scientific reports

Check for updates

OPEN

Spatial heterogeneity of air pollution statistics in Europe

Hankun He[1], Benjamin Schäfer[1,2,3]✉ & Christian Beck[1,4]

Air pollution is one of the leading causes of death globally, and continues to have a detrimental effect on our health. In light of these impacts, an extensive range of statistical modelling approaches has been devised in order to better understand air pollution statistics. However, the time-varying statistics of different types of air pollutants are far from being fully understood. The observed probability density functions (PDFs) of concentrations depend very much on the spatial location and on the pollutant substance. In this paper, we analyse a large variety of data from 3544 different European monitoring sites and show that the PDFs of nitric oxide (NO), nitrogen dioxide (NO_2) and particulate matter (PM_{10} and $PM_{2.5}$) concentrations generically exhibit heavy tails and are asymptotically well approximated by q-exponential distributions with a given width parameter λ. We observe that the power-law parameter q and the width parameter λ vary widely for the different spatial locations. For each substance, we find different patterns of parameter clouds in the (q, λ) plane. These depend on the type of pollutants and on the environmental characteristics (urban/suburban/rural/traffic/industrial/background). This means the effective statistical physics description of air pollution exhibits a strong degree of spatial heterogeneity.

Air pollution is among the highest contributors to premature death and disease worldwide, causing a significant number of deaths from stroke, lung cancer and heart diseases[1,2]. Besides human health, air pollution affects vegetation, natural ecosystems, climate change, the built environment and subsequently the economy[3]. In Europe, air pollution is the single largest environmental health risk[4–6] and its long-term effects are very serious. Although emissions and ambient concentrations have fallen steadily in Europe over the past few decades, as stated in Refs.[7,8], many European countries still exceed European Union's (EU) standards[7] and World Health Organization's (WHO) guidelines, see[9] for the levels of air pollutants in 2018. Two key air pollutants, namely, particulate matter (PM) and nitrogen oxides (NO_x), pose a considerable threat to the health of citizens. About 55,000 and 417,000 premature deaths in 41 European countries in 2018 were attributed to NO_2 and $PM_{2.5}$, respectively[3]. Pollutants such as ozone (O_3), sulfur dioxide (SO_2) and carbon monoxide (CO) negatively affect human health as well. Particularly, ground level O_3 has been estimated to have caused 20,600 premature deaths in Europe in 2018; this yearly number has risen by 20% since 2009[3]. In this paper we focus on two of the most dangerous pollutants, NO_x and PM, but the methodologies presented in our paper can be similarly applied to other substances.

The impact of air pollution on health does not only depend on the pollutant type but also on the type of surrounding environment, i.e. people living next to traffic-heavy roads or industries face higher exposure to air pollution. The EU[10] uses environmental surrounding types to classify air quality monitoring sites into traffic, industrial, background, urban, suburban and rural, based on predominant emission sources and building density. From a policy perspective, this allows for evaluating the effectiveness of measures targeting specific emissions sectors and assessing the impact of those associated pollutants which dominate the area surrounding a given monitoring station, such as for example traffic or industry (or their absence). Despite progress made by EU and UK policies[7,8] addressing all sectors to reduce emissions and protect citizens from pollutants, meeting the emission reduction commitments by 2030 remains a challenge[11]. Similar challenges to satisfy given policies are also existing in other parts of the world.

Having a thorough understanding of the time-varying statistics, i.e. of the entire probability density function (PDF) of air pollution, is crucial for policymakers involved in defining thresholds or reducing overall exposure to air pollution. It is also crucial for the construction of suitable statistical physics models. PDFs such as gamma, log-normal and Weibull distributions[12] have been widely used for fitting air pollutant concentration data. However, these distributions decay approximately like exponential functions at large values, while earlier investigations have found heavy tails in air pollution statistics[13], which are not well-described by the above

[1]School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK. [2]Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, 76344 Eggenstein-Leopoldshafen, Germany. [3]Faculty of Science and Technology, Norwegian University of Life Sciences, 1432 Ås, Norway. [4]The Alan Turing Institute, London NW1 2DB, UK. ✉email: benjamin.schaefer@kit.edu

distributions. Some recent studies[14–16] have explored the COVID-19 lockdown effects on air quality (in Europe and in megacities such as Delhi), focusing on comparing the PDFs or given moments of the PDFs before and during the lockdown. Superstatistical methods, originating from turbulence modelling[17] and applied to many fields[18,19], offer a powerful effective approach to describe the dynamics of air pollution assuming the existence of well-separated time scales[13]. Air pollutants such as NO_x have been dealt with success in the superstatistical approach, taking into account nonequilibrium situations with fluctuating variance parameters[13]. However, this approach has been verified for limited data sets only, chosen from the UK (London), and also only for a limited set of pollutants, mainly NO and NO_2. On a European scale, and for much larger data sets, it remains unclear whether heavy tails are generically observed and whether an effective superstatistical description is applicable. This is the topic of this paper.

The above consideration leads us to a problem that is of general interest for statistical physics approaches to environmental science. Can we apply standard methods from nonequilibrium statistical physics, such as the above-mentioned superstatistical methods, to environmentally relevant time series of pollution concentrations, and if yes, how can we extract the corresponding superstatistical parameters from the time series? And how spatially heterogeneous are the observed results? What are the values of the relevant superstatistical parameters for different air pollutants and different geographical environments? Furthermore, what are typical distributions of observed PDF fitting parameters for the large number of sites distributed across the European continent? These important types of questions relating to large ensembles of different measuring stations will be dealt with in the following.

The paper is organized as follows. First, we introduce our large data set involving 3544 measuring sites. Next, we investigate the relation between mean and standard deviation for the observed PDFs to clarify if the PDFs can be approximated by simple exponential distributions or if more complicated functions are needed. We then systematically investigate the PDFs of all sites, in particular the tail behaviour, and show that the tails are generally much better described by q-exponential functions with a given width parameter λ than by functions such as exponential and log-normal, meaning there is generically power-law decay.

Subsequently, we use the maximum likelihood estimation method (MLE) to extract the q and λ parameters for the best-fitting q-exponential distribution, and present plots of scattered points in the (q, λ) plane which exhibit interesting patterns for our large number of spatial locations investigated. Our main result is that air pollution statistics is extremely heterogeneous, with the local variations of best-fitting parameters spanning many orders of magnitudes. Our investigation is the first one that investigated this in a systematic way for very large ensembles of different measuring stations. We show that there is a complex pattern structure in the 2-dimensional (q, λ) parameter space that depends both on the pollutant type as well as on the classification type of the local surroundings.

Results

The data set considered. In this paper, we aim to conduct a large-scale statistical analysis of air pollutants, typically NO_x and PM, on a European scale. Technically, we access our air quality monitoring data from a large number of locations in Europe through the interface "Saqgetr", which is an R package available on the Comprehensive R Archive Network (CRAN)[20]. The vast majority of the accessible data are openly available from the European Commission's Airbase and air quality e-reporting (AQER) repositories[21,22]. To utilise the data efficiently, they have been processed into a harmonised form with consistent and careful treatment of the observations and metadata by Stuart K. Grange[20]. Concentration level readings and site environment types are the two key quantities we investigate. We import 9698 locations data throughout Europe within the time span of January 2017 to December 2021, recorded at 1-h intervals. To minimise the influences of seasonal fluctuations in time series, we eliminate sites whose data are too short, typically less than 1 year. We also exclude sites where a high percentage of measurements falls below the detection limit, since sites with clean air are not our primary analysis goal. Furthermore, we filter out sites whose data are corrupted, the used code and further details are described in the "Methods" section. We arrive at 3544 sites with data that meet our criteria before we proceed with our statistical analysis. Each data set contains at least 8760 (24×365) data points, up to about 43, 800 $(5 \times 24 \times 365)$ if the full 5 year period is available.

To provide a general overview of our analysed data, we show all the data sites' locations in Fig. 1a, as well as an example time series of a selected site: Bahnhofstrasse, Weiz in Austria for illustration purposes. Measured concentration time series and histograms are shown in Fig. 1b–e, for NO, NO_2, PM_{10}, and $PM_{2.5}$. NO_x and PM show seasonal cycles, i.e. during winter higher pollutant concentrations are more common. We also observe that for this example site the probability density of NO decays at a slower rate to zero than those of the other three pollutants. Apparently, typical distributions exhibit some heavy tails, which we will analyze in much more detail in the following.

Instead of considering the full distribution of each site's pollutants, in the following we will concentrate onto the tails. One reason for doing so is that we are particularly interested in the statistics of high pollution states, which are most damaging and described by the tails of the distribution.

With data sets from 3544 air pollution monitoring sites we require an efficient and context-based automated approach of analysing the sites, details are described in the "Methods" section. References[7,10] give more details on macro- and micro-scale sampling. In brief, stations are divided into three categories: traffic, industrial, and background based on predominant emission sources; the surrounding areas are classified as urban, suburban, or rural based on the density/distribution of buildings. Station types are combined with area types to provide an overall station classification, and we analyse our data conditioned on this station classification. We use the following definitions: "urban traffic": a site located in close proximity to a busy road in a continuously built-up urban area; "suburban/rural industrial": a site whose pollution level is influenced predominantly by emissions from an industrial area or an industrial source in largely built-up or remote areas; "rural background": a site

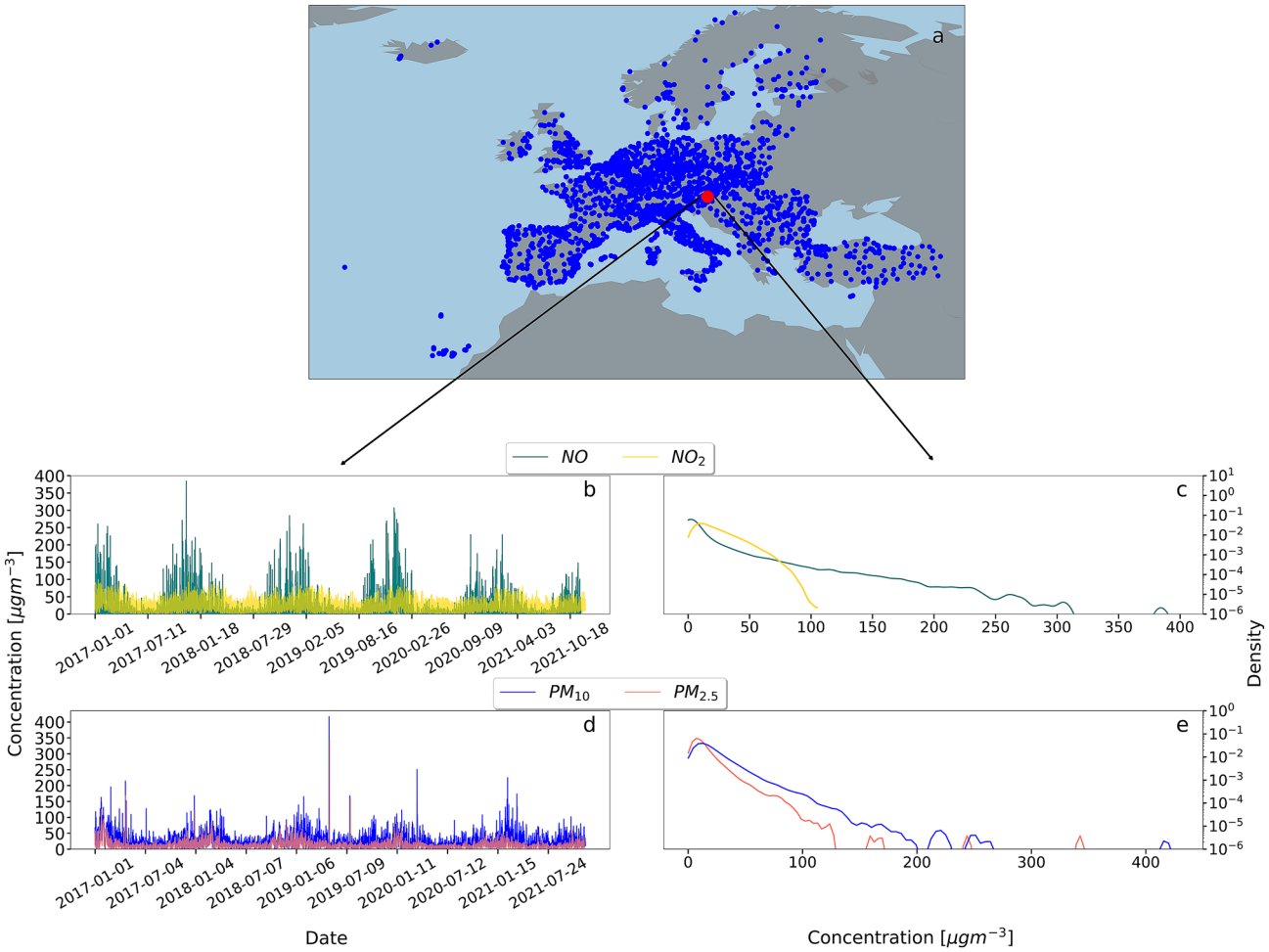

Figure 1. (**a**) Illustration of the available data sites on a map of Europe, with the red circle labeling our example site: Bahnhofstrasse, Weiz, Austria. Measured time series are shown in (**b**) and (**d**), the corresponding probability densities in (**c**) and (**e**). All pollutants display clear seasonality in their trajectories. Maps were created using Python 3 and geoplots.

whose pollution level is influenced by the combined contribution from all sources upwind of the station and not in built-up areas. Eventually, seven environmental area types "urban traffic", "suburban/rural traffic", "urban background", "suburban background", "rural background", "urban industrial" and "suburban/rural industrial" are used in our statistical analysis.

Checking exponentiality. Let us first consider the simplest possible hypothesis, namely that pollution concentrations follow an exponential distribution. In this case the PDF is given by

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}. \tag{1}$$

For exponential distributions one has the general fact that

$$\text{mean} \;=\; \text{standard deviation} \;=\; \frac{1}{\lambda}. \tag{2}$$

Thus, for each of our 3455 measuring stations we can easily test the hypothesis of an exponential distribution by plotting mean versus standard deviation for the measured data. If pollutants were to follow an exponential distribution, we would expect a clustering along the diagonal in such a plot. Stations with larger λ (smaller mean and variance) would correspond to cleaner air, and are expected to be found closer to the origin (0, 0) as compared to highly polluted locations. Our results are shown in Fig. 2.

The majority of the points are clustered above (b–d) or below (a) the diagonal lines, indicating deviations from an exponential distribution. These deviation patterns are different for each of the four substances.

Apparently, the PDFs of NO and NO_2 are very different, as the points scatter mainly below (NO) and mainly above (NO_2) the diagonal. The scattering plots for $PM_{2.5}$ and PM_{10} are more centered around the diagonal, but there are some unusual $PM_{2.5}$ states with large standard deviation and low means.

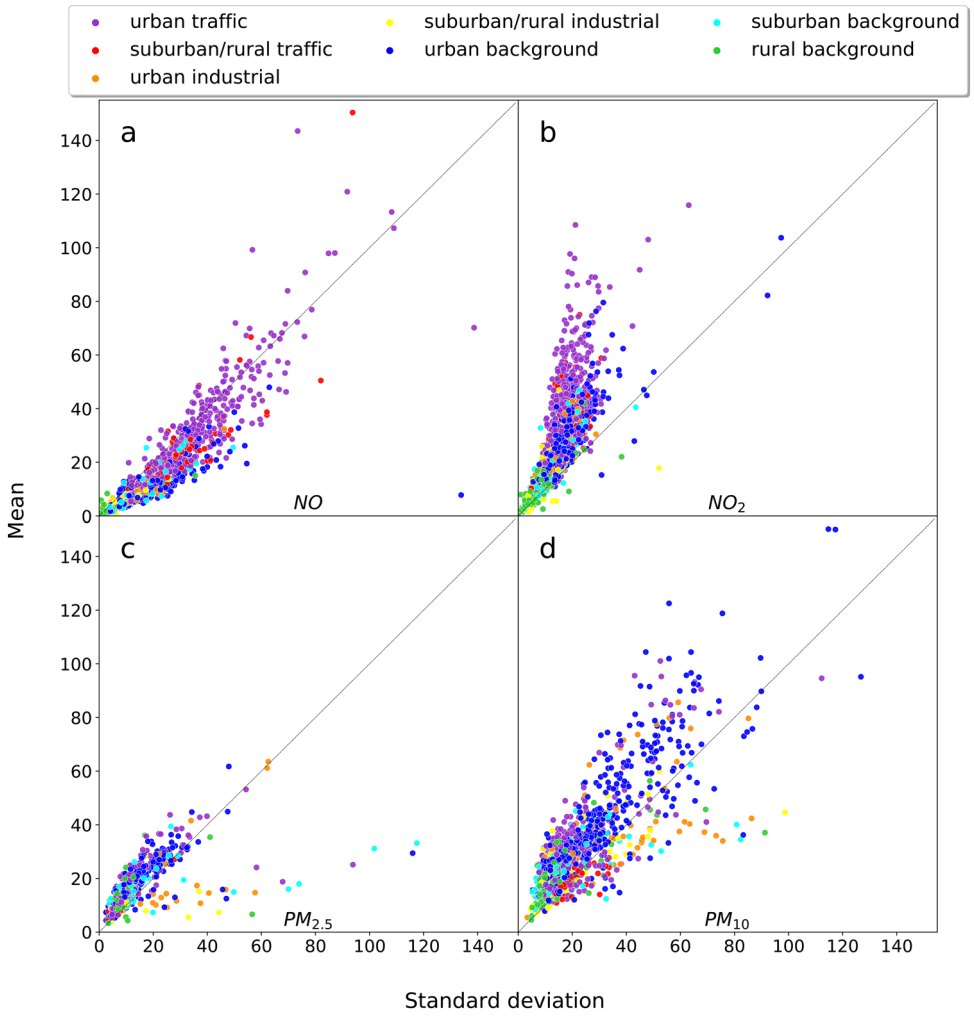

Figure 2. For each of the pollutants and measuring stations, we plot mean versus standard deviation. The area type surrounding the measuring station is color-coded. Data do not follow an exponential distribution, as evidenced by the fact that the majority of dots do not fall onto the diagonal lines. Different patterns are observed for the four different substances *NO*, NO_2, $PM_{2.5}$, PM_{10}. Green colors (rural stations with clean air) cluster near the origin. However, clustering of the same color patches is observed to be stronger for NO_x as compared to *PM*.

Generally, connected clusters of the same colors are more pronounced for *NO* and NO_2 as compared to *PM*. This could be attributed to the long-range transport of the *PM*-particles by moving air. Patterns and spatiotemporal scales for *PM* have been extensively discussed previously in[23–26]. The transport depends on weather conditions and removes the memory to the site where the particles were originally produced. Thus the colors are more mixed in our plots. As the weather patterns and the local meteorological features each contribute to the transport of *PM*-particles, the measuring site type has limited impact on the observed PDFs of *PM*. These observations motivate the usage of other statistical fitting functions explored in the next sections.

Fitting power-law tails for the data. As exponential tails apparently do not fit the data well, as illustrated by the deviations from the diagonal in Fig. 2, we now propose a different fitting function, motivated by many previous investigations in generalized versions of statistical mechanics[27]. This is a fitting by a so-called q-exponential, which asymptotically decays with a power-law exponent $-\frac{1}{q-1}$. q-exponentials better describe the high concentration tails of our data than other possible candidate distributions, see the detailed demonstration in the "Methods" section. The normalized PDF is defined as follows:

$$f_{q,\lambda}(x) = (2-q)\lambda[1-\lambda(1-q)x]^{\frac{1}{1-q}} \text{ for } 1-\lambda(1-q)x \geq 0, x > 0, \tag{3}$$

where q is the entropic index[27–29], λ is a positive width parameter and x, in our case, denotes the air pollutant concentration. Equation (3) contains the exponential distribution as a special case, namely for $q = 1$, as the q-exponential function, defined as $e_q(x) = [1+(1-q)x]^{\frac{1}{1-q}}$, converges to the exponential function in the limit $q \to 1$. For $q < 1$, $f_{q,\lambda}(x)$ lives on a finite support and becomes exactly zero above a critical value x, since,

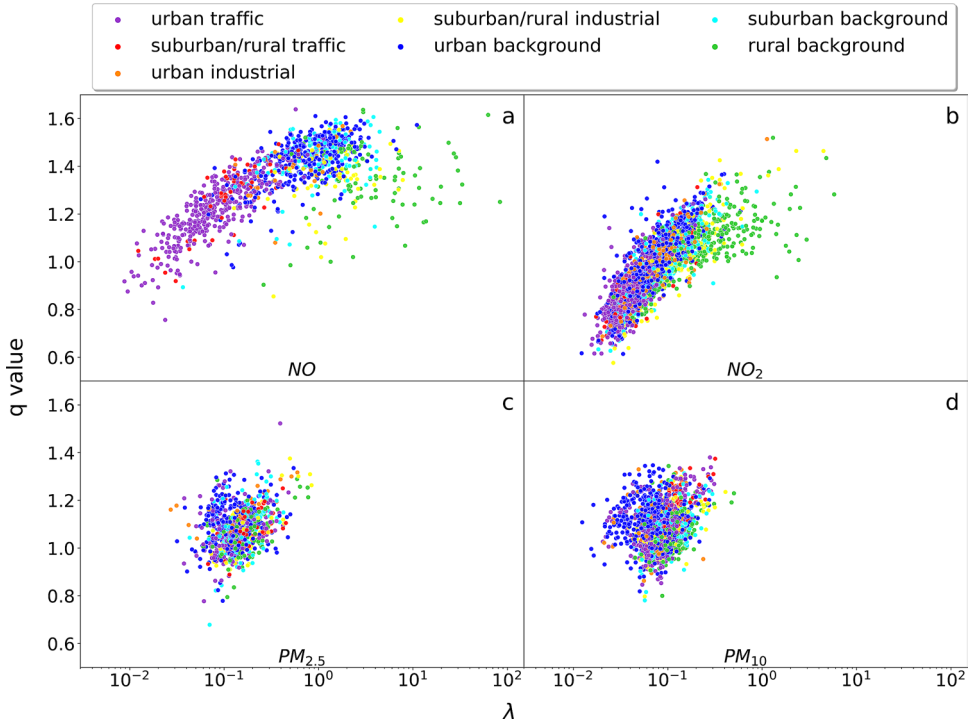

Figure 3. Best-fitting parameters of q-exponentials. We observe an increasing trend of q versus log λ for NO (**a**) and NO_2 (**b**), whereas a more disk-shaped pattern is observed for $PM_{2.5}$ (**c**) and PM_{10} (**d**). The environmental characterizations of the measuring stations are again encoded by colors. Again we observe correlated patches of a single given color for NO and NO_2, where for $PM_{2.5}$ and PM_{10} the pattern is more mixed.

by definition, $e_q(x) = 0$ for $1 - \lambda(1-q)x < 0$. In contrast, if $q > 1, 1 - \lambda(1-q)x > 0$, then Eq. (3) exhibits power-law asymptotic behavior.

The occurrence of q-exponentials with $q > 1$ in PDFs of complex systems is very well-motivated by superstatistical models[30,31]. In these types of models, one assumes a temporally fluctuating parameter λ for local exponential distributions as given in Eq. (1). These fluctuations of λ take place on a long time scale, much longer than local air pollution concentration fluctuations. The marginal distribution, obtained by integration over all possible values of λ, and describing the long-term behaviour of the air pollution concentration dynamics, is then a q-exponential, with

$$q = \frac{\langle \lambda^2 \rangle}{\langle \lambda \rangle^2}. \tag{4}$$

Here $\langle \cdots \rangle$ denotes the expectation with respect to the PDF of λ, see[30] for more details. Strictly speaking, a q-exponential is only obtained exactly if λ is Γ-distributed, but the general idea of superstatistics is that a parameter q can be defined by Eq. (4) for more general distributions different than the Γ distribution as well. The concept that wind changes and other effects (such as traffic fluctuations) can lead to a superstatistical dynamics for pollution concentrations was first worked out in[13], where further details can be found, in that case for the special example of NO and NO_2 concentrations as measured in London. Our investigation here is much more general, as we include data of thousands of measuring stations, and also investigate $PM_{2.5}$ and PM_{10} concentrations.

For all our 3544 measuring stations we extract histograms of the pollution concentration from the measured time series, and determine the best-fitting parameters q and λ for the given data set. More details on the numerical procedure are described in the Method section. Our results are shown in Fig. 3.

A truly surprising result of our analysis is the fact that we observe an immensely large range of values of the parameter λ for the best-fitting q-exponential as given in Eq. (3) for the various measuring stations. Note the logarithmic scale of the plots, the parameter λ can take on values as small as 10^{-2} up to values as large as 10^2, which spans four orders of magnitude. Typical q-values are in the range 0.8–1.4, but there are subtle differences between the various substances, with NO reaching large q-values such as 1.6, and NO_2 reaching small q-values such as 0.6 in the scattering plots. Also the shape of the scattering cloud of points is different for the different substances. For example, the typical range of λ for $PM_{2.5}$ and PM_{10} varies only by a factor 10, whereas for NO and NO_2 it varies by a factor 10^3. The scattering plot data look more spherically symmetric for $PM_{2.5}$ and PM_{10}, as compared to NO and NO_2.

Figure 3a,b, indicate a roughly linear approximate relationship between q and log λ for NO_x, with most of the points corresponding to traffic clustered at the left, while urban/suburban background points are in the middle part, and rural background points are scattered widely at the right. The PDF decay rate increases as λ increases

from highly polluted urban traffic sites to less polluted rural background areas. For $PM_{2.5}$ and PM_{10}, there is a different weak uphill relationship between q values and λ, as can be seen in Fig. 3c,d. The urban background points are reaching small λ values such as 0.01 for PM_{10}, and suburban/rural traffic, suburban/rural industrial and rural background points cluster on the right hand side with large λ. The attained range of λ values is smaller as compared to the case of NO_x, and the shape of possible values (q, λ) as displayed in the Figure is more spherical. The colors appear to be more randomly mixed.

The stronger colour mixing for $PM_{2.5}$ and PM_{10} can again be interpreted by the fact that by air movement transport the distributions cannot be uniquely identified with the original environmental types where the *PM*-particles were produced. The large scattering of parameters (q, λ) shows that for a given substance at a given environmental type there is not just one possible distribution, but a large range of possible distributions. These distributions may also vary in time, according to the weather conditions. Still, our scattering plots suggest that there is a *typical* range of parameter values for a given environmental type (e.g. lower λ and thereby higher mean values at urban traffic sites). The fact that there is a broad distribution of parameters is very much in line with the basic modelling assumption of superstatistics, in this case however applied to a spatial ensemble of different locations. There is a strong heterogeneity in space, meaning different spatial measuring locations have quite different PDFs. This spatial heterogeneity is a second effect, adding to the temporal heterogeneity of local exponentials, which effectively leads to q-exponentials at individual locations as explained above.

Spatial distribution plots of λ-values. Finally, we are interested in the PDFs of λ values for our fits of the various classified locations where the measurements are taken. We compare the summary statistics (such as distribution, range and quartiles) of λ for the four air pollutants with the aid of so-called violin plots, see Fig. 4. Within these, we visualise the distribution of λ using density curves, which correspond to the approximate frequency of data.

An interesting result of the violin plots shown in Fig. 4 is that the λ distributions extend to very large values, as indicated by the long extensions to the right for area types such as urban background, suburban/rural industrial and rural background. Additionally, the probability distributions of λ (represented by the "shapes" of the violins) exhibit nontrivial behaviour for some of the environmental site types. For example, for *NO* urban industrial sites there is a rather unusual pattern with several local maxima and minima. A much more generic pattern, with just a single broad maximum, is observed for sites which are suburban/rural industrial, as well as for those with a rural background, and this structure is there for all 4 different types of pollutants.

Another intriguing result is that the order of the median rankings (from low to high) for *NO* and NO_2 are almost the same, with the exception of a swap between urban industrial and urban background. For $PM_{2.5}$ and PM_{10} there are more swaps. Figure 4c,d, show that for $PM_{2.5}$ and PM_{10} the typical values of λ are smaller, below 0.9. A smaller λ indicates a more heavily polluted site. Furthermore, we observe in the case of *PM* only minor differences in the medians of λ for different environment categories. The reason for this is that the type of environment has a direct effect on NO_x concentrations while they have only a minor effect on *PM* since the particles travel and lose the memory of their environmental category. Nevertheless, suburban/rural traffic and rural background sites have the largest and second largest medians for λ, respectively.

As the environmental type did not emerge as a major factor impacting the distribution of *PM*, we also consider weather conditions, specifically wind speed, as a way to categorize and explain differences in *PM* distributions. National Oceanic and Atmospheric Administration (NOAA) Integrated Surface Database (ISD)[32] offers detailed surface meteorological data for over 35,000 locations across the globe. The worldmet[33] R package allows us to import hourly wind speed data. Each of the 3455 sites analysed is joined with its closest wind speed data for pollutant concentration measurement taken at the same time interval. See the Method section for the detailed processing steps. This data fusion allows for further classification of sites based on mean wind speed for which we utilize Beaufort's wind force scale[34], i.e. calm &light air/light breeze/gentle breeze/moderate breeze. Excluding sites with insufficient data, we obtain 1364 sites classified by wind force scale.

Taking wind speed into account, we conclude that differences in *PM* distributions are, at least partially, explained by different wind speeds, see Fig. 5. The violin plots for NO_2, $PM_{2.5}$ and PM_{10} all show a similar pattern: High wind speeds correspond to higher λs, which indicates a higher decay rate at heavy tails and, therefore, less pollution. This is coherent to the fact that stronger winds translate to a greater dispersion of particles. *NO*, by contrast, does not display a clear dependence on wind speed.

Discussion and conclusion

The use of q-exponentials for air pollution statistics has been previously advocated by Williams et al.[13], however that study was special since it was only looking at locations in Greater London, and the substances investigated were just *NO* and NO_2. In this paper, we have extended the statistical analysis to a much larger database, taking into account data from 3544 measuring stations, and analysing the statistics of $PM_{2.5}$ and PM_{10}, in addition to *NO* and NO_2. Naturally, for this vast amount of data novel methods needed to be developed for the fitting procedures, and novel graphical representations (scattering plots of parameter tuples) were used to illustrate the spatial heterogeneity of the results. Our main findings can be summarized as follows:

Firstly, we have clear evidence that generically PDFs of pollution concentrations do not decay in an exponential way. A much better fit is given by q-exponentials, which asymtotically decay as a power-law if $q > 1$, with exponent $-1/(q-1)$. Our analysis complements previous work, in which often different distributions have been used, including log-normal, gamma and Weibull distributions. Overall, we find that q-exponentials yield a better fit of the tails. The q-exponential is also a very plausible physical model, since—in the spirit of superstatistics—it simply arises from the agglomeration of many exponential distributions that have temporal fluctuations of the effective decay rate. In our investigation we also tested the other candidate distributions mentioned above and

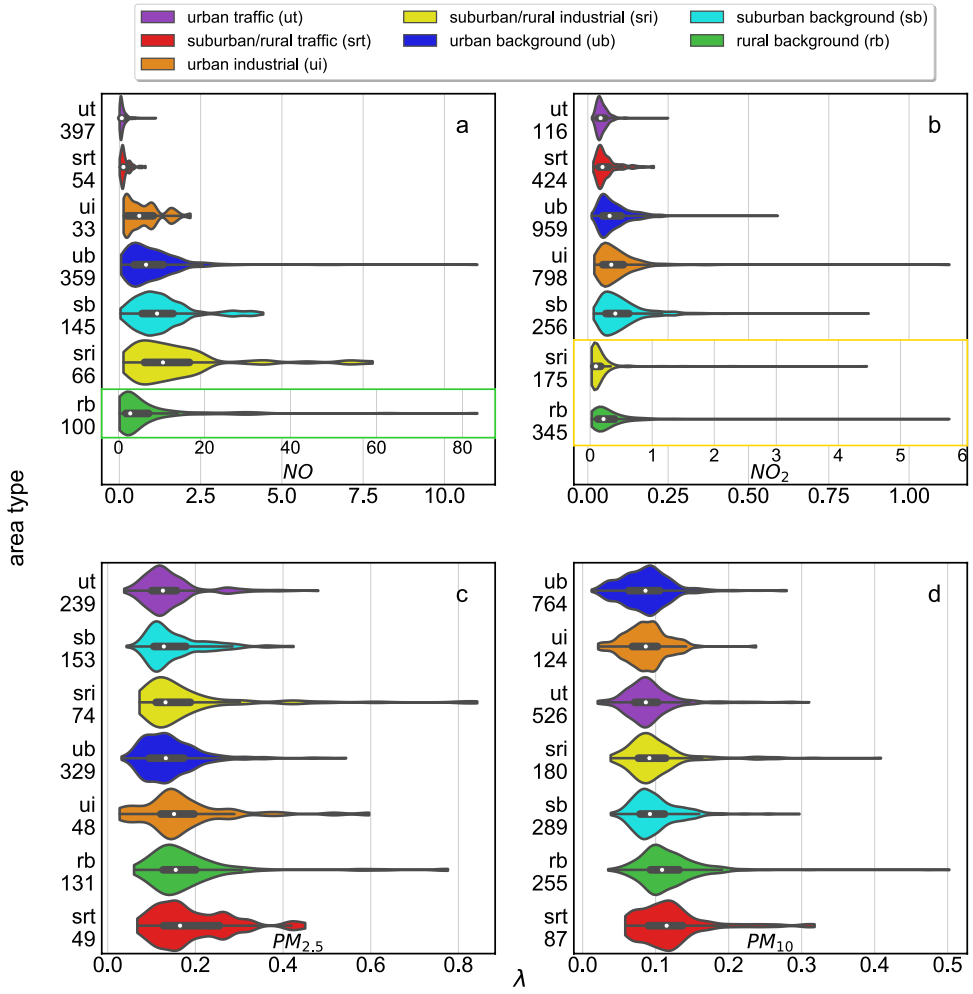

Figure 4. The violin plots show the distribution of λ for seven environment types as well as the median as a white dot, the interquartile range as a thick black bar, and the 95% confidence interval as a thin black bar within the colored violin. The environment types were ranked by medians of λ from lowest to highest. At each y-axis the number of sites evaluated in each category is reported. In the case of *NO* (**a**), a rescaling has been applied to capture the different scale for rural background (green). Likewise, there is also a rescaling for suburban/rural industrial (yellow) and rural background for NO_2.

found that they sometimes yield a good fit of the low-concentration behaviour, close to the maximum, but not of the tail behaviour, see our "Methods" section for details.

Secondly, q and λ, as obtained from the optimum fittings of data from 3544 measuring stations, exhibit interesting patterns in the (q, λ) plane. The shape of these regions is characteristic for each of the 4 substances investigated, with big differences between *NO*, NO_2 and the *PM* statistics. Recall that q contains information about the tail, i.e. extreme events and λ about the scale and thereby the mean pollution level, i.e. we distinguish thereby between regions with low and high average pollution and simultaneously regions with many or few extreme events.

Thirdly, environmental types, i.e. the surroundings of the measuring station, play an important role. We color-coded these different environments into 7 categories. For *NO* and NO_2 each category occupies a typical sub-region in the (q, λ) plane, whereas for $PM_{2.5}$ and PM_{10} the picture is more mixed. This can be explained by the fact that there is transport by moving air for the *PM*-particles, so that the memory to the environment of the station where the actual measurements are done is lost, i.e. these particles may travel quite a long way and many of them are not produced locally. The patterns and spatio-temporal scales for *PM*-dynamics have been extensively examined before in[23–26], the transport is dependent on the weather conditions and removes the memory of the original source site.

As a next step, it would be desirable to examine correlations between *PM* and NO_x concentrations and to include more environmental factors, including for example wind direction and surface temperature. While the current paper is focused on the methods, further statistical analysis could support policymakers to produce more precise rules and thresholds for individual types of environmental conditions and meteorological conditions, taking into account fluctuations and extreme events. Our analysis could also be extended to other substances,

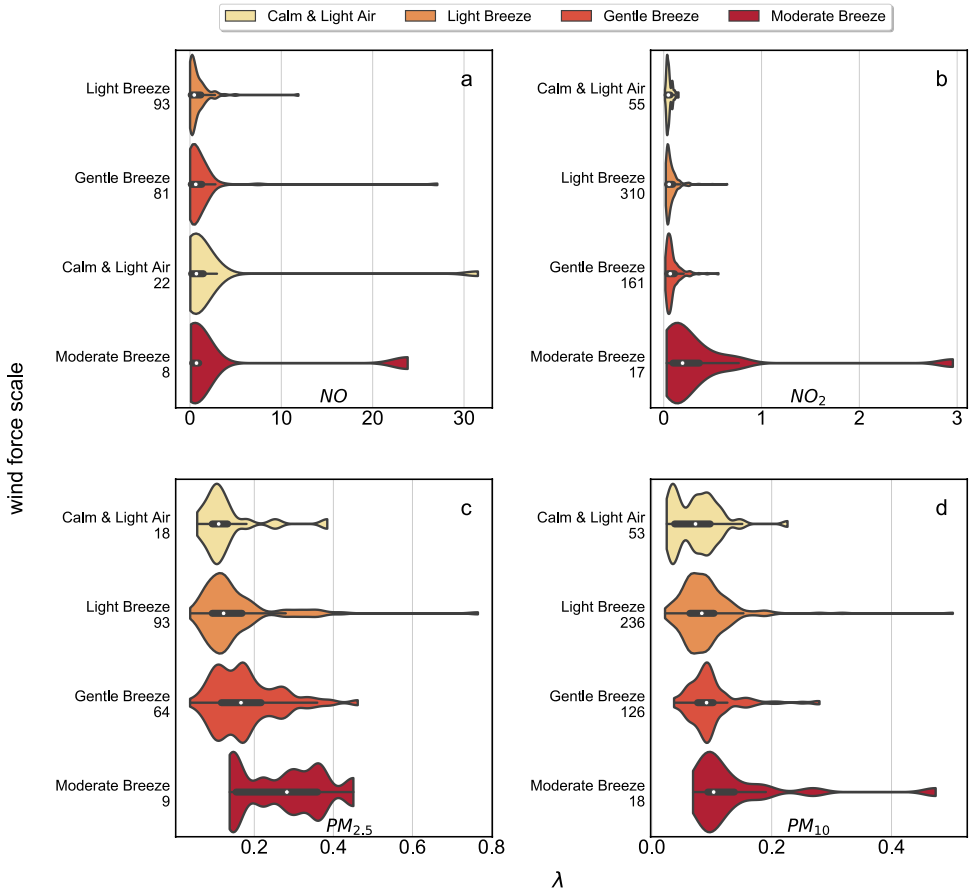

Figure 5. The scale of wind force has a large effect on air pollutants such as $PM_{2.5}$ (**c**), PM_{10} (**d**) and NO_2 (**b**). NO (**a**) exhibits limited spatial heterogeneity when wind speeds are low. The violin plots show the distributions of λ for four wind force scales. At each y-axis the number of sites evaluated in each category is reported. An overlaid box plot depicts the interquartile range and the central white dot indicates the median. The latter was used for ranking the wind force scale from lowest to highest. $PM_{2.5}$, PM_{10} and NO_2 rank in the same order.

such as sulfur oxide, carbon oxide and ozone, besides NO_x and PM. We stress again that the detailed description of the entire pollution spectrum, including the exact behaviour of the tails, seems critical here to better estimate the risks of very high pollution situations. Concluding, our analysis shows that there is strong heterogeneity in the data, and one needs to be careful because the PDFs vary strongly from one location to another.

Methods

Data processing. We import European air pollution data via "saqgetr"[20] R package to analyse air quality data—or more generally atmospheric composition data. The package provides users with fast access to thousands of sites' data from air quality networks, which are supported by Ricardo Energy & Environment. Additionally, we import surface meteorological data from NOAA ISD[32] via the "worldmet"[33] R package. The ISD contains detailed surface meteorological data for over 35,000 locations across the globe. We import all 9698 pollution monitoring sites' hourly data between 2017 and 2021. For each valid concentration measurement, we import the surface meteorological data from the site's closest weather station. Note that the pollutant station's nearest weather station cannot be farther or equal than 0.1 degrees latitude/longitude in our selection process. We save all data as individual .csv files. The raw data contain detailed information about the air pollution monitoring sites and their hourly measurements. From them, we select the following information:

1. Name and site code. Each site has a unique site code for identification and simplification in coding. (E.g. "gb0050a" for the Rosia Road in Gibraltar).
2. Longitude and latitude, which are used to show stations' locations on maps (such as Fig. 1).
3. NO, NO_2, $PM_{2.5}$ and PM_{10} pollutants' hourly concentration data.
4. Station and area types, which are combined for classifying sites.
5. Wind speed data, whose mean is used for classifying sites.

The data sets retrieved with the Saqgetr package contain several problems, for which our solutions are:

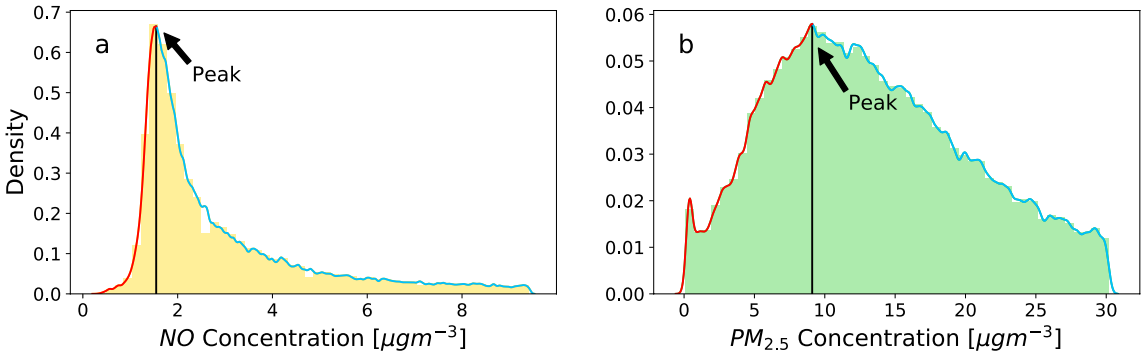

Figure 6. Determination of the lower cutoff at peak density (vertical line). We display the empirical PDFs of the *NO* (**a**) and $PM_{2.5}$ (**b**) concentrations. The empirical smoothed PDF estimates the densities of the data before (red) and after the peak (turquoise). To evaluate the tails, we only consider concentration values that are larger than the peak, see code for details.

1. We filter out sites with less than one year of *NO*, NO_2, $PM_{2.5}$ and PM_{10} pollution concentration data. We need a minimum number of 8760 data points (365 days × 24 h per day) for one pollutant for a meaningful statistical analysis and to avoid analysing a single season.
2. Measurements below the detection limit, including some that are even below zero, are usually replaced by the detection limit divided by two. We filter out sites with more than 15% of data below the detection limit. This is aligned with the recommendations of the US Environmental Protection Agency (US EPA)[35], which suggests that substitution may be a viable approach when up to 15% of the data cannot be detected. This steps is further justified as sites with extremely low pollution are not the main focus of this study.
3. We remove sites whose data is repeating in a single measurement at least 15% of the time. We assume that these sites either lack precision in measurement or contain too much corrupted data.

Fitting procedure. The filtered data sets were analysed in Python. First, we plot the PDF of each site's pollutant concentration level using a log scale. Then, we consider the distributions of the higher concentration tail by finding the maximum of the distribution and filtering out the smaller concentration distribution (left of the maximum of the distribution). One reason for doing so is that we are particularly interested in the high concentration tail behaviour. As an example, let us discuss the probability densities for low pollution concentrations for *NO* in Amstetten, Austria (Fig. 6a) and for $PM_{2.5}$ in Riadok, Slovakia (Fig. 6b). We determine the highest point of the kernel line estimate and its corresponding concentration level (black vertical line). The underlying density distribution of the red line created an increasing slope which we cut off. The underlying density distributions of the turquoise line and the higher concentration tail are the main focus of our analysis.

We fit the so-obtained data with q-exponential distributions derived via maximum likelihood estimation (MLE), and determine the best-fitting parameters q and λ. Shalizi[36] described methods for estimating the parameters of the q-exponential using MLE. We carry out the fit, using the *scipy.stats.rv_continuous.fit* functions of the scipy module[37].

While we already demonstrated deviations from exponential distributions in Fig. 2, we now further justify the use of q-exponential as our primary fitting method, and explain why it describes high concentration tails more effectively than most other fitting functions.

In previous analysis of air pollution, other functions such as log-normal, Weibull, and gamma have been widely employed[12,38–41]. These functions have a maximum quantitatively similar to the air pollutant concentration PDF curve, which also has a peak as shown in Fig. 6. We compare those possible candidate distributions with exponential and q-exponential distributions by fitting all five distributions on the concentration data obtained for all 3544 sites. For each fitting function, we obtain the optimal fitting parameters via MLE, i.e. maximizing the likelihood under the assumption that the data originates from the given distribution. Then, we compute the log-likelihood for each distribution so that we can simply compare log-likelihood numbers. The distribution with the highest log-likelihood is the best fit. We illustrate this in Fig. 7a, see also github for full technical details.

The q-exponential distributions using MLE fits generally show robust fitting results. The log-likelihood for the q-exponential in Fig. 7a is the highest of all the five fittings, making the observations most probable given the parameters. Visually, the q-exponential (purple line) with $q = 1.482$ describes the heavy concentration tails better than other fitting methods. Another, quite different, case with q-exponentials outperforming the other fits uses London Marylebone site's NO_2 concentration data as an example in Fig. 7b. The q-exponential with $q = 0.81$

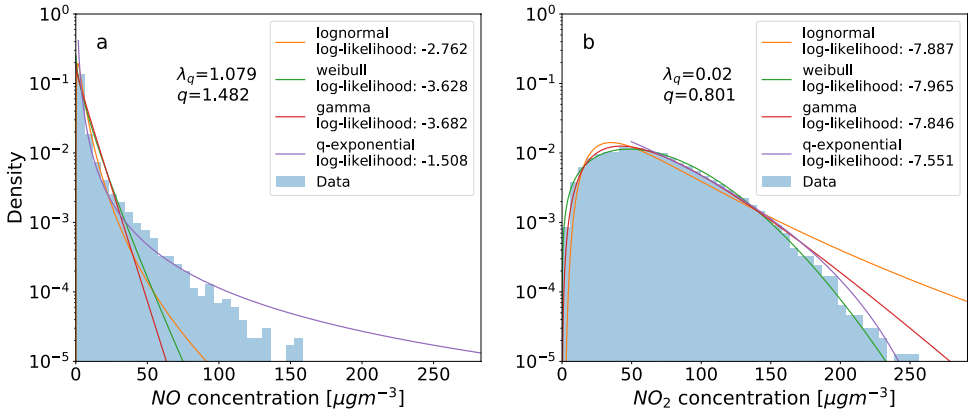

Figure 7. The tails of pollutant concentration PDFs are best described by q-exponentials. We display the PDF of NO, recorded at Amstetten, Austria (**a**) and NO_2, recorded at London Marylebone Road (**b**), together with the best log-normal (orange), Weibull (green), gamma (red) and q-exponential fits (purple), and note their respective log-likelihood values. Also displayed are the values of the q and λ_q parameters for the q-exponential distribution. The q-exponential curves upwards (U-shaped curve) for $q > 1$ (**a**), whereas it curves downwards (inverted U-shape curve) for $q < 1$ (**b**). The q-exponential fits the PDF tails best according to highest log-likelihoods.

again has the highest log-likelihood fit. When we apply q-exponential fits systematically to all sites, its mean (over all sites) log-likelihood is highest among the five considered distributions. Hence, we apply q-exponentials as our main method for analysing the tails of air pollution statistics. For more information, see the code on github.

Data availability

Data sets analyzed in the present study can be obtained via the saqgetr package from https://github.com/skgrange/saqgetr. They should be downloaded as csv files and named after their site codes. The code to generate the figures in the paper, as well as the implementation of the method for the data sets used in the paper, are available at https://github.com/hurst0415/Spatial-heterogeneity-of-air-pollution-statistics.

Received: 28 February 2022; Accepted: 5 July 2022
Published online: 16 July 2022

References

1. World Health Organization (WHO). How air pollution is destroying our health https://www.who.int/news-room/spotlight/how-air-pollution-is-destroying-our-health (2020).
2. Shah, A. S. *et al.* Short term exposure to air pollution and stroke: Systematic review and meta-analysis. *BMJ* **350**, h1295 (2015).
3. Ortiz, A. & Guerreiro, C. *Air Quality in Europe—2020 report*. EEA report (Publications Office of the European Union, 2020) https://www.eea.europa.eu/publications/air-quality-in-europe-2020-report. Accessed on 1 Mar 2022.
4. Health Effects Institute (HEI). State of Global Air 2019. *Special Report* (2019).
5. World Health Organization. Who Global Ambient Air Quality Database (Update 2018) *World Health Organization: Geneva, Switzerland* (2018).
6. GBD 2016 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet (London, England)* **390**, 1345 (2017).
7. European Union. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union* (2008).
8. European Union. Directive 2016/2284/EC of the European Parliament and of the Council of 14 December 2016 on the reduction of national emissions of certain atmospheric pollutants. *Official Journal of the European Union* (2016).
9. World Health Organization (WHO). *Air Quality Guidelines: Global Update 2005: Particulate Matter Ozone, Nitrogen Dioxide, and Sulfur Dioxide* (World Health Organization, 2006)
10. European Parliament. Directive 2011/24/EU of the European Parliament and of the Council of 9 March 2011 on the application of patients' rights in cross-border healthcare (2011). Accessed on 1 Mar 2022.
11. European Commission. The clean air programme. http://ec.europa.eu/environment/air/clean_air/index.htm (2013). Accessed 1 Mar 2022.
12. Lu, H.-C. Comparisons of statistical characteristic of air pollutants in Taiwan by frequency distribution. *J. Air Waste Manag. Assoc.* **53**(5), 608–616 (2003).
13. Williams, G., SchÃfer, B. & Beck, C. Superstatistical approach to air pollution statistics. *Phys. Rev. Res.* **2**, 013019. https://doi.org/10.1103/PhysRevResearch.2.013019 (2020).
14. Schäfer, B. *et al.* Covid-19 impact on air quality in megacities. arXiv preprint arXiv:2007.00755 (2020).
15. Baldasano, J. M. COVID-19 lockdown effects on air quality by NO2 in the cities of Barcelona and Madrid (Spain). *Sci. Total Environ.* **741**, 140353 (2020).
16. Environmental Research Group. The effect of COVID-19 lockdown measures on air quality in London (2020). Report (King's College London, 2020). https://assets.ctfassets.net/9qe8184l2nz4/2TM8WJUt2w1cHecdjkVIRQ/2e5a91667d676b3c63f1e748156b68c4/ERG_response_to_Defra.pdf.
17. Beck, C. Statistics of Three-Dimensional Lagrangian Turbulence. *Phys. Rev. Lett.* **98**, 064502. https://doi.org/10.1103/PhysRevLett.98.064502 (2007).

18. Metzler, R. Superstatistics and non-Gaussian diffusion. *Eur. Phys. J. Spec. Top.* **229**, 711–728 (2020).
19. Beck, C. *et al*. Nonextensive statistical mechanics, superstatistics and beyond: Theory and applications in astrophysical and other complex systems. *The European Physical Journal Special Topics* **229**, 707–709 (2020).
20. Grange, S. K. *saqgetr r package* (2019). https://github.com/skgrange/saqgetr. Accessed on 21 May 2022.
21. European Parliament, Council of the European Union. DIRECTIVE 2008/50/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 May 2008 on ambient air quality and cleaner air for Europe (2008). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32008L0050. Accessed on 1 Mar 2022.
22. European Commission. 2011/850/EU: Commission Implementing Decision of 12 December 2011 laying down rules for Directives 2004/107/EC and 2008/50/EC of the European Parliament and of the Council as regards the reciprocal exchange of information and reporting on ambient air quality (notified under document C(2011) 9068) (2011). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32011D0850. Accessed on 1 Mar 2022.
23. Prospero, J. M., Olmez, I. & Ames, M. Al and Fe in PM 2.5 and PM 10 suspended particles in South-Central Florida: The impact of the long range transport of African mineral dust (2001).
24. Uno, I. et al. Trans-Pacific yellow sand transport observed in April 1998: A numerical simulation. *J. Geophy. Res.* **106**, 18331–18344 (2001).
25. Bardouki, H. *et al*. Chemical composition of size resolved atmospheric aerosols in the eastern Mediterranean during summer and winter (2003).
26. Kallos, G., Papadopoulos, A., Katsafados, P. & Nickovic, S. Model Simulation and Results, Trans-Atlantic Saharan Dust Transport (2006).
27. Tsallis, C. Introduction to nonextensive statistical mechanics: Approaching a complex world. *Springer* **1**, 1–2 (2009).
28. Hanel, R., Thurner, S. & Gell-Mann, M. Generalized entropies and the transformation group of superstatistics. *Proc. Natl. Acad. Sci.* **108**, 6390–6394 (2011).
29. Jizba, P. & Korbel, J. Maximum entropy principle in statistical inference: Case for non-Shannonian entropies. *Phys. Rev. Lett.* **122**, 120601 (2019).
30. Beck, C. & Cohen, E. G. D. Superstatistics. *Phys. A* **322**, 267–275 (2003).
31. Beck, C., Cohen, E. G. D. & Swinney, H. L. From time series to superstatistics. *Phys. Rev. E* **72**, 056133 (2005).
32. Smith, A., Lott, N. & Vose, R. The integrated surface database: Recent developments and partnerships. *Bull. Am. Meteorol. Soc.* **92**, 704–708 (2011).
33. Carslaw, D. *Worldmet: Import surface meteorological data from NOAA integrated surface database (ISD)*. GitHub repository (2017). https://github.com/davidcarslaw/worldmet. Accessed on 21 May 2022.
34. World Meteorological Organization Commission for Maritime Meteorology. *The Beaufort Scale of Wind Force: (technical and Operational Aspects)*. 3 (WMO, 1970).
35. United States. Office of Environmental Information. Quality Staff. Guidance for Data Quality Assessment: *Practical Methods for Data Analysis EPA QA/G-9, (QA00 Update)* (U.S. Environmental Protection Agency) (2000). https://books.google.com.hk/books?id=G3DmHAAACAAJ. Accessed on 1 Mar 2022.
36. Shalizi, C. R. Maximum Likelihood Estimation for q-Exponential (Tsallis) Distributions (2007). math/0701854.
37. Virtanen, P. *et al*. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
38. Lu, H.-C. The statistical characters of pm10 concentration in Taiwan area. *Atmos. Environ.* **36**, 491–502 (2002).
39. Bencala, K. E. & Seinfeld, J. H. On frequency distributions of air pollutant concentrations. *Atmos. Environ.* **1967**(10), 941–950 (1976).
40. Achcar, J. A., Cepeda-Cuervo, E. & Rodrigues, E. R. Weibull and generalised exponential overdispersion models with an application to ozone air pollution. *J. Appl. Stat.* **39**, 1953–1963 (2012).
41. Marani, A., Lavagnini, I. & Buttazzoni, C. Statistical study of air pollutant concentrations via generalized gamma distributions. *J. Air Pollut. Control Assoc.* **36**, 1250–1254 (1986).

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie-Sklodowska-Curie Grant agreement No 840825. We also gratefully acknowledge funding from a 2022 QMUL Research England Policy Impact grant awarded to C.B. and funding from the Helmholtz Association under grant no. VH-NG-1727 to B.S.

Author contributions

H.H., B.S. and C.B. conceived the project, H.H. and B.S. processed the data, performed the data analysis and produced the plots. All authors interpreted the results, wrote and reviewed the manuscript.Overall B.S. and C.B. contributed equally.

Funding

Open Access funding enabled and organized by Projekt DEAL.

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.