Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: http://www.elsevier.com/locate/atmosenv

A strategy for modelling heavy-tailed greenhouse gases (GHG) data using the generalised extreme value distribution: Are we overestimating GHG flux using the sample mean?

M.S. Dhanoa [a], A. Louro [b], L.M. Cardenas [b], A. Shepherd [c,*], R. Sanderson [d], S. López [e,f], J. France [a]

[a] Centre for Nutrition Modelling, Department of Animal Biosciences, University of Guelph, Guelph, ON, N1G 2W1, Canada
[b] Rothamsted Research, North Wyke, Okehampton, Devon, EX20 2SB, UK
[c] Institute of Biological and Environmental Sciences, School of Biological Sciences, University of Aberdeen, 23 St Machar Drive, Aberdeen, AB24 3UU, UK
[d] Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Gogerddan, Aberystwyth, Ceredigion, SY23 3EB, UK
[e] Departamento de Producción Animal, Universidad de León, E-24007, León, Spain
[f] Instituto de Ganadería de Montaña, CSIC-Universidad de León, Finca Marzanas S/n, Grulleros, 24346, Leon, Spain

HIGHLIGHTS

- Using sample means may be overestimating GHG fluxes.
- GEV solves excessive skewness and kurtosis of greenhouse gas flux data.
- Strategy of options for analysing GHG data rather than black-box approach.
- CO_2 estimates from GEV less affected by data in the long tail than sample mean.
- CO_2 estimates from Box-Cox are more affected by long-tail data than from GEV.

ARTICLE INFO

Keywords:
Nitrous oxide
Carbon dioxide
Generalised extreme value
Finney correction
Heavy-tailed data
Skewness correction

ABSTRACT

In this study, we draw up a strategy for analysis of greenhouse gas (GHG) field data. The distribution of GHG flux data generally exhibits excessive skewness and kurtosis. This results in a heavy tailed distribution that is much longer than the tail of a log-normal distribution or outlier induced skewness. The generalised extreme value (GEV) distribution is well-suited to model such data. We evaluated GEV as a model for the analysis and a means of extraction of a robust average of carbon dioxide (CO_2) and nitrous oxide (N_2O) flux data measured in an agricultural field. The option of transforming CO_2 flux data to the Box-Cox scale in order to make the distribution normal was also investigated. The results showed that average CO_2 estimates from GEV are less affected by data in the long tail compared to the sample mean. The data for N_2O flux were much more complex than CO_2 flux data due to the presence of negative fluxes. The estimate of the average value from GEV was much more consistent with maximum data frequency position. The analysis of GEV, which considers the effects of hot-spot-like observations, suggests that sample means and log-means may overestimate GHG fluxes from agricultural fields. In this study, the arithmetic CO_2 sample mean of 65.6 (mean log-scale 65.9) kg CO_2–C ha^{-1} d^{-1} was reduced to GEV mean of 60.1 kg CO_2–C ha^{-1} d^{-1}. The arithmetic N_2O sample mean of 1.038 (mean log-scale 1.038) kg N_2O–N ha^{-1} d^{-1} was substantially reduced to GEV mean of 0.0157 kg N_2O–N ha^{-1} d^{-1}. Our analysis suggests that GHG data should be analysed assuming a GEV distribution of the data, including a Box-Cox transformation when negative data are observed, rather than only calculating basic log and log-normal summaries. Results of GHG studies may end up in national inventories. Thus, it is necessary and important to follow all procedures that contribute to minimise any bias in the data.

* Corresponding author.
E-mail address: anita.shepherd@abdn.ac.uk (A. Shepherd).

https://doi.org/10.1016/j.atmosenv.2020.117500
Received 30 November 2019; Received in revised form 21 March 2020; Accepted 10 April 2020

1. Introduction

Greenhouse gas (GHG) flux data from agricultural fields are difficult to measure precisely because of their inherent spatial and temporal variability. This variability comes from influencing factors such as soil moisture and underlying drainage, field aspect and slope, pH and field distribution of dung or fertilizer. Hot-spots, or rather hot-moments (recorded peaks are time peak rather than spatial peaks), in GHG data are a common occurrence and may cause much nuisance for data analysis (Dixon et al., 2010; Loick et al., 2017). As a result, data recorded on any time scale tend to include high and low peaks resulting in a skewed distribution.

Although GHG emissions information can be extended by computer simulation using soil biogeochemical cycling models, crucially the modelled data require field data for calibration and validation. Hence robust methods for analysis of field data are key to obtaining both accurate field data and simulated data.

A common method of analysis is to transform skewed data to a log-scale. However, as explained and illustrated in Dhanoa et al. (2016), skewness does not always mean a log-normal distribution. Skewness caused by a few extreme values or outliers may be handled by transforming data (Atkinson, 1982), e.g. using the Box and Cox (1964) system or the Finney (1941) correction. If there are many outliers and the data transformation option fails (Atkinson, 1982), the generalised extreme value (GEV) distribution offers an option. This is a very flexible distribution with only three parameters to estimate, sometimes referred to as the Fisher–Tippett distribution after its progenitors (Fisher and Tippet, 1928; Eastoe, 2017), though the common form used in several versions of the GEV follows McFadden (1978).

The GEV is a class of probability distribution, incorporating a heavy-tailed distribution (Evans et al., 2000), that can be fitted to GHG data in order to extract metrics such as the mean and standard deviation.

Plant traits are generally positively skewed, and usually log-transformed. Edwards et al. (2015) used GEV to determine the shape of seed mass distributions.

Küchenhoff and Thamerus (1996) used GEV in the extreme value analysis of Munich air pollution data. Ercelebi and Toros (2009) also used GEV to model Istanbul air pollution (in particular ozone [O_3], benzene [C_6H_6], nitric oxide [NO]). The interactions among these affect N cycling, e.g. [$NO + O_3 \rightarrow NO_2 + O_2$].

Recently, for modelling air pollution data, Korkmaz (2015) described the two-sided generalised Gumbel distribution, which is a special case of the GEV (Type I distribution). Martins et al. (2017) did extreme value modelling of air pollution data and compared results amongst two large urban regions of South America. Battista et al. (2016) used GEV to model urban concentrations of pollutants in the city of Rome (Italy).

GEV is often applied in climatology to changes in temperature and precipitation extremes occurring as the effect of an increase in GHGs, to characterise event magnitudes and frequencies (Kharin and Zwiers, 2004; Katz, 2010). Beniston (2004) analysed the 2003 heat wave data in Europe and showed an association with increased atmospheric GHG concentrations. Studies have so far tended to apply GEV to the climate effects of GHG, rather than the sampled measurements of GHGs themselves.

The purpose of this study is to assess the suitability of the GEV when analysing GHG data from agricultural fields, which often contain larger than expected extreme values forming a thick-tailed data distribution. Its purpose is also to show that the GEV method could minimise bias inherent in simple means of skewed GHG data, and to draw up a strategy for analysis of GHG field data.

Table 1
Chemical composition of applied slurry and digestate.

Property	Units	Slurry application	Digestate application
Dry matter	%	6.5	4.8
Density	kg l^{-1}	1.006	1.00
Ammonium, $NH_4^+ - N$	g kg^{-1} dry matter	18.5	97.3
Nitrate, $NO_3^- - N$	g kg^{-1} dry matter	0.0	0.0
Total N	% of dry matter	2.67	16.9
pH	–	7.30	8.16
Total carbon	% of dry matter	38.4	38.6
C:N ratio	–	14.4	2.3

2. Materials and methods

2.1. *Experimental design and data collection*

The data set originated from a study conducted at Rothamsted Research, North Wyke, Devon, UK (50° 46' 10'' N, 3° 54' 05'' W). The site is on a permanent grassland in a maritime climate (mean annual temperature 9.6 °C; mean annual precipitation 1056 mm).

Four treatments were tested: a) control with no nitrogen (N) fertilizer applications (CN); b) digestate from anaerobic treatment of food waste (DG); c) ammonium nitrate (AN); d) cattle slurry (SL) (Louro et al., 2013; Pezzola et al., 2012).

The soil is a silty clay loam, classed under the British soil classification as clayey typical non-calcareous pelosol of the Halstow series and a stagni-vertic cambisol.

The digestate (from Andigestion biogas plant in Holsworthy, UK) comprised food residues, liquid waste from abattoirs and municipal waste from an anaerobic fermentation cycle lasting 50 days. Cattle slurry was collected from a dairy farm nearby the study site and the applied ammonium nitrate comprised 34.5% N. Chemical composition of the slurry and digestate can be seen in Table 1.

Further information on soil characteristics and chemical composition of the materials applied can be found in Louro et al. (2013) and Pezzolla et al. (2012).

The four treatments were applied in a randomized block design with

Table 2
Glossary of input parameters calculated from GHG flux data.

GHG data parameter	Description	References
AM	Arithmetic sample mean	
γ	Skewness	
$\hat{\mu}$	Sample mean on the log-scale	Finney (1941)
$\hat{\sigma}^2$	Sample variance on the log-scale	Finney (1941)
y	Original data value	Box and Cox (1964)
c	Positive constant which when added to all dataset values makes all data above zero (only if negative values exist)	Box and Cox (1964)
λ	Transformation parameter to fit normal distribution curve	Box and Cox (1964)
ξ	Peak location parameter in the GEV function	Smith (1989)
η	Shape parameter in the GEV function	Smith (1989)
α	Scale parameter in the GEV function	Smith (1989)
Γ	Gamma function	Abramowitz and Stegun (2014)
u	Return period of peak level	Sexto et al. (2013)

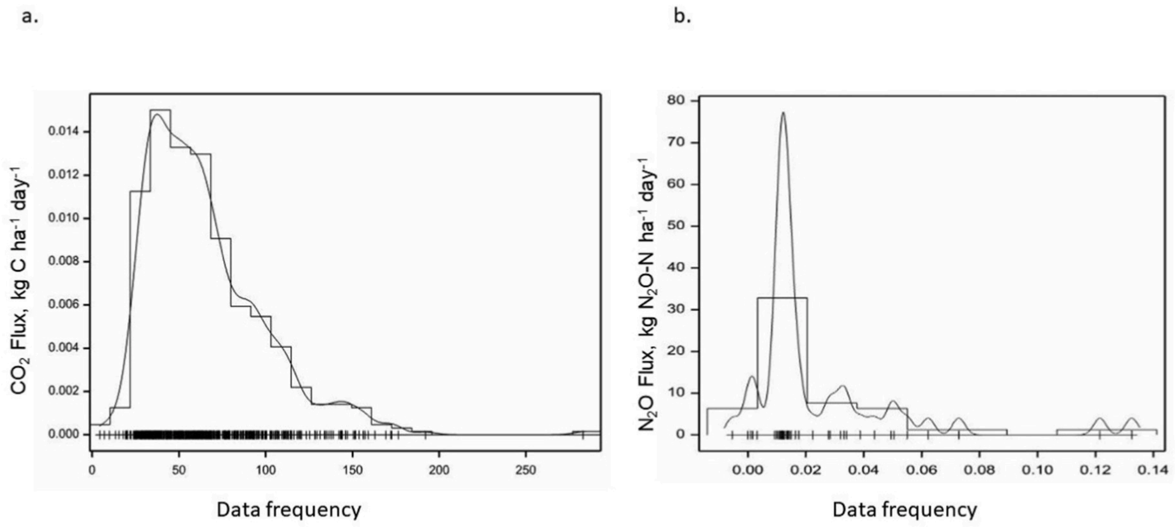

Fig. 1. Kernel density plot of (a.) CO_2 and (b.) N_2O flux data showing main area of data frequency and the composition of the heavy-tail.

three replicate plots per treatment on 8 September 2011, and applied to supply the equivalent rate of 80 kg N ha^{-1}. Nitrous oxide (kg N_2O–N ha^{-1} d^{-1}) and carbon dioxide (kg CO_2–C ha^{-1} d^{-1}) were measured in the 12 plots throughout 47 days between 12 September and 28 October 2011 using one dark non-transparent long-term chamber (LiCor 8100−104) per replicate plot connected to a photoacoustic infrared gas monitor (Lumasense Technologies, INNOVA model 1412i) and an infrared gas analyser (LI-COR Lincoln, Nebraska USA, model LI-8100A). The flux was collected daily from the 12 chamber readings at 11:00 am. There were 12 sets of data each with 47 observations.

2.2. Analysis of GHG data with generalised extreme value (GEV) distribution

A glossary of input parameters required for this study is listed in Table 2.

A kernel density plot (Sheather and Jones, 1991) for CO_2 and N_2O fluxes, showing the position of observation frequency and the nature of skewness, is given in Fig. 1. This illustrates that the processes which cause GHGs to produce apparent outliers combine to give data in a heavy-tailed distribution (also known as thick-tailed, long-tailed, fat-tailed, etc.). When such data are summarised, non-robust statistics such as the sample mean can be highly inflated. The classic approach to deal with a skewed distribution is to check if it follows the log-normal distribution. This is usually done by transforming the data to the log-scale and then testing whether the transformed data follow the normal distribution. One complication comes if the original data contains zeroes or negative values. In this case a positive constant equal to the sample minimum must be added to make all data positive, and one must be added where zero values are present. Once the constant has been applied, the data can be transformed to the log-scale (see Dhanoa et al., 2016 for further information on log transformation). It is worth noting here that the back transformed value of the mean estimated from the log-scaled data is not the same as the calculated arithmetic mean on the original scale, rather the geometric mean. To calculate the mean on the original scale the Finney correction must be applied. Finney (1941) showed that

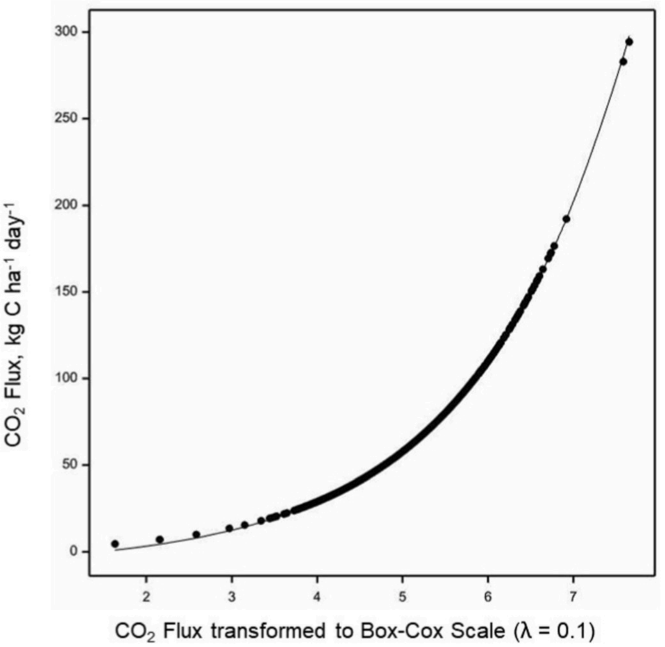

Fig. 2. Empirical relationship ($y = A + B\,R^{y*}$) between CO_2 flux data on the original scale and the corresponding values scaled according to the Box and Cox (1964) transformation with $\lambda = 0.1$.

$$AM = e^{\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}} = e^{\widehat{\mu}}\, e^{\frac{\widehat{\sigma}^2}{2}} \qquad (1)$$

where AM is the arithmetic sample mean on the original scale and $\widehat{\mu}$ and $\widehat{\sigma}^2$ are the estimates of the sample mean and the variance on the log-scale, respectively. Any constant applied prior to logarithmic

Table 3
Sample statistics for CO_2 flux data (kg CO_2–C ha^{-1} d^{-1}) over 47 days from three replicate plots of each of four treatments comprising digestate (DG), cattle slurry (SL), control (CN) and ammonium nitrate (AN).

Treatment-Plot	Sample Mean	Sample SD	Skewness	Kurtosis
DG-1	82.71	58.860	2.071	4.383
DG-2	85.67	44.654	0.400	−0.875
DG-3	65.52	30.580	0.711	−0.506
SL-1	50.56	15.525	0.465	−0.597
SL-2	70.34	36.105	0.882	−0.394
SL-3	73.01	28.340	0.773	−0.171
CN-1	56.85	33.904	1.532	2.377
CN–2	58.19	22.458	1.094	1.007
CN-3	35.07	18.287	1.431	2.540
AN-1	70.24	31.530	0.375	−0.645
AN-2	67.24	18.762	−0.139	−0.379
AN-3	72.07	33.038	0.687	0.406

Overall	Mean = 65.62
	Median = 58.39
		35.399	1.688	5.555

transformation should be subtracted.

An alternative option is to use a data transformation system such as the Box-Cox transformation (Box and Cox, 1964):

$$y^{*} = \frac{(y+c)^{\lambda} - 1}{\lambda} \tag{2}$$

where y^* is the transformed value, y is the actual value, c is the positive constant added to make all data above zero and λ the transformation parameter, enabling the best approximation of a normal distribution curve. To perform this transformation, the value of λ must be estimated first. The algorithm to estimate λ [by numerical search] usually fails in the presence of negative values, so it is prudent to add a suitable constant as detailed above if negative or zero values are present. Having estimated the value of λ and checked if the transformed data follow a normal distribution, one now has the task of transforming the mean estimate back to the original scale. There are not many validated methods, with the exception of the method detailed by Taylor (1985), to perform back-transformation from the Box-Cox scale. However, for convenience, the empirical exponential regression relationship ($y = A + B\,R^{y^*}$) may be used to convert Box-Cox scale quantity y^* to the original scale quantity y and subtracting any constant applied if necessary. This tends to be a good nonlinear relationship for CO_2 flux (see Fig. 2) and for N_2O flux.

The median value of a sample can be a robust statistic but it will be influenced by the presence of a heavy- or long-tail. However, there are some heavy-tailed distributions (Evans et al., 2000) such as extreme value Type I (Gumbel), Type II (Fréchet) or Type III (reverse Weibull) distributions, all of them special cases of the GEV distribution (Coles, 2001). Rather than focussing on these three special cases, GEV is used generally in the data analysis presented here. The shape parameter η allows fitting of this distribution to a variety of data histogram shapes. In heavy-tailed distributions the sample mean is pulled away from the majority of the data values and can be greatly overestimated. Fitting the GEV ensures that the mean estimate represents the majority of the data and thus mitigates overestimation bias. From the estimate of the GEV shape parameter η, one can see which thick-tail type distribution describes the data best. A more important outcome is the estimate of data average (μ) that is relatively free of the effect of data in the long tail. The estimate of μ is calculated as a function of the GEV parameters ξ, η and α (eq. (3) and eq. (4)).

Table 4
Mean plot values for CO_2 flux data (kg CO_2–C ha^{-1} d^{-1}) over 47 days from three replicate plots of each of four treatments comprising digestate (DG), cattle slurry (SL), control (CN) and ammonium nitrate (AN) estimated on the log-scale, Box-Cox scale (Box and Cox, 1964) and by generalised extreme value analysis (GEV).

Treatment-Plot	Mean log-scale[a]	Mean Box-Cox scale[b]	GEV		
			ξ	η	α
DG-1	81.58	70.49	54.20	0.371	26.332
DG-2	87.21	74.93	65.55	−0.050	37.020
DG-3	65.68	59.71	48.60	0.260	19.977
SL-1	50.64	48.51	43.78	−0.075	13.061
SL-2	70.38	63.12	50.66	0.291	21.909
SL-3	73.13	68.51	59.54	0.054	21.171
CN-1	56.70	49.93	39.89	0.265	18.901
CN-2	58.27	54.76	48.15	0.016	16.890
CN-3	35.86	31.19	27.19	0.011	13.504
AN-1	71.03	63.75	57.04	−0.123	27.561
AN-2	67.67	64.60	61.17	−0.338	19.038
AN-3	72.82	65.44	57.46	−0.048	27.178
Overall	65.89	58.53	48.94	0.116	23.670

[a] Finney (1941) correction applied.
[b] Transformed back from Box-Cox scale using an empirical regression relationship, viz. $y = A + B\,R^{y*}$ where $y = CO_2$ flux and $y* =$ flux on Box-Cox scale with $\lambda = 0.1$.

Table 5
Sample statistics for N_2O flux data (kg N_2O–N ha^{-1} d^{-1}) over 47 days from three replicate plots of each of four treatments comprising digestate (DG), cattle slurry (SL), control (CN) and ammonium nitrate (AN).

Treatment-Plot	Sample Mean	Sample SD	Skewness	Kurtosis
DG-1	1.045	0.0277	2.391	6.013
DG-2	1.046	0.0185	1.206	1.464
DG-3	1.036	0.0120	1.020	3.140
SL-1	1.031	0.0047	−0.152	−0.524
SL-2	1.051	0.0391	1.982	3.023
SL-3	1.038	0.0123	−0.312	1.255
CN-1	1.032	0.0108	2.007	4.709
CN-2	1.033	0.0082	1.062	1.984
CN-3	1.026	0.0050	−1.435	3.065
AN-1	1.042	0.0165	1.137	0.326
AN-2	1.033	0.0081	0.181	−0.473
AN-3	1.039	0.0126	1.562	2.538

Overall	Mean = 1.038
	Median = 1.033
		0.0187	3.457	1.766

GEV is a simple three parameter probability function with cumulative distribution function (Evans et al., 2000), $F(x)$, and probability density function, $f(x)$, defined as follows.

The cumulative distribution function for the GEV (Smith, 1989; Martins et al., 2017) is given by

$$F(x) = \exp\left(- \left[1 - \eta \frac{(x-\xi)}{\alpha} \right]^{\frac{1}{\eta}} \right) \text{ for } \eta \neq 0 \text{ and } \alpha > 0, \tag{3}$$

with ξ being the data peak location parameter, η the shape parameter and α the scale parameter. In this functional form, $\eta < 0$ indicates a Fréchet distribution and $\eta > 0$ a reverse Weibull (Eastoe, 2017). The limiting value at $\eta = 0$ is the Gumbel distribution. Parameter ξ is related to the position of the majority of data peak similar to the geometric mean or mode position in skew distributions.

The corresponding formula for the probability density function (Singh, 1998) is

$$f(x) = \frac{1}{\alpha}\left[1 - \frac{\eta}{\alpha}(x-\xi)\right]^{\frac{1-\eta}{\eta}} \exp\left(- \left[1 - \frac{\eta}{\alpha}(x-\xi)\right]^{\frac{1}{\eta}} \right) \tag{4}$$

From these references the mean (μ), standard deviation (σ) and skewness (γ) of the GEV distribution can be calculated:

$$\widehat{\mu} = \widehat{\xi} + \frac{\widehat{\alpha}}{\widehat{\eta}}\left[1 - \Gamma(1+\widehat{\eta})\right]$$

$$\widehat{\sigma} = \hat{\alpha}\hat{\eta}\sqrt{\Gamma(1+2\widehat{\eta}) - \left[\Gamma(1+\widehat{\eta})\right]^2}$$

$$\widehat{\gamma} = \mathrm{sgn}(\widehat{\eta}\frac{-\Gamma(1+3\widehat{\eta}) + 3\Gamma(1+\widehat{\eta})\Gamma(1+2\widehat{\eta}) - 2\left[\Gamma(1+\widehat{\eta})\right]^3}{\left[\Gamma(1+2\widehat{\eta}) - \left[\Gamma(1+\widehat{\eta})\right]^2\right]^{3/2}} \tag{5}$$

Symbol Γ denotes the gamma function (Abramowitz and Stegun, 2014).

The quantile function [the inverse of $F(x)$] of the GEV distribution is:

$$F^{-1}(u) = \xi + \frac{\alpha}{\eta}\left[1 - \left[-\ln(u)\right]^{\eta}\right] \text{ with } 0 < u < 1 \tag{6}$$

When the interest is to estimate re-occurrence of (say) the maximum of a particular pollutant, then the value $F^{-1}(1-u)$ is the return level associated with the return period $1/u$ (Sexto et al., 2013).

3. Results and discussion

The nature of GHG data is such that any spot value may be not representative of the flux size in a particular agricultural field. This is why the data in this study were collected every day over a period of 47 days. However, this extra time dimension creates a need to summarise data so the treatments may be compared by simple analysis of variance (ANOVA) based on the statistical design. Alternatively, a repeated measurement ANOVA of the design may be carried out without summarizing the data and a simple randomised block ANOVA using meaningful summary statistics is also desirable. For this purpose, the time course profile may be modelled if a suitable model is identifiable, otherwise calculating the area under the curve can be a good surrogate summary.

To understand the averaging problem, the sample average (implicitly assuming a normal distribution), log-normal based mean, Box-Cox transformation based mean and mean from the fit of GEV distribution were considered. This exercise was completed with CO_2 flux data (Tables 3 and 4). The N_2O flux data (Tables 5 and 6) had very small scale size observations and both positive and negative values. Thus, the algorithm to estimate λ did not converge to a satisfactory solution. Even to calculate the mean via the log-scale, it was necessary to use $\ln(x + c)$ with $c =$ absolute value of the minimum (N_2O flux) + 1.0. Because of this difficulty the results for N_2O flux amended as above are included. From the parameters of the GEV distribution, the GEV mean for CO_2 flux was estimated to be 60.1 kg CO_2–C ha^{-1} d^{-1} shown in Table 7 (note that (μ – ξ) is the contribution from data in the heavy tail). Similarly, the estimated GEV mean for N_2O flux (net of the added constant of 1.0203) was 0.0157 kg N_2O–N ha^{-1} day^{-1} ({1.036–1.0203}; Table 8).

Table 6
Mean plot values for N_2O flux data (kg N_2O–N ha^{-1} d^{-1}) over 47 days from three replicate plots of each of four treatments comprising digestate (DG), cattle slurry (SL), control (CN) and ammonium nitrate (AN) estimated on the log-scale, Box-Cox scale (Box and Cox, 1964) and by generalised extreme value analysis.

Treatment-Plot	Mean log-scale[a]	Mean Box-Cox scale[b]	GEV		
			ξ	η	α
DG-1	1.0450	1.059	1.033	0.278	0.0124
DG-2	1.0458	1.061	NA	NA	NA
DG-3	1.0358	1.053	1.031	−0.112	0.0107
SL-1	1.0307	1.049	NA	NA	NA
SL-2	1.0513	1.064	NA	NA	NA
SL-3	1.0377	1.055	1.034	−0.331	0.0128
CN-1	1.0320	1.050	1.027	0.089	0.0068
CN-2	1.0326	1.050	1.029	−0.062	0.0038
CN-3	1.0259	1.044	NA	NA	NA
AN-1	1.0421	1.058	NA	NA	NA
AN-2	1.0333	1.051	1.030	−0.242	0.0078
AN-3	1.0394	1.056	1.034	0.058	0.0086
Overall	1.0376	1.037	1.030	0.072	0.0108

NA = Not available, distribution did not fit.

[a] Finney (1941) correction applied.

[b] Transformed back from Box-Cox scale using an empirical regression relationship, viz. $y = A + B\,R^{y*}$ where $y = N_2O$ flux (with added constant) and $y^* =$ flux on Box-Cox scale with $\lambda = -4.0$.

3.1. Carbon dioxide flux data

The example data employed here demonstrate a heavy tailed distribution as shown by skewness of 1.688 ± 0.104 and kurtosis of 5.555 ± 0.208 due to the presence of excessive hot-moments or extreme values (see Fig. 3). This feature of data distributions means non-robust statistics such as the arithmetic mean will be biased positively. When examined on the log-scale the data were still non-normal. Similarly, data on the Box-Cox Scale with $\lambda = 0.1$ did not become normal. However, when

Table 7
CO_2 flux sample mean and mean and standard deviation as calculated from the parameters of the GEV distribution.

Treatment-Plot	Sample Mean	GEV	
		$\widehat{\mu}$	$\widehat{\sigma}$
DG-1	82.71	62.06	25.245
DG-2	85.67	88.84	50.949
DG-3	65.52	55.94	20.197
SL-1	50.56	52.35	18.672
SL-2	70.34	58.26	21.773
SL-3	73.01	70.68	25.432
CN-1	56.85	46.78	19.057
CN-2	58.19	57.63	21.213
CN-3	35.07	34.84	17.081
AN-1	70.24	76.73	42.851
AN-2	67.24	81.60	53.569
AN-3	72.07	74.49	37.260
Overall	65.62	60.14	26.679

Table 8
N_2O flux sample mean and mean and standard deviation calculated from the parameters of the fitted GEV distribution (flux data used include the added constant 1.02029 to overcome negative and zero values in the original data).

Treatment-Plot	Sample Mean	GEV	
		$\widehat{\mu}$	$\widehat{\sigma}$
DG-1	1.045	1.037	0.0124
DG-2	1.046	NA	NA
DG-3	1.036	1.038	0.0163
SL-1	1.031	NA	NA
SL-2	1.051	NA	NA
SL-3	1.038	1.047	0.0350
CN-1	1.032	1.031	0.0079
CN-2	1.033	1.034	0.0095
CN-3	1.026	NA	NA
AN-1	1.042	NA	NA
AN-2	1.033	1.037	0.0158
AN-3	1.039	1.038	0.0103
Overall	1.038	1.036	0.0127
		Net 0.0157	

NA = Not available, distribution did not fit.

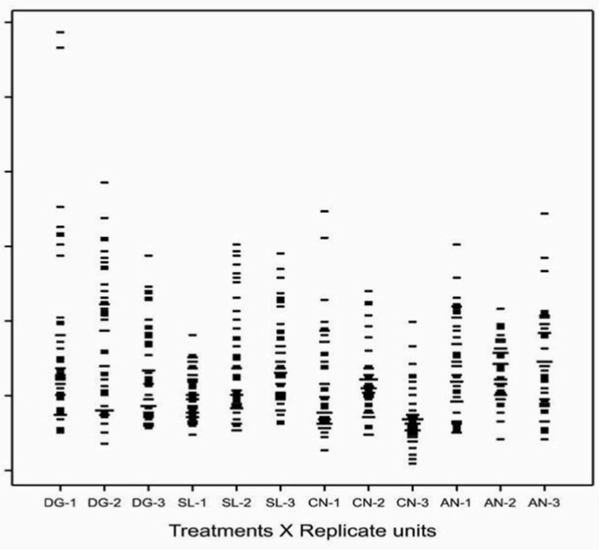

Fig. 3. CO_2 flux data showing observations contributing to skewness and heavy tail.

using the generalised extreme value distribution, considering the probability plot (Atkinson, 1985), the data were found to be consistent with that distribution (Fig. 4).

When analysing CO_2 flux data, the generalised extreme value distribution successfully fitted individual treatments and overall (GEV parameters given in Table 4) and it provided a better description of the data compared to the normal, log-normal and Box-Cox transformed data. It therefore seems GEV is a viable option to analyse long-tailed or heavy-tailed GHG data. The analysis of CO_2 data shows that fitting the GEV distribution can reduce bias from the sample mean estimate (Table 7) and the standard deviation is also smaller. From the fitted parameters of the GEV distribution (Table 4), the mean μ and standard deviation σ were calculated (Table 7).

3.2. *Nitrous oxide flux data*

Nitrous oxide flux data appear very different across the 12 plots in this study. Values range from high positive values to negative values (Fig. 5). Thus, the data for N_2O flux were much more complex than CO_2 flux data due to the presence of negative fluxes. These data form mixtures of distributions. The graphical test (Atkinson, 1985) showed that even GEV was not able to fit to the individual plot data sets entirely satisfactorily (Fig. 6) despite the addition of a constant of 1.0203 (i.e. 1 + minimum absolute data value) to the data. From the GEV parameter estimates in Table 6, the estimates of mean μ and standard deviation σ were calculated (Table 8).

4. Conclusions

This study shows that when analysing GHG data from agricultural fields, detailed analysis is required before proceeding to the application of a suitable methodology. Black-box or default statistics such as simple sample mean can give biased estimates. It is prudent to test implicit distributional assumptions in order to identify an appropriate methodology.

From the above, we can draw up a general strategy for GHG field data analysis:

1. Check the distribution of the data and see if it is normally distributed.
2. Check for presence of hot-moments or outliers and deal with them if present (see Dhanoa et al., 2016 for various tests).

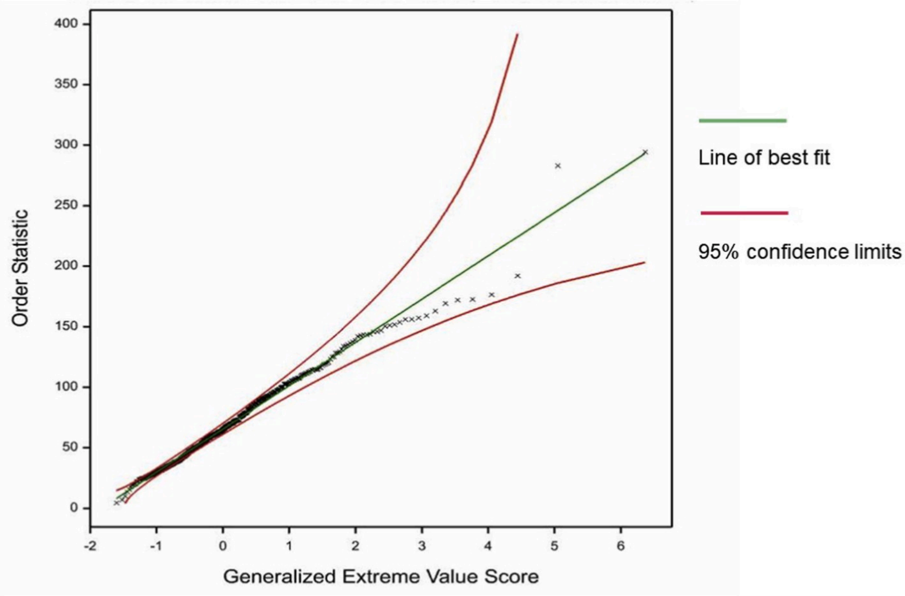

Fig. 4. Probability plot when modelling CO_2 flux data using the GEV distribution.

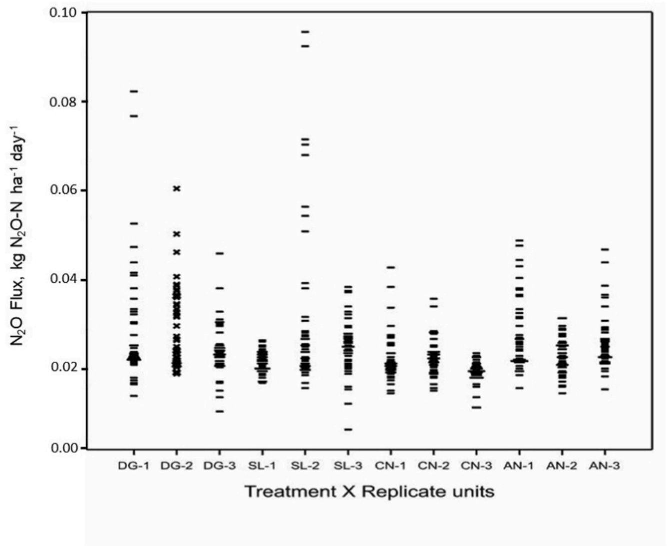

Fig. 5. N_2O flux data showing observations that contribute to skewness and a heavy tail.

3. If the data distribution appears to be skewed, consider if the data are expected to follow a log-normal distribution. Check the distribution after logarithmic transformation. If data observations include negative and/or zero values, then $\ln(x + c)$ should be used with value of constant c such that all data observations are positive. As explained above, when converting back any log-scale statistics on to the original scale the Finney (1941) correction must be applied (Dhanoa, 2017) and any constant that was added must be subtracted.
4. If the majority of the data appear to be normal apart from a few outliers, then the Box-Cox transformation may be considered. When $\lambda = 0.0$, logarithmic transformation is indicated otherwise use the Box-Cox scale as described above. Again, add a constant c to make all data positive.
5. However, if the distribution tail is long with many divergent observations in that tail, then the option of the generalised extreme value distribution may be relevant.

In the case of GHG data studies, many of the results may end up in national inventories. Thus, it is necessary and important to follow all procedures that contribute to minimise any bias in the data summaries, to enable meta-analysis and other statistical comparisons of treatments and studies provide suitable measure(s) of uncertainty.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

M.S. Dhanoa: Conceptualization, Methodology, Formal analysis, Writing - original draft. **A. Louro:** Resources, Data curation. **L.M. Cardenas:** Funding acquisition, Writing - original draft, Writing - review & editing. **A. Shepherd:** Writing - original draft, Writing - review & editing. **R. Sanderson:** Writing - review & editing. **S. López:** Writing -

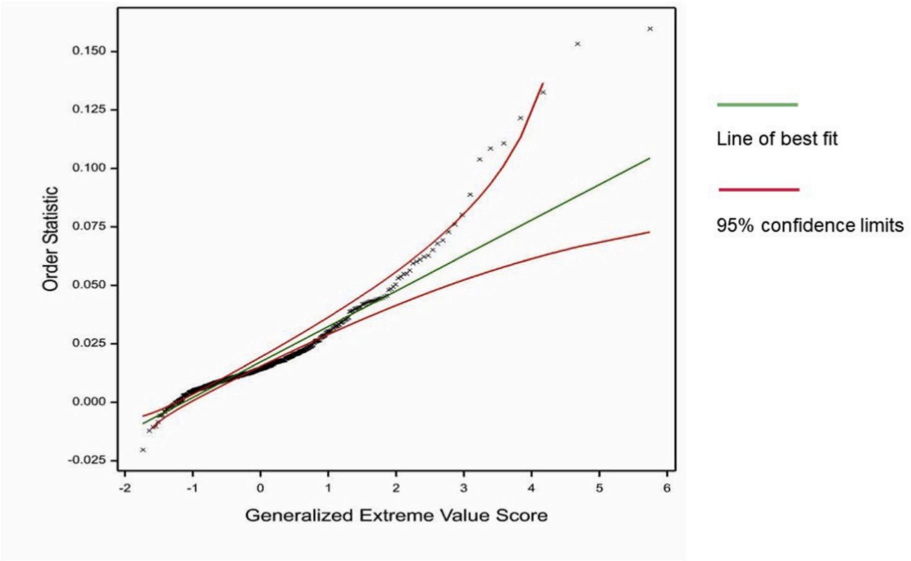

Fig. 6. Probability plot when modelling N_2O flux data using the GEV distribution.

review & editing. **J. France:** Methodology, Writing - review & editing.

Acknowledgements

The work was supported by the Biotechnology and Biological Sciences Research Council (BB/P01268X/1, BBS/E/C/000I0320).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2020.117500.

References

Abramowitz, M., Stegun, I.A., 2014. Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables. Martino Fine Books, Eastford, CT, USA.

Atkinson, A.C., 1982. Regression diagnostics, transformations and constructed variables. J. Roy. Stat. Soc. B 44, 1–36.

Atkinson, A.C., 1985. Plots, Transformations and Regression. Oxford University Press, Oxford, UK.

Battista, G., Pagliaroli, T., Mauri, L., Basilicata, C., de Lieto Vollaro, R., 2016. Assessment of the air pollution level in the city of Rome (Italy). Sustainability 8, 838–852.

Beniston, M., 2004. The 2003 heat wave in Europe: a shape of things to come? An analysis based on Swiss climatological data and model simulation. Geophys. Res. Lett. 31, L02202.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations (with discussion). J. Roy. Stat. Soc. B 26, 211–246.

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag, London, UK.

Dhanoa, M.S., Sanderson, R., López, S., Kebreab, E., France, J., 2016. Consequences of metabolic scaling and log-scale allometry on means and variances and parameter estimates from Type I and Type II linear regression models. e-Planet 14 (1), 1–10.

Dhanoa, M.S., 2017. Cinderella's transformation. Significance 14, 46.

Dixon, E.R., Blackwell, M.S.A., Dhanoa, M.S., Berryman, Z., de la Fuente Martinez, N., Junquera, D., Martinez, A., Murray, P.J., Kemp, H.F., Meier-Augenstein, W., Duffy, A., Bol, R., 2010. Measurement at the field scale of soil delta^{13}C and delta^{15}N under improved grassland. Rapid Commun. Mass Spectrom. 24, 511–518. https://doi.org/10.1002/rcm.4345.

Eastoe, E., 2017. Extreme value distributions. Significance 14, 12–13. https://doi.org/10.1111/j.1740-9713.2017.01014.

Edwards, W., Moles, A.T., Chong, C., 2015. Generalised extreme value distributions provide a natural hypothesis for the shape of seed mass distributions. PLoS One 10 (4), e0121724. https://doi.org/10.1371/journal.pone.0121724.

Ercelebi, S.G., Toros, H., 2009. Extreme value analysis of Istanbul air pollution data. CLEAN- Soil Air Water 37, 122–131.

Evans, M., Hastings, N.A.J., Peacock, J.B., 2000. Statistical Distributions, 3rd ed. John Wiley & Sons, New York.

Finney, D.J., 1941. On the distribution of a variate whose logarithm is normally distributed. J. Roy. Stat. Soc. B 7, 155–161.

Fisher, R.A., Tippett, L.H.C., 1928. Limiting forms of the frequency distributions of the largest or smallest member of a sample. In: Proceedings of the Cambridge Philosophical Society, vol. 24, pp. 180–190.

Katz, R.W., 2010. Statistics of extremes in climate change. Clim. Change 100, 71–76.

Kharin, V.V., Zwiers, F.W., 2004. Estimating extremes in transient climate change simulations. J. Clim. 18, 1156–1173. https://doi.org/10.1175/JCLI3320.1.

Korkmaz, M.C., 2015. Two-sided generalised Gumbel distribution with application to air pollution data. Int. J. Stat. Distrib. Appl. 1, 19–26.

Küchenhoff, H., Thamerus, M., 1996. Extreme value analysis of Munich pollution data. Environ. Ecol. Stat. 3, 127–141.

Loick, N., Dixon, E., Abalos, D., Vallejo, A., Matthews, P., McGeough, K., Watson, C., Baggs, E., Cardenas, L.M., 2017. "Hot spots" of N and C impact nitric oxide, nitrous oxide and nitrogen gas emissions from a UK grassland soil. Geoderma 305, 336–345.

Louro, A., Sawamoto, T., Chadwick, D., Pezzolla, D., Bol, R., Baez, D., Cardenas, L., 2013. Effect of slurry and ammonium nitrate application on greenhouse gas fluxes of a grassland soil under atypical South West England weather conditions. Agric. Ecosyst. Environ. 181, 1–11.

Martins, L.D., Wikuats, C.F.H., Capucim, M.N., de Almeida, D.S., da Costa, S.C., Albuquerque, T., Carvalho, V.S.B., de Freitas, E.D., Andrade, M. de F., Martins, J.A., 2017. Extreme value analysis of air pollution data and their comparison between two large urban regions of South America. Weather Clim. Extremes 18, 44–54.

McFadden, D., 1978. Modeling the choice of residential location. Transportation Research Record 673, 72–77.

Pezzolla, D., Bol, R., Gigliotti, G., Sawamoto, T., Louro-López, A., Cardenas, L., Chadwick, D., 2012. Greenhouse gas (GHG) emissions from soils amended with digestate derived from anaerobic treatment of food waste. Rapid Commun. Mass Spectrom. 26, 2422–2430.

Sexto, B.M., Vaquera, H.H., Arnold, B.C., 2013. Use of the Dagum distribution for modelling tropospheric ozone levels. J. Environ. Stat. 5 (5), 1–11.

Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Stat. Soc. B 53, 683–690.

Singh, V.P., 1998. Generalized extreme value distribution. In: Entropy-Based Parameter Estimation in Hydrology. Water Science and Technology Library, Vol 30. Springer, Dordrecht, The Netherlands, pp. 169–183.

Smith, R.L., 1989. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. Stat. Sci. 4, 367–377.

Taylor, J.M.G., 1985. Measures of location of skew distributions obtained through Box-Cox transformations. J. Am. Stat. Assoc. 80 (390), 427–432. https://doi.org/10.1080/01621459.1985.10478135.