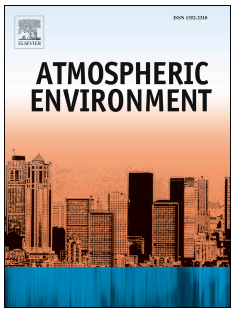


A strategy for modelling heavy-tailed greenhouse gases (GHG) using the generalised extreme value distribution: Are we overestimating GHG flux using the sample mean?

M.S. Dhanoa, A. Louro, L.M. Cardenas, A. Shepherd, R. Sanderson, S. Lopez, J. France

PII: S1352-2310(20)30237-5

DOI: https://doi.org/10.1016/j.atmosenv.2020.117500

Reference: AEA 117500

To appear in: *Atmospheric Environment*

Received Date: 30 November 2019

Revised Date: 21 March 2020

Accepted Date: 10 April 2020

Please cite this article as: Dhanoa, M.S., Louro, A., Cardenas, L.M., Shepherd, A., Sanderson, R., Lopez, S., France, J., A strategy for modelling heavy-tailed greenhouse gases (GHG) using the generalised extreme value distribution: Are we overestimating GHG flux using the sample mean?, *Atmospheric Environment* (2020), doi: https://doi.org/10.1016/j.atmosenv.2020.117500.

Dhanoa, M.S.: Conceptualization, Methodology, Analysis, Writing- original draft preparation. **Louro, A.**: Resources, Data curation, **Cardenas, L.M.**: Funding acquisition, writing mid-draft preparation, reviewing, editing. **Shepherd, A.**: Writing- final draft preparation, wrote response to peer review **: Sanderson, R**: Reviewing and editing. **Lopez, S.:** Reviewing, Editing, **France, J.**: Methodology, Reviewing, Editing.

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

M.S. Dhanoa[1], A. Louro[2], L.M. Cardenas[2], A. Shepherd[3]*, R. Sanderson[4], S. Lopez[5,6] and J. France[1]

[1] Centre for Nutrition Modelling, Department of Animal Biosciences, University of Guelph, Guelph ON, N1G 2W1, Canada.

[2] Rothamsted Research, North Wyke, Okehampton, Devon, EX20 2SB, UK.

[3] Institute of Biological and Environmental Sciences, School of Biological Sciences, University of Aberdeen, 23 St Machar Drive, Aberdeen AB24 3UU, UK

[4] Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Gogerddan, Aberystwyth, Ceredigion, SY23 3EB, UK.

[5]Departamento de Producción Animal, Universidad de León, E-24007 León, Spain

[6]Instituto de Ganadería de Montaña, CSIC-Universidad de León, Finca Marzanas s/n, Grulleros, 24346 Leon, Spain

*Corresponding author

ABSTRACT

In this study, we draw up a strategy for analysis of GHG field data. The distribution of greenhouse gas (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 mean 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 value 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 65.62 (mean log-scale 65.89) kg CO_2-C ha^{-1} d^{-1} was reduced to GEV mean of 60.14 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 reduced to GEV mean of 0.01571 kg N_2O-N ha^{-1} d^{-1}. Our analysis suggests that GHG data should be analysed using the GEV method, including a Box-Cox transformation when negative data is present, 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 help minimise any biases in the data.

Keywords: nitrous oxide; carbon dioxide; generalised extreme value; Finney correction; heavy-tailed data; skewness correction.

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 are a cause of 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 requires field data for calibration and validation. Hence robust methods for analyses 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, 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, 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 distributions, incorporating a heavy-tailed distribution, that we can fit 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. [$NO + O_3 \rightarrow N_2O + O_2$].

Recently, for modelling air pollution data, Korkmaz (2015) described two-sided generalised Gumbel (TSGG) distribution, which is a special case of 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. Beniston (2004) analysed the 2003 heat wave data in Europe and showed an association with enhanced atmospheric GHG concentrations. Battista et al (2016) used GEV to model urban concentrations of pollutants (legislated under the Directive 2008/50/EC) in the city of Rome (Italy).

GEV is often applied to climatology. It is applied 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). 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 eliminate bias inherent in simple means of skewed GHG data, and to draw up a strategy for analysis of GHG field data.

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 (50:46:10N, 3:54:05W). 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 similar silty clay loam used for the digestate sampling is a clayey non□calcareous Pelostagnogley of the Hallsworth series.

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 d (or days).

Cattle slurry was collected from a dairy farm nearby the study site and spread at a rate of 80 kg total N ha^{-1}.

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 three replicate plots per treatment on 8 September 2011, and applied to supply the equivalent rate of 80 kg N ha^{-1} on each occasion. 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 d 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 11:00H. 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 Figure 1. This illustrates that the processes which cause GHGs to produce apparent outliers combine to give data a heavy-tailed distribution (also known as thick-tailed, long-tailed, fat-tailed, etc.). When one summarises such data, 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 one must add a positive constant equal to the sample minimum to make all data positive and one must also add 1.0 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 anti-logged value of the mean estimate 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

$$AM = e^{\hat{\mu} + \frac{\hat{\sigma}^2}{2}} = e^{\hat{\mu}}\, e^{\frac{\hat{\sigma}^2}{2}} \qquad \text{eq. 1}$$

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

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

$$y^* = \frac{(y+c)^{\lambda}-1}{\lambda} \qquad \text{eq. 2}$$

Where y^* is the transformed data value, y is the actual data 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 follows 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 ($y' = A + B\,R^x$) may be used to convert Box-Cox scale quantity x 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 Figure 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 many heavy-tailed distributions such as extreme value Type I (Gumbel) and extreme value Type II (Weibull) among others. These

Type I and II along with Type III (Frèchet) are 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 positive shape parameter indicates the Frèchet distribution, zero the Gumbel, and negative the Weibull (Eastoe, 2017). 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 η and α (eq. 3 and eq. 4). Parameter α is the position of the majority of data peak similar to the geometric mean or mode position in skew distributions.

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

The cumulative distribution function for the GEV (Smith, 1986; Martins et al. 2017) is given by

$$F(x) = exp\left\{-\left[1+\eta\frac{(x-\xi)}{\alpha}\right]^{\frac{1}{\eta}}\right\} \text{ when } \eta \neq 0$$

$$F(x) = exp\left\{-exp\left[-\frac{(x-\xi)}{\alpha}\right]\right\} \text{ when } \eta = 0 \qquad \text{eq. 3}$$

with ξ being the data peak location parameter, η the shape parameter and α the scale parameter.

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

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

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

$$\hat{\xi} = \hat{\mu} - \frac{\hat{\alpha}}{\hat{\eta}}\{1 - \Gamma(1+\hat{\eta})\}$$

$$\hat{\alpha} = \frac{\hat{\sigma}\,\hat{\eta}}{\{\Gamma(1+2\hat{\eta}) - [\Gamma(1+\hat{\eta})]^2\}^{1/2}}$$

$$\hat{\gamma} = sign(\hat{\eta})\frac{-\Gamma(1+3\hat{\eta})+3\Gamma(1+\hat{\eta})\Gamma(1+2\hat{\eta})-2[\Gamma(1+\hat{\eta})]^3}{\{\Gamma(1+2\hat{\eta})-[\Gamma(1+\hat{\eta})]^2\}^{3/2}} \quad \text{eq. 5}$$

Symbol Γ denotes the gamma function.

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

$$F^{-1}(u) = \zeta - \frac{\alpha}{\eta}[1 - \{-log\, u\}^{-\eta}] \text{ with } 0 < u < 1 \quad \text{eq. 6}$$

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 not be representative of the flux size in a particular agricultural field. This is why the data in this study was collected every day over a period of 47 d. 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 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 60.14 kg CO_2-C ha^{-1} d^{-1} shown in Table 7 (note that $(\mu - \xi)$ is the contribution from data in the heavy tail). Similarly, the estimated GEV mean for N_2O flux (net of the added constant of 1.02029) was 0.01571 kg N_2O-N ha^{-1} day^{-1} ({1.036 − 1.02029}; Table 8).

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 Figure 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 using the generalised extreme value distribution, the data were found to be consistent with that distribution (Figure 4).

When analysing CO_2 flux data, the generalised extreme value distribution fitted successfully to 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 lower. 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 (Figure 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. Even GEV did not fit to the individual plot data sets entirely satisfactorily (Figure 6) despite the addition of a constant of 1.02029 (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).

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).

3. If the data distribution appears to be skew, 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 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 observations in it, 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 help minimise any biases in the data summaries, to enable meta-analysis and other statistical comparisons of treatments and studies provide suitable measure(s) of uncertainty.

ACKNOWLEGEMENTS

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

REFERENCES

Atkinson, A.C. (1982) Regression diagnostics, transformations and constructed variables (with discussion). *Journal of the Royal Statistical Society, Series B*, 44, 1-36.

Atkinson, A.C. (1985) *Plots, Transformations and Regression*. Oxford University Press, Oxford.

Battista, G., Pagliaroli, T., Mauri, L., Basilicata, C., and de Lieto Vollaro, R. (2016) Assessment of the Air Pollution Level in the City of Rome (Italy). *Sustainability,* 8 (9): 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. *Geophysical Research Letters,* 31, L02202

Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26, 211-46.

Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London, 224pp.

Dhanoa, M.S., Sanderson, R., Lopez, S., Kebreab, E. and 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(2), 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 Communications in Mass Spectrometry* 24, 511-518. http://dx.doi.org/10.1002/rcm.4345

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

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

Ercelebi, S. G. and Toros, H. (2009) Extreme Value Analysis of Istanbul Air Pollution Data. *CLEAN - Soil Air Water* 37(2):122−131

Evans, M., Hastings, N.A.J. and Peacock, J.B. (2000) *Statistical Distributions* (3rd Edition). John Wiley & Sons, New York.

Finney, D.J. (1941) On the distribution of a variate whose logarithm is normally distributed. J*ournal of the Royal Statistical Society Series* B, 7, 155-161.

348 Fisher, R.A. and Tippett, L.H.C. (1928) Limiting forms of the frequency distributions of the
349 largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical*
350 *Society* 24, 180–90.
351
352 Katz RW (2010) Statistics of extremes in climate change. *Climate Change*, 100, 71–76.
353
354 Kharin, V.V. and Zwiers, F.W. (2004) Estimating Extremes in Transient Climate Change
355 Simulations. *Journal of Climate*, 18, 1156-1173. https://doi.org/10.1175/JCLI3320.1.
356
357 Korkmaz, M. C. (2015) Two-sided generalised Gumbel distribution with application to air
358 pollution data. *International Journal of Statistical Distributions and Applications*, 1 (1), 19-
359 26.
360
361 Küchenhoff, H. and Thamerus, M. (1996) Extreme value analysis of Munich pollution data.
362 *Environmental and Ecological Statistics*, 3 (2), 127-141.
363
364 Loick, N., Dixon, E., Abalos, D., Vallejo, A., Matthews, P., McGeough, K., Watson, C.,
365 Baggs, E., Cardenas, L.M. (2017) "Hot spots" of N and C impact nitric oxide, nitrous oxide
366 and nitrogen gas emissions from a UK grassland soil. Geoderma, 305 (2017) 336–345
367
368 Louro, A., Sawamoto, T., Chadwick, D., Pezzolla, D., Bol, R., Baez, D. and Cardenas, L.
369 (2013) Effect of slurry and ammonium nitrate application on greenhouse gas fluxes of a
370 grassland soil under atypical South West England weather conditions. *Agriculture,*
371 *Ecosystems and Environment*, 181, 1-11.
372

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 and Climate Extremes*, **18**, (December), Pages 44-54

McFadden, Daniel (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. and Chadwick, D. (2012) Greenhouse gas (GHG) emissions from soils amended with digestate derived from anaerobic treatment of food waste. *Rapid Communications in Mass Spectrometry*, 26, 2422-2430.

Sexto, B. M., Vaquera, H.H. and Arnold, B. C. (2013) Use of the Dagum distribution for modelling tropospheric Ozone levels, *Journal of Environmental Statistics*, 5 (5), 1-11.

Sheather, S.J. and Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

Singh, V.P. (1998) Generalized extreme value distribution. In *Entropy-Based Parameter Estimation in Hydrology*, vol. 30, pp169-183. Springer, Dordrecht, the Netherlands.

Figure 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.

Figure 2. Empirical relationship ($y' = a + B\ R^x$) between CO_2 flux data on the original scale and the corresponding values scaled according to the Box-Cox (1964) transformation with $\lambda = 0.1$

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

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

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

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

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	-	7.3	8.16
Total carbon	% of dry matter	38.4	38.6
C:N ratio	-	14.4	2.3

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)
η	data peak location parameter	GEV, Smith, 1986
ξ	data distribution shape parameter	GEV, Smith, 1986
α	scale parameter	GEV, Smith, 1986
Γ	Gamma function	Singh, 1998
u	Return period of peak level	Sexto et al. 2013

Table 3. Sample statistics for CO_2 flux data (kg CO_2-C ha^{-1} d^{-1}) over 47 d 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

Table 4. Mean plot values for CO_2 flux data (kg CO_2-C ha^{-1} d^{-1}) over 47 d 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*	Mean Box-Cox scale**	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

* Finney (1941) correction applied.

** Transformed back from Box-Cox scale using an empirical regression relationship, viz. $y' = a + B R^x$ where y = CO_2 flux and x = 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 d 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	0.01868	3.457	1.766

Table 6. Mean plot values for N_2O flux data (kg N_2O-N ha^{-1} d^{-1}) over 47 d 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*	Mean Box-Cox scale**	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	1.03767	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

* Finney (1941) correction applied.

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

NA = Not available, distribution did not fit.

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

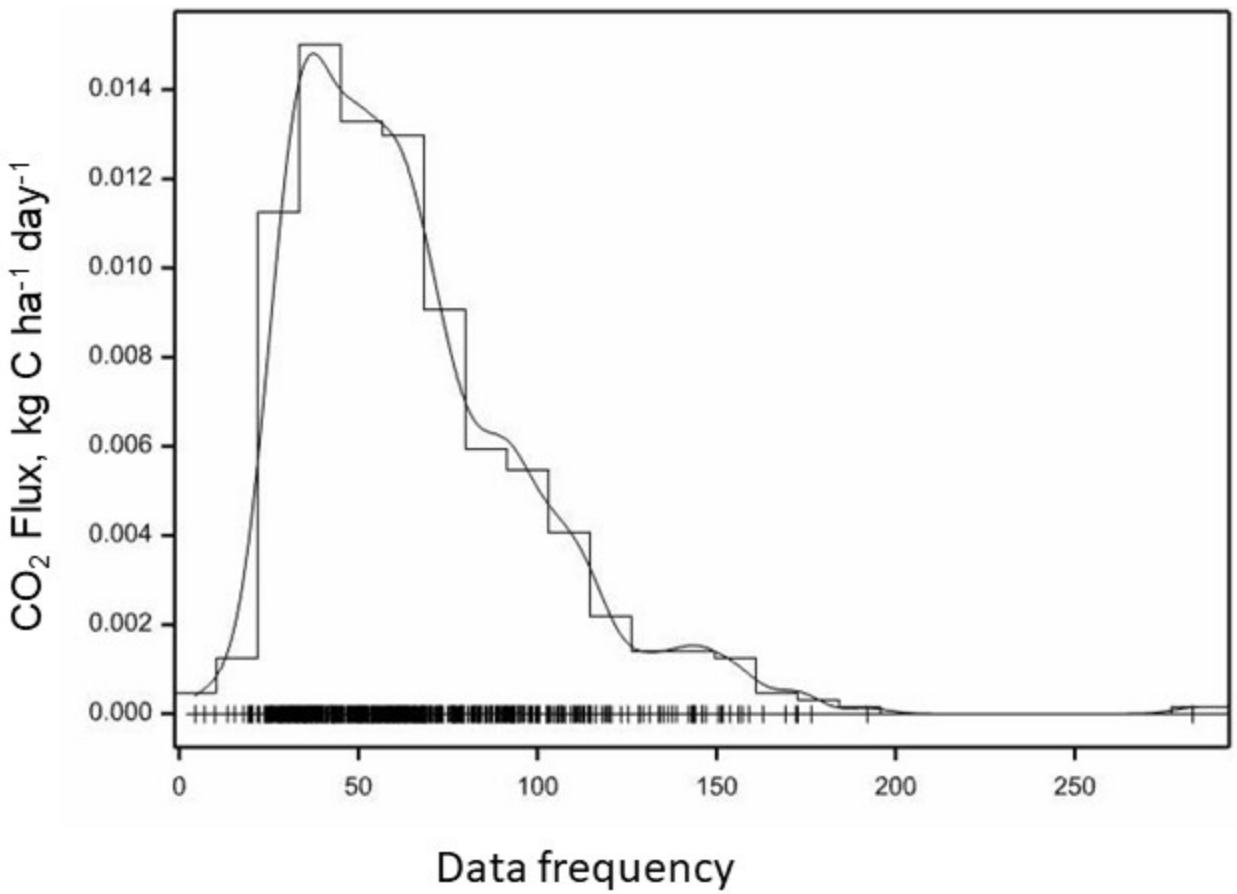
Treatment-Plot	Sample Mean	GEV	
		$\hat{\mu}$	$\hat{\sigma}$
DG-1	82.71	62.06	25.245
DG-2	85.67	88.84	50.949
DG-3	65.52	55.94	20.197
SL-1	50.56	52.35	18.672
SL-2	70.34	58.26	21.773
SL-3	73.01	70.68	25.432
CN-1	56.85	46.78	19.057
CN-2	58.19	57.63	21.213
CN-3	35.07	34.84	17.081
AN-1	70.24	76.73	42.851
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).

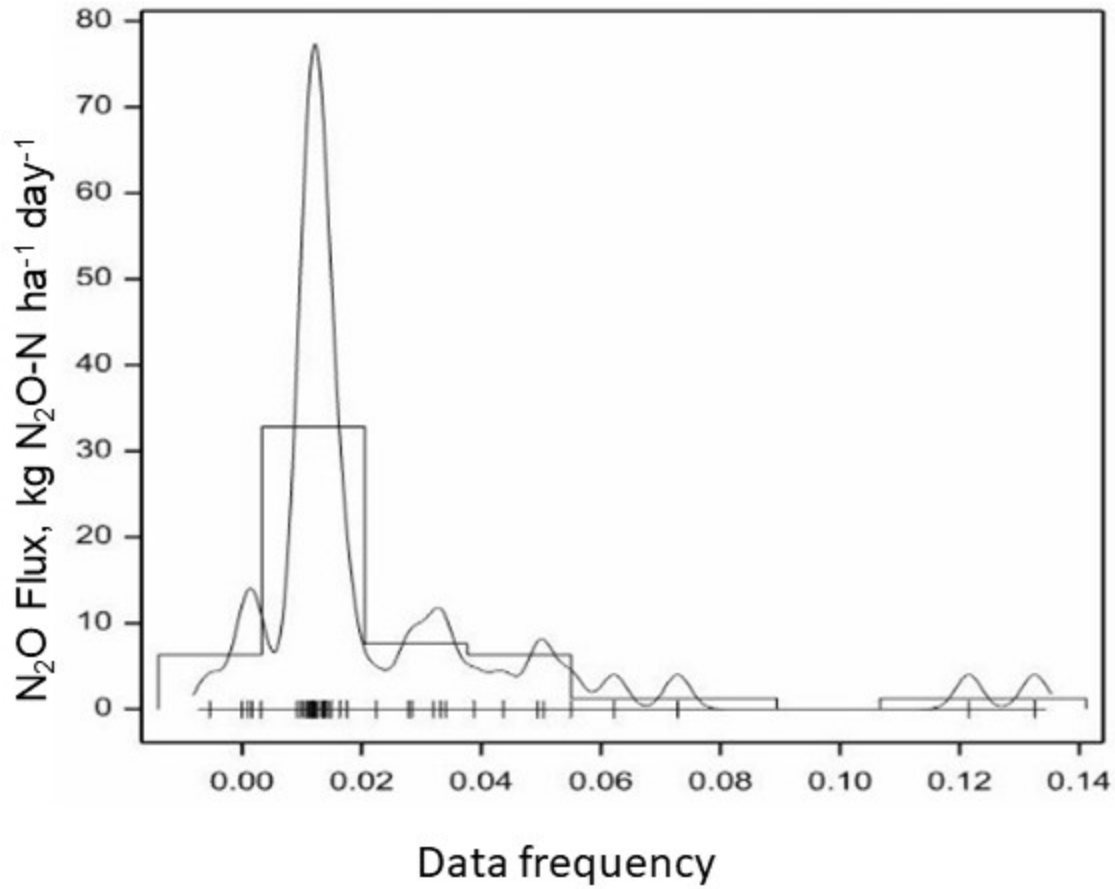
Treatment-Plot	Sample Mean	GEV	
		$\hat{\mu}$	$\hat{\sigma}$
DG-1	1.045	1.037	0.0124
DG-2	1.046	NA	NA
DG-3	1.036	1.038	0.0163
SL-1	1.031	NA	NA
SL-2	1.051	NA	NA
SL-3	1.038	1.047	0.0350
CN-1	1.032	1.031	0.0079
CN-2	1.033	1.034	0.0095
CN-3	1.026	NA	NA
AN-1	1.042	NA	NA
AN-2	1.033	1.037	0.0158
AN-3	1.039	1.038	0.0103
Overall	1.038	1.036	0.0127
		Net 0.01571	

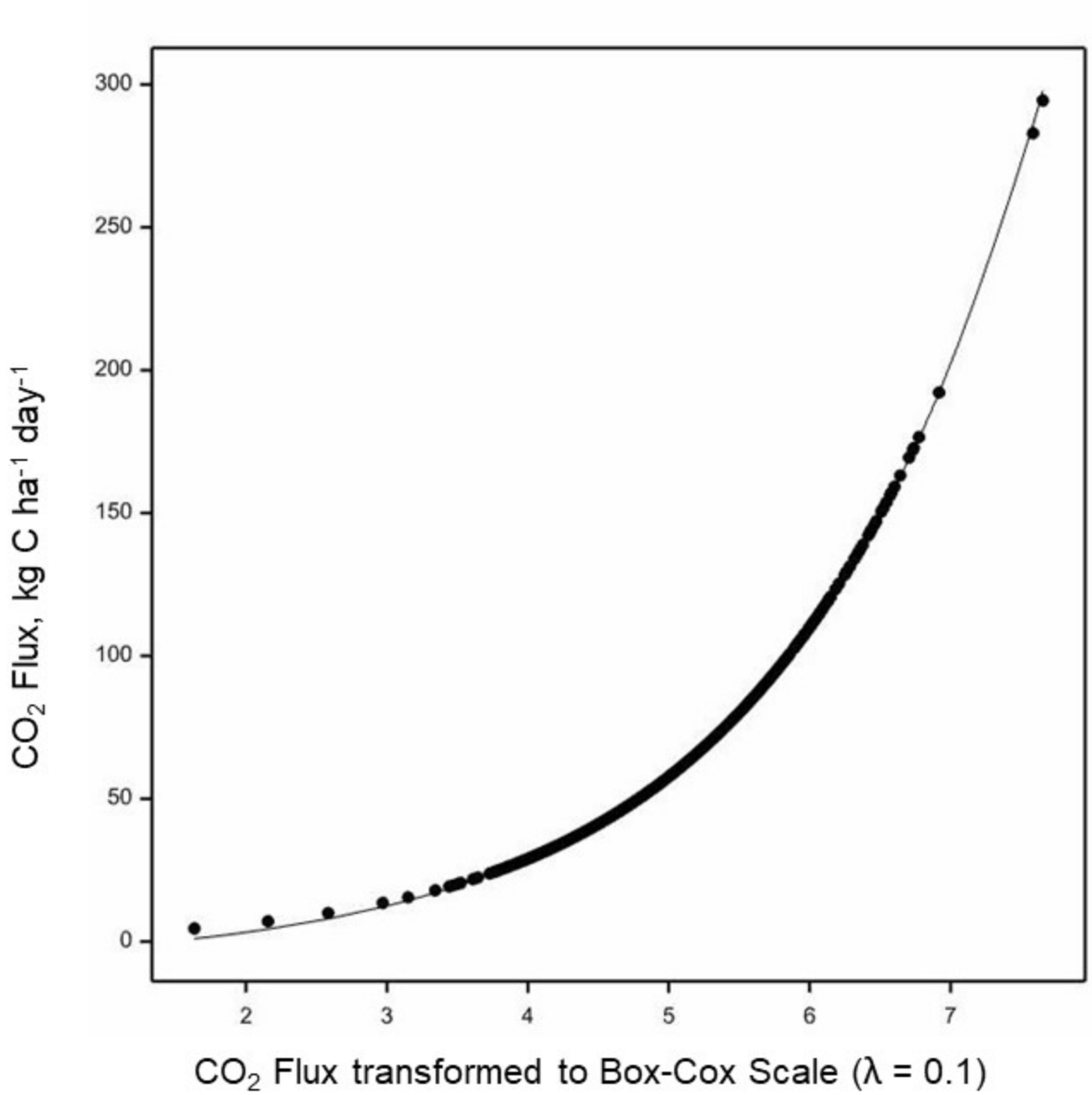
NA = Not available, distribution did not fit.

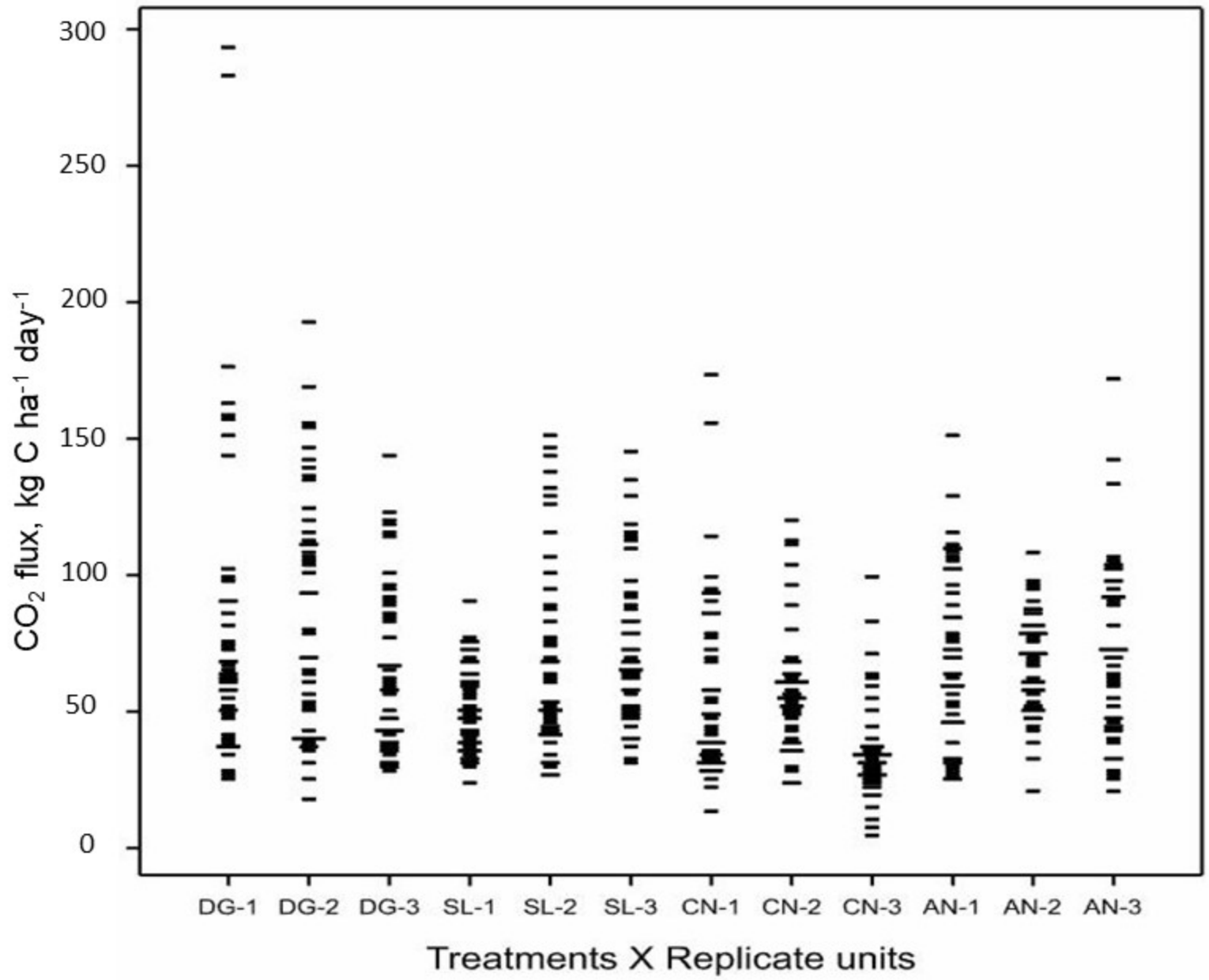
a.

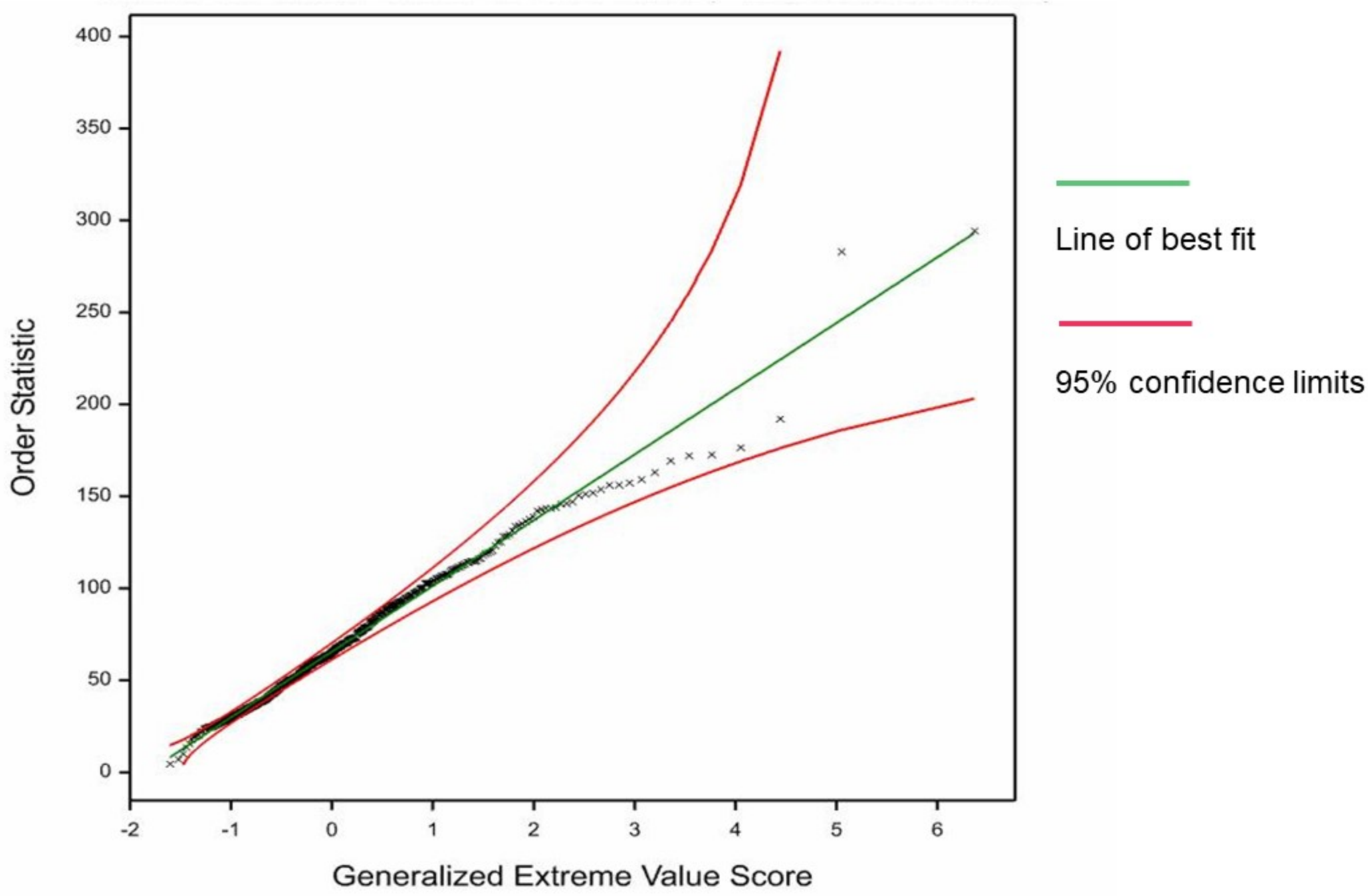

0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
CO_2 Flux, kg C ha^{-1} day^{-1}
0
50
100
150
200
250
Data frequency

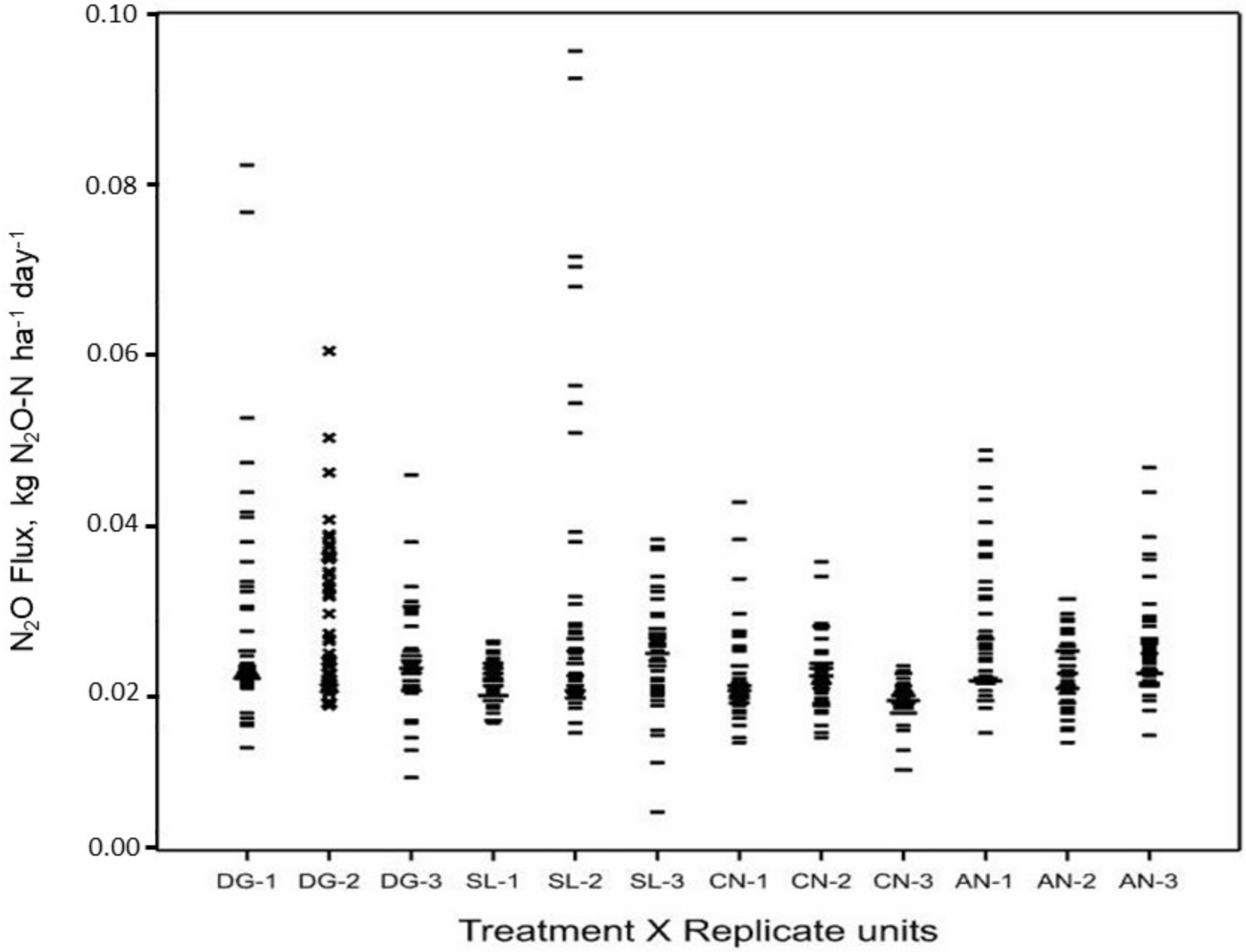
b.

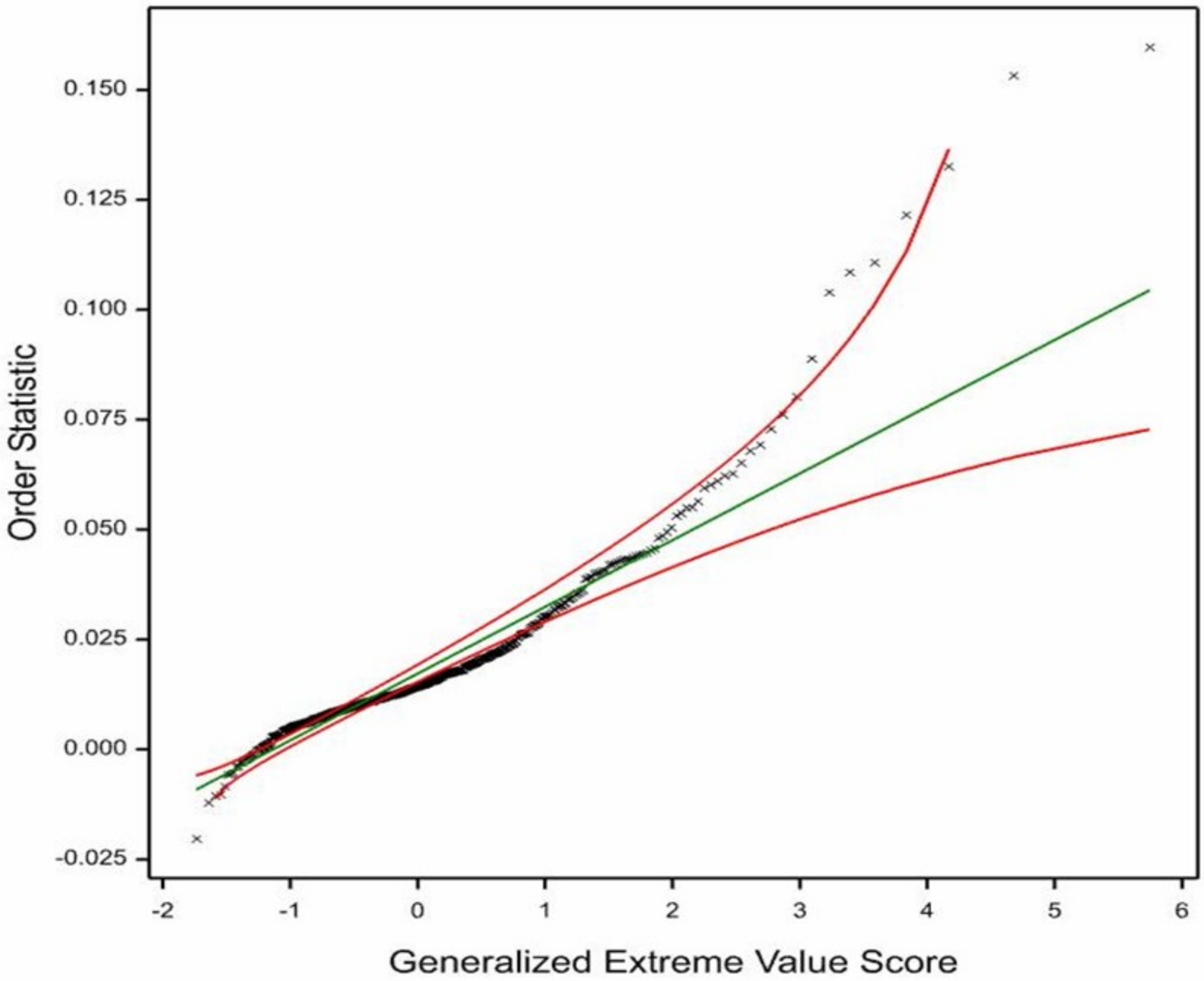

80
70
60
50
40
30
20
10
0
N_2O Flux, kg N_2O-N ha^{-1} day^{-1}
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Data frequency


300
250
200
150
100
50
0
2
3
4
5
6
7
CO_2 Flux, kg C ha^{-1} day^{-1}
CO_2 Flux transformed to Box-Cox Scale (λ = 0.1)


300
250
200
150
100
50
0
CO_2 flux, kg C ha^{-1} day^{-1}
DG-1
DG-2
DG-3
SL-1
SL-2
SL-3
CN-1
CN-2
CN-3
AN-1
AN-2
AN-3
Treatments X Replicate units


Order Statistic
400
350
300
250
200
150
100
50
0
-2
-1
0
1
2
3
4
5
6
Generalized Extreme Value Score
Line of best fit
95% confidence limits


0.10
0.08
0.06
0.04
0.02
0.00
N2O Flux, kg N2O-N ha-1 day-1
DG-1
DG-2
DG-3
SL-1
SL-2
SL-3
CN-1
CN-2
CN-3
AN-1
AN-2
AN-3
Treatment X Replicate units



Line of best fit

95% confidence limits

Highlights:

- Using sample means may be overestimating GHG fluxes
- GEV solves excessive skewness and kurtosis of greenhouse gas flux data
- Strategy ofoptions 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

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: