scientific reports

OPEN

Statistical characterization of airplane delays

Evangelos Mitsokapas[1,2], Benjamin Schäfer[1,2✉], Rosemary J. Harris[1] & Christian Beck[1]

The aviation industry is of great importance for a globally connected economy. Customer satisfaction with airlines and airport performance is considerably influenced by how much flights are delayed. But how should the delay be quantified with thousands of flights for each airport and airline? Here, we present a statistical analysis of arrival delays at several UK airports between 2018 and 2020. We establish a procedure to compare both mean delay and extreme events among airlines and airports, identifying a power-law decay of large delays. Furthermore, we note drastic changes in plane delay statistics during the COVID-19 pandemic. Finally, we find that delays are described by a superposition of simple distributions, leading to a superstatistics.

The aviation industry was a rapidly growing sector until recently, prior to the current COVID-19 pandemic. Economic growth led to higher average yearly distances travelled, as well as higher air traffic volumes, robustly observed among several regions worldwide until 2019[1,2]. But both the ongoing pandemic[3] and also the push towards more renewable options in aviation[4] may induce a considerable change in the industry in the future. This makes the industry a very interesting object to study as it transforms.

As a passenger, an important benchmark for evaluating travel options, e.g. in terms of airports, airlines or even modes of transportation (train vs plane) is the punctuality of each option. In particular, flight delays severely decrease customer satisfaction[5] and might lead to customers choosing a different airport or airline, in the long term. Generally, it is important to quantitatively understand delay-risks both in terms of the expectation values but also in terms of the extreme events, i.e. quantifying how likely a very early or very late arrival is.

The study of delays in aviation is already an active field of research. Previous, simple, investigation frameworks to classify and categorize delays have been proposed[6] but mostly rely on mean values. In other cases, stochastic models of plane delays[7] were developed either without considering the corresponding probability distributions or assuming simple Normal or Poisson distributions[8]. More recent work also includes the application of machine learning techniques to aviation data, e.g. via recurrent neural networks[9]. One problem of any data-driven approach is that many articles on aviation research solely rely on proprietary data: In a recent review investigating 200 research articles, 68% were based on proprietary data[10]. Hence, to enable the broader applicability of machine learning applications, more publicly available data are still required.

To quantify delay statistics, we will go beyond the often-used averages of delays[6] and instead investigate the entire probability density function of delays at a given airport. Thereby, we consider all possible delay values, from highly negative delays (i.e. flights arriving significantly earlier than their scheduled arrival time) to severely positively delayed flights. These delay distributions are influenced by many different aspects, including random events, congestion, delay propagation between airports[11,12] and (for long-haul flights on large scales) the topological structure of the worldwide air transportation network[13,14]. To explain the emergence of heavy tails in a local distribution, i.e. extreme deviations from the mean, we will utilize superstatistical modelling[15]. Such an approach has been successfully applied in transport before, for modelling train delays[16]; it has also attracted recent interest when describing fluctuations in the energy system[17] and air pollutant concentrations[18] and it has been extended to the general framework of diffusing diffusivities in nonequilibrium statistical physics and biologically inspired physics[19–21].

In this article, we present new data collected from 2018 to 2020 at several UK airports, with a particular focus on Heathrow, being the most important international hub in the UK. The data were publicly available from the arrival information of each airport, given out on their websites each day but had to be collected and processed for further usage. While the past arrival data can no longer be accessed via the airport websites, all collected data have been uploaded in a repository, see "Methods". We analyse the full probability density of delay distributions and introduce certain performance indices to describe these distributions, such as the mean delay, the exponential decay rate of negative delays, and the power-law exponent of large positive delays. These indices are then compared for the different UK airports and the different airlines operating at these airports, to understand the

[1]School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK. [2]These authors contributed equally: Evangelos Mitsokapas and Benjamin Schäfer. ✉email: b.schaefer@qmul.ac.uk
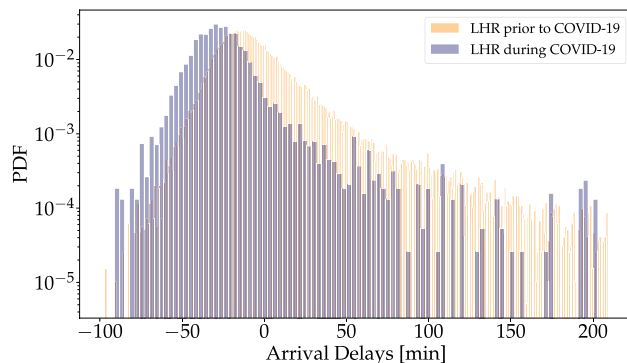
1

Figure 1. Flight delays follow a broad distribution with large negative and positive delays. We display LHR delay histograms prior to and during the COVID-19 pandemic, both normalized. As the COVID-19 LHR data set is significantly smaller in size, compared to the regular LHR data set, it contains many gaps, where no data were recorded. The COVID-19 data set is significantly shifted towards the left (smaller delays) as compared to the pre-pandemic time.

main features of the delay statistics (such as frequency of extreme delays, average delay per airport or per airline, etc) in a more systematic way. Finally, we deal with a theoretical model to explain features of the delay statistics. We show that the power law of large positive delays can be linked to a superposition of exponential delays with a varying decay parameter, in a superstatistical approach. Conversely, negative delays (early arrivals) do not exhibit any power laws but simply behave in an exponential way, with extremely early arrivals exponentially unlikely. Throughout this article, we assume that passengers prefer to arrive as early as possible, i.e. with as little positive and as much negative delay as possible.

New data

We collected flight details from a number of different airports. For the purposes of this article, we have taken into consideration the top five UK airports, in order of passenger traffic[22], namely: London Heathrow Airport (LHR), London Gatwick Airport (LGW), London Luton Airport (LTN), London Stansted Airport (STN) and Manchester Airport (MAN). For a period of time lasting between Autumn 2018 and Spring 2019, we collected a combined total of approximately two-hundred and twenty thousand (2.2×10^5) flight-arrivals from all five airports mentioned above. Furthermore, we continued collecting flight-information from London Heathrow during the 2020 COVID-19 pandemic, to illustrate the effect the lockdown had on the delay distribution. For each flight, we recorded the airline company operating the flight along with the corresponding flight number, departure and arrival airports, as well as scheduled and actual landing times. The delay is then computed simply as the difference between an aircraft's scheduled arrival time and its actual arrival time. Note that airlines and airports presumably have some freedom in setting the scheduled arrival time, potentially influencing the average "delay" (average difference between scheduled and actual arrival). We made all collected data publicly available. For details of the data processing and availability, see "Methods".

The main body of our data (about 85%) is sourced from London Heathrow, making it the chief focus of our analysis simply due to its size. London Heathrow is an international airport operating flights of 80 different airlines in total, which fly to 84 different countries around the world, as of 2019[22]. Of course, in addition there are domestic flights within the UK. The passenger nationalities are 48% European and UK and 52% from the rest of the world. It is the busiest airport in Europe by passenger traffic[22].

The empirical probability density function (PDF) of all delays is a key characteristic to monitor, see Fig. 1 for all Heathrow delays. There, we compare the data collected from 2018 to 2019 with more recent data collected during the 2020 COVID-19 pandemic (during the first lockdown in Spring to Summer 2020), which led to a drastic reduction in air transport[23,24]. There are two interesting observations: Firstly, the delay statistics under COVID-19 are shifted to the left, indicating overall smaller delays (including more negative delays); secondly, the general shape of the distribution does not change drastically. In particular, we observe a fast decay of the PDF of negative delays on the left side and a much slower decay of the PDF on the right side for positive delays. In the following sections, we will analyse this behaviour in much more detail.

Quantifying delay statistics

Starting from a histogram of the flight delays, we derive three indices/measures to quantify flight delay distributions: Mean delay, exponent of left exponential and power-law exponent of right q-exponential, as explained below in detail. We will use the LHR data previous to any COVID-19 influence as our main example.

As a first step, we split the full histogram at its peak value into two histograms, a left flank of predominantly negative delays and a right flank of predominantly positive delays, see Fig. 2. Based on the shape of the empirical distributions, we use exponentials and q-exponentials as fitting functions, see also "Methods" for details. Splitting the histogram has two advantages: Firstly, the analysis of each flank is much simpler than the analysis of the full aggregated data. Secondly, a given stakeholder might be particularly interested in positive rather than negative delays, or vice versa.
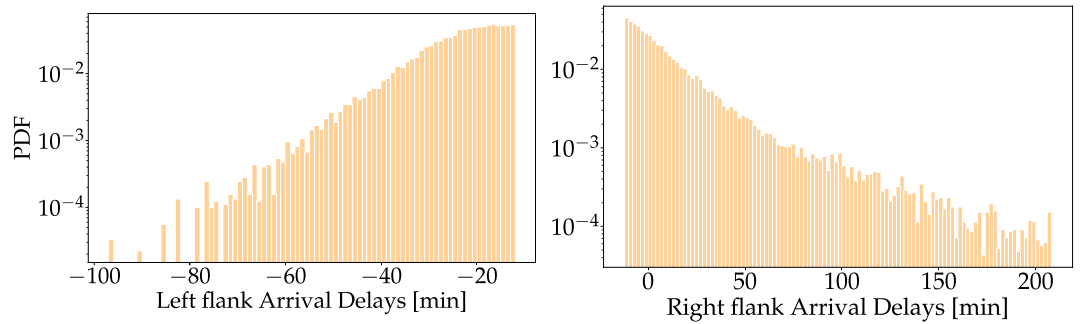
Figure 2. Splitting the full distribution at the peak leads to two easier-to-fit flanks. Left: Negative delays decay approximately linearly in the log-scale and thereby suggest an exponential fit (1). Right: Positive delays display substantial heavy tails and thereby suggest the usage of a q-exponential function (2).
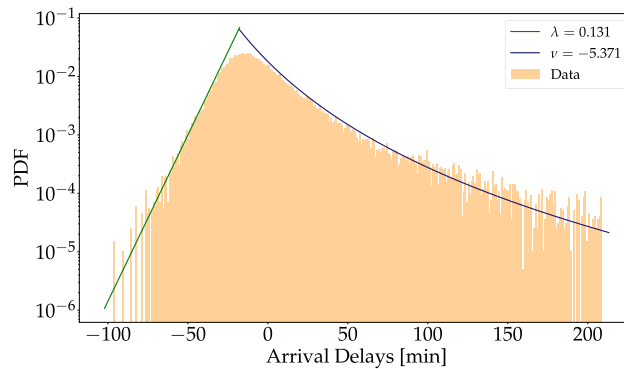

Figure 3. Exponential (green) and q-exponential (blue) theoretical distributions capture the empirical distribution. The fits are obtained via the MLE method, see "Methods" for fitting details. To complement the over-estimated "peak" (tent-like shape) we introduce the mean delay μ index

The left flank is observed to be well approximated by an exponential function of the form

$$p(t_L; \lambda) = \lambda e^{-\lambda t_L}, \quad \lambda > 0, \tag{1}$$

where t_L are the rescaled arrival delays on the left flank, see "Methods" for details. The exponent λ here quantifies the exponential decay of the probability of early arrivals. Therefore, a large λ implies that few flights arrive very early while a small λ indicates that very large negative delays are observed. Since we assume that passengers prefer to arrive as early as possible, a small λ indicates good performance.

The right flank of the delay distribution obeys a power law, i.e. a slow decay of $p \sim t^\nu$, with ν negative. To quantitatively describe the right flank, we use a q-exponential function[25] of the form

$$p(t_R; q, \lambda_q) = (2 - q)\lambda_q \left[1 + (q - 1)\lambda_q t_R \right]^{\frac{1}{1-q}}, \tag{2}$$

where t_R are the rescaled arrival delays on the right flank, see "Methods" for details. The power-law exponent, i.e. the rate at which the probability density decays for high (positive) delay values, is given by $\nu := 1/(1 - q), 1 < q < 2$. Note that the scale parameter $\lambda_q > 0$ is relevant for the precise fit but does not impact the power-law exponent ν. Since the power-law decay is controlled by the value q, we utilize q to characterize the right flank. Contrary to the left-flank exponential decay, good performance is indicated by the absolute value of the right-flank power law exponent ν being large. The reason is that large (absolute) values of ν imply a rapid decay of the probability density of positive delays, i.e. fewer extreme events of very delayed arrivals.

Finally, we note that the two flanks describe the tails of the distribution well, but overestimate the height of the peak, i.e. the most likely value, see Fig. 3. To include more information on the most frequent delays, we complement the two previous fits by using the mean delay μ as a third index. Here we interpret a small positive μ, or a negative μ (indicating early arrival), as desirable for passengers. In the case of LHR, the three delay indices that we introduced are $\lambda = 0.131, \mu = -5.06$ and $\nu = -5.371$. We also introduce a continuous smooth fitting function for the full range in the "Connecting the flanks" section.

Note that the mean value μ can be easily manipulated by airline companies by scheduling flight arrival times later then actually needed, hence always causing a negative mean delay, which may artificially improve their performance. On the contrary, the tail behavior truthfully represents the extreme event statistics for both positive and negative delays and cannot be easily manipulated by the operators.
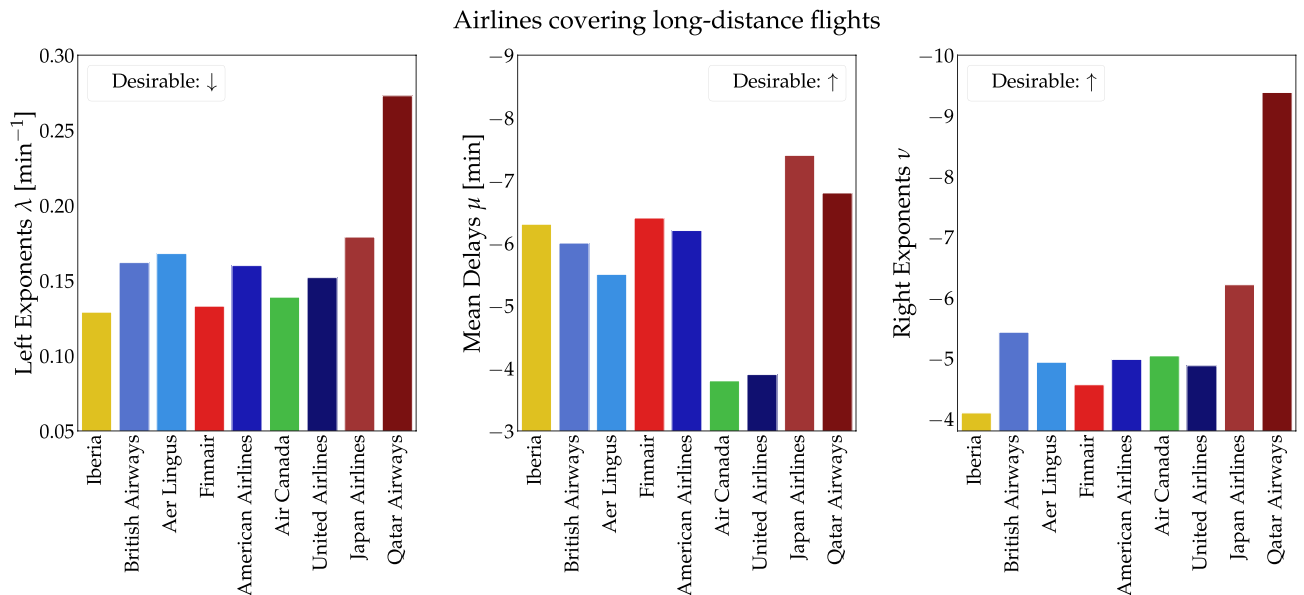
Figure 4. International airlines appear to differ substantially in their three delay indices. We plot the left-side (negative) delay exponential decay, right-side (positive) delay power-law decay and the mean delay. Arrows indicate whether a small or large value is desirable.

Comparison of airports and airlines

We here use the previously developed framework to quantify and compare delay statistics for different airlines and airports. Intuitively, we expect that long-distance flights would, on average, yield more extreme early or late arrivals, compared to the corresponding short-distance ones. Thus, we distinguish between short-distance airlines, covering mostly domestic and European destinations, and airlines that include long-distance, international destinations, as well as destinations within Europe. We first compute the three indices λ, μ, ν for each of those airline groups and then compare full airport statistics, aggregating all airlines.

There are several factors impacting the delay distribution for each airport or airline: Airline policies, flight routes, technical defects or issues with documentation contribute to 27% of all delays[26]. Specifically, overseas flights are more sensitive to wind (head wind or tail wind), as well as unstable weather conditions (storms, fog) and military exercises. Airlines operating international flights, as illustrated in Fig. 4, exhibit considerable variations in their flight delay indices. Note that a low left exponent λ may be regarded as a desirable property (flights often arrive very early) while good performance is definitely indicated by low mean μ and right exponent ν (low mean delay and few very late arrivals). Since the latter two quantities tend to be negative, their absolute values should be large. Comparing the airlines, we observe a "grouping" behaviour for some of the carriers. On the one hand, airlines having a blend between short-distance (e.g. domestic or EU) and overseas destinations, such as Iberia, British Airways (BA), Aer Lingus and Finnair, appear to follow a similar trend for each index. On the other hand, airlines that do not possess such a spread of destinations tend to perform well only in some of the indices. As an illustrative example, we choose Air Canada and United Airlines: Although both their left and right exponents are in a similar range to the other airlines, their mean delays are substantially less negative than those of their competitors.

Characterization of short-distance flights shows a strong grouping of the delay behavior for some airlines. As seen in Fig. 5, comparison of five of the largest low-cost domestic and European providers, reveals a systematic similarity between Wizz Air, easyJet and Ryanair. All three airlines manage to perform well in the left exponent metric, maximizing early arrivals, while they maintain an acceptable negative average delay (with easyJet obtaining the lowest value here). Again, they are characterized by similar right-exponents, translating to a certain share of overall late arrivals. Furthermore, Jet2 outperforms all other short-distance airlines in λ left-exponents and mean delays. Finally, Vueling resembles Wizz Air and Ryanair values in the λ and μ metrics but seems to have less late arrivals as per its high right exponent ν.

Comparing the long distance airlines with the short-distance ones, we notice some differences: Airlines covering long distances tend to display lower (more desirable) left exponents as well as more negative mean delays. Meanwhile, the right exponent behavior is similar between the two groups with Vueling and Qatar Airlines as the "outliers" in their respective categories. Whether this behavior is due to company policies or flight distance remains a question for future research.

Studying the indices for individual airports yields interesting insights as well. Airports populated by airlines flying mainly to domestic and EU destinations, such as LTN and STN, have a mixed score in both early and late arrivals, with an approximately net zero mean delay, see Fig. 6. On the one hand, STN is characterized by the minimum λ value, showing the best performance in early arrivals in the group of airports, while LTN attains the maximum value. On the other hand, it can be seen that LTN scores the best ν value while STN lies very slightly above the group median ν. Interestingly, mean delays at MAN airport are net zero, contrary to LHR and LGW where arrivals are scheduled in such a way that the mean delay is negative. Furthermore, MAN seems to have a

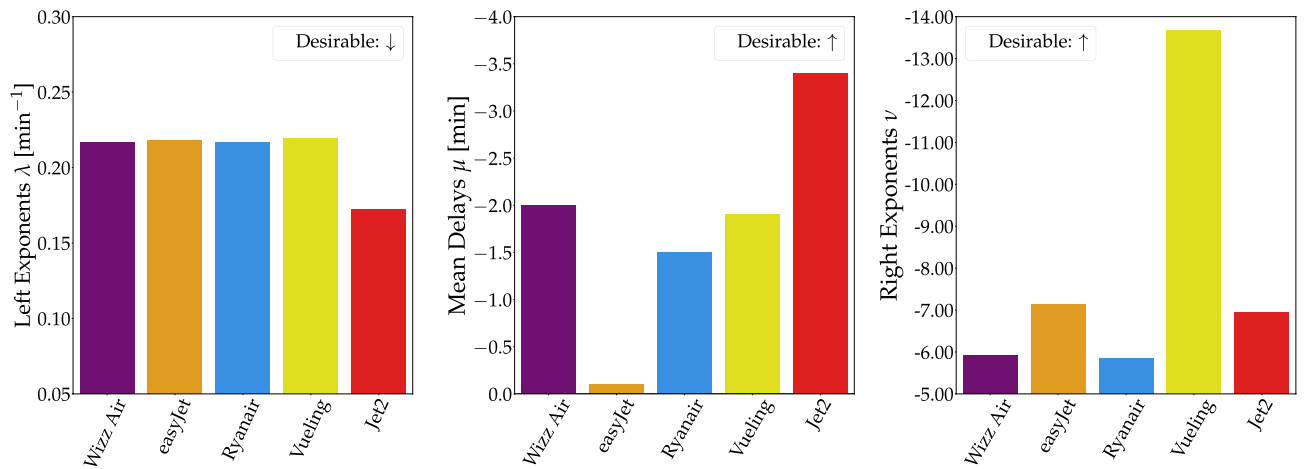Airlines not covering long-distance flights

Figure 5. Delay indices for low-cost airlines not covering long-distance flights. Wizz Air, easyjet, Ryanair and Vueling share the largest λ index (early arrivals). Jet2 has the lowest mean delay μ and Vueling is characterized by the lowest ν index (late arrivals).
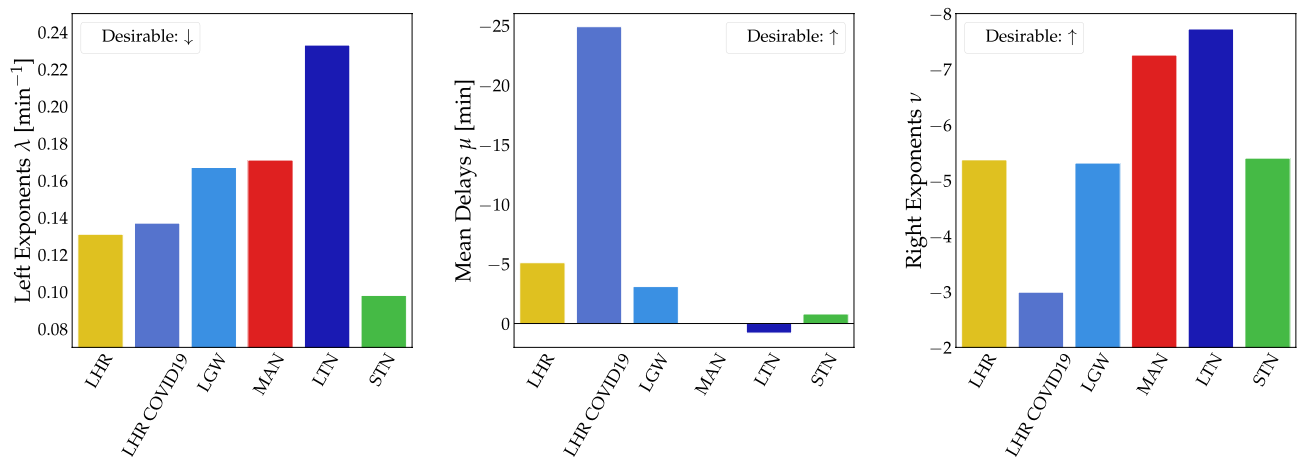
UK Airports Three Measures

Figure 6. Airports appear to differ substantially in the three delay metrics. Airports that serve mostly domestic and European destinations, such as LTN and STN, behave differently from international airports such as LHR, LGW and MAN.

similar performance to LGW in the early arrivals index, having a slightly worse score, but does attain the second best value when compared from the perspective of extreme positive delays. International airports LHR and LGW (with the exception of LHR COVID-19) tend to cluster around similar values for all delay indices.

LHR during the COVID-19 pandemic outperforms all airports on the mean delay index by a large margin. Indeed focussing in on LHR, we see a clear difference between the time prior to the pandemic ($\mu_{\text{LHR}} \approx -5\,\text{min}$) and during the pandemic ($\mu_{\text{LHR COVID19}} \approx -25\,\text{min}$). The reason behind this is that the dramatic reduction of flight traffic worldwide saw many flights arriving too early. Interestingly, the left exponent, i.e. the decay of early arrivals, did not change substantially, compared to LHR under business-as-usual conditions since the shape of the delay distribution on the left did not change much but was only shifted to more negative values. The right flank behaves quite differently: Both business-as-usual and LHR during the COVID-19 pandemic, recorded relatively heavily delayed flights, which arrived more than 3 hours late (see also Fig. 1). The right index reveals the likelihood of these extreme events. In the case of LHR under COVID-19, the low mean delay suggests early arrival but relative extreme events are still present and hence the right exponent reveals this poor performance.

Notice that we cannot fully exclude a sampling bias of the airline analysis due to the different number of flights recorded for each airport: For a given airline, e.g. BA, we use all flights at all airports in our data set. However, since we recorded more total flights in LHR, the BA distribution is influenced more by the LHR data than by other airports.
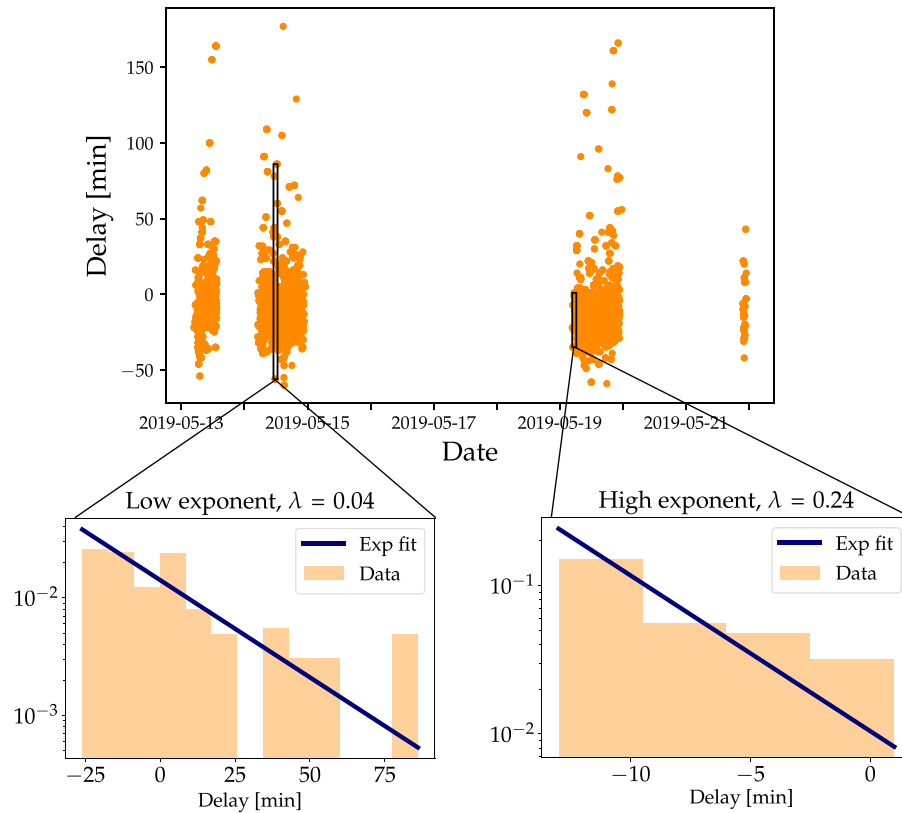
Figure 7. We analyse the full time series of plane delays and extract a time window during which we observe locally exponential distributions. These local distributions can decay slowly or fast, i.e. the rate λ is fluctuating.

Superstatistical modelling of delays

As we have seen previously, the right flank of the delay statistics exhibits heavy tails and is well-described by a q-exponential. Let us now explore a potential explanation for this particular distribution by employing the framework of superstatistics[15,27,28]. Superstatistics is relevant when an aggregated system (e.g. a long time series) displays heavy tails, but the system may then be disentangled into many smaller sub-parts (e.g. short time periods of the trajectory). These sub-parts then are no longer heavy-tailed but follow a simple local distribution, for example an exponential or a Gaussian. This idea has been successfully applied, for example, to train delays[16], electric power systems[17] and intermittent wind statistics[29].

Assuming for now that the right-flank delays are indeed q-exponentially distributed and follow a superstatistics, we should be able to observe "local" exponential densities, with a decay parameter λ. Superimposing all these λ, we get a q-exponential if the λ themselves follow a χ^2-distribution:

$$f(\lambda) = \frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{n}{2\lambda_0}\right)^{\frac{n}{2}}\lambda^{\frac{n}{2}-1}e^{-\frac{n\lambda}{2\lambda_0}}. \tag{3}$$

Here n denotes the number of degrees of freedom characterizing the fluctuations in λ and λ_0 is the sample mean of λ. Indeed, choosing an appropriate time scale to separate the trajectory (see next paragraph), the heavy tails of the delay distributions vanish and instead the distributions are well described by simple exponential functions, see Fig. 7.

Let us explain how to extract the relevant time scale T on which we locally observe exponential distributions. Since we know that an exponential distribution has a kurtosis of $\kappa_{\text{exponential}} = 9$, we test time windows of different size $\Delta\tau$ and compute the local average kurtosis[15] as

$$\bar{\kappa}(\Delta\tau) = \frac{1}{\tau_{\max} - \Delta\tau}\int_0^{\tau_{\max}-\Delta\tau} d\tau_0 \frac{\langle(u-\bar{u})^4\rangle_{\tau_0,\Delta\tau}}{\langle(u-\bar{u})^2\rangle^2_{\tau_0,\Delta\tau}}, \tag{4}$$

where $\tau_{\max}$ is the length of the time series u and $\bar{u}$ is the mean of the time series. We denote by $\langle\dots\rangle_{\tau_0,\Delta\tau}$ the expectation formed for a time slice of length $\Delta\tau$ starting at τ_0. For the LHR data, we compute the local kurtosis and thereby determine the long time scale: $\bar{\kappa}(T) = 9$, for $T \approx 1.55h$, see Fig. 8.

Next, let us carry out an important consistency check: As explained above, the mixing of numerous local exponential distributions with exponents following a χ^2-distribution leads to a q-exponential. Now, we can make a histogram of the λ-distribution and fit it with a χ^2- and an inverse χ^2-distribution. Then, we derive the
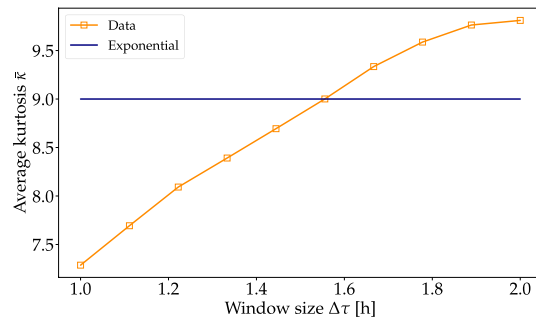
Figure 8. The average kurtosis $\bar{\kappa}$ of the data set is plotted as a function of the time window $\Delta\tau$ in hours (yellow). The intersection between the horizontal line at $\bar{\kappa} = 9$ (the kurtosis of an exponential distribution) and the $\bar{\kappa}$ vs Δt curve gives the optimal value for Δt; we find $T \approx 1.55$ hours.
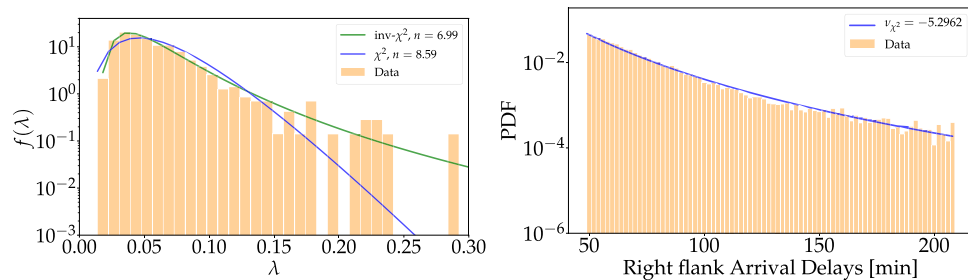

Figure 9. Applying superstatistics leads to consistent results. Left: We extract the distribution of local exponents and compare them to a χ^2 and inverse χ^2 fit (based on the method of least squares). Right: Using the previously derived χ^2 distribution, we again derive a q-exponential with right exponent $\nu_{\chi^2} \approx -5.296$, compared to the fitted one of $\nu_{MLE} \approx -5.371$. We note that the power-law decay of the data is well captured by the q-exponential induced by the χ^2-distribution. The blue curve is scaled to the same amplitude as the data for visual guidance.

q-exponential from the fitted χ^2-distribution and compare it with the direct fit of the q-exponential and the original data. This is illustrated in Fig. 9.

We note that the empirical λ-distribution is slightly better fitted by an inverse χ^2- than a χ^2-distribution, as also observed in other application areas[18,30]. Overall, the superstatistical description seems consistent, given the short time series of flight delays under consideration. The q-exponential derived from the χ^2 tends to overestimate the PDF at low values, which is understandable as we also exclude them for the fitting of the q-exponential via MLE (see "Methods"). Still, the tail behavior of the q-exponential based on the χ^2 matches the real data and the MLE fit nicely. This means the observed power laws of the right flanks are essentially explained by a suitable superstatistics which describes changes in the microvariables on a time scale of $T \approx 1.5$ hours.

Connecting the flanks

So far, we focused our attention on describing and fitting the tail aspects of the distribution, namely the left, approximately exponential, flank and the right, approximately q-exponential, flank. Both these functions combined overestimate the peak of the distribution and hence, we also included the mean delay as the final metric in our framework. Now, let us consider how the two tail distributions could be merged in one smooth-fitting function.

First, we note that the so far mostly ignored central part of the delay distribution can be approximated by a Gaussian distribution, based on the parabola shape in the log-scale plots. We use this insight to propose the following continuous fitting function

$$
p(t) = \begin{cases} A_e \exp\left(-\lambda\sqrt{C + (t - t_{\text{peak}})^2}\right), & t < t_{\text{peak}} \\ A_q \exp_q\left(-\lambda_q\sqrt{C + (t - t_{\text{peak}})^2}\right), & t \geq t_{\text{peak}} \end{cases} \tag{5}
$$

with $\exp_q(t) = (2 - q)\lambda_q\left[1 + (q - 1)\lambda_q t\right]^{\frac{1}{1-q}}$ being the q-exponential function. Here, A_e and A_q are amplitudes, C is a curvature parameter, describing the approximately Gaussian part in the center, t_{peak} is the delay at the peak of the delay distribution, where we split into left and right flanks and t is the delay value, see "Methods" for fitting details and code.

The resulting fit is a smooth function, covering the full delay range, see Fig. 10. Since the new curvature parameter C also influences the general shape, the new values for q and λ, now named $\tilde{q}$ and $\tilde{\lambda}$, are slightly different from the ones solely focusing on the tails (empirically we tend to observe a slight reduction in λ and
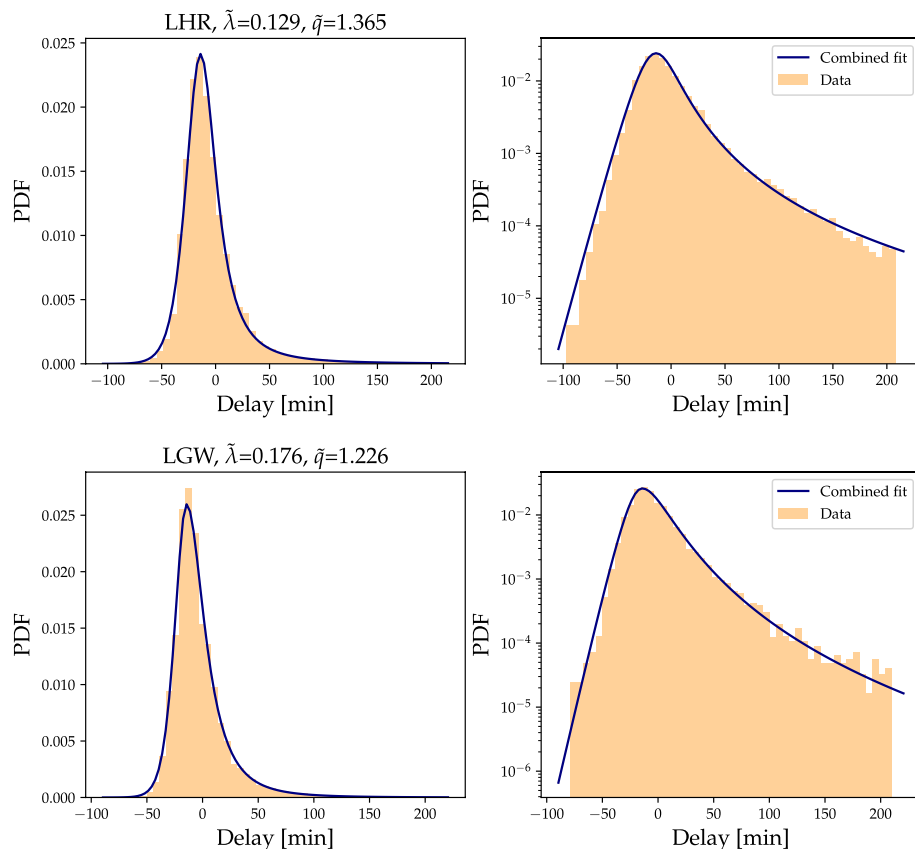
Figure 10. Using the approximately Gaussian shape in the center, we smoothly combine left and right flank fits into one coherent fit of the full delay data set. To emphasize the quality of the fit, we display both a linear (left) and logarithmic (right) scale of the PDF for LHR (top) and LGW (bottom), the two airports with the most flights in our data set.

increase in q). Still, the general observations using the delay indices and comparing airlines, such as in Figs. 4, 5 and 6, remain mostly unchanged. Equation (5) provides an alternative approach to the three delay indices introduced so far. If one is interested in describing the full distribution as accurately as possible, we recommend using Eq. (5). Meanwhile, to compare performance of individual airlines or to obtain a general impression of the delay distribution, the three delay indices are a simplified framework, allowing easy and robust estimation and comparison. Finally, note that the full curve is not strictly a probability density function as we did not enforce that its integral equals one. While theoretically making it easier by reducing the number of parameters, that would make the fitting more difficult in practice as the integrals cannot be evaluated analytically by hand and impose additional constraints during the fitting. Also note that our observed flight delays are constrained to the finite interval $[-100, 210]$, whereas the fitting function is defined on $[-\infty, \infty]$, which makes the normalization outside the interval ambiguous.

Discussion and conclusions

In summary, we have analysed a newly obtained data set of plane delays for various British airports, which contains tens of thousands of flights, aggregated over multiple months. We believe this is a substantial improvement on some earlier studies which, to the best of our knowledge, only investigated a few days of measurements and a couple of thousand flights, thereby greatly underestimating the contribution of the tails to the probability distribution[31]. Interestingly, we find that all investigated airports and even individual airlines at each airport follow a qualitatively similar distribution, namely an approximately exponential decay on the left flank (of negative delays) and a slowly decaying power law on the right flank (of positive delays). To characterize these distributions and systematically compare airlines and airports, we have developed a framework to quantify delay performance. Critically, we do not merely use the mean delay but also consider extreme events of both positive and negative delays via their respective flanks in the empirical probability distribution. Applying this newly developed framework, we find substantial differences between airlines serving short and long-distance routes.

We offer an explanation for the emerging power law on the right flank via superstatistics: The local q-exponential distribution with its heavy tails seems to arise from many superimposed exponential distributions. In particular, we identify the long time scale T as approximately 1.5 hours, during which delays fall off exponentially. Comparing to other superstatistical results[27,28], we note the relevance of both χ^2-distributions and inverse-χ^2

-distributions for the scale parameter, similar to the ones observed in air pollution or cancer[18,30], stressing again the universality of superstatistics. Finally, we propose a continuous function to capture the full delay statistics. While this introduces additional parameters and the superstatistical theory mentioned previously can no longer be used to rigorously derive the fitting function, this fit does describe the full distribution with high accuracy.

Our framework of three delay indices to characterize flight delay distributions can be applied quite generally to measure the punctuality of flights, going beyond an analysis based on just the mean. Crucially, while airlines or airports might be able to "game" the system of mean delays, this is not possible with the left and right exponents. Companies could shift their flight schedule, i.e. announce intentionally that flights will take longer than they do in practice, and thereby systematically record early arrivals so pushing their mean delay to negative values. However, such a procedure would still leave the remaining two indices (left and right exponent) untouched so that they provide a stable way of measuring performance.

One remarkable result is the impact of the global pandemic of COVID-19 on the delay statistics. Heathrow (LHR) under COVID-19 conditions (travel restrictions, quarantine upon arrival, etc) displays an impressively low mean delay, while the left flank decay was mostly unchanged. Interestingly, LHR still experienced some relatively heavily delayed flights during the COVID-19 pandemic, which leads to pronounced heavy tails towards the right and thereby a poor performance in the right exponent. These observations indicate that in different (COVID-19) situations and given fewer flights, airports can perform better in some aspects (e.g. mean delay) than under business-as-usual conditions, while other observables (extreme delays) can still be improved. Aside from the upsides of COVID-19-related lockdown measures on air quality[32,33] or CO_2 emissions[34], we find that having fewer flights also improves delay statistics.

We have assumed throughout this article that negative delays are preferred by all passengers. However, some passengers might value arrival at exactly the predicted time more highly than arriving early. This would change the interpretation of the left index slightly: Instead of desiring low exponents, airlines and airports should aim for high exponents. Similarly, the absolute value of the delay should be zero, i.e. arrival on time should be the default. Regardless of preference, the indices, as introduced, provide a sufficient framework to measure the delay performance.

In the future, we would like to apply our framework to delay statistics at other airports in different countries, and investigate how delays are related to geographical distance of the flights. In particular it would be interesting to see how our three indices differ between years, countries and so on. From a more fundamental perspective, we aim to further understand correlations in the flight delays. Preliminary indications from the British data are that on "typical" days correlations decay quickly but on some "exceptional" days (perhaps those where external factors affect many flights) the autocorrelation function can settle on a non-zero value for some time and many flights have long delays which contribute to the tail of the probability density function. Long-range temporal correlations and memory effects have been studied in many other physical and non-physical systems[35,36]; modelling such effects here is challenging, since the build-up of delays at one airport may be influenced by earlier flights to and from completely different airports, but practically important since controlling the "cascading" of delays would lead to a significantly improved passenger experience. In this way, future investigations could take into account spatio-temporal information from the entire worldwide air transportation network. More concretely, our data set could be expanded in type of information as well as volume. First, it would be interesting to also study departure delays, in addition to the arrival delays studied here. Furthermore, we could explicitly include flight duration and distance and investigate correlations between delays and flight distance/duration for many different airports in the world.

Methods

Data processing.
As we mentioned in the main text, for each flight, we recorded the airline company operating the flight, the flight number, the departure and arrival airports as well as the scheduled and actual landing times, as provided on the airport web page. The data was cleaned and organized according to the delay, computed as the difference between scheduled arrival time and actual arrival time for each flight. We kept data for each arrival airport as well as a summary of the overall delays, independent of the arrival airport. A "negative" delay occurs when the actual aircraft arrival is earlier than the expected one, according to the scheduled timetable. After examining the data it became evident that a reasonable cut-off point as to how early or late an aircraft can arrive at the designated airport should be implemented. This prevents over-representation of individual extreme events in the resulting probability distributions. We decided that the delays (in minutes) would have to be contained in the interval $[-100, 210]$.

Theoretical distribution fitting.
Here we explain the fitting procedure in more detail. We approximate the empirical distribution of the left flank, where negative delays are dominant, with an exponential distribution of the form

$$p(t_L; \lambda) = \lambda e^{-\lambda t_L}, \quad \lambda > 0. \tag{6}$$

As we have seen in the main text, the observed distribution curves towards a Gaussian distribution around the peak value and thereby deviates from an exponential distribution. Hence, we restrict our fitting to values deviating from the central area as follows. Let t_{peak} be the delay at which the distribution reaches its highest PDF value and t_{min} the smallest delay we observe. Then, we restrict our exponential fit to any delay falling in the interval $[t_{min}, t_{peak} - 0.3|t_{min} - t_{peak}|]$, where $|\ldots|$ indicates the absolute value. Following this restriction, we define the left flank delay values as

$$t_L = -t + t_{\text{peak}} - 0.3|t_{\min} - t_{\text{peak}}|, \quad t \in [t_{\min}, t_{\text{peak}} - 0.3|t_{\min} - t_{\text{peak}}|]. \tag{7}$$

We now turn to the right flank of the empirical distribution, i.e. the portion of the data set that constitutes the majority of the positive delays. The q-exponential is much better at incorporating parts of the Gaussian central distribution on the right-hand side than the exponential distribution is on the left flank. Hence, we only exclude the smallest 10% of the data, i.e. we consider delays t in the interval interval $[t_{\text{peak}} + 0.1|t_{\max} - t_{\text{peak}}|, t_{\max}]$, where $t_{\max}$ is the highest delay observed. Hence the right-flank delays to be fitted are defined as

$$t_R = t - t_{\text{peak}} - 0.1|t_{\max} - t_{\text{peak}}|, \quad t \in [t_{\text{peak}} + 0.1|t_{\max} - t_{\text{peak}}|, t_{\max}]. \tag{8}$$

Our theoretical distribution choice is now a q-exponential

$$p(t_R; q, \lambda_q) = (2 - q)\lambda_q \big[1 + (q - 1)\lambda_q t_R\big]^{\frac{1}{1-q}}, \tag{9}$$

with parameters λ_q and q. It has been shown that q-exponentials and q-Gaussians arise from maximizing Tsallis entropy[25].

Note that both t_L and t_R are defined such that they start at 0 and continue towards positive values to keep the fitting functions easier.

These two functions (exponential and q-exponential) are fitted to the data using a maximum likelihood estimate (MLE), i.e. maximizing the Likelihood $L(\boldsymbol{\theta}, \mathbf{x})$. Here, $\mathbf{x}$ indicates the data we wish to fit and $\boldsymbol{\theta}$ the set of parameters that are being optimized. The likelihood of a parameter setting $\boldsymbol{\theta}$ on a given one-dimensional data set $\mathbf{x} = x_1, x_2, \ldots, x_N$ is computed as

$$L(\boldsymbol{\theta}, \mathbf{x}) = \prod_{i=1}^{N} p(x_i, \boldsymbol{\theta}), \tag{10}$$

with probability density function $p(x_i, \boldsymbol{\theta})$, dependent on the parameters $\boldsymbol{\theta}$. Technically, we carry out the MLE using the *scipy.stats* module in python with custom PDFs, see also Code availability (below) for a link to the code.

Fitting the smooth combined function.   To obtain a smooth fit, combining both flanks, we employ the following procedure. We first estimate the exponential decay rate λ based on the lowest 70% of negative delays, then estimate q and the q-exponential decay rate λ_q based on almost the full right-hand side of the histogram. This is identical to the procedure for the individual flanking fits. Next, we estimate the central curvature C, which we assume to be identical for both intervals, and the amplitudes A_e and A_q, as well as λ_q using least squares fitting. While carrying out this least-square fit, we also allow the parameters q and λ to vary slightly from the MLE-optimal value determined earlier, while all other parameters are not bounded. The reason to allow any variance is to ensure a continuous fit while keeping the change from the optimal MLE parameters small. Empirically, we find that restricting $0.95\, q_{\text{MLE}} \le \tilde{q} \le 1.15\, q_{\text{MLE}}$ and $0.95\, \lambda_{\text{MLE}} \le \tilde{\lambda} \le 1.05\, \lambda_{\text{MLE}}$ yields the best results. Technically, we use the *scipy.stats* module to perform the MLE fits and the least-square fit; continuity is ensured using constraints in the *symfit* package.

Airline data.   In Figs. 4 and 5 we compared several airlines. Let us briefly list how many flights we analysed to derive our delay indices: For the short-distance airlines "Wizz Air": 2428, "easyJet": 15449, "Ryanair": 13488, "Vueling": 1034, "Jet2": 1215; for the other airlines we have "Iberia": 12892, "British Airways": 38257, "Aer Lingus": 7331, "Finnair": 8560, "American Airlines": 23119, "Air Canada": 7247, "United Airlines": 6797, "Japan Airlines": 5966, "Qatar Airways": 5935. For all airlines we have at least 1000 flights and often several thousand flights.

Data availability
The original data of airport arrivals has been uploaded to an open repository: https://osf.io/snav9. All data that support the results presented in the figures of this study are available from the authors upon reasonable request.

Code availability
Python code to reproduce figures, perform the fits and extract the delay indices, is also uploaded here: https://osf.io/snav9/.

References
1. Hakim, M. M. & Merkert, R. The causal relationship between air transport and economic growth: Empirical evidence from South Asia. *J. Transp. Geogr.* **56**, 120 (2016).
2. Brida, J. G., Monterubbianesi, P. D. & Zapata-Aguirre, S. Exploring causality between economic growth and air transport demand for Argentina and Uruguay. *World Rev. Intermodal Transp. Res.* **7**, 310 (2018).
3. Suau-Sanchez, P., Voltes-Dorta, A. & Cugueró-Escofet, N. An early assessment of the impact of COVID-19 on air transport: Just another crisis or the end of aviation as we know it?. *J. Transp. Geogr.* **86**, 102749 (2020).
4. Kuhn, H., Falter, C. & Sizmann, A. Renewable energy perspectives for aviation, in *Proceedings of the 3rd CEAS Air and Space Conference and 21st AIDAA Congress* 1249–1259, Venice (2011).
5. Efthymiou, M., Njoya, E. T., Lo, P. L., Papatheodorou, A. & Randall, D. The impact of delays on customer's satisfaction: An empirical analysis of the British Airways on-time performance at Heathrow airport. *J. Aerosp. Technol. Manag.* **11**, 1–13 (2019).

6. Rebollo, J. J. & Balakrishnan, H. Characterization and prediction of air traffic delays. *Transp. Res. C Emerg. Technol.* **44**, 231 (2014).
7. Rosenberger, J. M. *et al.* A stochastic model of airline operations. *Transp. Sci.* **36**, 357 (2002).
8. Mueller, E. & Chatterji, G. Analysis of aircraft arrival and departure delay characteristics, in *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, 5866 (2002).
9. Gui, G. *et al.* Flight delay prediction based on aviation big data and machine learning. *IEEE Trans. Vehic. Technol.* **69**, 140 (2019).
10. Li, M. Z. & Ryerson, M. S. Reviewing the datas of aviation research data: Diversity, availability, tractability, applicability, and sources. *J. Air Transp. Manag.* **75**, 111 (2019).
11. Fleurquin, P., Ramasco, J. J. & Eguiluz, V. M. Systemic delay propagation in the US airport network. *Sci. Rep.* **3**, 1159 (2013).
12. Pyrgiotis, N., Malone, K. M. & Odoni, A. Modelling delay propagation within an airport network. *Transp. Res. C Emerg. Technol.* **27**, 60 (2013).
13. Verma, T., Araújo, N. A. & Herrmann, H. J. Revealing the structure of the world airline network. *Sci. Rep.* **4**, 1 (2014).
14. Guimera, R., Mossa, S., Turtschi, A. & Amaral, L. N. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci.* **102**, 7794 (2005).
15. Beck, C., Cohen, E. G. D. & Swinney, H. L. From time series to superstatistics. *Phys. Rev. E* **72**, 056133 (2005).
16. Briggs, K. & Beck, C. Modelling train delays with q-exponential functions. *Phys. A Stat. Mech. Appl.* **378**, 498 (2007).
17. Schäfer, B., Beck, C., Aihara, K., Witthaut, D. & Timme, M. Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics. *Nat. Energy* **3**, 119 (2018).
18. Williams, G., Schäfer, B. & Beck, C. Superstatistical approach to air pollution statistics. *Phys. Rev. Res.* **2**, 013019 (2020).
19. Metzler, R. Superstatistics and non-Gaussian diffusion. *Eur. Phys. J. Spec. Top.* **229**, 711 (2020).
20. Chubynsky, M. V. & Slater, G. W. Diffusing diffusivity: A model for anomalous, yet Brownian, diffusion. *Phys. Rev. Lett.* **113**, 098302 (2014).
21. Itto, Y. & Beck, C. Superstatistical modelling of protein diffusion dynamics in bacteria. *J. R. Soc. Interface* **18**, 20200927 (2021).
22. UK Civil Aviation Authority, UK Airports—Annual Statements of Movements, Passengers and Cargo—Table 09 (2019) [Online; Accessed 10 Sept 2020].
23. Makortoff, K. Heathrow cargo flights rise 500% as airport restyles itself as "vital airbridge" (2020) [Online; Accessed 19 Aug 2020].
24. Nižetić, S. Impact of coronavirus (COVID-19) pandemic on air transport mobility, energy, and environment: A case study. *Int. J. Energy Res.* **44**, 10953 (2020).
25. Tsallis, C. Possible generalization of Boltzmann–Gibbs statistics. *J. Stat. Phys.* **52**, 479 (1988).
26. EUROCONTROL, Delays—three questions and many answers (2018) [Online; Accessed 10 Sept 2020].
27. Beck, C. Dynamical foundations of nonextensive statistical mechanics. *Phys. Rev. Lett.* **87**, 180601 (2001).
28. Beck, C. & Cohen, E. G. D. Superstatistics. *Phys. A Stat. Mech. Appl.* **322**, 267 (2003).
29. Weber, J. *et al.* Wind power persistence characterized by superstatistics. *Sci. Rep.* **9**, 1 (2019).
30. Chen, L. L. & Beck, C. A superstatistical model of metastasis and cancer survival. *Phys. A Stat. Mech. Appl.* **387**, 3162 (2008).
31. Caccavale, M. V., Iovanella, A., Lancia, C., Lulli, G. & Scoppola, B. A model of inbound air traffic: The application to Heathrow airport. *J. Air Transp. Manag.* **34**, 116 (2014).
32. Shrestha, A. M. *et al.* Lockdown caused by COVID-19 pandemic reduces air pollution in cities worldwide, EarthArxiv (2020).
33. Schäfer, B. *et al.* Covid-19 impact on air quality in megacities, arXiv preprint. arXiv:2007.00755 (2020).
34. Le Quéré, C. *et al.* Temporary reduction in daily global CO_2 emissions during the COVID-19 forced confinement. *Nat. Clim. Change* **10**, 647–653 (2020).
35. Rangarajan, G. & Ding, M. (eds) *Processes with Long-Range Correlations: Theory and Applications, Lecture Notes in Physics* Vol. 621 (Springer, 2003).
36. Beran, J., Feng, Y., Ghosh, S. & Kulik, R. *Long-Memory Processes: Probabilistic Properties and Statistical Methods* (Springer, 2013).

Acknowledgements

Author contributions

E.M., B.S., and C.B. conceived and designed the research. E.M. collected the data, E.M. and B.S. analysed the data and produced the figures. R.J.H. and all other authors contributed to discussing and interpreting the results and writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.