Chapter 8
Uncertainty Quantification for Entity-Scale Greenhouse Gas Emissions

Authors:

F. Jay Breidt, Colorado State University and NORC at the University of Chicago (Lead Author)
Stephen M. Ogle, Colorado State University

Suggested chapter citation: Breidt, F. J., and S.M. Ogle. 2024. Chapter 8: Uncertainty quantification of greenhouse gas emissions. In Hanson, W.L., C. Itle, K. Edquist. (eds.). *Quantifying greenhouse gas fluxes in agriculture and forestry: Methods for entity-scale inventory*. Technical Bulletin Number 1939, 2nd edition. Washington, DC: U.S. Department of Agriculture, Office of the Chief Economist.

Table of Contents

Chapter 8 Uncertainty Quantification for Entity-Scale Greenhouse Gas Emissions 8-1

8.1 Introduction 8-5

8.1.1 Overview of Methods for Predicting GHG Emissions 8-5

8.1.2 Decision Tree for Classifying Emission Source Methods for UQ 8-6

8.1.3 Organization of the Chapter 8-7

8.2 Overview of UQ 8-7

8.2.1 Sources of Uncertainty in Entity-Scale GHG Prediction 8-7

8.2.2 UQ via PDFs 8-8

8.2.3 General Principle of Propagating Uncertainty Via Monte Carlo 8-8

8.2.4 Recommendations for Summarizing Monte Carlo Output 8-10

8.2.5 Numerical Example of Monte Carlo Analysis 8-12

8.2.6 Special Case: Right-Triangular Distribution 8-15

8.3 Step-by-Step Guidance for UQ 8-16

8.3.1 Explicit Model-Based Methods 8-16

8.3.2 Explicit Measurement-Based Methods 8-17

8.3.3 Implicit Model-Based Methods 8-19

8.4 Extension of Monte Carlo for Unknown Activity Data Inputs 8-20

List of Figures

Figure 8-1. Decision Tree to Choose the Type of Method for a Source Category 8-7

Figure 8-2. UQ via Monte Carlo Analysis 8-9

Figure 8-3. Histogram From M = 10,000 Monte Carlo Draws From a Normal Distribution (Curved Dashed Line), with True Percentiles Plus Estimates and Confidence Intervals 8-13

Figure 8-4. Relative Uncertainty for Total Emissions, Measured as Percent Coefficient of Variation, Decreases as the Number of Entities in the Sum Increases, Provided Those Entities Do Not Have Perfectly Correlated Emission Factors 8-15

Figure 8-5. Right-Triangular PDF for GHG Emission With Known Activity, a, and Lower Bound of $\mathbf{1} - \boldsymbol{\alpha}\mathbf{100}\%$ Prediction Interval 8-16

Acronyms, Chemical Formulae, and Units

CO_2	carbon dioxide
dbh	diameter at breast height
GHG	greenhouse gas
LME	linear mixed effect
MVN	multivariate normal
N_2O	nitrous oxide
PDF	probability density function
PSU	primary sampling unit
UQ	uncertainty quantification

8. Uncertainty Quantification for Entity-Scale Greenhouse Gas Emissions

8.1 Introduction

This chapter provides an overview of options to quantify uncertainty for the emissions estimation methods provided in previous chapters of this report.

8.1.1 Overview of Methods for Predicting GHG Emissions

If greenhouse gas (GHG) emissions were measured at the entity scale, the only uncertainty would be due to the measurement process. But, in nearly all cases, the emissions are instead estimated by calculation methods. These methods vary in complexity, but all are functions of activity data inputs and emission factors.

- The simplest way to predict a GHG emission would be to multiply a known entity-scale activity data input by an entity-scale emission factor or set of factors. This is possible with some methods in this report; in those cases, the uncertainty in emission factors can be quantified and is provided in the description of the method. Examples include liming and carbon dioxide (CO_2) emissions, indirect soil nitrous oxide (N_2O) emissions, and non-CO_2 emissions from field burning of agricultural residues.
- The most complex methods described in this report involve models with many parameters that represent biogeochemical processes; for these methods, it is not feasible to derive uncertainty in the individual parameters. Uncertainty is instead quantified based on comparisons of model-based predictions to field measurements. Examples include cropland and grassland soil carbon stock changes and direct soil N_2O emissions, which are predicted with the DayCent ecosystem model.

Uncertainty quantification (UQ) in entity-scale GHG prediction is the formal process of describing the likelihood of different possible emissions, given what is known and what is unknown at the entity scale. In this report, the activity data inputs are assumed to be known, without uncertainty, at the entity scale based on the operator's knowledge about management of the system (i.e., assumed to be certain). Extensions to unknown activity data inputs are briefly discussed in section 8.4 for cases where the operator is not sure about the management activity.

Though activity data inputs are typically known without uncertainty at the entity scale, GHG emissions remain uncertain because they are not measured, and because they are determined by many factors that the prediction method does not fully capture. For example, suppose activity is measured by the size of a herd of cows. But cows naturally vary in their physiology due to breed, gender, age, and other factors, and this variation is affected by management practices and environmental factors (e.g., weather, pasture, or range conditions). Unless all these effects are incorporated into a perfect scientific model, the GHG emissions from this herd remain uncertain. In general, uncertainty in this report arises from uncertain emission factors: that is, it arises because the methods do not address all the relevant, naturally varying effects that determine the conversion of entity-scale activity to GHG emissions.

In this report, uncertainty in GHG emissions is quantified via a probability density function (PDF), described in further detail in section 8.2.2. Uncertainty in the emission factors, which is also based on PDFs for the factors, needs to be propagated through the method to determine the final PDF of

GHG emissions for the entity. One standard approach for propagating uncertainty is a Monte Carlo analysis, described in section 8.2.3.

8.1.2 Decision Tree for Classifying Emission Source Methods for UQ

The complexity of the Monte Carlo analysis for propagation of uncertainty is determined in large part by the method. Figure 8-1 presents two ways methods are classified in this report.

First, methods are either model-based or measurement-based:

- For **model-based** methods, uncertainty at the entity scale is fully described by PDFs for emission factors (available elsewhere in this report), with no entity-scale measurements required to determine the PDF. Examples of model-based methods include liming and urea CO_2, soil N_2O and soil carbon methods.
- In **measurement-based** methods, PDFs for emission factors and parameters are estimated from measurements at the entity scale (typically from a sample), and the resulting PDFs introduce uncertainty into the emissions estimate. As an example, a random sample of trees on a woodlot could have its volume characteristics measured to represent the entire woodlot and the growth over time, resulting in a PDF as described in the methods for woody biomass carbon stocks for cropland and grazing land.

Second, model-based methods are either implicit or explicit:

- **"Implicit"** means there is no PDF directly on model parameters, possibly due to the number of parameters or the complexity of the model. (It is theoretically possible to quantify the uncertainty in parameters for a complex model as a joint probability distribution, which would then be classified as an explicit method, but this is not the case for the complex methods included in this report). Implicit methods rely on an empirical method (a statistical model) for comparing measurements to model outputs. The implicit method should also control independent variables (covariates), that may explain some of the uncertainty in GHG emissions, and correlations (e.g., spatial autocorrelation), induced by the design of the studies performed to obtain the measurements. Examples of implicit methods include soil carbon stock changes and direct soil N_2O emissions, which are predicted with the DayCent ecosystem model.
- For **explicit** methods, PDFs are derived directly on parameters, which are typically emission factors; for those cases, this report provides the PDFs with each source category. Explicit methods usually have relatively few parameters and relatively simple model structure. Examples include liming and CO_2 emissions, indirect soil N_2O emissions, and non-CO_2 emissions from field burning of agricultural residues.

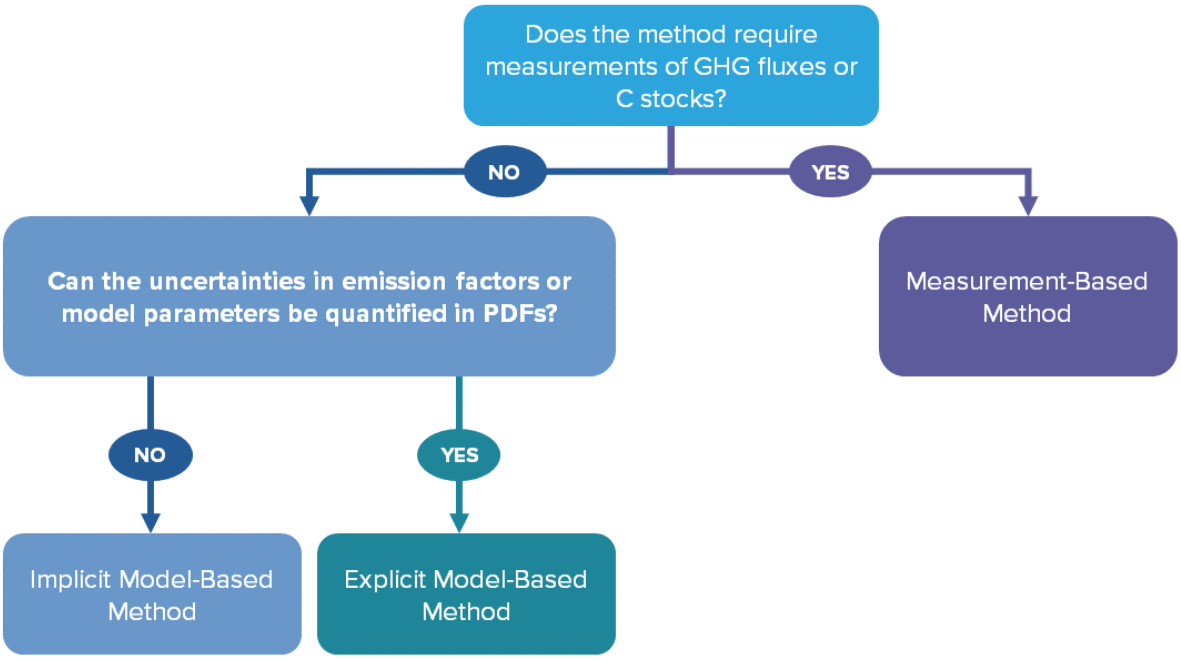

Figure 8-1. Decision Tree to Choose the Type of Method for a Source Category (See Section 8.3 for Error Propagation Methods for Each Type)

8.1.3 Organization of the Chapter

- Section 8.2 gives an overview of UQ, including general principles for describing uncertainty via PDFs, propagating uncertainty via Monte Carlo methods, summarizing Monte Carlo output, and interpreting the summaries.
- Section 8.3 provides step-by-step guidance for UQ with explicit model-based methods (section 8.3.1), explicit measurement-based methods (section 8.3.2), and implicit model-based methods (section 8.3.3).
- Section 8.4 describes extensions of the Monte Carlo analysis for unknown activity data inputs.

8.2 Overview of UQ

8.2.1 Sources of Uncertainty in Entity-Scale GHG Prediction

Suppose $\mu(a, f)$ is the true entity-scale GHG emission given known activity data inputs a and known emission factors f. Let $m(A, F)$ denote the GHG prediction method output for uncertain activity data inputs A and unknown emission factors F. Then the difference between the prediction from the method using these unknown inputs and the true GHG emission can be written as

$$m(A,F) - \mu(a,f) = \{m(A,F) - m(a,F)\} + \{m(a,F) - m(a,f)\} + \{m(a,f) - \mu(a,f)\}.$$

- The first term, $m(A,F) - m(a,F)$, is due to unknown activity data inputs and is assumed to be zero in this report. (See section 8.4.)
- The second term, $m(a,F) - m(a,f)$, is due to uncertain emission factors, the dominant source of uncertainty for most sources in this report. Uncertainty due to uncertain emission

factors is quantified by creating PDFs and using Monte Carlo analysis to propagate the uncertainty through the method to the GHG emission.

- The last term, $m(a, f) - \mu(a, f)$, is model uncertainty due to misspecification (e.g., incompleteness) of the scientific model. In this report, the explicit methods focus only on the dominant sources of uncertainty given current scientific understanding; they do not include model uncertainty. The implicit method does include model uncertainty, because it is an empirical method that compares model predictions to emissions observations.

8.2.2 UQ via PDFs

For this report, uncertainty in a generic quantity Y is described with a PDF $p_Y(y)$, which is a function that takes a possible value y of Y and returns a nonnegative "probability density." This probability density is not itself a probability, but the integral of the PDF over a specified interval of values from a to b is the probability that the random quantity Y takes on a value between a and b:

$$\mathrm{P}[a \leq Y \leq b] = \int_a^b p_Y(y)\, dy.$$

PDFs are provided for the emission factors and other calculation variables for some of the source categories, and PDFs for the emission estimates are generated using the UQ methods for the source categories.

A simple example of UQ for an explicit, model-based method that predicts GHG emissions as $G = m(A, F)$ would be $G = m(A, F) = A \times F$, where A represents one or more entity-level activity data inputs, and F represents one or more entity-level emission factors. The entity-level activity data are known ($A = a$ for some specified value(s), a). The entity-level emission factors are unknown, and their uncertainty is described by one or more given PDFs, $p_F(f)$, in the methods report. For simplicity, the report considers a single activity data input and emission factor.

Because the entity-level emissions depend on at least one unknown input, G is unknown and its uncertainty is described by a PDF, $p_G(g)$. This PDF for G is produced by "propagating" the uncertainty in the emission factor through the method. For this report, the error is propagated using a Monte Carlo approach, as discussed in section 8.2.3.

8.2.3 General Principle of Propagating Uncertainty Via Monte Carlo

Monte Carlo analysis is a principled and straightforward approach to uncertainty propagation. It generates a large number of replicates (e.g. 10,000 replicates) of the possible GHG emissions. This analysis is typically performed using statistical software. Random numbers are selected for the emission factors based on the PDF and used with the activity data to estimate GHG emissions. This process is replicated many times and then the GHG emissions PDF and its properties (e.g., mean, variance, and the median and other percentiles) are estimated using statistical techniques.

Figure 8-2 presents a generalized process, see sections 8.3.1 through 8.3.3 for specific steps based on this method.

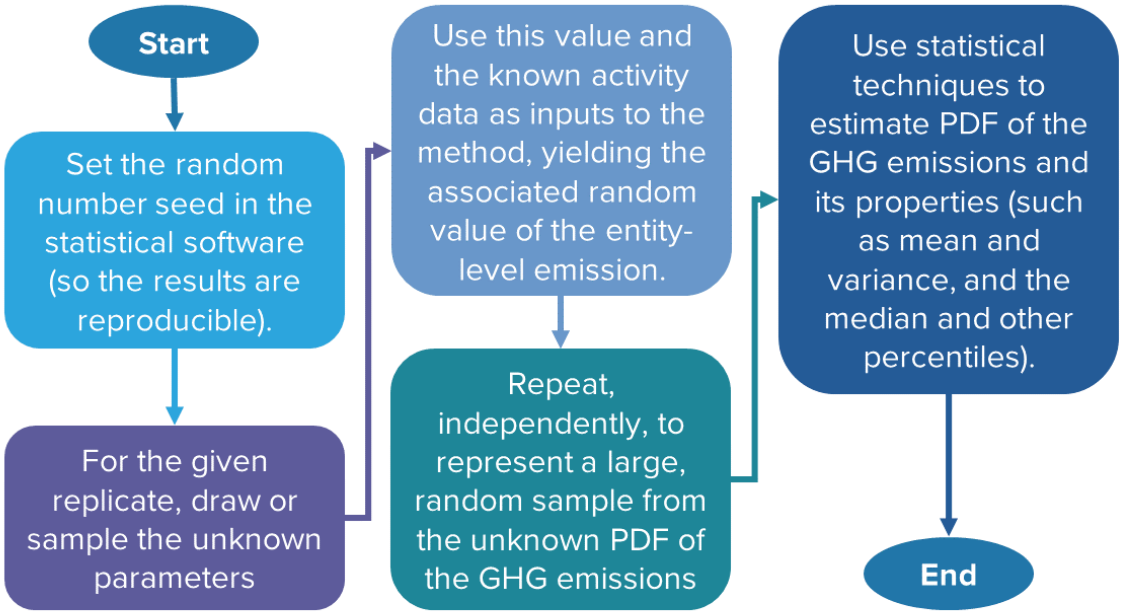

Figure 8-2. UQ via Monte Carlo Analysis

Because it relies on a random sample, the Monte Carlo analysis introduces a new source of uncertainty, which has nothing to do with the original uncertainty in GHG emissions. However, the Monte Carlo uncertainty can be made as small as desired in approximating the unknown PDF $p_G(g)$ because the sample size M can be made large, limited only by computing time.

Increasing M does not decrease the uncertainty about GHG emissions, but simply gives a more precise estimate of the PDF for the GHG emissions. Uncertainty about the entity-level GHG emissions would only be reduced by directly measuring entity-level GHG emissions, by measuring or otherwise reducing uncertainty about entity-level emission factors, or by improving the scientific model.

The Monte Carlo approach has several strengths. First, it is transparent because it does not involve complicated mathematical derivations. Second, it is readily transferable across methods, as it is a general-purpose approach, regardless of the complexity of the method. Third, it is easily adaptable as new information becomes available. For example, if a new source of uncertainty in the method is identified and its PDF is developed, or if the PDF is refined for a known source of uncertainty in the method, the Monte Carlo analysis is easily updated to reflect this new information.

The Monte Carlo approach can also be used to propagate uncertainty when emission predictions are summed across different sources, **provided the uncertainties in those predictions are independent.** For example, doing so would be reasonable if the underlying data used to derive the estimates are independent between source categories—and not reasonable if the underlying data are the same for source categories. Monte Carlo methods can be adapted to handle the uncertainty in sums of predictions across different sources that cannot be regarded as independent, but this is beyond the scope of this chapter.

8.2.4 Recommendations for Summarizing Monte Carlo Output

The following provides an overview of how to summarize Monte Carlo output. Note that statistical software typically provides Monte Carlo analyses summary plots and information.

1. Plot the Monte Carlo approximation to the PDF, either as a histogram of the data set $\{G_r\}_{r=1}^{M}$ or as a smoothed version of the histogram, via a kernel density estimator. Check that values along the horizontal axis are plausible values of entity-level GHG emissions, with higher density corresponding to more plausible values.
2. Estimate and report a central value of the GHG emissions PDF.

   While the mode, or most frequent value, is one standard measure of central tendency, it is not readily estimated by the Monte Carlo approach this report describes for UQ, and is not recommended for most PDFs encountered in GHG uncertainty computations. (The exception is right-triangular PDFs, described in section 8.2.6.)

   Another standard measure of central tendency is the mean. While the theoretical mean of the GHG emissions PDF is readily estimated by the empirical average of the Monte Carlo replicates, use the median. The theoretical median is defined for continuous $p_G(g)$ as the value $\theta_{0.5}$ such that:

$$0.5 = \mathrm{P}[G \le \theta_{0.5}] = \int_0^{\theta_{0.5}} p_G(g)\, dg;$$

   The median cuts off $0.5 \times 100\%$ of the probability in the PDF, so it is the 50th percentile. Other percentiles (2.5th and 97.5th) are used in determining a prediction interval for the GHG emissions from the entity, so choice of the median implies that a common set of estimation methods can be used to summarize the Monte Carlo results. Also, the median is insensitive to skewness and heavy tails, unlike the mean, and generally simple to understand.

   To estimate the median and other percentiles, first sort the Monte Carlo replicates to obtain the order statistics:

$$G_{(1)} \le G_{(2)} \le \cdots \le G_{(r)} \le \cdots \le G_{(M)},$$

   The parentheses in the subscripts denote sorted data. Then choose the value in the "middle" of the sorted list by picking the order statistic with index equal to ceiling$(0.5M)$: $\hat{\theta}_{0.5} = G_{(\mathrm{ceiling}(0.5M))}$. For example, choose $G_{(500)}$ if $M = 1{,}000$ or $G_{(501)}$ if $M = 1{,}001$.[1]

   The empirical median is the Monte Carlo estimate of the theoretical median, $\theta_{0.5}$. Similarly, other percentiles are defined as the values θ_q that cut off $q \times 100\%$ of the probability in the PDF,

$$q = \mathrm{P}\left[G \le \theta_q\right] = \int_0^{\theta_q} p_G(g)\, dg.$$

   To estimate each percentile, choose the corresponding empirical percentile: the qMth order statistic in the sorted list, rounding up if qM is not an integer:

[1] Another standard definition of the empirical median takes the unique middle value if M is odd and the average of the two middle values if M is even, but for the large values of M used in Monte Carlo analysis, this distinction is not important. This report uses the definition above for consistency with other percentiles.

$$\hat{\theta}_q = G_{(\text{ceiling}(qM))}.$$

3. Report estimates of the 2.5th and the 97.5th percentiles, because these theoretical quantities satisfy the following probability equation for the entity-level GHG:

   $$0.95 = 0.975 - 0.025 = \text{P}[G \leq \theta_{0.975}] - \text{P}[G \leq \theta_{0.025}] = \text{P}[\theta_{0.025} \leq G \leq \theta_{0.975}].$$

   Estimating the theoretical percentiles with the corresponding empirical percentiles,

   $$(\hat{\theta}_{0.025}, \hat{\theta}_{0.975}) = \left(G_{(\text{ceiling}(0.025M))}, G_{(\text{ceiling}(0.975M))}\right),$$

   yields a Monte Carlo 95-percent prediction interval for the entity-level GHG. That is the probability that the true entity-level GHG emission G lies between $\hat{\theta}_{0.025}$ and $\hat{\theta}_{0.975}$ is approximately 0.95.

To summarize, (1) plot the Monte Carlo approximation to the PDF, typically as a histogram; (2) compute and report a measure of central tendency, i.e., the empirical median; then (3) compute and report an approximate 95-percent prediction interval by using the empirical 2.5th and 97.5th percentiles.

Box 8-1. Assessing the Precision of Monte Carlo Estimates

Because the empirical median and other percentiles are estimates from the Monte Carlo sample, they have their own uncertainties, which can be made smaller by increasing the Monte Carlo sample size, M. That is, if the Monte Carlo analysis were repeated, the estimated median and other estimated percentiles would change, due to the random sampling, but the amount of possible change will be small for a larger M. The amount of possible change in the estimated percentiles can be quantified from the same Monte Carlo sample used to estimate the percentiles, by computing 95-percent confidence intervals for the percentiles. These confidence intervals use standard statistical large-sample approximations, which are excellent for the large values of M in typical Monte Carlo analysis.

These confidence intervals would usually not be reported: they are used only by the analyst to assess the precision of the Monte Carlo estimates. If the intervals are deemed to be too wide, the Monte Carlo analysis would be expanded by increasing the value of M.

Theoretical percentiles θ_q are estimated via order statistics (empirical percentiles), $\hat{\theta}_q$, as described above. Confidence intervals for theoretical percentiles are obtained by choosing pairs of order statistics, as follows. First, choose the index of the lower order statistic, rounding down to get an integer:

$$L = \text{floor}\left\{0.5 + (Mq) - 1.96\sqrt{Mq(1-q)}\right\}.$$

Second, choose the index of the upper order statistic, rounding up to get an integer:

$$U = \text{ceiling}\left\{0.5 + (Mq) + 1.96\sqrt{Mq(1-q)}\right\}.$$

Finally, the confidence interval for the percentile θ_q is the pair of order statistics, $\left(G_{(L)}, G_{(U)}\right)$.

For example, consider the theoretical 2.5th percentile, $\theta_{0.025}$, and suppose $M = 10{,}000$. Then $Mq = 250$, so the empirical percentile is $\hat{\theta}_{0.025} = G_{(250)}$, and the indices for the confidence interval for $\theta_{0.025}$ are

$$L = \text{floor}\left\{0.5 + (10{,}000 \times 0.025) - 1.96\sqrt{10{,}000 \times 0.025(0.975)}\right\} = 219$$

and

$$U = \text{ceiling}\left\{0.5 + (10{,}000 \times 0.025) + 1.96\sqrt{10{,}000 \times 0.025(0.975)}\right\} = 282.$$

This translates to 95 percent confidence that the theoretical 2.5th percentile, $\theta_{0.025}$, lies between the order statistics $G_{(219)}$ and $G_{(282)}$ obtained in the Monte Carlo simulation with $M = 10{,}000$ replicates. If this interval is too wide for sufficient precision, simply increase the Monte Carlo sample size.

Similar computations can be conducted for the upper endpoint of the prediction interval, $\theta_{0.975}$, or for the median, $\theta_{0.5}$.

8.2.5 Numerical Example of Monte Carlo Analysis

To illustrate the Monte Carlo analysis, consider an example of an explicit, model-based method that predicts GHG emissions as $G = m(A, F) = A \times F$, with the activity data input known to be $A = 10$ and with the unknown emission factor F described by a normal PDF with theoretical mean, $\mu_F = 3$ and variance, $\sigma^2 = 1$:

$$p_F(f) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(f-3)^2\right\}.$$

For this example, the PDF of G is also normal, with theoretical median $\theta_{0.5} = 30$, theoretical 2.5th percentile $\theta_{0.025} = 10.4$, and theoretical 97.5th percentile $\theta_{0.975} = 49.6$ and the theoretical quantities estimated using 10,000 replications of the Monte Carlo analysis ($M = 10{,}000$). Random emission factors $F_1, F_2, \ldots, F_M$ drawn independently from the normal distribution with mean (3) and variance (1), help compute the simulated emissions ($G_1 = 10F_1, G_2 = 10F_2, \ldots, G_M = 10F_M$)

To summarize the Monte Carlo draws:

1. Plot the histogram, as shown in figure 8-3. In a Monte Carlo analysis, the true PDF of the GHG emissions (G) would be unknown, but it is known in this illustration and is plotted in the figure as a dashed, bell-shaped curve. The histogram is an excellent approximation to the true PDF.
2. Compute and report the empirical median as a measure of central tendency. For any Monte Carlo sample of size $M = 10{,}000$, the empirical median will be the order statistic with index equal to ceiling $(0.5M) = 5{,}000$. For the Monte Carlo simulation used in this illustration, the empirical median is

   $$\hat{\theta}_{0.5} = G_{(5{,}000)} = 29.93,$$

   This is very close to the theoretical median $\theta_{0.5} = 30$. The theoretical median is plotted as a vertical dashed line and the empirical median is plotted as a vertical solid line in the center of figure 8-3. The two lines are nearly coincident and difficult to distinguish visually.
3. Compute and report a 95-percent prediction interval for G, using the empirical 2.5th percentile and the empirical 97.5th percentile:
4. Empirical 2.5th percentile $\hat{\theta}_{\text{ceiling}(0.025M)} = G_{(250)} = 10.37$.
5. Empirical 97.5th percentile $\hat{\theta}_{\text{ceiling}(0.975M)} = G_{(9{,}750)} = 49.37$.

6. Ninety-five percent of all GHG emissions are expected to fall between these bounds. These empirical bounds are close to the true theoretical percentiles of $\theta_{0.025} = 10.4$ and $\theta_{0.975} = 49.6$. The theoretical 2.5th percentile is plotted as a vertical dashed line and the empirical 2.5th percentile is plotted as a vertical solid line on the left of figure 8-3. The theoretical 97.5th percentile is plotted as a vertical dashed line and the empirical 97.5th percentile is plotted as a vertical solid line on the right of figure 8-3. In each case, the estimates and theoretical values are difficult to distinguish visually.

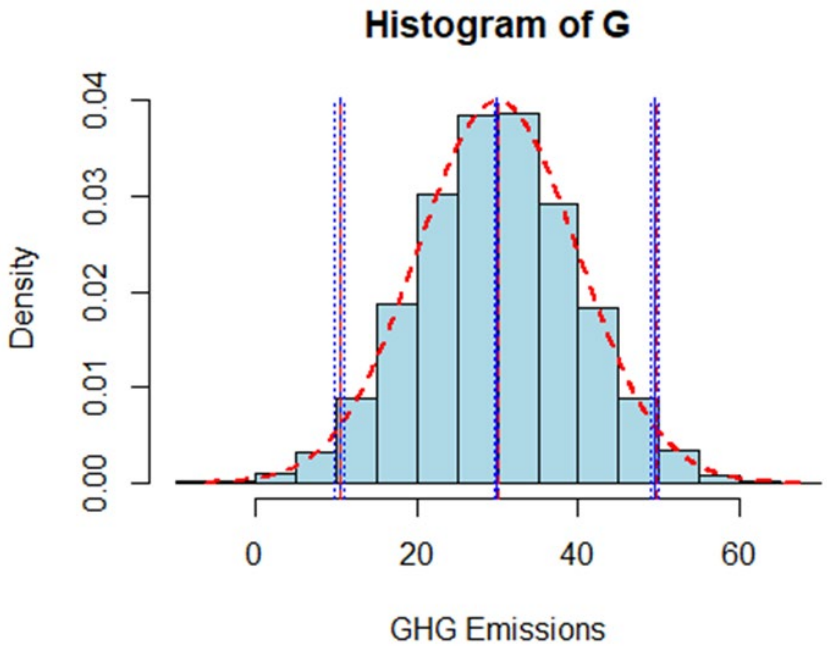

Figure 8-3. Histogram From *M* = 10,000 Monte Carlo Draws From a Normal Distribution (Curved Dashed Line), with True Percentiles Plus Estimates and Confidence Intervals

If the Monte Carlo analysis were repeated, the estimated median and 2.5th and 97.5th percentiles would change but would not change by much if M is large. To determine if $M$is large enough (e.g., 10,000 replications) use the Monte Carlo sample to compute 95-percent confidence intervals corresponding to each estimated percentile, as shown in box 8-1: the width of these confidence intervals gives an idea of expected variation if the Monte Carlo were repeated. If the intervals are sufficiently narrow, conclude that the Monte Carlo sample size is sufficient.

A 95-percent confidence interval for the median is the pair of order statistics with indices:

$$L = \text{floor}\{0.5 + (0.5M) - 1.96\sqrt{M(0.5)(0.5)}\} = 4{,}902$$

$$U = \text{ceiling}\{0.5 + (0.5M) + 1.96\sqrt{M(0.5)(0.5)}\} = 5{,}099.$$

The 95-percent confidence interval for the median from the Monte Carlo sample is:

$$\left(G_{(4{,}902)}, G_{(5{,}099)}\right) = (29.69, 30.15),$$

This shows that the theoretical median is precisely estimated. The confidence interval is plotted with a pair of vertical dotted lines in the center of figure 8-3.

For the 2.5th percentile, the 95 percent confidence interval uses the indices $L = 219$ and $U = 282$, so the confidence interval is (9.77, 10.88). The confidence interval is plotted with a pair of vertical dotted lines on the left of figure 8-3.

For the 97.5th percentile, the confidence interval uses the indices $L = 9{,}719$ and $U = 9{,}782$, so the confidence interval is (48.99, 49.85). The confidence interval is plotted with a pair of vertical dotted lines on the right of figure 8-3.

The confidence intervals for the median and 2.5th and 97.5th percentiles show that with $M = 10{,}000$, each theoretical percentile is precisely estimated. If the intervals were judged to be insufficiently narrow, the Monte Carlo analysis could be repeated with a larger value of M.

Box 8-2. Potential Reduction in Uncertainty With Aggregation Across Entities

Uncertainties are often large at the entity scale, and carbon programs need ways to manage the risk associated with this uncertainty. Aggregation across entities is one way to reduce those uncertainties.

Consider the simplest version of an explicit GHG emissions model, in which the emissions are computed as $G_j = a_j F_j$, where $a_j > 0$ is the known activity data for entity j and F_j is the unknown emission factor for entity j. The uncertainty in the emission factor is reflected in a PDF with mean μ and variance σ^2. Important here is the coefficient of variation, defined as the standard deviation of emissions relative to expected emissions, in percent for the total emissions over n entities:

$$cv = \frac{\sqrt{\mathrm{Var}\left(\sum_{j=1}^{n} a_j F_j\right)}}{\mathrm{E}\left[\sum_{j=1}^{n} a_j F_j\right]} \times 100\%,$$

If $n = 1$, this expression becomes

$$cv = \frac{\sqrt{\mathrm{Var}(a_1 F_1)}}{\mathrm{E}[a_1 F_1]} \times 100\% = \frac{\sqrt{a_1^2 \sigma^2}}{a_1 \mu} \times 100\% = \frac{\sigma}{\mu} \times 100\%.$$

As n increases, the variance increases, but so does total emissions; therefore, relative uncertainty as measured by cv decreases. The amount of decrease depends on the amount of correlation among emission factors on different entities, $\mathrm{Corr}(F_j, F_k)$ for $j \neq k$.

Entity-level emission factors are unlikely to be identical due to natural variation from entity to entity. Nearby entities with similar geographic characteristics and similar management practices might be expected to have more similar emission factors, and hence higher correlation, than entities that are more "distant" in terms of entity-level conditions and practices. For simplicity, assume all the entities that are combined have the same amount of correlation with each other, $\mathrm{Corr}(F_j, F_k) = \rho$ for $j \neq k$. The most extreme versions of this assumption are $\rho = 1$, so that entities have perfectly correlated emission factors, and $\rho = 0$, so that entities have uncorrelated emission factors. The true correlations are likely to vary across pairs of entities, with some higher and some lower values.

Under the assumption of constant correlation, it can be shown that

$$cv = \frac{\sigma}{\mu}\left\{\rho + \frac{\left(\sum_{j=1}^{n} a_j^2\right)/n}{\left(\sum_{j=1}^{n} a_j/n\right)^2}\frac{(1-\rho)}{n}\right\}^{1/2} \times 100\%.$$

If $\rho = 1$, then the entities have perfectly correlated emission factors, and the relative uncertainty never decreases: it equals $(\sigma/\mu) \times 100\%$ for any number of entities. In all cases with $\rho \neq 1$, the relative uncertainty decreases as the number of entities in the sum increases, with the greatest decrease when the entities have uncorrelated emission factors.

Figure 8-4 shows the coefficient of variation as a function of ρ and number of entities, for a simulated example in which the activity data are simulated as normal random variables with mean 10 and standard deviation 1 and then treated as fixed and known, while the random emission factors have mean $\mu = 10$ and standard deviation $\sigma = 5$. The coefficient of variation for a single entity, or any number of perfectly correlated entities, is then $(\sigma/\mu) \times 100\% = 50\%$. For all other cases, the coefficient drops below 50%, with the greatest decrease when the entities' emission factors are uncorrelated.

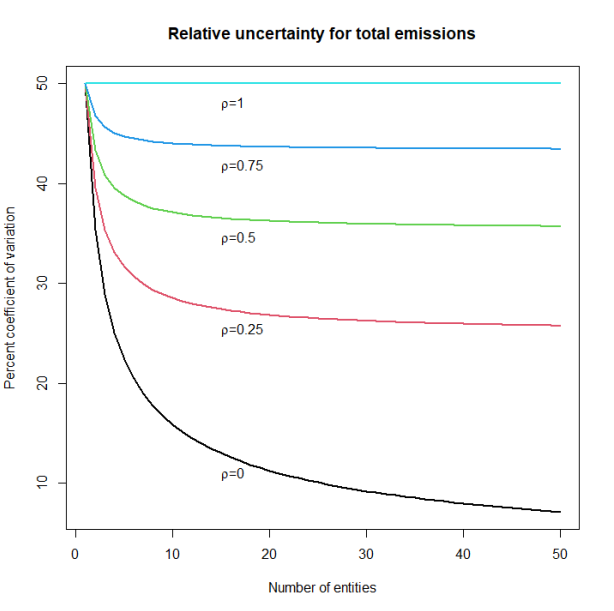

Figure 8-4. Relative Uncertainty for Total Emissions, Measured as Percent Coefficient of Variation, Decreases as the Number of Entities in the Sum Increases, Provided Those Entities Do Not Have Perfectly Correlated Emission Factors

8.2.6 Special Case: Right-Triangular Distribution

For some sources (e.g., urea CO_2), the uncertainty is described with a right-triangular PDF, which describes all possible values of the emission factor as lying between zero and some maximum value, ϕ, with PDF that increases linearly from zero at zero to $2/\phi$ at ϕ. Mathematically, the PDF is $p_F(f) = 2f/\phi^2$ for $0 \leq f \leq \phi$, otherwise $p_F(f) = 0$. If the GHG emission is $G = aF$ for some known activity data input, a, then the PDF of G can be derived directly, rather than via Monte Carlo. The resulting PDF is $p_G(g) = 2g/(a\phi)^2$ for $0 \leq g \leq a\phi$ and $p_G(g) = 0$ elsewhere. This PDF is shown in figure 8-5.

For this right-triangular PDF, the standard prediction approach is to use the mode, $a\phi$, instead of the mean or median. The Monte Carlo approach is not used to determine a prediction interval. Instead, the prediction interval is determined analytically as $\left(\sqrt{\alpha}a\phi, a\phi\right)$. The probability that the GHG emission falls in this interval is then the difference in area between the large triangle and the small triangle in figure 8-5, or $1 - \left(\frac{1}{2}\right)\left(\sqrt{\alpha}a\phi\right)\frac{2\sqrt{\alpha}}{a\phi} = 1 - \alpha$. For $\alpha = 0.05$, this yields a 95-percent prediction interval.

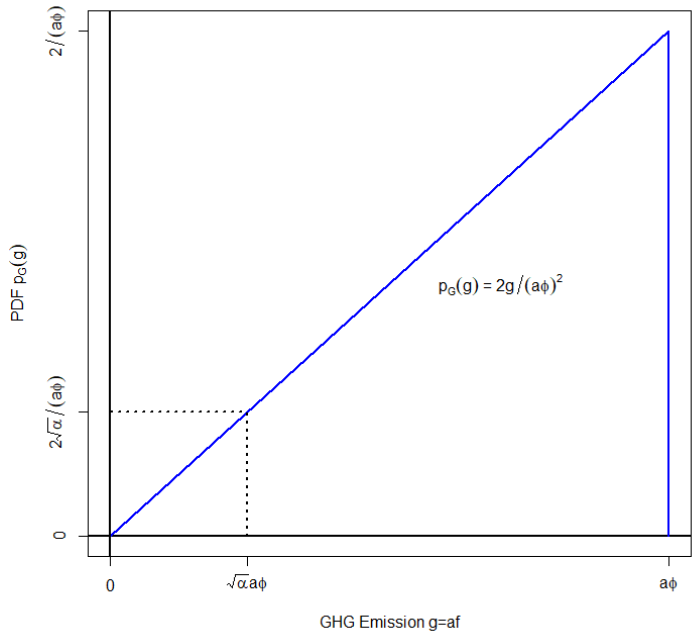

Figure 8-5. Right-Triangular PDF for GHG Emission With Known Activity, a, and Lower Bound of $(1-\alpha)100\%$ Prediction Interval

8.3 Step-by-Step Guidance for UQ

8.3.1 Explicit Model-Based Methods

For explicit model-based methods, PDFs can be placed directly on parameters, which are typically emission factors, and no entity-scale measurements are needed to determine the relevant PDFs. Instead, these PDFs are provided in the methods description for each source. If these PDFs are not right-triangular, use a Monte Carlo approach as described in section 8.2.3:

1. Start by setting the random number seed in the statistical software, so that results are reproducible.
2. For the rth replicate, select a random draw F_r of the unknown emission factor(s) from the relevant PDFs. In models with multiple factors or parameters, select random draws from the joint probability distribution of the factors or parameters. For example, multiple factors or parameters that have a multivariate normal as their joint distribution will be specified in terms of a mean vector and a covariance matrix. If a joint probability distribution is not otherwise specified, then randomly select values from each of the PDFs for the factors or parameters. This selection implies that the factors or parameters are independent, and their joint distribution is the product of the individual PDFs.

3. Use these random values and the known activity data as inputs to the method, yielding the rth random value $G_r = m(a, F_r)$ of the entity-level emission.
4. Repeat, independently, for $r = 1,2, \ldots, M$. The resulting M Monte Carlo replicates $\{G_r\}_{r=1}^{M}$ represent a large, random sample from the unknown PDF $p_G(g)$.
5. Summarize the Monte Carlo results based on the median and 95-percent prediction interval, as described in section 8.2.4.

8.3.2 Explicit Measurement-Based Methods

For measurement-based methods, this report does not directly provide PDFs for emission factors; they are instead estimated from measurements at the entity scale. Typically, these measurements are taken only on a sample, so some uncertainty is introduced. For example, a random sample of trees on a woodlot could have its volume characteristics measured to represent the entire woodlot and the growth over time, resulting in a PDF.

PDFs for these explicit measurement-based methods will be context-specific, but the general approach of Monte Carlo UQ will still apply. Because the unknown emission factor will typically rely on both model parameters estimated from sources outside the entity and entity-level measurements, denote the unknown emission factor by:

$$F = h(\theta, \kappa)$$

where:

$h()$	=	a known function
θ	=	one or more unknown model parameters that are estimated from scientific studies external to the entity
κ	=	one or more unknown entity-level characteristics

Because the model parameters are often estimated by regression or other statistical techniques, it is reasonable to treat the PDF for the unknown θ parameters as multivariate normal (MVN) with mean vector μ_θ and variance-covariance matrix Σ_θ. The estimates of μ_θ and Σ_θ are obtained from this methods document, using information from scientific studies that are independent of the entity.

In many cases, the unknown entity-level characteristics κ will be estimated based on measurements obtained from a sample. Standard probability sampling designs include all units in the population of interest in a "sampling frame" and have positive and known probabilities of selection. These sampling designs lead to approximately normally distributed estimates of κ in moderate to large sample sizes, under very mild conditions on the characteristics of the measurements. There is no need for the original measurements to be normal or close to normal: the measurements could be binary, or counts, or right-skewed continuous. It is therefore reasonable to treat the PDF for the unknown entity-level characteristics κ as MVN with mean vector μ_κ and variance-covariance matrix Σ_κ. The covariances in Σ_κ are usually not zero because estimated characteristics that use the same sample are correlated.

The estimates of μ_κ and Σ_κ are obtained from entity-level measurements and the sampling design that leads to the measurements. Methods of estimation for different designs are well-documented. Statistical software (including SAS, Stata, SPSS, and the "survey" package in R) can provide estimates of the mean vector and covariance matrix given basic information on the sampling design, including:

- Unique stratum identifiers (if any), which are disjoint subpopulations that cover the population and from which independent samples are selected;
- Unique identifiers of primary sampling units (PSUs) which are the units initially sampled from the frame, even if there are subsequent stages of selection; and
- Sampling weights, which are the inverses of the sample inclusion probabilities.

A complete description of estimation and variance estimation for various sampling designs is beyond the scope of this chapter.

In explicit model-based methods, the Monte Carlo analysis begins by sampling $F_1, F_2, \ldots, F_M$ independently from a given PDF, $p_F(f)$. For the explicit measurement-based methods of this section, use a Monte Carlo analysis as described in section 8.2.3. See box 8-3 for a sample calculation:

1. Start by setting the random number seed in the statistical software, so that results are reproducible.
2. For the rth replicate, sample θ_r independently from MVN$(\mu_\theta, \Sigma_\theta)$, sample κ_r independently from MVN$(\mu_\kappa, \Sigma_\kappa)$, and compute $F_r = h(\theta_r, \kappa_r)$.
3. Use these random values and the known activity data as inputs to the method, yielding the rth random value $G_r = m(a, F_r)$ of the entity-level emission.
4. Repeat, independently, for $r = 1,2, \ldots, M$. The resulting M Monte Carlo replicates $\{G_r\}_{r=1}^{M}$ represent a large, random sample from the unknown PDF $p_G(g)$.
5. Summarize the Monte Carlo results based on the median and 95-percent prediction interval, as described in section 8.2.4.

Box 8-3. Example of Explicit Measurement-based Method

Equation 3-6 (in chapter 3) describes aboveground woody tree biomass stock, a key determinant of the unknown emission factor, as:

$$h(\theta, \kappa) = \{\beta_0(average\ stems\ per\ plot) + \beta_1(average \ln(dbh)\}(\#plots\ per\ ha)(area\ in\ ha)$$

Table 3-6 (in chapter 3, provided with relevant entries below) presents $\theta = (\beta_0, \beta_1)$ for various taxa. In this example, $\kappa = (average\ stems\ per\ plot, average\ ln(dbh))$ is unknown and is estimated at the entity scale from a sample of plots (where dbh is the diameter at breast height).

Group	Taxon	95% Confidence Interval	β_0	β_1
Conifer	Abies, 0.35 spg[a]	±20%	–2.3123	2.3482

[a] spg is the specific gravity of wood on a green volume to dry-weight basis

The above table is not a complete replication of table 3-6 in chapter 3, only relevant information for the example in this chapter.

To determine the MVN PDF for θ, use the 95-percent confidence intervals in table 3-6, expressed as plus or minus some percentage. For a parameter β with estimated value b and 95-percent confidence interval $\pm d100\%$, where:

- Variance = $\left(\frac{bd}{1.96}\right)^2$
- Standard deviation = $\frac{|b|d}{1.96}$

Therefore, the corresponding PDF for β_0 is normal with mean -2.3123 and standard deviation $\frac{|-2.3123|(0.2)}{1.96} = 0.235949$.

Similarly, the corresponding PDF for β_1 is normal with mean 2.3482 and standard deviation $\frac{|2.3482|(0.2)}{1.96} = 0.239612$.

Table 3-6 does not provide covariances between estimated parameters. One conservative approach then is to maximize the variance of the emission factor by assuming the correlation between the estimates is either perfectly negative (if β_0 and β_1 have opposite signs) or perfectly positive (if β_0 and β_1 have the same signs). This assumption implies that the covariance is as shown in the equation below, where $\Sigma_{\theta,11}$ and $\Sigma_{\theta,22}$ are the variances:

$$\Sigma_{\theta,12} = \Sigma_{\theta,21} = sign(\beta_0\beta_1)\left(\Sigma_{\theta,11}\right)^{1/2}\left(\Sigma_{\theta,22}\right)^{1/2}$$

These computations imply that the PDF for θ is:

$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} \sim MVN\left(\begin{bmatrix}-2.3123\\ 2.3482\end{bmatrix}, \begin{bmatrix}(0.235949)^2 & (-1)(0.235949)(0.2396122)\\ (-1)(0.235949)(0.2396122) & (0.2396122)^2\end{bmatrix}\right).$$

To determine the MVN PDF for κ in this example of woody tree biomass stock, sampling design, plus all measurements obtained from the sample, are required. Then this information helps estimates of the mean vector μ_κ and variance-covariance matrix Σ_κ.

In this example, one sample would be used to obtain estimates of various characteristics, e.g., average stems per plot for different taxa and average $ln(dbh)$ for different taxa. These estimates will be dependent, and proper estimation of Σ_κ will account for this dependence.

8.3.3 Implicit Model-Based Methods

Implicit model-based methods do not rely on any entity-scale measurements to determine emission factors. Their uncertainty is fully described with PDFs given elsewhere in this report. But those PDFs are not specified directly on model parameters, typically due to the complexity of these models, which represent biogeochemical processes. Instead, uncertainty is quantified based on comparisons of model-based predictions to field measurements from experimental studies (not from the entity under consideration). Examples include soil carbon stock changes and direct soil N_2O emissions, which are predicted with the DayCent ecosystem model and compared to experimental results from long-term field experiments to quantify uncertainty in model structure and parameterization.

The comparison of model predictions to field measurements uses a statistical model to account for independent variables (covariates) to explain some of the uncertainty in GHG emission predictions and to account for the correlations among measurements from the field experiments. The standard statistical model for this empirical method is a linear mixed effect (LME) model, with fixed effects to account for covariates and with random effects to account for spatial and temporal correlations. The implication of this statistical model at an entity scale is that the GHG emissions are modeled as:

$$G = \mu(A, F) + x^{\top}\beta + b$$

where:

$\mu(A, F)$ = the output of the model with known activity data inputs A and with emission factors F that are implicitly defined

$x^{\top}$ = a vector of known covariates at the entity scale (such as soil texture, management practice, climate variables, and related information about the management system and environmental conditions)

β = a vector of unknown fixed effect regression coefficients that have been estimated from the long-term field experiments

b = sum of one or more random effects that represent field-to-field variation that is not explained either by the model or by the fixed effects

Based on the estimation from the field experiments, the uncertainty in the fixed effects is described with a MVN PDF, with mean vector $\hat{\beta}$ and covariance matrix $\hat{\Sigma}$ from the fit of the LME. The uncertainty in the random effects is described with a normal PDF with mean 0 and with variance $\hat{\tau}^2$ equal to the sum of the estimated variances of all the random effects that are summed to create b.

For an entity with known activity data inputs A and known covariates $x^{\top}$, Monte Carlo UQ then proceeds with the following steps:

1. Start by setting the random number seed in the statistical software, so that results are reproducible.
2. For the rth replicate, draw a MVN random vector $\beta^{(r)} \sim MVN(\hat{\beta}, \hat{\Sigma})$, and select a normal random variable(s) $b^{(r)} \sim MVN(0, \hat{\tau}^2)$.
3. Compute $G^{(r)} = \mu(A, F) + x^{\top}\beta^{(r)} + b^{(r)}$.
4. Repeat, independently, for $r = 1,2, \ldots, M$. The resulting M Monte Carlo replicates $\{G_r\}_{r=1}^{M}$ represent a large, random sample from the unknown PDF $p_G(g)$.
5. Summarize the Monte Carlo results based on the median and 95-percent prediction interval, as described in section 8.2.4.

8.4 Extension of Monte Carlo for Unknown Activity Data Inputs

This chapter assumes activity data inputs are known at the entity scale. If these inputs are subject to some uncertainty, that uncertainty should be quantified with an appropriate PDF, $p_A(a)$. Assuming the uncertainty in the activity data is independent of the uncertainty in the emission factors, the Monte Carlo approach extends in a straightforward way. Proceeding as in section 8.2.3, generate a large number, M, of replicates of the possible GHG emissions with the following steps:

1. Start by setting the random number seed in the statistical software, so that results are reproducible.
2. For the rth replicate, draw a random activity data input A_r from the PDF $p_A(a)$and draw a random emission factor F_r from the PDF $p_F(f)$.
3. Use these random values as inputs to the method, yielding the rth random value $G_r = A_r F_r$ of the entity-level emission.
4. Repeat, independently, for $r = 1,2, \ldots, M$. The resulting M Monte Carlo replicates $\{G_r\}_{r=1}^{M}$ represent a large, random sample from the unknown PDF $p_G(g)$.

5. Summarize the Monte Carlo results based on the median and 95-percent prediction interval, as described in section 8.2.4.