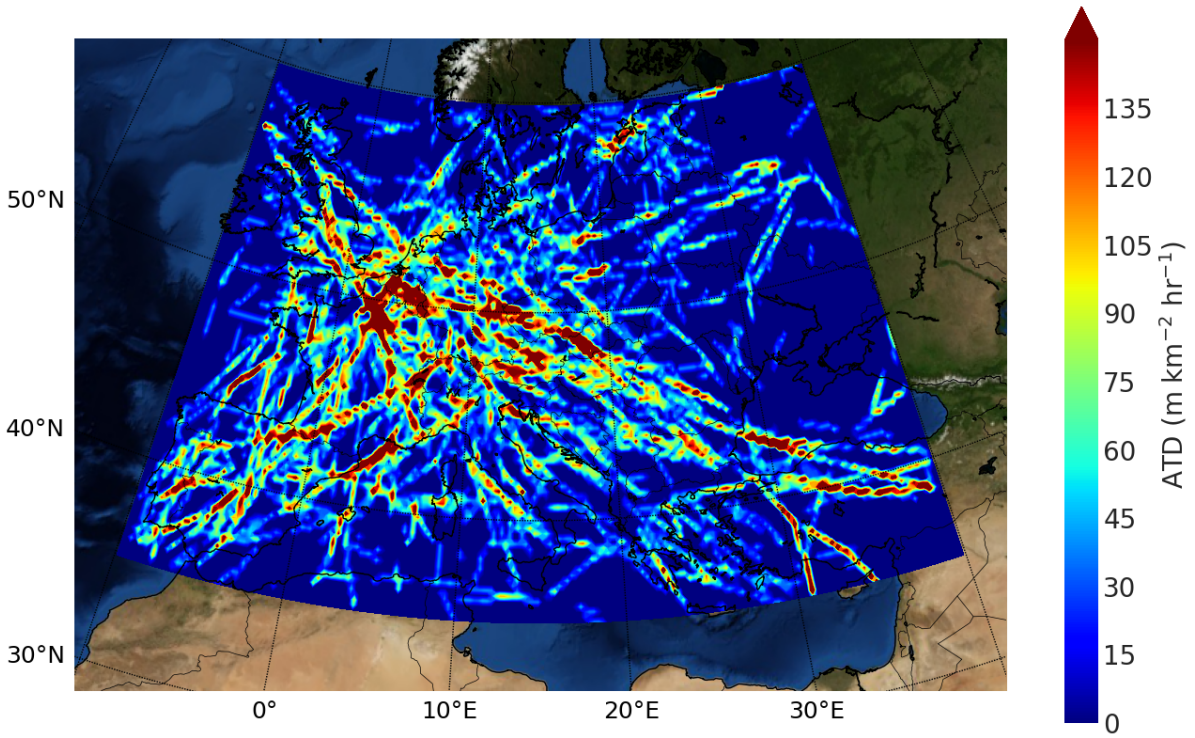


Figure 3.1. Air Traffic Density shown for the 1st of March 2015 between 12 PM and 12:15 PM. Units are in m km^{-2} hr^{-1}. The data has been gridded on a $0.25^\circ \times 0.25^\circ$ grid.

A more accurate approach which would better fit the analysis is to parametrize air traffic by taking into account the engine exhaust properties. The reasoning behind this is that large air vehicles expelling more particles into the air may have a larger impact on cirrus properties than smaller vehicles. Based on the aircraft type the engine types could be estimated, which in turn could yield the fuel flow (in kg/s). The total fuel combustion within a box each hour (i.e. kilograms of fuel burnt per km^2 per hour) could be computed by looking at the time spent for each aircraft in the respective atmospheric box and multiplying this with its fuel flow, assuming usual cruising conditions.

The flowchart of the fuel flow retrieval algorithm is shown in Figure 3.2. The Aircraft Performance Database [5] maintained by EUROCONTROL is a user interface where aircraft details including performance and technicalities can be found. Aircraft can be found based on ICAO code. Using the ICAO codes reported in the *Flights* data sets, the corresponding aircraft metadata can be scraped off the website using a URL retrieval algorithm. A Natural Language Processing (NLP) algorithm has been built, as technical aircraft details are given in varying text formats. This algorithm is given the task to extract 1) the number of engines and 2) the engine identification number for each aircraft. The engine types, now converted into parsed strings free of special characters, are algorithmically searched for in the ICAO Aircraft Emissions Databank [9] - a data set in CSV format containing a great collection of engine types and their specifications, amongst which engine fuel flow. Extracting the respective fuel flow from the database yields upon multiplication with the number of engines the total fuel flow of the aircraft.

Moreover, Figure 3.2 shows at each principal edge an example for March 2015 of the total number of aircraft types that are included at that point in the analysis. The pre-processed data set contains 130 unique aircraft types, from with 115 are found in the Aircraft Performance Database. Subsequently, seven aircraft types are lost when trying to extract the number of engines and engine type from the metadata, which is associated with the performance of the NLP algorithm. The weakest link within the system is there where the engine fuel flow should be derived from the ICAO Aircraft Emissions Databank. Here about half of the total number of aircraft engines are not recognized, which makes the model incorporating fuel flow not yet implementable. The analysis will resume with the ATD parametrization based on Euclidean distance. Incorporation of fuel burn is left for further research.

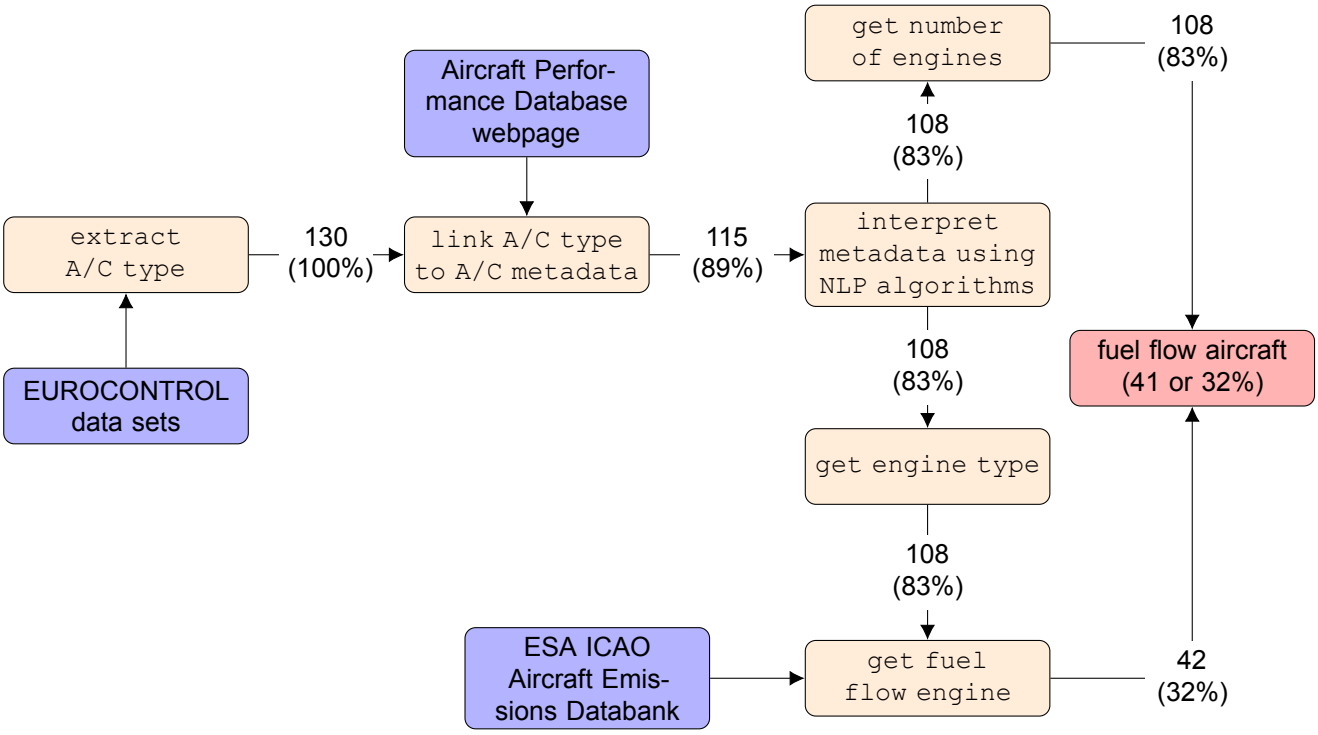


Figure 3.2. Flowchart of retrieval algorithm fuel flow. The numbers at the arrows show the total number of aircraft types contained in the analysis for March 2015. The percentages are taken w.r.t. the total of 130 aircraft types.

Part of the Dataframe for March 2015 is shown in Figure 3.3, which shows for different aircraft types the engine metadata that was found by the algorithm on the Aircraft Performance Database ("*engine type*" column). The columns "*nr engines*" and "*engine info*" are the output of the NLP algorithm trying to identify the number of engines and the engine type.

Index	AC_type	engine_type	nr engines	engine_info
0	B738	2 x 117 kN CFMI CFM56-7B turbofans.	2	117kNCFMICFM567Bturbofans
1	B752	2 x 162.8 kN P&W PW2037	2	1628kNPWPW2037
2	B772	2 x 332 kN P&W PW 4074	2	332kNPWPW4074
3	B77L	2x General Electric GE90-110B1 110.100lbf(490k…	2	GeneralElectricGE90110B11101001bf490kNGE
4	B763	2 x 281.6kN P&W PW4062	2	2816kNPWPW4062
5	A388	4 x 311kN R-R Trent 900	4	311kNRRTrent900
6	B77W	2 x General Electric GE90-115B 115,000 lbF (510 kN).	2	GeneralElectricGE90115B1150001bF510kN
7	B788	2x 280kN General Electric GEnx	2	280kNGeneralElectricGEnx
8	A321	2 x 133kN CFM56-5B1	2	133kNCFM565B1
9	A319	2 x 98kN CFM56-5A4	2	98kNCFM565A4
10	E190	General Electric CF34-10E		

Figure 3.3. Dataframe showing a couple of aircraft types and their corresponding number of engines and engine types as obtained from the algorithm.

3.2. Time Series Extrapolation of Air Traffic

Section 3.1 described how air traffic is parametrized. In Section 2.1.1 the poor data quality of the crowdsourced data sets was mentioned, that span the years 2019 and 2020. Clearly, the data quality is insufficient to be able to attain high-resolution ATD mapping on acceptable confidence levels for

those data sets. The way this is handled is by constructing a linear regression model shown in Figure 3.4, built upon the number of monthly flights and corresponding mean ATD for the years 2015-2018, as those quantities are expected to be closely related to each other,

The ATD has been determined for each month on a monthly time window[1], thereby preventing the risk of not detecting flight paths crossing several windows (see Section 2.1.2). The flight levels were taken between 6 and 14 km altitude. The number of monthly flights is the filtered residue of the *Flights* data sets. The Ordinary Least-Squares (*OLS*) regression line with a 95% confidence interval is shown along with the data points, and the squared correlation coefficient R^2 of the bi-variate data is 0.86. The algorithmic output discussed in Section 2.1.2, that yields an approximation of the number of flights in the eight months spanning 2019 and 2020, are used as input in the created linear regression model to approximate the ATD for those years.

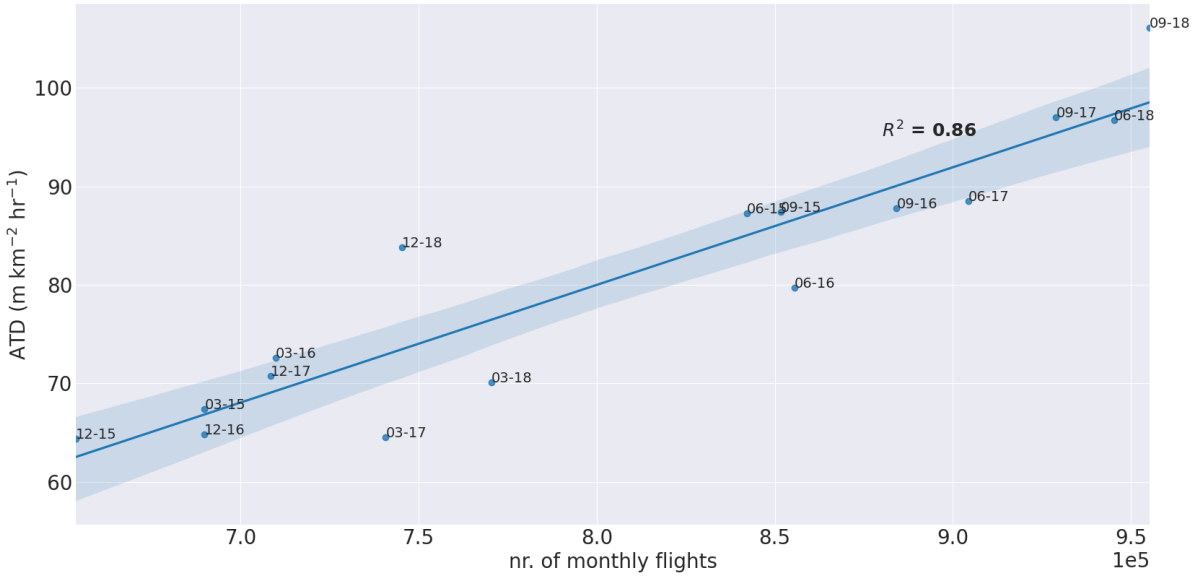


Figure 3.4. ATD vs nr of flights for each month till 2018.

3.3. Data-Based Assessment of Meteorological Effects

An essential notion when investigating cirrus clouds is their dependence on ambient conditions. In one or another way the main geophysical variables that are of influence - air temperature, relative humidity (*RH*) and pressure - should be regarded. In data-based research Machine Learning (*ML*) algorithms can provide an outcome for confounding variables that are not of direct interest for the research, as those mentioned. Two ML models are built that could fit its purpose: Logistic Regression and Random Forest. One of those models should provide a filter to locations where cirrus cannot form under any air traffic conditions, and that are therefore regarded as noise in the analysis.

3.3.1. Logistic Regression

The probability P of cirrus formation depends particularly on whether the ambient temperature exceeds a threshold temperature T_{thr}, where T_{thr} is a function of relative humidity h and pressure level p and engine exhaust properties e [15]. Exhaust properties are for now disregarded as no sufficient data on the exhaust properties is available. Looking at Figure 1.1 there appears to be, given h, a near-linear decision boundary as to where cirrus contrails may form. In (nearly) all cases where upper-tropospheric or lower-stratospheric air reaches supersaturation ($h \geq 100\%$), temperatures are sufficiently low for water droplets to freeze and as the surrounding air is sufficiently damp, formed ice crystals keep growing by absorbing water from their direct environment. This atmospheric state is referred to as ice-supersaturated.

For simplicity, cirrus occurrence within a certain atmospheric box could be coded as a binary response,

[1] Computationally, this is also more efficient.

i.e. 1 for cirrus and 0 for no cirrus. If conditions make it impossible for AIC to form and prevail, those locations will result in noise in the analysis, since the binary output will be insensitive to the features. In other words, those locations can be expected to be indifferent to ATD regarding contrails. The aim is to identify the decision boundary where grid cells should be excluded from the analysis. Due to the binary mapping of the response variable, a logistic regression (*LR)* model could be used. Using LR, the probability P of favorable persistent contrail conditions can be written as:

$$P = \frac{exp(b_0 + b_1 T + b_2 h)}{1 + exp(b_0 + b_1 T + b_2 h)}, \text{ or} \tag{3.2a}$$

$$\log\left(\frac{P}{1-P}\right) = b_0 + b_1 T + b_2 h. \tag{3.2b}$$

An important underlying assumption, which might be off, that can be seen from Equation 3.2b is that the logit of the probability, $\log\left(\frac{P}{1-P}\right)$, is a linear function temperature and RH. It is expected that ATD also influences the probability of cirrus occurrence. However, the location of the decision boundary is expected not to be sensitive to this boundary, as it is particularly constraint by meteorological variables.

For training and testing the model, data from January 2015 is used as it is not part of the overall time series analysis. Data is gridded onto the aforementioned $0.25° \times 0.25°$ grid, and divided into 11 vertical layers based on the available pressure levels from ERA5 (Section 2.4) in the upper troposphere and lower stratosphere. The mid-layer pressure reported by CALIPSO CALIOP data is used to allocate each detected cirrus cloud to its corresponding vertical layer. The data is randomly split into a train and test set (50%-50%) and the model accuracy is assessed for both the train and test set as a function of a cut-off probability, which bisects the outcome space (e.g. for a cut-off probability of 0.5, all probabilities of 0.5 of higher for cirrus occurrence are classified as ”1”, the remaining as ”0”). This is shown in Figure 3.5. The importance of the model performance on the test set resides in the fact that the model should be able to distinguish locations where conditions are right for cirrus to form from locations where cannot form on unseen data used in the analysis.

Also the recall score[2] and precision score[3] are shown for various thresholds. A constant predictor line is added, which illustrates the accuracy if the model would always predict the dominant class (no cirrus, in 92% of all cases). The model accuracy does not beat the constant dummy predictor, and the precision score is low. This might be related to data quality, too poor spatial resolution (particularly in the vertical) and other variables as advective processes and cirrus lifetime. The highly unbalanced nature of the data classes might also play a role in the low precision scores, as those models tend to have a bias towards the majority class. At the moment when the accuracy score reaches the accuracy of the dummy classifier, around a cut-off of 0.9, all boxes are classified as ”no cirrus”, meaning the precision score becomes undefined (division by zero). This shows why the accuracy metric is not per definition an adequate measure for the model performance, especially for highly unbalanced data: despite the accuracy being highest at a cut-off of 0.9, the model is essentially useless. It is much more desirable for the precision score to be high, as this would mean a great proportion of the locations recognized by the model as potential cirrus location are in fact locations where cirrus occurs. A final reflection on the LR model performance is that the linear assumption mentioned before might be violated.

[2]The number of correctly predicted cirrus cloud occurrences divided by the actual number of cirrus cloud occurrences.
[3]The number of correctly predicted cirrus cloud occurrences divided by the total number of predicted cirrus cloud occurrences.

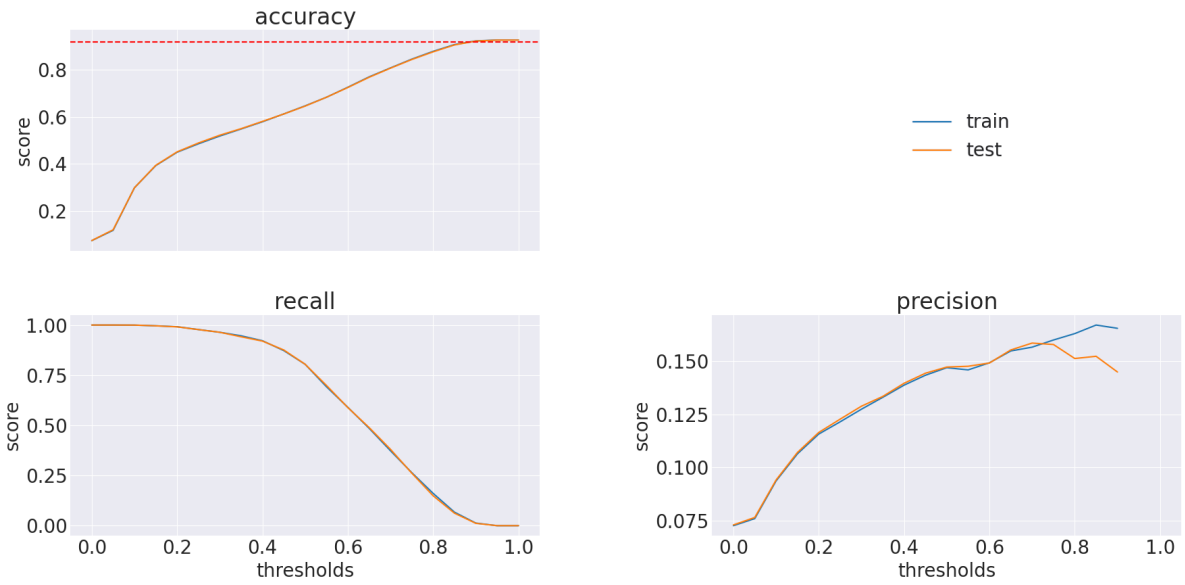


Figure 3.5. Logistic Regression performance on test data on the *accuracy* metric (correctly classified instances divided by the total), the *precision* metric and the *recall* metric.

3.3.2. Random Forest

To by-pass the parametric nature of the LR model, decision trees are considered due to their non-parametric property. The algorithm of a decision tree works by partitioning the feature space of a data set into subsets, where at each time the partition is performed on one variable. This is usually done in such a way that each subset is as homogeneous as possible. The Random Forest (*RF*) method works by combining an ensemble of decision trees into one model. As RF is a non-parametric approach, the major concern when employing this model is the tuning of the hyperparameters such as tree depth.

The air space is again binned into 11 layers based on data availability of ERA5 Reanalysis (centered at pressure levels of 100, 125, 150, 175, 200, 225, 250, 300, 350, 400 and 450 hPa). Gathering data for each defined $0.25° \times 0.25°$ bin over 11 pressure levels results in a data set of around 1 million data points for Jan 2015.

In order to see how well a RF model is able to predict the outcome based on the predictors, the data set for January 2015 is again randomly split into a train and a test set. The model is run for 500 trees, where a balanced class weight is assigned to both classes in order to account for the class imbalance and the potential bias this generates in the model. In addition, the data set is reconstructed by oversampling the minority class such that the new class balance would be two-third cirrus and one-third no cirrus. This is done in another attempt to handle the large class imbalance, potentially leading to model bias towards the majority class. Both duplicating existing instances from the minority class and generating new data belonging to the minority class (using *SMOTE* [2]) were done. Figure 3.6 shows that the oversampling procedure led to an improvement of the model performance on the training data, but the performance on test data is low. The model performance on the train set were assessed using the Out-of-Bag (*OOB*) score using a cross-validation approach. Due to the non-parametric nature of the data, the training accuracy performance is well above the constant predictor line (in red), in contrast with the LR model. The threshold on x again indicates the cut-off probability. Comparing the test accuracy to the constant predictor line that corresponds to the test set (red solid line), the test accuracy never exceeds this line. It would be particularly desirable for the precision score to be higher for the test set, as this is the principal indicator for the model quality when adopted within this research. For high cut-off probabilities the precision shoots up rapidly, albeit the size of the predicted cirrus (1s) class becomes very small. This is also reflected into the very low recall score. The oversampling procedure led to an improvement of the train performance, while generally decreasing the test performance. Finally, exclusion of data on pressure levels where very little cirrus occurs does not lead to significant improvements in the model performance.

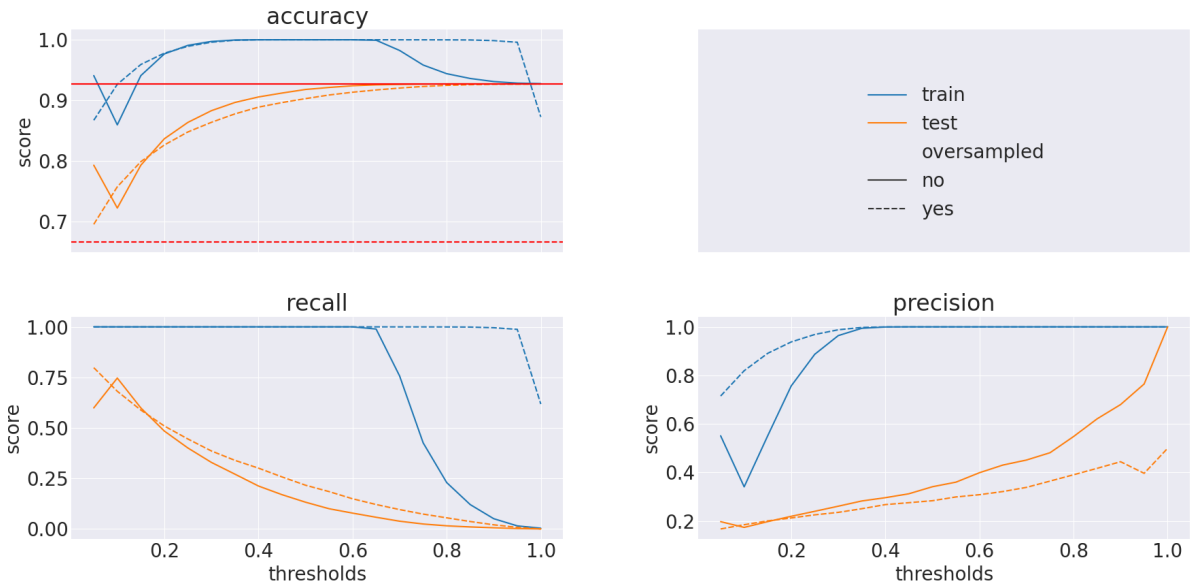


Figure 3.6. Performance of Random Forest model, measured by the accuracy, recall and precision.

Considering the rather flawed performance of both the LR and RF model on unseen data, the decision is made to not incorporate those models in producing the research outputs. Instead, the intuitive binning practise based on air supersaturation is applied, accompanied with meteorological statistics and a fraction related to the number of Ice Supersaturation Regions (*ISSRs*) [14]. This latter quantity is computed by the dividing the number of regions where supersaturated conditions apply ($h \geq 100\%$) by the total number of regions.

3.4. CALIPSO Time Series

The long-term (2015-2020) cirrus cover time series over the ROI is constructed by averaging all available CALIPSO data for each month matching the air traffic data availability, that is, for March, June, September and December. All data wrangling steps which were described in Section 2.2 are applied to each data set, keeping only mid and high confidence data. The altitude at which cirrus occurs according to the CALIPSO products are used to match the cirrus cloud with the corresponding meteorological conditions that apply within that atmospheric layer, using ERA5 Reanalysis data. This methodology allows for a construction of CALIPSO time series within both supersaturated and sub-saturated air separately. The purpose of doing this is to have some kind of incorporation of meteorological conditions that could otherwise mask out the effect of air traffic on cirrus cover. Despite being a rather crude method, it is within the time frame of the research the only viable method after omitting the LR and RF model from the analysis.

Following up on Section 2.2, each CALIPSO overpass during all included months can be allocated to a one-hourly time window. The aggregation of all overpasses is illustrated in the polar barchart in Figure 3.7. Apparent here is the clustered appearance of CALIPSO overpasses in time, one being around midnight and the other around noon. This offers the opportunity to cluster CALIPSO cirrus data into daytime and nighttime, noting that some potential variability between daytime and nighttime cover might be induced by air traffic. A drawback of this semi-diurnal overpass cycle is that the monthly mean cirrus cover magnitudes, that will be retrieved from those products, might be a poor reflection of the actual mean cirrus cover during that month as the two diurnal ranges might not reflect the daily mean levels (e.g. if cirrus cover around midnight and noon are significantly higher on average than during the rest of the day).

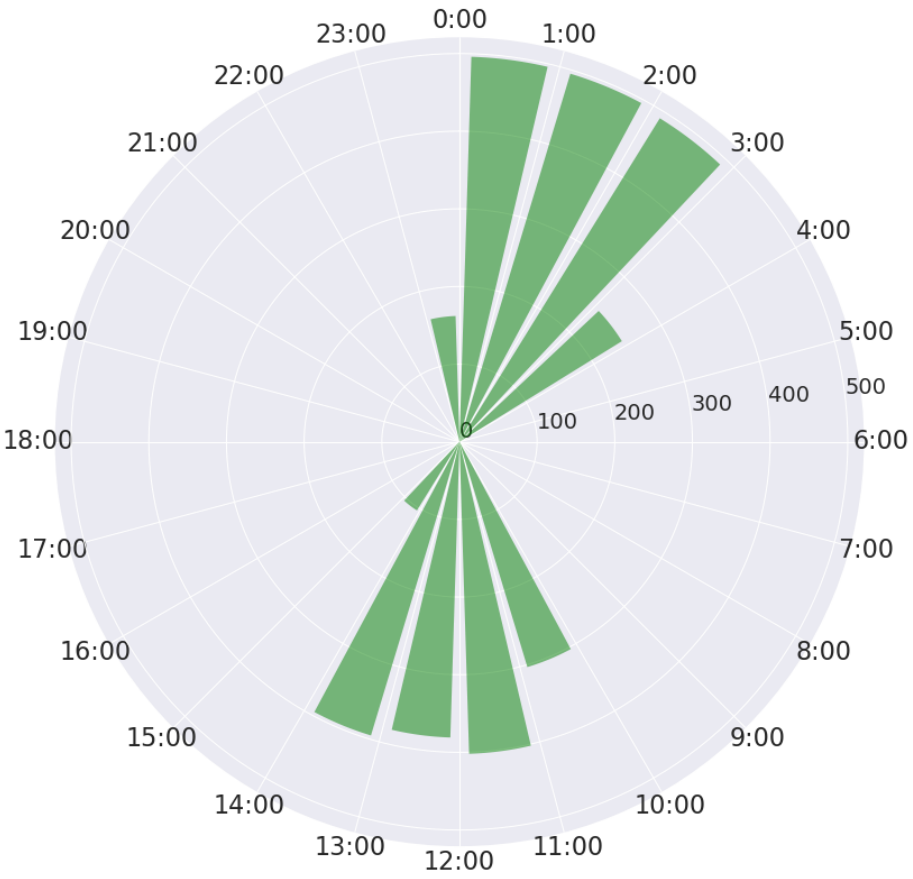


Figure 3.7. CALIPSO overpass frequency

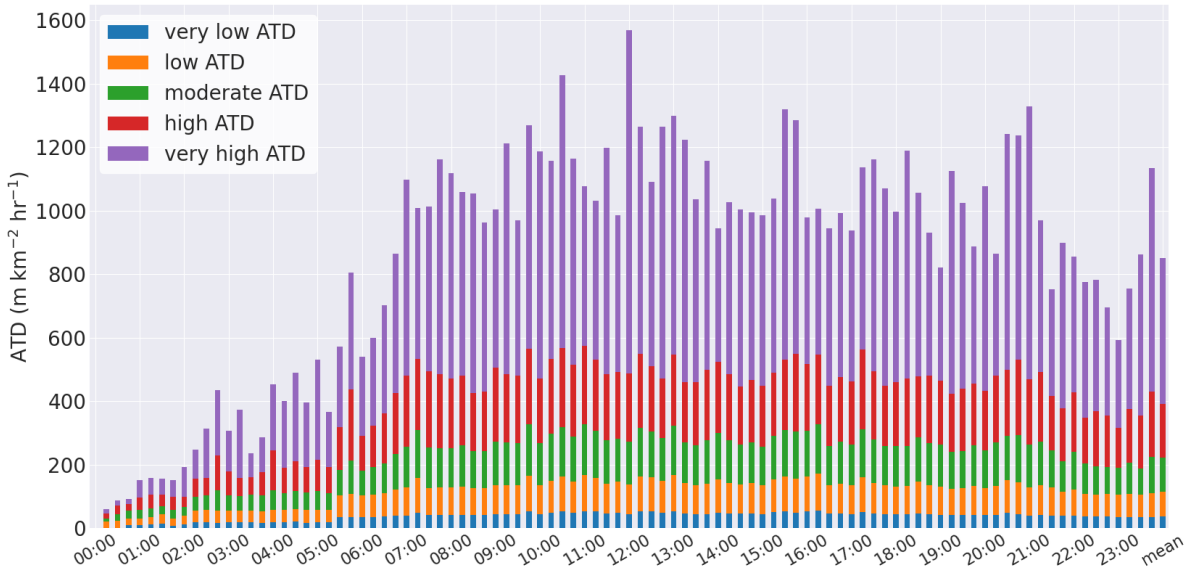
3.5. High-density Grid Cell Binning

This section focuses on the second research question, that 1) does not require data over such a large time span on low temporal resolution, and 2) allows for a statistical assessment on the relation between cirrus cover and ATD on shorter timescales (tens of minutes). Using the CLAAS 2.1 product cirrus cover could be extracted for an entire month, for which March 2015 has been chosen as all required data products are available for this month. Upon geospatially mapping cirrus cover into a $0.25° \times 0.25°$ grid for the entire month, the change in cirrus cover at each location between sampling periods (15 mins) is computed, which comes down to roughly 60 million data points (spatial grid dimension of 200×100, temporal dimension of 744).

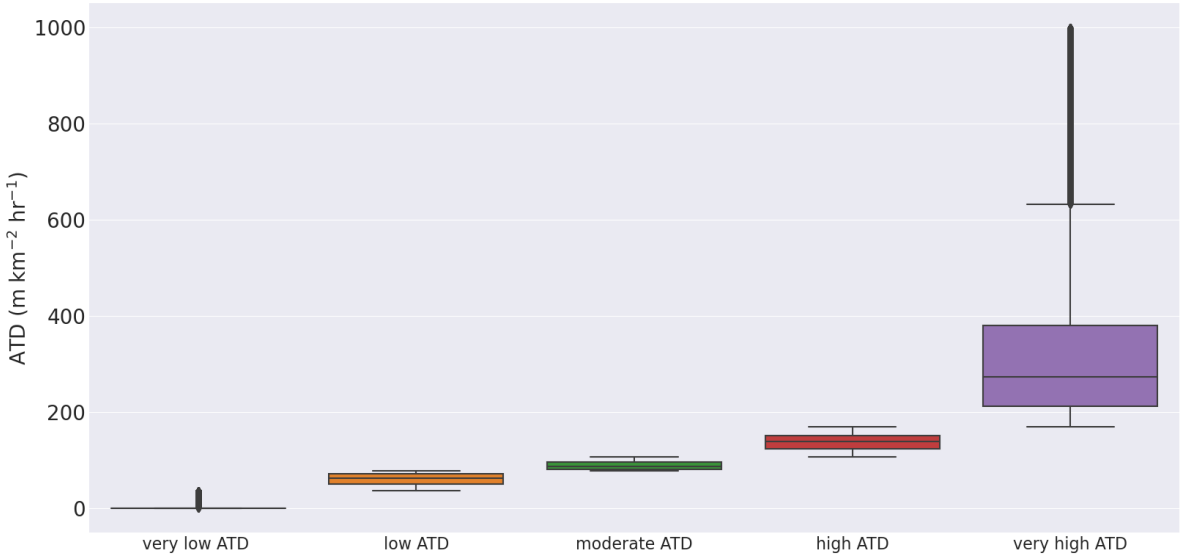
The ATD is computed on the same temporal frequency of 15 mins, albeit the sampling time of this parameter is shifted 15 minutes backwards compared to the time over which the change in cirrus is computed. This is done to allow for any potential premature contrails to form and grow, hence assessing the response of cirrus cover on ATD.

After accounting for meteorological variability the data sets are binned based on ATD using the 5-point Likert scale - "*very low ATD*", "*low ATD*", "*moderate ATD*", "*high ATD*" and "*very high ATD*". The Jenks optimization method, also called Jenks natural breaks classification method, is used to achieve this. The optimization method seeks for natural splits in uni-variate data whereby it tends to minimize the intracluster variability and maximize the intercluster variability. The method is not highly scalable, meaning it could not be implemented on the entire data set. Instead, the algorithm is run on each separate time window as shown in Figure 3.8a, which shows for every 15 minutes over an average day the stacked barchart of cluster assignments. The class boundaries are averaged over all times to obtain the final five clusters that are applied to the entire data set. Boxplots for each cluster are shown in Figure 3.8b. Looking at the number of classified outliers in the highest cluster, choosing more clusters might

be ideal. For ease of interpretability however those five clusters are maintained. Furthermore, including more bins in the higher range of ATD strongly reduces the number of cells included in the respective bins and concomitantly the confidence levels. As with K-means clustering, a drawback of the Jenks optimization method is that the number of clusters should be given to the algorithm, a choice that in most cases might be arbitrary. Using the bins shown in Figure 3.8b the cirrus cover at corresponding locations will be binned, and their statistics will be assessed.



(a) Jenks method applied to each time window averaged over March 2015.



(b) Side-by-side boxplots of the different clusters based on ATD. Magnitude shows the ATD. Outliers are points that are 1.5 times above (below) the inter-quartile range past the highest (lowest) quartile.

Figure 3.8. Cluster method of data from March 2015 based on ATD.

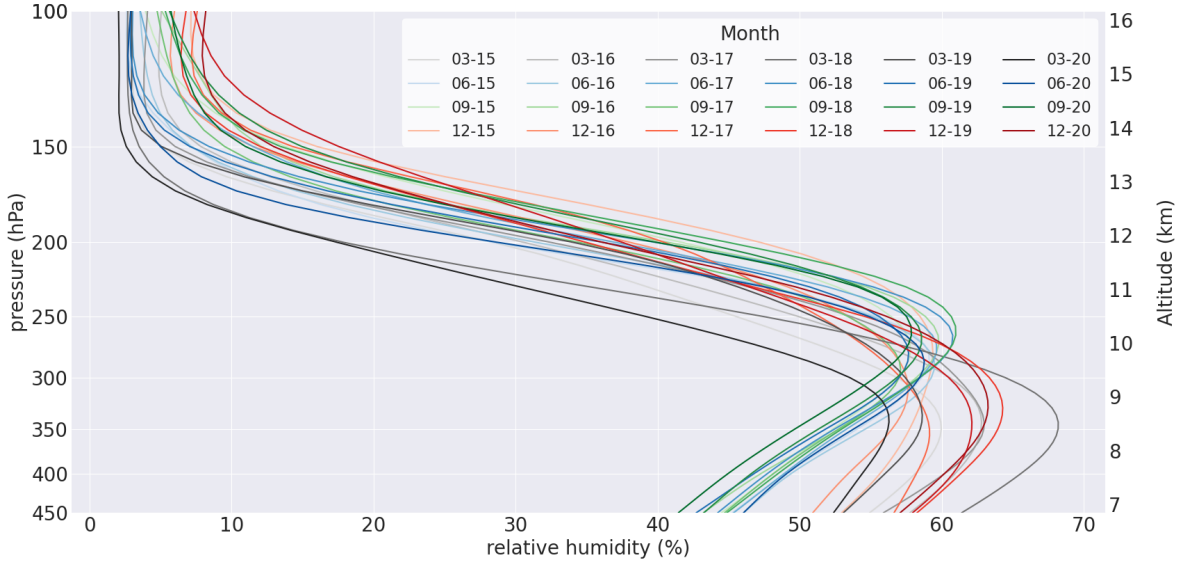
4

Results

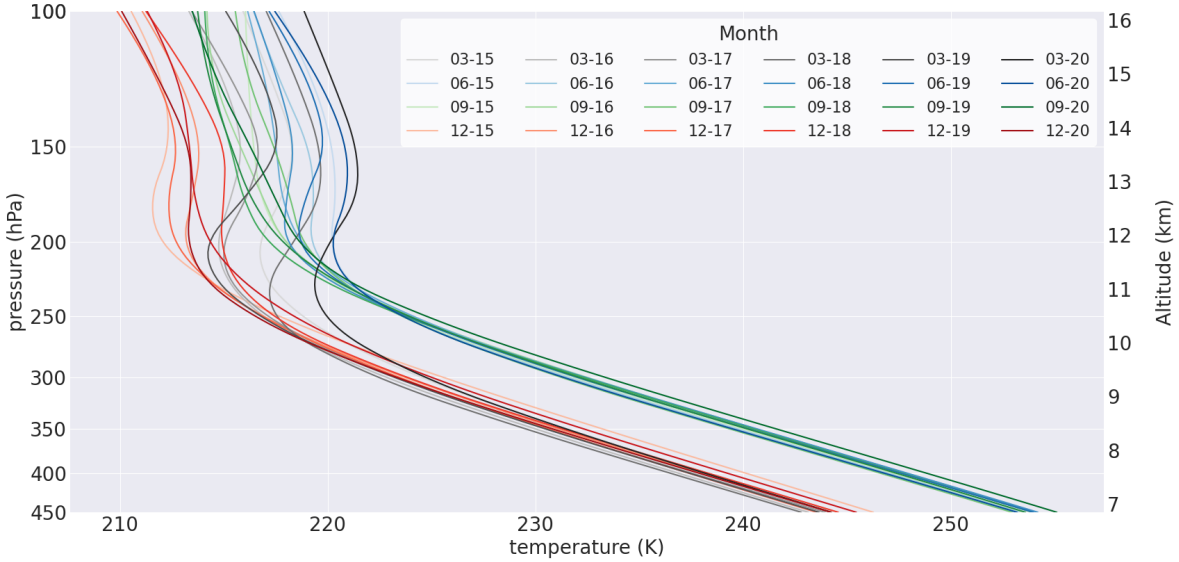
4.1. Long-Term Time Series Analysis

The monthly cirrus cover determined by CALIPSO, together with the monthly ATD and meteorological variables, have been determined within each of the 11 pre-specified vertical layers. These data are spatially averaged (in the horizontal) and temporally averages (over one month). The vertical RH profiles from Figure 4.1a have been constructed based on 11 pressure levels by cubic splines. The figure shows a peak in RH in the upper troposphere, and a strong decline in RH with increasing height afterwards. This is a clear tropopause signature, consistent with literature. The RH maxima are located at lower altitudes during winter than during summer, which could be related to a seasonal vertical shift of the tropopause. The tropopause is located lower during winter than during summer, potentially impeding moisture to rise further upwards. Also, the maxima are higher in magnitude during winter. March 2020 was an exceptional year as the RH maximum was even lower than usual during the summer months, while the temperature above 250 hPa was also anomalously high.

The vertical temperature profiles from Figure 4.1b are piece-wise linear regression graphs based on temperature data attained on 11 pressure levels. Piece-wise linear regression is chosen as temperature is expected to vary at a constant lapse rate with height. The temperature decreases with height at a seemingly constant lapse rate in the upper troposphere, consistent with literature, and the upper tropospheric mean temperature in the summer months are clearly offset compared to winter months with about 10K. In the tropopause and lower stratosphere the temperature profiles appear not to be strongly affected by the seasonal cycle. However, a discrepancy with my expectations arises at lower stratospheric temperatures, where temperature first increases with height as expected and subsequently drops again. This points to some bias in the data within the lower stratosphere.



(a) Vertical profile of relative humidity.

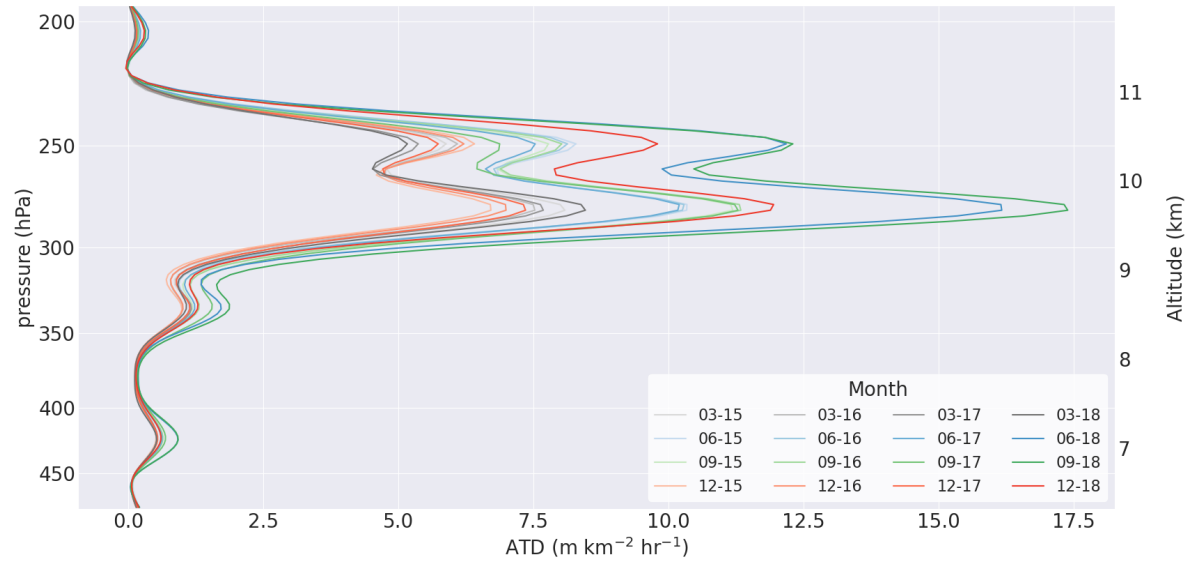


(b) Vertical profile of temperature.

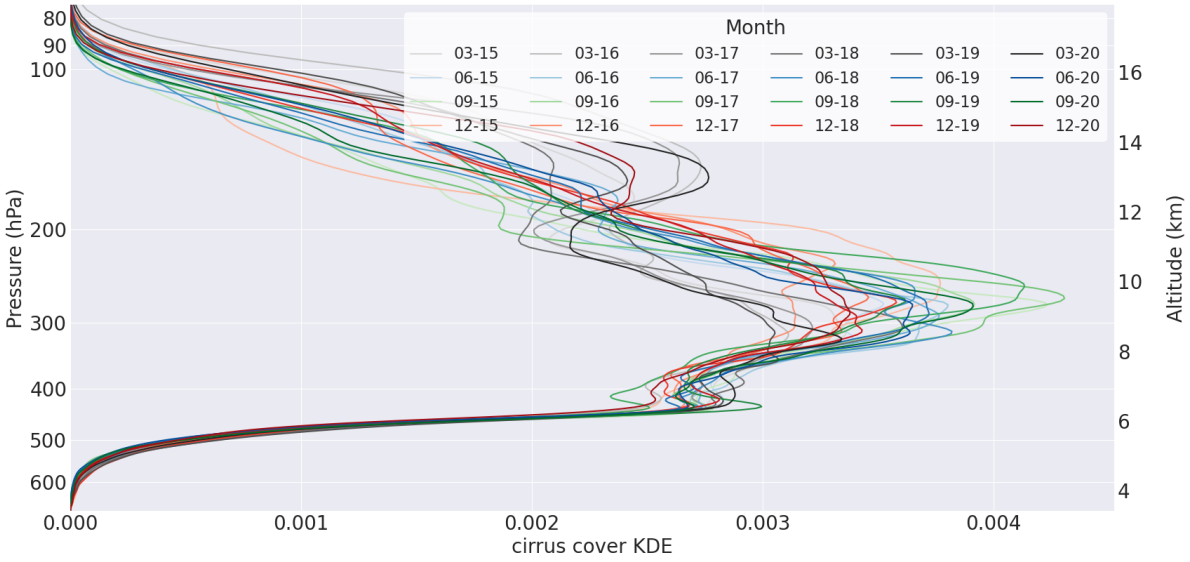
Figure 4.1. Vertical temperature and RH profiles for each month included in the analysis. Data from ERA5 reanalysis. The color aesthetic is used to distinguish different months. The opacity is used to indicate the year, with lower transparency for more recent years.

In Figure 4.2b the vertical ATD profiles are shown. Clearly most air traffic is flying between 9.5 and 10.5 km altitude, indifferent of the specific months. What is apparent from this figure is that air traffic increases in magnitude over the course of four years, consistent with Figure 1.2.

The vertical cirrus cloud distribution derived from CALIPSO is shown in Figure 4.2b. Consistent with literature the vast majority of cirrus clouds occur between 6 and 14 km height. The Kernel Density Estimate (*KDEs*) emphasize that the highest proportion of cirrus is occurring around 10 km altitude, where the vertical ATD maxima (see Figure 4.2b) coincide with the RH maxima. This is most evident for the summer months (June and September). During winter months the vertical cirrus distribution between 6 and 14 km is more uniform, possibly because of tropopause propagation downwards and relatively lower ATD at those altitudes where the RH attains its maximum. The validity of this proposition should be further researched.



(a) Vertical profile of ATD.



(b) Vertical profile of cirrus cover.

Figure 4.2. Vertical ATD and cirrus profiles for each month included in the analysis. Data from EUROCONTROL and CALIPSO. The color aesthetic is used to distinguish different months. The opacity is used to indicate the year, with lower transparency for more recent years.

Figure 4.3 shows a heat map of monthly average cirrus cover for all months included in the analysis. Months are shown on the vertical, years on the horizontal. There is a consistent pattern that cirrus cover is higher during the winter months (March and December) than during summer months, with a low anomaly for March 2020 and in December 2016, and a high anomaly for September 2017. This general pattern could be highly associated with meteorological variables, particularly RH and temperature, as shown in Figure 4.1. Over the evaluated time span the figure provides no indication of an upward trend in cirrus cover due to air traffic increase. Mean cirrus cover was lowest over 2020 when the COVID-19 outbreak took place, even though the difference with e.g. 2016 is small. A low maximum in RH while the temperature above 250 hPa was also anomalously high, could be, in combination with a strong decrease in air traffic (see Figure 4.4), explain the clear drop in cirrus cover seen in Figure 4.3.

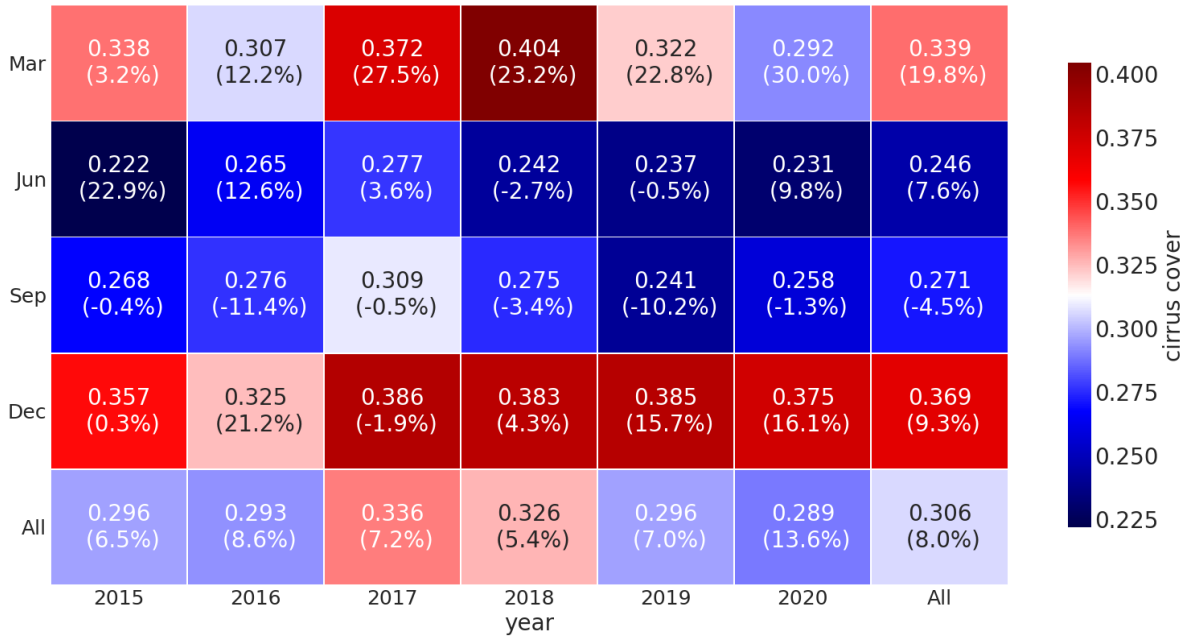


Figure 4.3. Heat map of cirrus cover retrieved from CALIPSO CALIOP data. Time series runs from 2015 till 2020. The percentages show the difference between daylight cirrus cover and night cover. The cover during daylight is taken as reference.

Dividing the CALIPSO data into two chunks for each month according to Figure 3.7 gives a comparison of cirrus cover during the day and during the night. Figure 4.3 shows this day-night difference percentage-wise, taking the day cover as reference (i.e. positive percentage means higher cover during the nights than during the day). It is evident from this figure that cirrus cover is generally higher during the night than during the day (respective means of 0.320 and 0.295 or 8% difference). For all years in September the opposite is the case. Apart from the possible linkage with synoptic-scale meteorology, the role of air traffic cannot be precluded based on this analysis, as AIC should predominantly occur during daytime. In fact, out of the four considered months, ATD is found highest in September (Figure 4.4) whilst the geopotential height of the RH maxima in those months (Figure 4.1a) closely coincides with the predominant flight level (Figure 4.2a).

Figure 4.4 shows the time series analysis on ATD, where 2015-2018 have been computed implicitly at each CALIPSO overpass location and averaged for the entire month, and the values in 2019 and 2020 are retrieved from the linear regression approach explained in Section 3.2. The linear regression model appears to underestimate the ATD for those two years, which could be linked partially to a bias in the linear regression model, but to a larger extent to the data quality for those years. A later publication of the full European air traffic data files for March 2019 on the EUROCONTROL website made it possible to evaluate the true number of flights that had been taken place during this month. From here it results that the estimated number of flights for this month using the inferior data set is 89% of the true value, i.e. an under-estimation of the number of flights of 11%.

From Figure 4.4 it can be seen that the ATD is consistently higher during summer than during winter (excluding 2020). This seeming anti-correlation between ATD and cirrus cover can be explained by the dominant effect of meteorology, overshadowing the signatures left by air traffic. This is partially confirmed by Figure 4.1 which shows consistently lower upper tropospheric temperatures (between approximately 450 hPa and 200 hPa) in winter, and in December also higher relative humidity, both favorable for cirrus to form. One feature that is evident from Figure 4.4 is the increase in air traffic between 2015 and 2018, followed by a sharp decline in 2020 due to the COVID-19 pandemic.

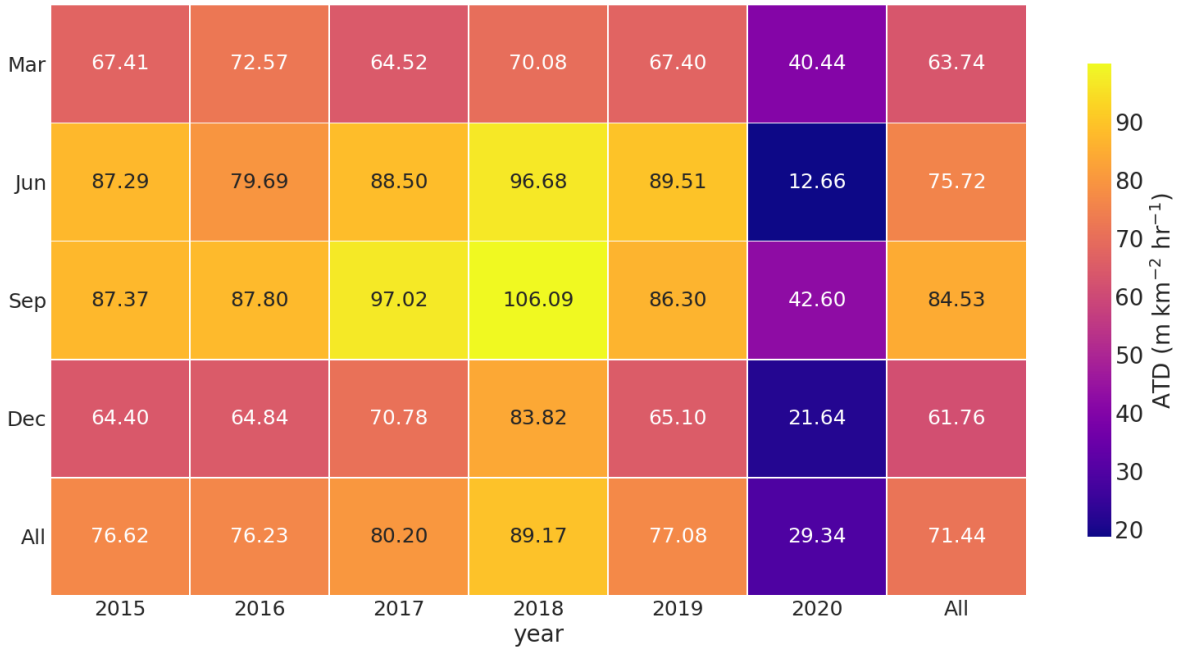


Figure 4.4. ATD time series 2015-2020.

In an attempt to eliminate a part of the variability induced by meteorological conditions, a raw data grouping is done based on air saturation level. A comparison of monthly cirrus cover between supersaturated and sub-saturated air is shown in Figure 4.5, where the data has been split based on daytime. I.e. the left boxplot in Figure 4.5 shows the percentage difference in cirrus cover between sub-saturated air and supersaturated air during daytime, while the boxplot shows the difference during nighttime. Clearly there is a higher spread in differences over night than during the day, and the difference of the means is statistically significant at a 99% confidence level with respective means of 3.7% and 5.1%. As diurnal meteorological cycles are very weak in the upper troposphere and lower stratosphere, a plausible explanation for the observed pattern from Figure 4.5 is a signature of air traffic flying during the day in supersaturated air (increasing the difference in cover between supersaturated and sub-saturated regions, compared to the night where air traffic is absent).

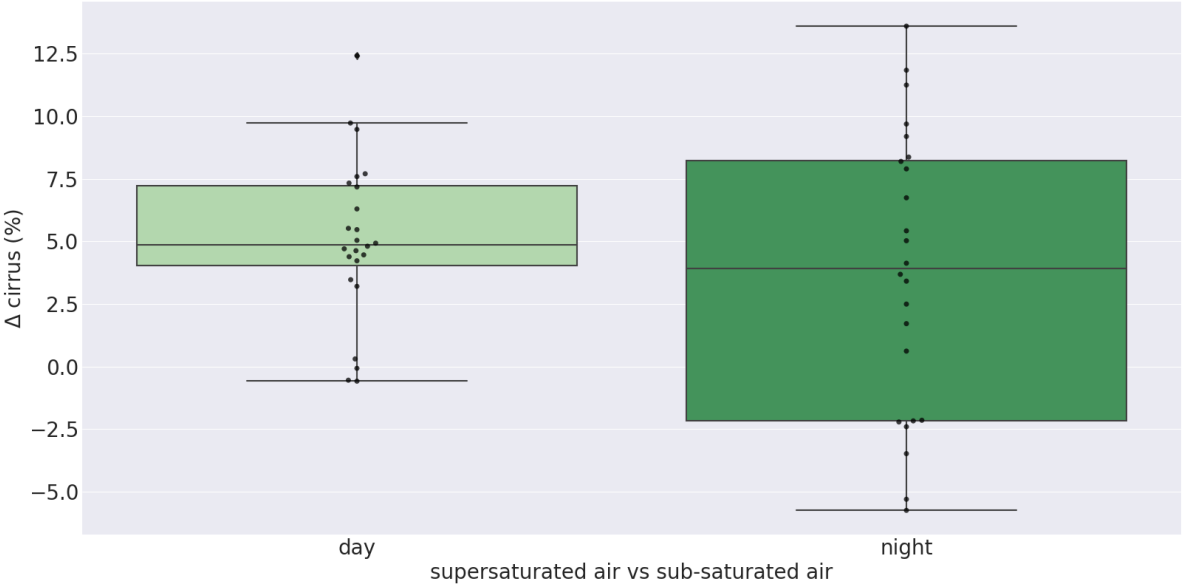
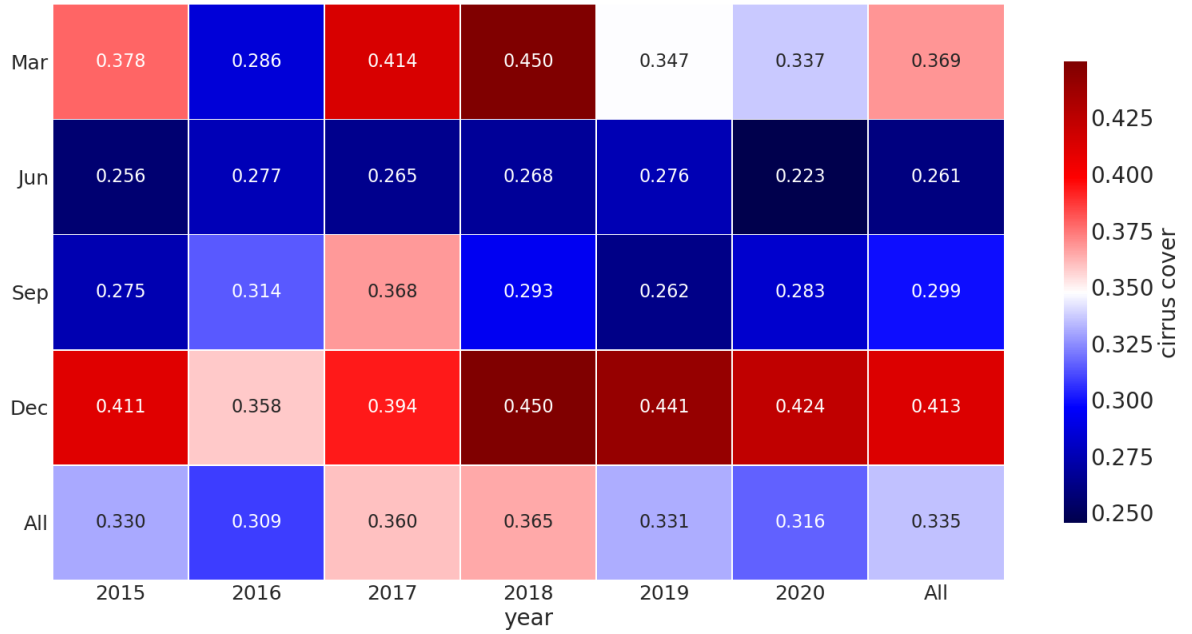
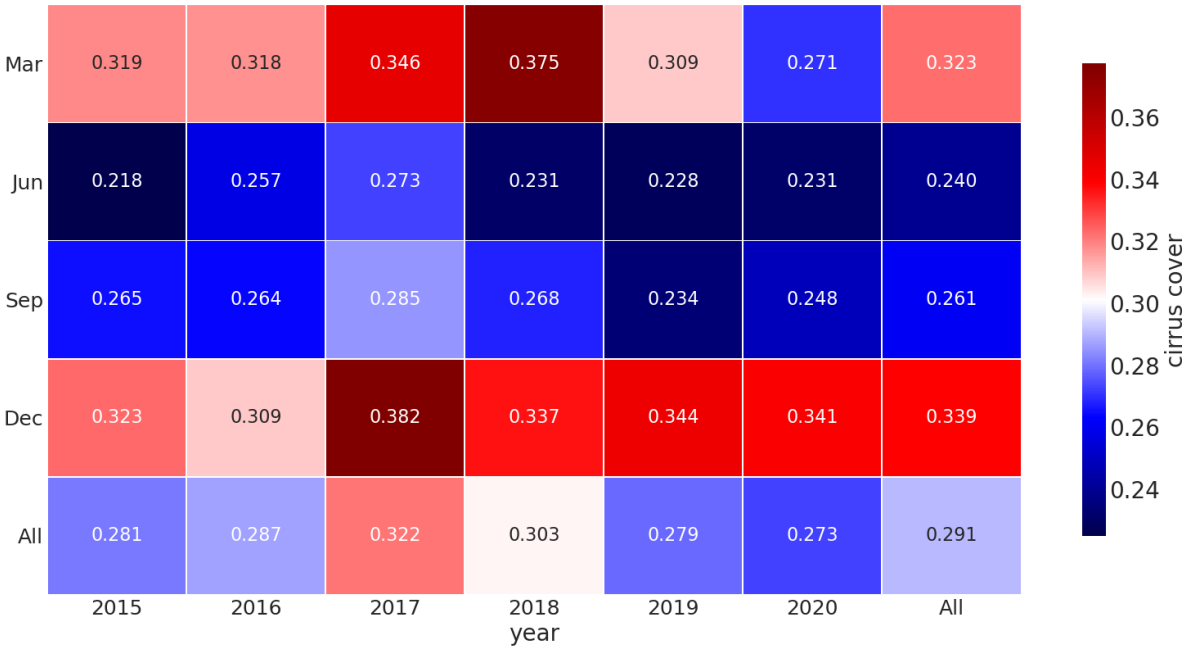


Figure 4.5. Side-by-side boxplot of difference in cirrus cover between sub-saturated and supersaturated air, subset on day and night cover.

The time series for cirrus cover in supersaturated and sub-saturated regions is shown in Figure 4.6, where the cirrus time series are shown for a) supersaturated and b) sub-saturated air. These figures show that there is a clear difference between cirrus cover in supersaturated air in comparison with sub-saturated air. Comparing March 2016 with March 2020 which show relatively low cirrus cover compared with other years in March, Figure 4.6 shows that those anomalies take place in a different meteorological domain - for March 2016 the cover in supersaturated air is anomalously low, while for March 2020 this is the case in sub-saturated air. Further research is needed on this. Yet for both saturation levels no statistically significant correlation with air traffic is found.



(a) Heat map of cirrus cover retrieved from CALIPSO CALIOP data in supersaturated air.



(b) Heat map of cirrus cover retrieved from CALIPSO CALIOP data in subsaturated air.

Figure 4.6. Heat maps showing the cirrus cover over Europe in a) supersaturated air and 2) in sub-saturated air. Time series runs from 2015 till 2020.

To further elaborate on the question what is causing the observed fluctuations in monthly mean cirrus cover, the fraction is ISSR regions is calculated for each month, which is shown in Figure 4.7. Also the ISSR fractions show a significant difference between summer and winter. It could be seen that the relatively low (high) mean cirrus cover in March 2020 (September 2016) correlates with a relatively low (high) ISSR fraction. The same applies to December 2016. The R^2 between ISSR and CALIPSO monthly cloud cover is 0.857, therefore explaining about 86% of the observed cirrus cover variance.

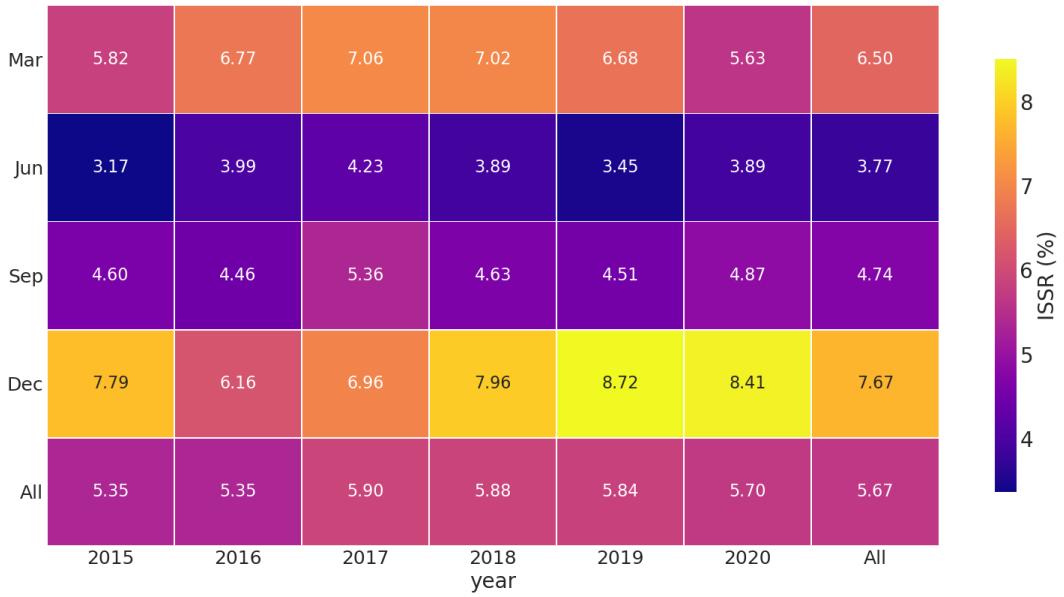


Figure 4.7. Heat map of the percentage of ice-supersaturated regions over the location where CALIPSO has overpassed.

4.2. Short-Term Time Series Analysis using SEVIRI

Spatially averaged flight data from March 2015, evaluated every 15 mins have been plotted in Figure 4.8 against time. The daily air traffic cycle is clearly visible with peaks during daytime around 100 m km^{-2} hr^{-1}, dropping to around 10 m km^{-2} hr^{-1} at night. Another cycle with a repeat period of 7 days can be deduced, which relates to more scheduled flights during the weekends, with the highest peaks attained on Sundays and the lowest on Tuesdays. The peaks show a nearly perfect sinusoidal pattern.

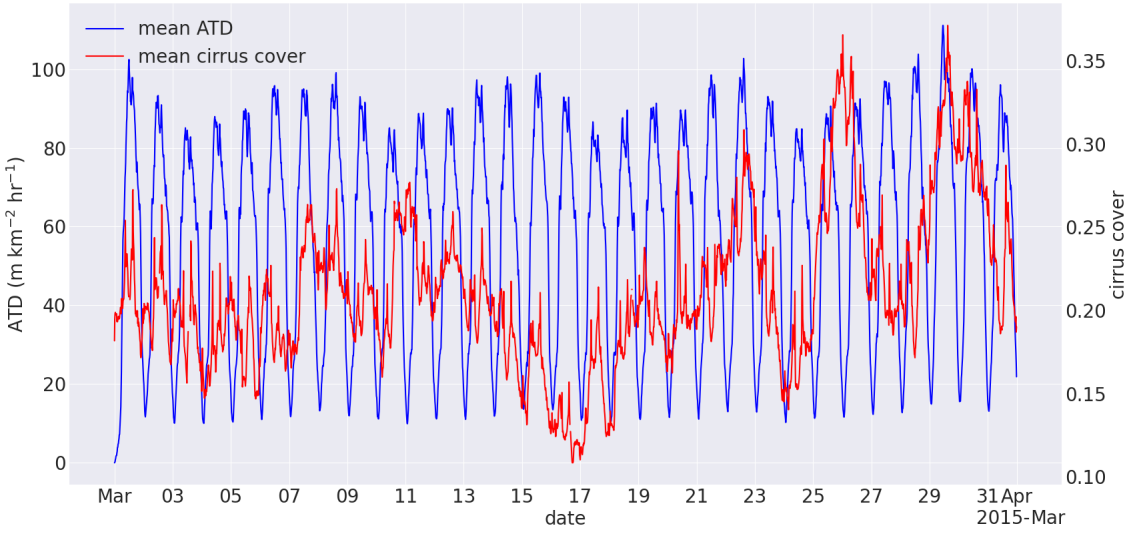


Figure 4.8. Time series of ATD and cirrus cover for March 2015.

The spatially averaged cirrus cover shown in red results as a much noisier signal with some apparent synoptic variability over time scales of a couple of days to a couple of weeks. This synoptic variability cannot be explained by air traffic, and appears to be dictated by synoptic meteorological conditions. This is illustrated in Figure 4.9, where the fraction of ISSRs is plotted against time. On the background the mean cirrus cover is shown. The mean cirrus cover strongly follows the ISSR pattern, and the latter explains over half ($R^2 = 0.52$) of the cirrus cover variability.

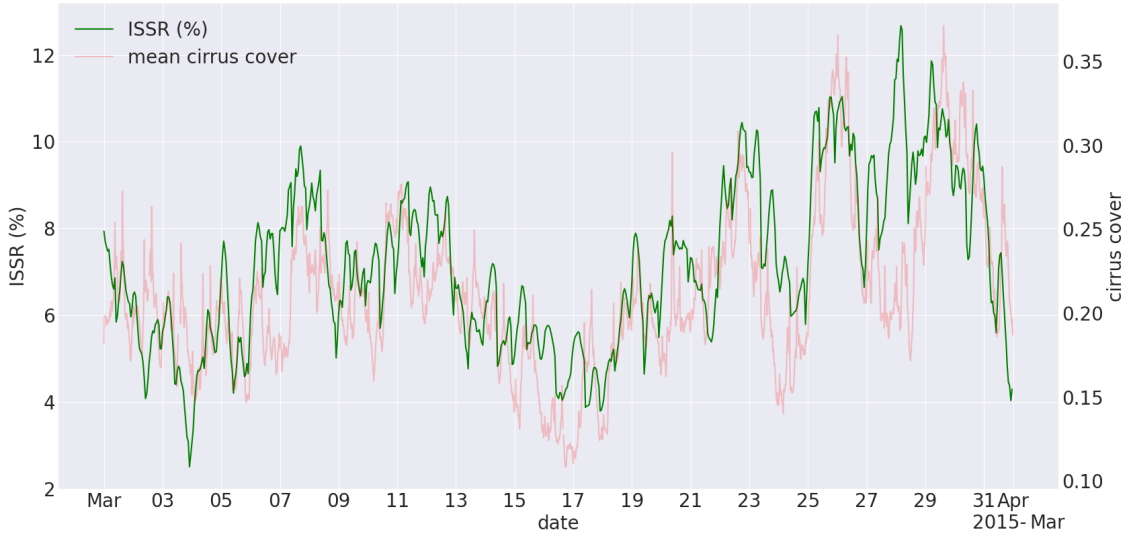


Figure 4.9. Ice supersaturated regions (ISSR) plotted as a percentage of all regions for March 2015. In the background the cirrus cover trend is plotted.

The monthly time series data is used in Figure 4.10 to construct the mean daily cycle of cirrus and ATD, along with their respective 95% confidence intervals. This figure clearly shows that cirrus cover experiences a large drop around noon, spanned by two sharp peaks of which the late-afternoon peak is highest. Those peaks appear consistently over the course of March 2015 and could not be entirely explained by meteorology. Diurnal temperature and RH variability are negligible in conjunction with ISSR fractions. This raises the question to what extend air traffic is related to those peaks, which is plausible looking at the ATD cycle, although this leaves the ”noon-drop” unexplained.

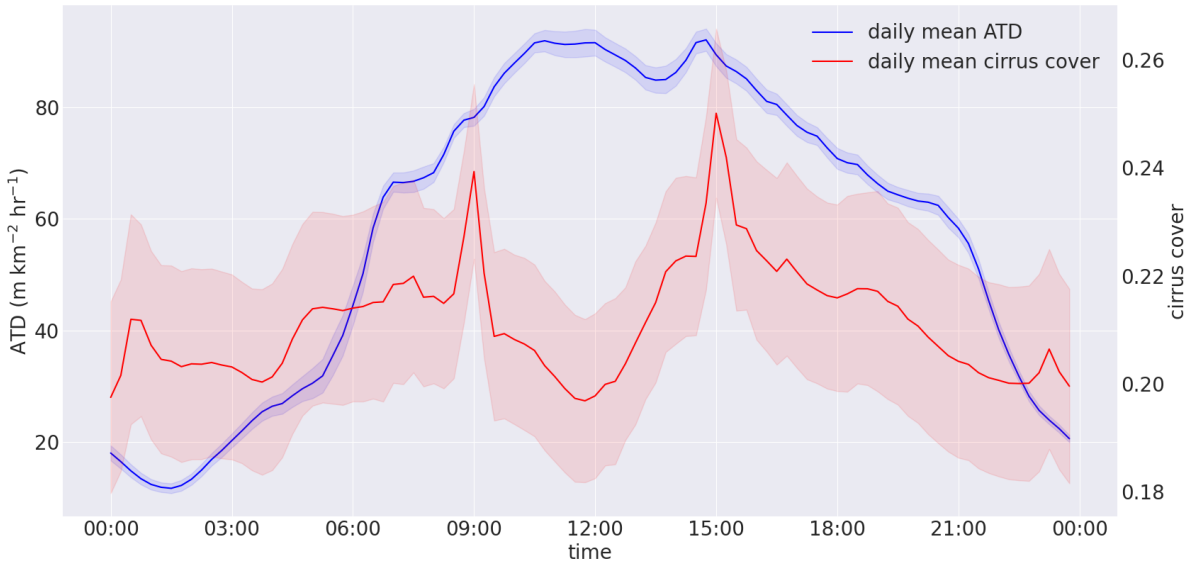


Figure 4.10. Mean daily cycle of ATD and cirrus cover for March 2015.

4.3. Short-Term Assessment of Cirrus Formation

The method explained in Section 3.5 is to bin all data in a preset number of classes based on ATD that are algorithmically determined by solving a maximization problem. Subsequently the classes are subset into supersaturated and sub-saturated regions as was done in the previous section, and the mean percentage point changes in cirrus cover over one sampling period (15 mins) for each subset is determined. This is shown in Figure 4.11. The error bars indicate the 95% confidence intervals for each class mean based on the bootstrapping technique, with 1,000 abstractions.

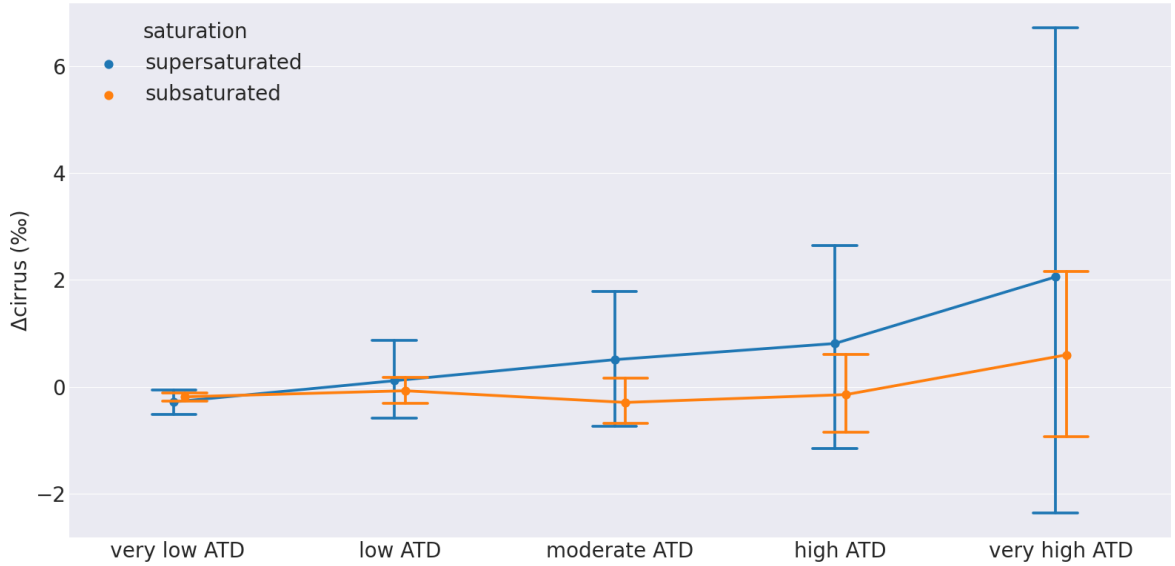


Figure 4.11. The mean cirrus cloud cover increment in permilles within the five different Air Traffic Density bins. Different plots are shown for sub-saturated ($RH < 100\%$) and supersaturated ($RH \geq 100\%$) air. Error bars show the 95% confidence intervals based on a bootstrapping technique with $n = 1000$.

Both trends from Figure 4.11 are tested for significance using a t-test with the significance level set at 0.05. Strong evidence is found for a significant positive correlation between ATD and cirrus cover change under supersaturated conditions, with a slope of 11.2 percentage points per km km^{-2} hr^{-1} and a p-value of $6 \cdot 10^{-5}$. In sub-saturated air however, no highly convincing support for a trend was found with a p-value of 0.06.

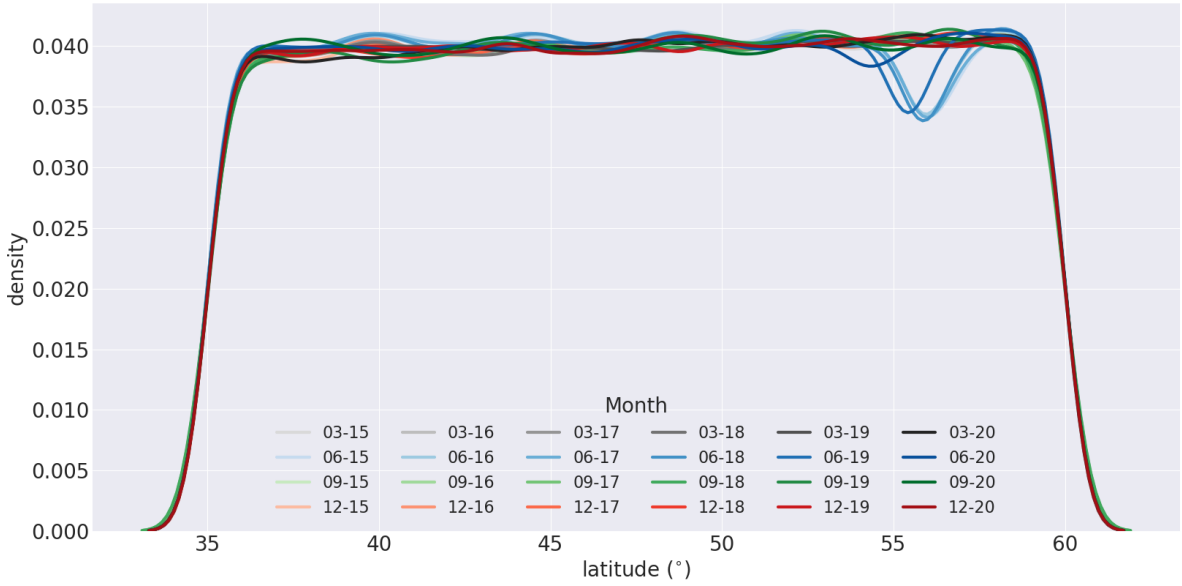
4.4. Verification & Validation

A couple of verification and validation procedures have been done to assess the coherency of the different data sources used, as well as given some interpretation to the reliability of any observed trends. The generalizability of cirrus cover over the entire ROI using CALIPSO is looked at, which was done in the temporal domain previously by looking at the hourly overpasses of CALIPSO (Figure 3.7). Here a similar analysis is done, but now for the spatial (lon-lat) domain. Furthermore, a high-level comparison of the Meteosat product with CALIPSO is made, and the correspondence between ERA5-reported temperatures and satellite-reported temperatures (CALIPSO) is looked at.

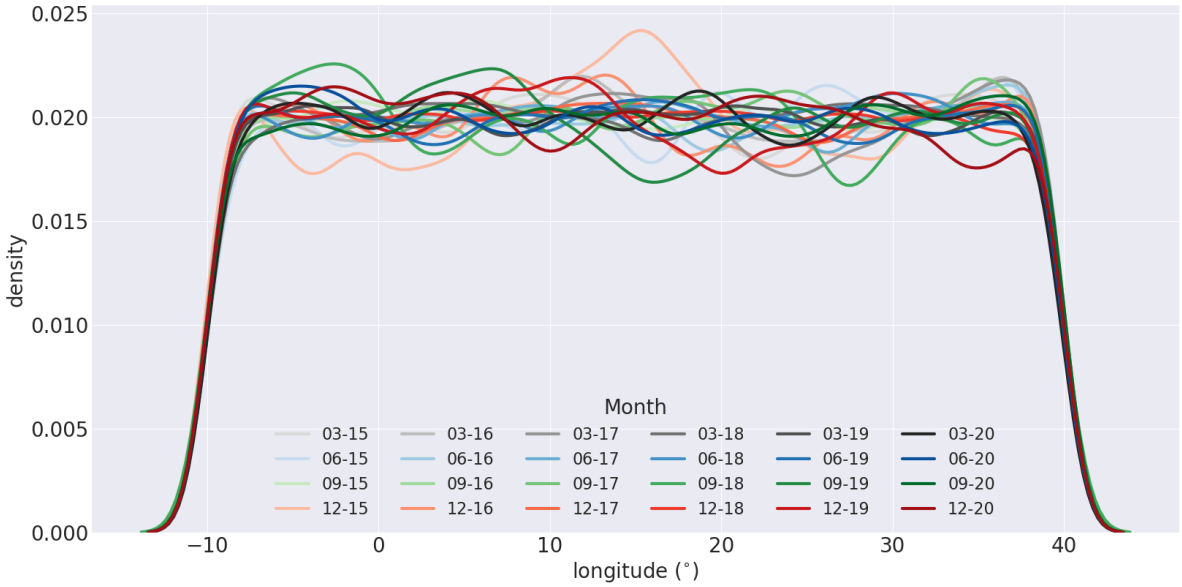
4.4.1. Overpass Frequency Distribution of CALIPSO Satellite

Important for the comparability of retrieved cirrus cover between months is the coherency in spatial coverage of the satellite for each month. Figures 4.12a and 4.12b show for each month the KDE of the CALIPSO overpass frequency along the latitudinal and longitudinal axis, respectively. The latitudinal coverage amongst all months appear more coherent than the longitudinal coverage, which can be understood by the high-inclination angle (97°) of the CALIPSO satellite. Hence during nearly each overpass the entire latitudinal domain is covered. This is not the case for the longitudinal axis, which is being covered for a relatively narrow band at each overpass. An apparent feature from Figure 4.12a are the months of June (blue curves) that show a drop in overpass frequency around 55° latitude. The cause of those deviations, and whether this has a significant influence on the results obtained, should

be further investigated.



(a) KDE for the latitudinal coverage of CALIPSO.



(b) KDE for the longitudinal coverage of CALIPSO.

Figure 4.12. Kernel Density Estimate for the spatial coverage of CALIPSO overpasses for all months included in the analysis.

4.4.2. Detection Accuracy of SEVIRI

In order to assess the reliability of the SEVIRI passive sensor product related to cirrus cloud detection, particularly those clouds with a small optical depth, the product is compared with the CALIPSO product over March 2015. All CALIPSO daytime overpasses, totalling 55, are converted into a gridded L3 representation on the $0.25° \times 0.25°$ grid and correlated with corresponding Meteosat observations. The result of this analysis is shown in Table 4.1. In 69% of the cases Meteosat detection corresponds with CALIPSO. Also, Meteosat recognized cirrus overcast in 42.2% of the cases, while for CALIPSO this is 41%. Despite the fact that in most cases both products agree on the occurrence of cirrus, one would expect more cirrus detection by the most sensitive sensor (CALIOP). The outcome of this analysis could however be influenced by the gridding procedure, as the data presented are a discretized version of the raw satellite data. The mean cirrus cover given by both products are 0.34 (34%) for CALIPSO and

0.22 (22%) for Meteosat, a significant difference.

		CALIPSO		
		no cirrus	cirrus	*total*
Meteosat	no cirrus	2647	897	*3544*
	cirrus	997	1589	*2586*
	total	*3644*	*2486*	*6130*

Table 4.1. Cross-tabulation of CALIPSO CALIOP cirrus detection vs Meteosat SEVIRI. Data has been taken at each CALIPSO overpass during March 2015.

Two histograms of the classes shown in blue and red are shown in Figure 4.13. The hypothesis is that COT from the blue class should be lower than those from the red class, as for the blue class the active sensor is able to detect the clouds whereas the passive sensor is not. This is confirmed by looking at Figure 4.13, where the normalized probability density for optically very thin clouds is significantly higher for the blue than for the red class.

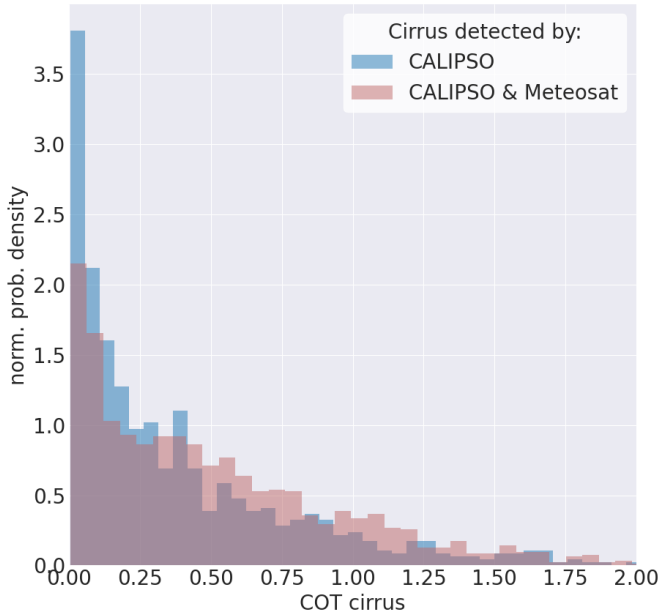


Figure 4.13. Comparison of COT between CALIPSO and Meteosat.

4.4.3. ERA5 Temperature versus Satellite-Observed Temperature

From the CALIPSO backscatter signal additional features like mid-layer cloud temperature can be extracted as already mentioned. Those temperatures however can only be extracted when a backscatter object, such as a cloud, is present. Therefore this data source for temperature, that could potentially be used as an indicator for ambient air temperature, is merely used as a validation tool. Even though cloud temperature might slightly deviate from the surrounding temperature, this temperature gradient is assumed small enough to ignore it for now.

Figure 4.14 shows a scatterplot of retrieved mid-layer cloud temperatures from CALIPSO and the corresponding temperature reported by ERA5. For this again the ROI has been divided into 11 pressure layers, each centered at the corresponding ERA5 pressure levels. The temperature offset with the layer temperature from ERA5 Reanalysis resulted in a mean error of 6.3 K, and the squared correlation coefficient $R^2 = 0.66$. It should be emphasized that the temperatures from ERA5 Reanalysis are column-averaged values over a height of, depending on the location, 25 or 50 hPa, and might hence on itself yield slightly different values. Despite the fact that the CALIPSO-reported temperatures cannot

be used for the entire analysis, this verification method serves as another indication of biases in using different data sources.

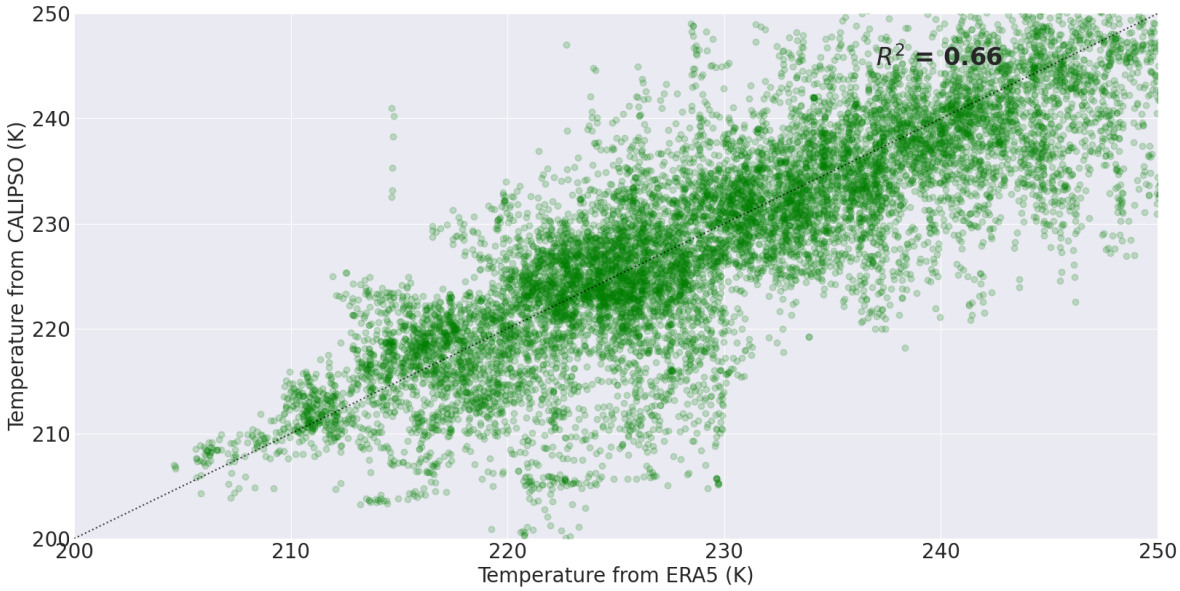


Figure 4.14. Comparison temperature reported by ERA5 and mid-layer cloud temperature reported by CALIPSO.

5

Conclusion & Recommendations

This report has given an overview of cirrus cover properties in European airspace in conjunction with air traffic, by combining multiple data sources into a big data approach. It has been shown that air traffic is subject to a consistent rise in volume of about 4% a year within the investigated airspace, leaving the question whether this gets reflected into cirrus cloud cover, the latter having relevance for Earth's radiative budget. No indication was found using CALIPSO that mean cirrus cover changes along with air traffic over the period 2015-2020.

Possible signatures of air traffic were found when subsetting atmospheric regions on saturation level and making inter-comparisons. There was found a statistically significant difference between the mean difference in cirrus cover over supersaturated and sub-saturated regions during daytime and nighttime. For March 2015-2020, March 2020 showed lowest cirrus cover (29.2% compared to 33.9% on average) while the difference in cirrus cover between day and night (30%) was highest during this month compared to all other months. Regarding the first observation, this could be to a large extent caused by deviating meteorological conditions, as RH was found to be significantly lower than usual in the upper troposphere, while the temperature was slightly higher than usual above 10 km. Regarding the large offset between daytime and nighttime cover, there might be a possibility that this is induced by reduced air traffic during this month related to COVID-19. Further analysis on this is required. Also the general abundance of cirrus overnight compared to daytime is apparent, whilst this pattern seems to be reversed during summer months, particularly in September. It cannot be excluded from this analysis that there is a seasonality involved in this feature. Concurrently, the limited diurnal variability in CALIPSO overpass times induces a problem regarding generalizability. This is exemplified by the observed "noon-drop" in daily mean cirrus cover for March 2015, which shows that the measured cirrus cover might be highly dependent on the satellite overpass time, even when averaging over a longer time period.

Looking at changes in cirrus cover over time scales of tens of minutes resulted in a convincing relationship with ATD in supersaturated regions, a relationship that was not found as evidently in sub-saturated regions. From this and the long-term analysis it is concluded that lifetime of contrails play a key role here, as correlations between air traffic and cirrus cover appear mostly on shorter timescales. The signatures might be seen back on longer timescales when applying models that properly take into account the variability in meteorological conditions. The LR and RF model investigated performed too poorly to implement in this research, but may be improved upon by including different variables, and by diving deeper into the hyperparameters involved. The appearance of cirrus resulted to correlate strongly with the fraction of ISSRs, both from the short-term and the long-term analysis.

Further elaboration on the research questions require higher data quality, 1) on the side of air traffic data for the years 2019 and 2020, and 2) on the linkage between aircraft metadata and engine fuel flow. In addition, the latter could be improved by further development of the NLP algorithm identifying the engine type and linking this with fuel flow. Parametrizing air traffic with incorporation of exhaust properties will increase the accuracy and relevance of the research, as engine properties are under

ungoing development.

Finally, some results point at possible biases in the meteorological data used, e.g. the temperature profiles having a positive lapse rate in the lower stratosphere and a mean error of 6.3K when comparing to CALIPSO. A good starting point would be to further discretize the analysis in the vertical by gathering meteorological data on more pressure levels. The cirrus detection capabilities of CALIPSO and Meteosat were confirmed to be largely offset, with CALIPSO detecting 34% cirrus cover versus 22% for Meteosat in March 2015 and both sensors disagreeing on the presence of cirrus in 31% of the cases. This illustrates the improvements that yet can be made regarding remote-sensed cirrus detection on a temporal and spatial high resolution.

6

Acknowledgements

First of all, I would like to thank the two supervisors, Dr. Mel Chekol and Prof. dr. Maarten Krol, that were involved in this project. Their feedback during each meeting (each fortnight) is much appreciated. A special thanks to Dr. Mel Chekol who has guided me through the project application process, despite and due to the lack of available resources at the time, and who has offered to supervise me when no other supervisor from the Department of Information and Computing Sciences was found.

Furthermore, I would like to thank Jan-Fokke Meirink from the KNMI for his reflectance on adequate data sources for this project, and for directing me to the CM SAF (EUMETSAT) CLAAS-2.1 data product. Another thanks goes to Ing. Carel van der Werf from Utrecht University, who has been my contact point concerning use of the Gemini cluster.

Finally, I would like to thank my university peer group and peer group moderator Drs. Yo-Yi Pat, study advisor at Utrecht University, for their mental support and (digital) companionship during the sometimes monotonous working days at home.

Bibliography

[1] Predicting Contrails Using an Appleman Chart. https://web.archive.org/web/20170612181029/http://science-edu.larc.nasa.gov/contrail-edu/resources-activities-appleman_student.php. Online; accessed: Mar 16 2021.

[2] Jason Brownlee. SMOTE for Imbalanced Classification with Python. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/. Online; accessed: Apr 05 2021.

[3] NASA Langley Research Center Atmospheric Science Data Center DAAC. CALIPSO LID L2 HDF4 Datafiles – Version 3.40. https://eosweb.larc.nasa.gov/project/CALIPSO/CAL_LID_L2_01kmCLay-ValStage1-V3-40_V3-40, 2020. Online; accessed: Feb 21 2021.

[4] EuroControl. R&D Data Archive. https://www.eurocontrol.int/dashboard/rnd-data-archive, 2020. Online; accessed: Jan 28 2021.

[5] EuroControl. Aircraft Performance Database. https://contentzone.eurocontrol.int/aircraftperformance/default.aspx?, 2021. Online; accessed: Feb 24 2021.

[6] Stephan Finkensieper, Jan-Fokke Meirink, Gerd-Jan van Zadelhoff, Timo Hanschmann, Nikolaos Benas, Martin Stengel, Petra Fuchs, Rainer Hollmann, Johannes Kaiser, and Martin Werscheck. Claas-2.1: Cm saf cloud property dataset using seviri. *Satellite Application Facility on Climate Monitoring*, Edition 2.1, 2020. doi: 10.5676/EUM_SAF_CM/CLAAS/V002_01.

[7] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N Thépaut. Era5 hourly data on pressure levels from 1979 to present. copernicus climate change service (c3s) climate data store (cds). 2018. doi: 10.24381/cds.bd0915c6.

[8] Yulan Hong and Guosheng Liu. The characteristics of ice cloud properties derived from cloudsat and calipso measurements. *Journal of Climate*, 28:3880—3901, 2016.

[9] International Civil Aviation Organization (ICAO). ICAO Aircraft Engine Emissions Databank. https://www.easa.europa.eu/domains/environment/icao-aircraft-engine-emissions-databank#group-easa-downloads. Online; accessed: Mar 02 2021.

[10] Bernd Kärcher. Formation and radiative forcing of contrail cirrus. *Nature Communications*, 9, 2018. doi: 10.1038/s41467-018-04068-0.

[11] Hermann Mannstein and Ulrich Schumann. Aircraft induced contrail cirrus over europe. *Meteorologische Zeitschrift*, 14:549–554, 2005. doi: 10.1127/0941-2948/2005/0058.

[12] Elena Mazareaunu. Year-on-year passenger traffic change of commercial airlines in Europe from 2013 to 2021. https://www.statista.com/statistics/658701/commercial-airlines-passenger-traffic-change-europe/. Online; accessed: Feb 10 2021.

[13] OpenFlights. Airport, airline and route data. https://openflights.org/data.html. Online; accessed: Feb 23 2021.

[14] Philipp Reuter, Patrick Neis, Susanne Rohs, and Bastien Sauvage. Ice supersaturated regions: properties and validation of era-interim reanalysis with iagos in situ water vapour measurements. *Atmos. Chem. Phys.*, 20, 2020. doi: 10.5194/acp-20-787-20201.

[15] Ulrich Schumann. On conditions for contrail formation from aircraft exhausts. *Meteorologische Zeitschrift*, 5:4–23, 1996.

[16] Matthias Schäfer, Martin Strohmeier, Vincent Lenders, Ivan Martinovic, and Matthias Wilhelm. Bringing up opensky: A large-scale ads-b sensor network for research. *In Proceedings of the 13th IEEE/ACM International Symposium on Information Processing in Sensor Networks (IPSN)*, pages 83–94, 2014. doi: 10.1109/IPSN.2014.6846743.

A

Feature Classification Flag Encoding

Bits	Field Description	Bit Interpretation
1-3	Feature Type	0 = invalid (bad or missing data) 1 = "clear air" 2 = cloud 3 = tropospheric aerosol 4 = stratospheric aerosol 5 = surface 6 = subsurface 7 = no signal (totally attenuated)
4-5	Feature Type QA	0 = none 1 = low 2 = medium 3 = high
6-7	Ice/Water Phase	0 = unknown / not determined 1 = ice 2 = water 3 = oriented ice crystals
8-9	Ice/Water Phase QA	0 = none 1 = low 2 = medium 3 = high
10-12	Feature Sub-type	
	If feature type = tropospheric aerosol, bits 10-12 will specify the aerosol type	0 = not determined 1 = clean marine 2 = dust 3 = polluted continental/smoke 4 = clean continental 5 = polluted dust 6 = elevated smoke 7 = dusty marine
	If feature type = cloud, bits 10-12 will specify the cloud type.	0 = low overcast, transparent 1 = low overcast, opaque 2 = transition stratocumulus 3 = low, broken cumulus 4 = altocumulus (transparent) 5 = altostratus (opaque) 6 = cirrus (transparent) 7 = deep convective (opaque)
	If feature type = stratospheric aerosol, bits 10-12 will specify stratospheric aerosol type.	0 = invalid 1 = PSC aerosol 2 = volcanic ash 3 = sulfate/other 4 = elevated smoke 5 = spare 6 = spare 7 = spare
13	Cloud / Troposheric Aerosol / Stratospheric Aerosol QA	0 = not confident 1 = confident
14-16	Horizontal averaging required for detection (provides a coarse measure of feature backscatter intensity)	0 = not applicable 1 = 1/3 km 2 = 1 km 3 = 5 km 4 = 20 km 5 = 80 km

Figure A.1. Table used to decode *Feature Classification Flags* 16 bit integer from CALIPSO products. Retrieved from `https://www-calipso.larc.nasa.gov/resources/calipso_users_guide/data_summaries/vfm/index_v420.php`.

B

Major Airport Hubs for March 2015

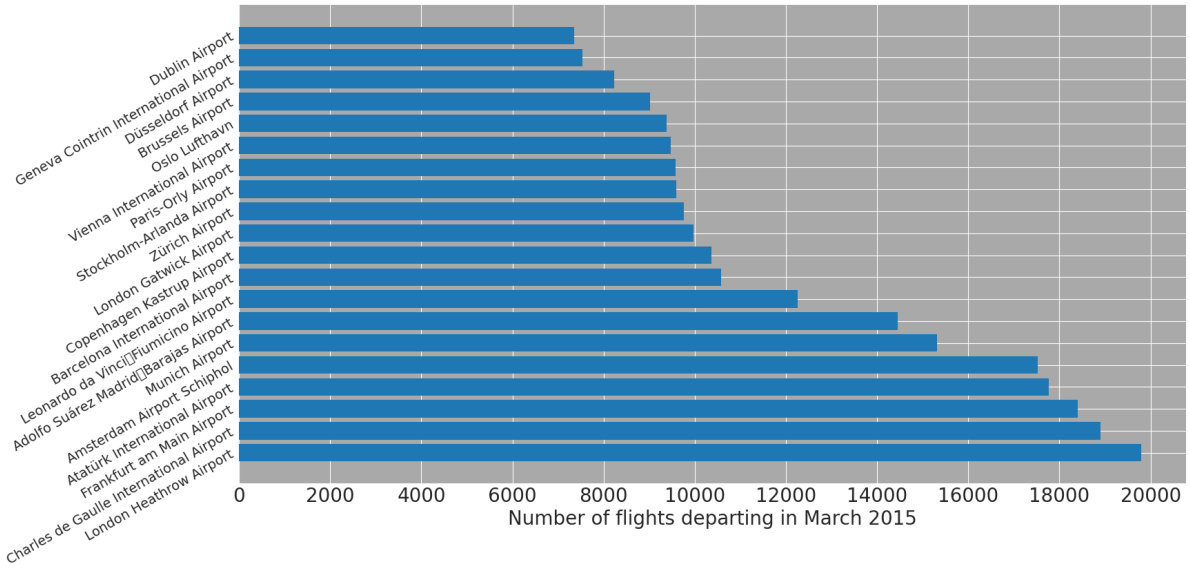


Figure B.1. Major airport hubs during March 2015.