An Assessment of the Global Load of PFCAs

John Chamberlin, Anfernee Ragin, and Cody Walls

Dr. John Washington, US EPA Ecosystems Research Division

**Table of Contents**

Purpose… pg. 2

Executive Summary… pg. 3

Data Summary… pg. 5

Analysis… pg. 7

Results/Conclusion… pg. 15

Appendix… pg. 16

**List of Figures/Tables**

Figure 3.1 pg. 7

Figure 3.2 pg. 8

Figure 3.3 pg. 8

Figure 3.4 pg. 9

Figure 3.5 pg. 9

Figure 4.1 pg. 12

Table 4.a pg. 12

Table 4.b pg. 14

Table 4.c pg. 16

Figure 4.2 pg. 14

Figure 4.3 pg.14

**I. Introduction/Purpose**

Working with data collected by Dr. John Washington, a research chemist at the US EPA National Exposure Research Laboratory, we are evaluating the environmental distribution of perfluorocarboxylic acids (PFCAs) throughout the world. PFCAs are synthetic compounds with excellent hydrophobic and oil-repelling properties, which lends them to numerous industrial and consumer applications. Stain-resistant carpet, microwavable popcorn bags, and the manufacture of teflon all involve PFCAs in some form. In response to studies that suggested potential toxicity of PFCAs1, the chemical industry largely discontinued their use in favor of polymeric forms called fluorotelomer-based polymers (FTPs). FTPs are essentially chains of parallel PFCAs linked together on a chemical backbone. While FTPs do gradually degrade over time into PFCAs, the industry estimates a half life on the order of 1000 years. However, recent research has observed much more rapid degradation, with half-lives estimated as 33-112 years2. The current study is partially motivated by these findings.

The dataset provided by Dr. Washington consists of concentrations for nine different PFCA compounds from soil samples from 62 sites around the world. Soil samples were analyzed in triplicate and the concentrations reported in mean picograms per gram. The soil samples themselves were collected by colleagues and associates of Dr. Washington, who attempted to choose collection sites that were as undisturbed by human activity as possible. The objective of the analysis is to characterize the baseline distribution for three PFCA compounds - perflurooctanoic acid (PFOA), perflurodecanoic acid(PFDA), and perfluorododecanoic (PFDoDA) acid, hereafter referred to as C8, C10, and C12 respectively, based on the number of carbon atoms contained in each molecule.

Dispersal of PFCAs in the environment occurs through two distinct routes. The first is direct release and spread of PFCAs from initial manufacturing and deterioration of PFCA-containing products. The second is indirect, and proceeds through degradation of FTPs into their constituent PFCAs. FTPs experience a complex breakdown pathway which involves a volatile alcohol intermediate which can disperse throughout the atmosphere. Further degradation converts these into PFCAs, which accumulate on the surface of the earth. Assuming random dispersal of PFCAs generated by this process, distance effects are not relevant. Thus, the baseline values may be normally distributed. In contrast, sites which experience direct contamination should be log­-normally distributed because concentration decays rapidly with distance from source. Thus, identification of baseline samples will require estimation of the parameters of these two distributions.

**II. Executive Summary**

Perfluorinated Carboxylic Acids are ubiquitous in the environment. Analysis of their distribution in worldwide soil samples may inform estimates of the breakdown rate of fluorotelomer-based polymers, which are composed of chemical linkages of PFCAs. Breakdown of FTPs results in dispersal of PFCAs via gaseous intermediates, culminating in a world wide baseline level of contamination. PFCAs also disperse directly, such as from manufacturing centers and from the deterioration of consumer products. These two dispersal routes produce distinct distributions which combine to form the global distribution - the lowest values are from indirect contamination, the highest are from direct contamination, and somewhere in the middle the distributions contribute roughly the same amount to the measured contamination level.

Our analysis attempted to identify this middle point. To do so, we validated certain assumptions dictated by the mixed-distribution model, in particular the normality of both groups. The baseline distribution was initially presumed to be normally distributed. However, qq-plots revealed that a lognormal distribution is at least as effective a descriptor of the variation within the lower half of the data points. Thus, we used a mixture-model algorithm to estimate the parameters of the two best-fitting log-normal distributions. The model finds a critical point of approximately 270 pg/g of PFOA, which means that this is the point which is equally likely to originate from either of the best-fitting distributions.

**III. Data Summary**

1. **Initial Data**

Initially, we were given 62 data points that represent locations spanning all seven continents and 38 countries/states/provinces. All though the data spans all continents, more than half the data points – 33 to be exact – are located in North America, 26 being in the United States. This implies that the PFCA is concentrated consistently in the U.S. than in other countries let alone continents.

The main dataset contains 12 variables for each of the 62 data points. The first variable, Sample ID, gives a brief abbreviation for the data point based on the location. Each sample ID is 4 characters – the first 2 abbreviate the continent that the location is in, while the last two is the number corresponding to the continent’s frequency count. For example, there are 33 data points in North America, whose abbreviation is NA. The sample IDs for these data points are labeled NA01 through NA33. There are 10 data points in Europe (EU), so the sample IDs are labeled EU01 to EU10. The second variable, Location, lists the city, state/province (if in U.S. or Canada), and continent of each data point. The third variable lists the latitude and longitude of the location. The 4th-12th variables represent the values of each PFCA from C6 through C14. Each variable is labeled based on the name of the chemical: PFHxA, PFHpA, PFOA, PFNA, PFDA, PFUnDA, PFDoDA, PFTrDA, and PFTeDA.

At first the rows of each raw dataset, including the main, were separated by continent, and some rows were used to list a continent or C6-C14, while the other variables were left empty. In order to read the data effectively using software, we organized the data to our preference by separating only the necessary variables – sample IDs, location, latitude, longitude and each PFCA into variables on a clean, separate table. Then we used statistical software to read this table and separate the C8, C10, and C12 values for further analysis.

1. **Primary Response Variables**

The three variables of interest for our analysis are C8, C10, and C12, which are PFOA, PFDA, and PFDoDA respectively. Each value is the mean of three soil samples in the given location, and is rounded to the nearest hundredth of a decimal. The goal of our analysis is to differentiate between sample sites with direct, local effects of PFCA exposure and those without any direct effects. Thus, the response variable is the concentration in picograms per gram for each of three PFCA compounds (C8, C10, and C12). To assess this will require the identification of a cutoff point where sites can be determined to be no longer experiencing the effects of direct exposure to PFCA compounds.

1. **Exploratory Data Analysis**

Figure 3.1 is a display of the distribution of concentrations of chemicals C8, C10, and C12 divided between the seven continents, which provides a reference as to the difference in the scale of concentrations between the three chemicals of interest. Additionally, the stark difference in concentrations between the more and less “developed” continents can also be seen in this figure. North America, Europe and Asia have the highest concentrations of PFCAs, while the levels in Africa, Antarctica and South America are not significantly high.

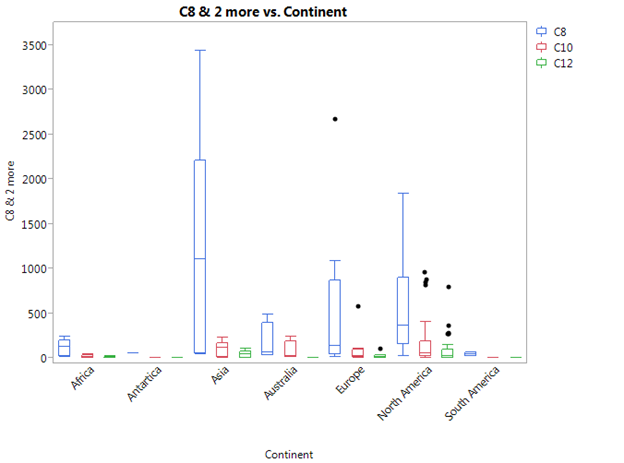


              Figure 3.1. Boxplots of C8,C10,and C12 by continent

Characterize Missing Data

For the compound C8, there are no values described as <LOQ or <MDL, and consequently no missing values. For the compound C10, there are two values described as <MDL, and three values described as <LOQ, for a total of five missing values. For the compound C12, there are thirteen values described as <MDL and fourteen values described as <LOQ for a total of twenty seven missing values.

In order to deal with the missing data points, we will need to determine what the minimum level of detection for the methods used in collection of this data. Because this was not provided for us by Dr. Washington, we will instead take the lowest value observed for compounds C10 and C12 and then divide them by two. The value that results will be used in the place of zero for those missing data points, so that we may log transform and carry out other statistical procedures unhindered by missing data. So, this results in the values being:

Lowest Value Observed for C10: 2.58 / 2 = 1.29

Lowest Value Observed for C12: 9.65 / 2 = 4.825

Thus, all the five missing values for C10 will be replaced with 1.29 and the twenty-seven missing values for C12 will be replaced with 4.825.

Data Transformation

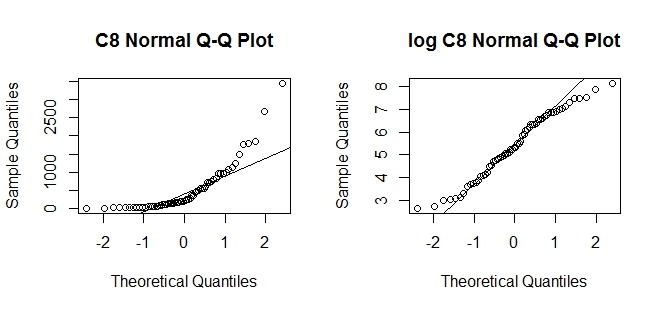
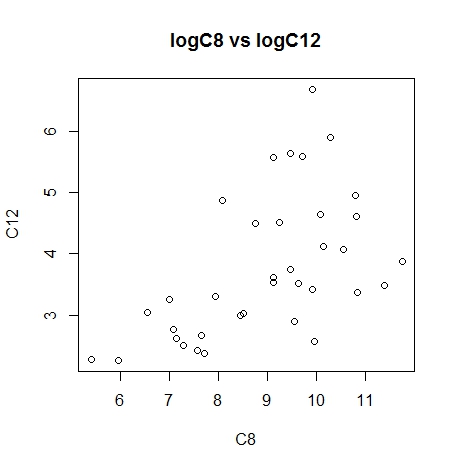
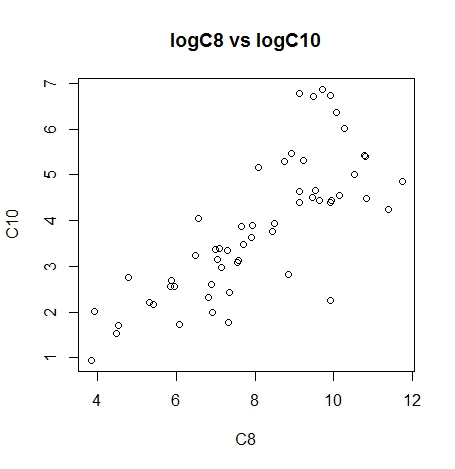
****

Figure 3.2 and 3.3. C8 and log(C8) Normal Q-Q Plot

Plots Showing Relationships of Interest



Figures 3.4 and 3.5 Correlation between L8 and L10, L8 and L12 respectively

It is worth noting that the log transformed C8 and C10 values have a correlation coefficient of r = .811 and the log transformed C8 and C12 values have r = .704. The fact that these values are highly correlated proves to be useful for making insights about C10 and C12 based on C8. The problem of missing data in those variables can be somewhat mitigated by using information relevant to C8.

Dealing with Missing Data and Other Questions of Interest

In order to deal with the missing data points, we determined the minimum level of detection for the methods used in collection of this data. Because this was not provided for us by Dr. Washington, we instead took the lowest value observed for compounds C10 and C12 and then divided them by two. The value that results was then used in the place of zero for those missing data points, in order that we may log transform or carry out any other statistical procedures unhindered by missing data.

Log transformation

After replacing <LOQ values and viewing the distributions of C8, C10, and C12 concentrations, it was clear that the data was extremely spread out, and could benefit from log transformation. We created three new variables L8, L10, and L12 that contain the corresponding log transformed values for each compound. We summed these three variables (L8-12) to create a logsum indicator variable that could serve to separate the data into two subsets. However, analysis of this variable sorted in ascending order or by continent did not reveal two distinct subsets or any natural cutoff points between certain “low exposure” and “high exposure” locations. This lead us to believe that perhaps there were two overlapping distributions for the low exposure and high exposure locations.

**IV. Analysis**

Validating Assumptions

If there are not two groups of data or there are actually more than two groups, such as baseline, medium exposure and elevated exposure - our analysis will be flawed. However, from our initial exploration of the data, it appears that there is, in fact, two fairly distinct groups so our assumption is valid. The problem in this case becomes how to adequately define and determine those two groups. Dr. Washington worked on the assumption that the “baseline” data was log normally distributed around some value, so we will assume this as well.

Furthermore, we must also make the assumption that once the data with elevated concentrations has been removed from consideration, the remaining data has not been affected by any sort of local, direct source. Because we have no way of verifying that, this assumption must be accepted as true also. Finally, we must assume that the locations from which we have non-elevated concentrations are good estimates of the true concentration level of the rest of the continents or even their surrounding area. Once again, because we have no way of scientifically verifying whether the sample is representative, we must accept this assumption on the basis of Dr. Washington’s claims.

Mixture Model

We used software to fit two separate normal distributions to the log-transformed C8 data. The R package ‘mixtools’ allows for the estimation of the parameters of the two best-fitting normal distributions for the data. The function tries out different values for the mean and shape of each distribution until it reaches two normal distributions that approximately fit the data. After estimating the parameters of the two distributions, we can find the midpoint value at which both distributions are an equally likely source for that sample. The results of the algorithm performed on the log transformed C8 data are shown in Figure 4.1 below.

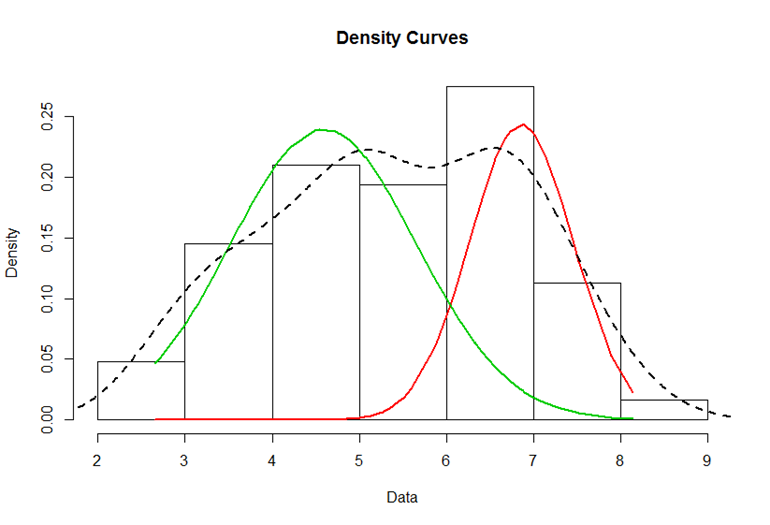
****

Figure 4.1. Density Curves of Estimated Log Normal C8 Distributions

After using the Expectation Maximum algorithm for two normal distributions from the R mixtools package, we were given estimates for the means and standard deviations of the “Baseline” and “Elevated” groups of log(C8) concentrations :

|  |  |  |
| --- | --- | --- |
|  | Baseline | Elevated |
| Mean Estimate | 4.590 | 6.857 |
| Standard Deviation Estimate | 1.066 | 0.591 |

Table 4.a. Mean and Standard Deviation Estimates from EM algorithm

This assumes that both the Baseline and Elevated C8 groups are log normally distributed. Using the standard z approximation table, it is basically a matter of trial and error to find the point which has equal probability of being in each distribution. The result of that trial and error is a value in terms of log(C8) equal to 6.048, which will serve as the cutoff between the two log normal distributions. When exponentiated to C8 scale, we arrive at a cutoff value of 423.18. When applied to the C8 data, we were able to obtain a Baseline group of 39 values and an Elevated group of 23 values.

Simulation of 10,000 Mixture Models

While these initial analyses yielded explicit results, we were concerned that our method was not robust enough. We needed to demonstrate that the mixture model gave a significantly better description of the data than a single distribution alone. So in order to increase our confidence in the estimates for the means and standard deviations of the two log-normal distributions, we conducted a simulation. Depending on the random seed generated, we had occasionally obtained different results using our EM algorithm. However, by running the algorithm with 10000 different random seeds, we were able to compare our initial results to the average result of the 10000 simulations.

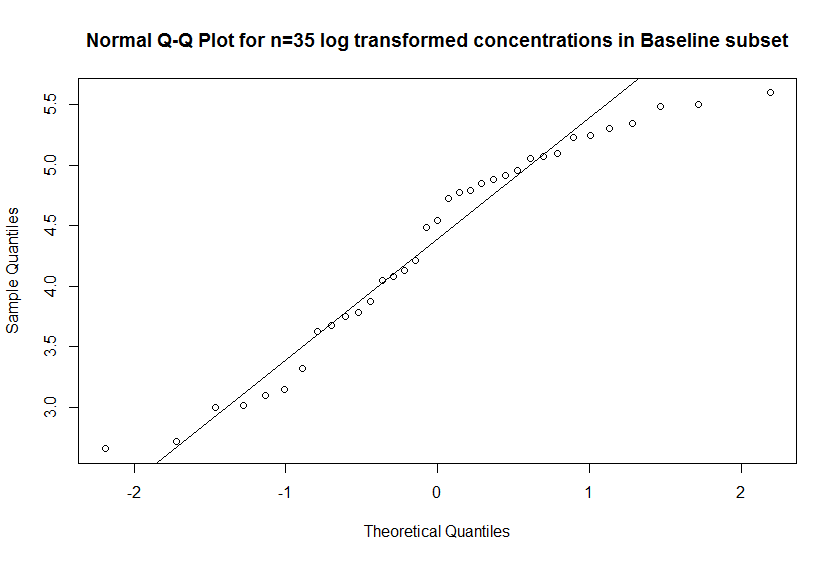
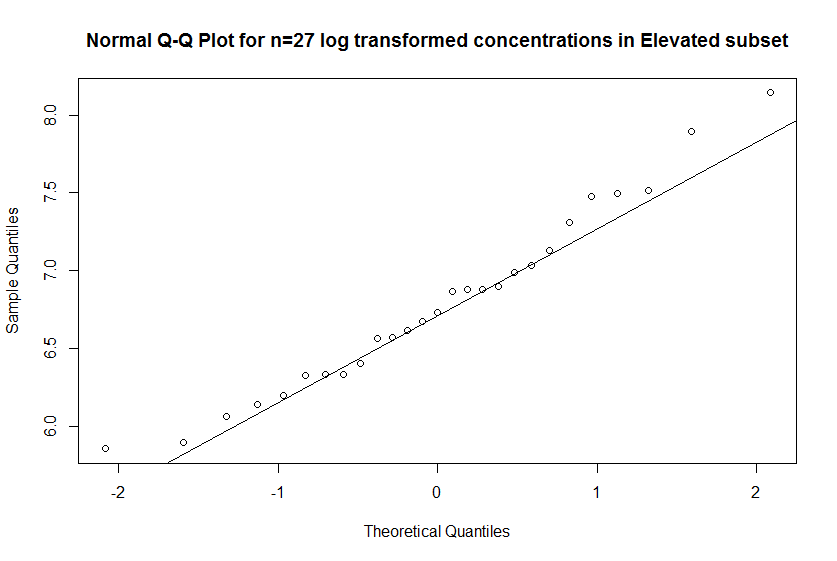
The result reported above seemed to be the most common output of the normal distribution parameters, though over the course of the 10,000 simulations, a few other outputs were obtained. We stored the 10,000 responses for mu1, mu2, sigma1, and sigma2 in their own respective vectors, and used the mean value for each of those four variables to parameterize another set of log normal distributions for C8. Average estimates for the means and standard deviations of the “Baseline” and “Elevated” groups of log(C8) concentrations are given in Table 4.b below:

|  |  |  |
| --- | --- | --- |
|  | Baseline | Elevated |
| Mean Estimate | 4.549 | 6.447 |
| Standard Deviation Estimate | 0.927 | 0.730 |

Table 4.b. Mean and Standard Deviation Estimates from EM Simulation

Using the average output of the EM algorithm, we find a value of 5.611 in terms of log(C8), which will serve as the cutoff between the two log normal distributions. In C8 scale, this is a concentration of 273.42 pg/g. This lower cutoff value resulting from the simulation study reduces the size of the Baseline subset to 35 locations, with an Elevated subset of 27 locations.

The qq-plots of these two subsets are shown below in Figures 4.2 and 4.3, indicating the relative normality of the two log transformed groups of C8 concentrations.

****

Figures 4.2 and 4.3. QQ-Plots for Baseline and Elevated subsets of L8 data

Though the groupings above seem adequate, upon closer inspection of the remaining locations in the baseline data, there appears to be a substantial outlier. The location with Sample.ID NA12 contains the highest value for all three variables of interest out of all of the 35 Baseline locations. Considering the C8 value for NA12 is so close to the new cutoff (270.2 < 273.2) and the C10 and C12 values for NA12 are both among the ten highest in the dataset, it seems reasonable to amend the cutoff point to C8 = 270.

This resulting subset of 34 locations from our original dataset of 62 points borders on not being log-normally distributed for C8 when the Shapiro-Wilkes test for normality is performed (p-value = .03183), though we believe that this grouping will give us the most appropriate estimate for the Baseline values. C10 and C12 are similarly not log normally distributed for the Baseline subset, though we would not expect them to be in lieu of the replaced <LOD values. Despite this, we will use the 34 baseline locations to calculate log scale point estimates for the population mean of each Baseline group and form confidence intervals centered around those values for our estimates of the global PFCA load.

V. Primary Results

Our primary objective is to distinguish the locations with high levels of PFCA exposure, in particular the compounds C8, C10 and C12, from locations with low levels of exposure. We named these two subsets of the data the ‘Baseline’ and ‘Elevated’ groups. After log transforming our data, we found that the best way to accomplish the task of separating the data into two groups was to treat each group as having a log-normal distribution, and attempt to estimate a midpoint between those two distributions. The bimodally distributed log(C8) data allowed us to use the Expectation Maximum algorithm to estimate a mean and standard deviation for two normal distributions within log(C8). After extensive repetition of our method and the exclusion of a value that was on the fence of our cutoff, we determined the value of C8 = 270pg/g as the dividing line between the Baseline and Elevated groups.

Given the skewed nature of the data in general and the uncertainty of some values in C10 and C12, it is not possible for us to divide the C10 and C12 with any comparable method.This in conjunction with the high correlation found between our log transformed response variables (Fig. 3.2,3.3) makes it reasonable for us to use the same grouping of Baseline locations for C8, C10, and C12. We calculated the point estimates for the log transformed PFCA concentrations within the Baseline grouping of 34 locations and followed up by creating 95% confidence intervals in log scale and after being exponentiated for C8, C10, and C12. All of the intervals were for the time being constructed using the T-distribution and are shown in Table 4.c below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | C8 | C10 | C12 |
| Log-scale 95% CI | (4.0,4.61) | (1.95,2.75) | (1.74,2.13) |
| 95% CI | (54.8,100.4) | (7.04,15.62) | (5.68,8.42) |

Table 4.c Confidence Interval in terms of PFCA concentration and log(PFCA Concentration)

VI. Conclusions

We have determined an approximate global load for the concentration of key variables C8, C10, and C12. We used confidence intervals constructed on our log transformed baseline subset of each of these variables as a means of placing bounds on our point estimates. While different confidence interval methods still need to be considered for estimating the global load, these results seem to be appropriate and useful. It is our hope that these global load estimates will give the EPA a more complete understanding of the rate at which FTPs degrade. Over the next few decades, governments may then be able to take action to stop the spread of these compounds if necessary.

**Citations**

1. Perfluoroalkyl acids: what is the evidence telling us? Betts KS.  Environ. Health Perspect. 2007 115 (5): A250–6. doi:10.1289/ehp.115-a250. PMC 1867999. PMID 17520044.
2. Decades-Scale Degradation of Commercial, Side-Chain, Fluorotelomer-Based Polymers in Soils and Water. John W. Washington, Thomas M. Jenkins, Keegan Rankin, and Jonathan E. Naile

Environmental Science & Technology 2015 49 (2), 915-923. DOI: 10.1021/es504347u