Air Pollutants Exposure Model Documentation (APEX, Version 5.2)

Volume II: Technical Support Document

EPA-452/R-19-005b
October 2019

Air Pollutants Exposure Model Documentation (APEX, Version 5.2) Volume II: Technical Support Document

U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Health and Environmental Impacts Division
Research Triangle Park, NC

DISCLAIMER

This document has been prepared at least partially by ICF (under EPA Contract No. EP-W-12-0101 [WA 4-55]). It has been subject to the Agency's review, and has been approved for publication an an EPA document. Mention of trade names or commercial products is not intended to constitute endorsement or recommendation for use.

CONTACT

Questions should be addressed to John E. Langstaff, U.S. Environmental Protection Agency, C504-06, Research Triangle Park, North Carolina 27711 (email: langstaff.john@epa.gov).

ACKNOWLEDGEMENTS

The primary author of this document is Graham Glen (ICF). It includes contributions from Melissa Nysewander, Luther Smith, and Casson Stallings (while at Alion Science and Technology, Inc.); Stephen Graham, Kristin Isaacs, Tom McCurdy (retired), and John Langstaff (EPA); and by ICF.

CONTENTS

LIST OF TABLES

LIST OF EXHIBITS

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 TRIM and the APEX Model

The Air Pollutants Exposure model (APEX) is part of EPA's overall Total Risk Integrated Methodology (TRIM) model framework (EPA, 1999). TRIM is a time-series modeling system with multimedia capabilities for assessing human health and ecological risks from hazardous and criteria air pollutants; it is being developed to support evaluations with a scientifically sound, flexible, and user-friendly methodology. The TRIM design includes three modules:

- Environmental Fate, Transport, and Ecological Exposure module (TRIM.FaTE);
- Human Exposure-Event module (TRIM.Expo); and
- Risk Characterization module (TRIM.Risk).

APEX is designed to estimate human exposure to criteria and air toxic pollutants at local, urban, and regional scales. The current release of the model is APEX5. Note that APEX has been extensively reviewed. Any changes to the computer code may lead to results that cannot be supported by this documentation. Model enhancements, bug fixes, and other changes are occasionally made to APEX, and thus users are encouraged to revisit the website https://www.epa.gov/fera/human-exposure-modeling-air-pollutants-exposure-model for notices of these changes.

1.2 Scope and Organization of This Document

The documentation of the APEX model is currently divided into two volumes. *Volume II: Technical Support Document* (this document) is intended to be a reference on the scientific basis of the APEX model. The scientific background, original references, and equations for the APEX model algorithms are included in this volume. Topics covered include: the methods implemented in APEX for sampling probability distributions, calculating microenvironmental concentrations, modeling ventilation, estimating exposure and dose, and assembling composite activity diaries. Other model algorithms, such as those for generating the study area and the simulated population, are also described.

Volume I: User's Guide, is designed to be a hands-on guide to using APEX. It is applicable to all levels of expertise—from novice to advanced—and focuses on how to run the APEX computer model, develop the appropriate input files, and interpret the model output files. A more complete introduction to APEX can be found in CHAPTER 1 of *Volume I*.

Nomenclature. The terms below are used throughout this guide.

- Diary: a set of events or activities (e.g., cooking, sleeping) for an individual in a given time frame (e.g., a day)
- Air quality district: the geographical area represented by a given set of ambient air quality data (either based on a fixed-site monitor or output from an air quality model)
- Event: an activity (e.g., cooking) with a known starting time, duration, microenvironment, and location (usually home or work)

- Microenvironment: a space in which human contact with an environmental pollutant takes place
- Profile: a set of characteristics that describe the person being simulated (e.g., age, gender, height, weight, employment status, whether an owner of a gas stove or air conditioner)
- Sector: the basic geographical unit for the demographic input to and output from APEX (usually census tracts)
- Study area: the geographical area modeled
- Study area population: total population of persons who live in the study area
- Meteorological zone: the geographical area represented by a given set of meteorological data (either based on a meteorological station or output from a meteorological model

The labeling conventions below are used in this document.

- *Input and Output File Names* are in italics title case (in some cases, key terms are also introduced in italics, not capitalized, within a paragraph)
- ***Model Variables*** are in bold italics
- ***KEYWORDS***, which are used in the input files to identify variables and settings, are given in uppercase bold italicsee
- `Input and output file excerpts` are in a box surrounded by a single line
- `Courier (fixed space)` font is used for folder names, paths, and system commands outside of APEX
- This document also contains references to the APEX model code. Specifically, the discussions of the model algorithms include mention to the module and function or subroutine in which they are implemented. The code locations are given in bold non-italic text in the format **Module:Subroutine** or **Module:Function.**

CHAPTER 2. OVERVIEW OF MODEL DESIGN AND ALGORITHMS

This chapter provides a brief outline of the key modeling steps, logic processes, and databases used in APEX.

APEX is designed to simulate population exposure to criteria and air toxic pollutants at local, urban, and regional scales. The user specifies the geographic area to be modeled and the number of individuals to be simulated to represent this population. APEX then generates a personal profile for each simulated person that specifies various parameter values required by the model. The model next uses diary-derived time/activity data matched to each personal profile to generate an exposure event sequence (also referred to as "activity pattern" or "composite diary") for the modeled individual that spans a specified time period such as one year. Each event in the sequence specifies a start time, the duration of an exposure, a geographic location, a microenvironment, and an activity. Probabilistic algorithms are used to estimate the pollutant concentration and ventilation (respiration) rate associated with each exposure event. The estimated pollutant concentrations account for the effects of ambient (outdoor) pollutant concentration, penetration factor, air exchange rate, decay/deposition rate, and proximity to emission sources, depending on the microenvironment, available data, and the estimation method selected by the user. The ventilation rate is derived from an energy expenditure rate estimated for the specified activity. Because the modeled individuals represent a random sample of the population of interest, the distribution of modeled individual exposures can be extrapolated to the larger population.

The model simulation includes up to seven steps, shown below.

1. Characterize the study area: APEX selects sectors (e.g., census tracts) within a study area—and thus identifies the potentially exposed population—based on the user-defined center and radius of the study area and availability of air quality and weather input data for the area. Alternatively, the user may specify a list of tracts or counties that comprise the study area.
2. Generate simulated individuals: APEX stochastically generates a sample of simulated individuals based on the census data for the study area and human profile distribution data (such as age-specific employment probabilities). The user can specify the size of the sample. The larger the sample, the more representative it is of the population in the study area (but also the longer the computing time).
3. Construct a sequence of activity events: APEX constructs an exposure event sequence (activity pattern) spanning the period of simulation for each of the simulated persons, based on the supplied Consolidated Human Activity Database (CHAD, McCurdy et al., 2000; EPA, 2009) data, although other data could be used.
4. Estimate energy expenditures and ventilation: APEX constructs a time-series of energy expenditures for each profile based on the activity event sequence. If necessary, these expenditures can be adjusted to maintain physiological realism and then used to estimate a number of ventilation metrics that are later used in estimating dose and in identifying an active subpopulation of active persons for use in creating exposure summary tables.
5. Calculate timestep concentrations in microenvironments for each pollutant: APEX enables

the user to define microenvironments that people in a study area would visit (e.g., by grouping location codes included in the supplied CHAD database). The model then calculates timestep concentrations of each pollutant in each of the microenvironments for the period of simulation, based on the user-provided ambient air quality data. All of the timestep concentrations in the microenvironments are re-calculated for each simulated individual.

6. Calculate exposures for each pollutant: APEX assigns a concentration to each exposure event based on the microenvironment occupied during the event and the person's activity. These values are averaged by timestep and by clock hour to produce a sequence of timestep and hourly-average exposures spanning the specified exposure period (typically one year). These hourly values may be further aggregated to produce daily, monthly, and annual average exposure values.
7. Calculate doses: APEX optionally calculates timestep, hourly, daily, monthly, and annual average dose values for each of the simulated individuals.

The model simulation continues until exposures are calculated for the user-specified number of simulated individuals. Figure 2.1–Figure 2.3 present these steps within a schematic of the APEX model design. The following chapters provide additional detail on the algorithms used in each of the above simulation steps.

The above steps are largely self-contained in the APEX computer code and do not depend on subsequent steps. For example, the generation of simulated individuals (step 2) is independent of any other profile characteristics or modeling results. This means that the profile variables do not depend on the diaries assigned to that profile or to the properties of the microenvironments for that profile. The assignment of diaries to the profile (step 3) depends on the profile variables but not on the microenvironments, the exposure, the dose, or the properties of any other profile. The calculation of microenvironment concentrations (step 4) can depend on the profile variables through the use of conditional variables, but in APEX, this step cannot depend on the contents of the selected activity diaries. Conceptually, this means that the microenvironments essentially have an existence of their own that is independent of the activities of the profile. For example, activities such as smoking and cooking can "occur" in a residence even when the person being profiled is not at home. In reality, the activities of the profiled person can have some effect on the microenvironments they visit, but this is not captured in the present version of the model. However, through the judicious use of source terms in microenvironments, APEX can simulate changes in concentrations due to the presence of the person (e.g., a "personal cloud" effect).

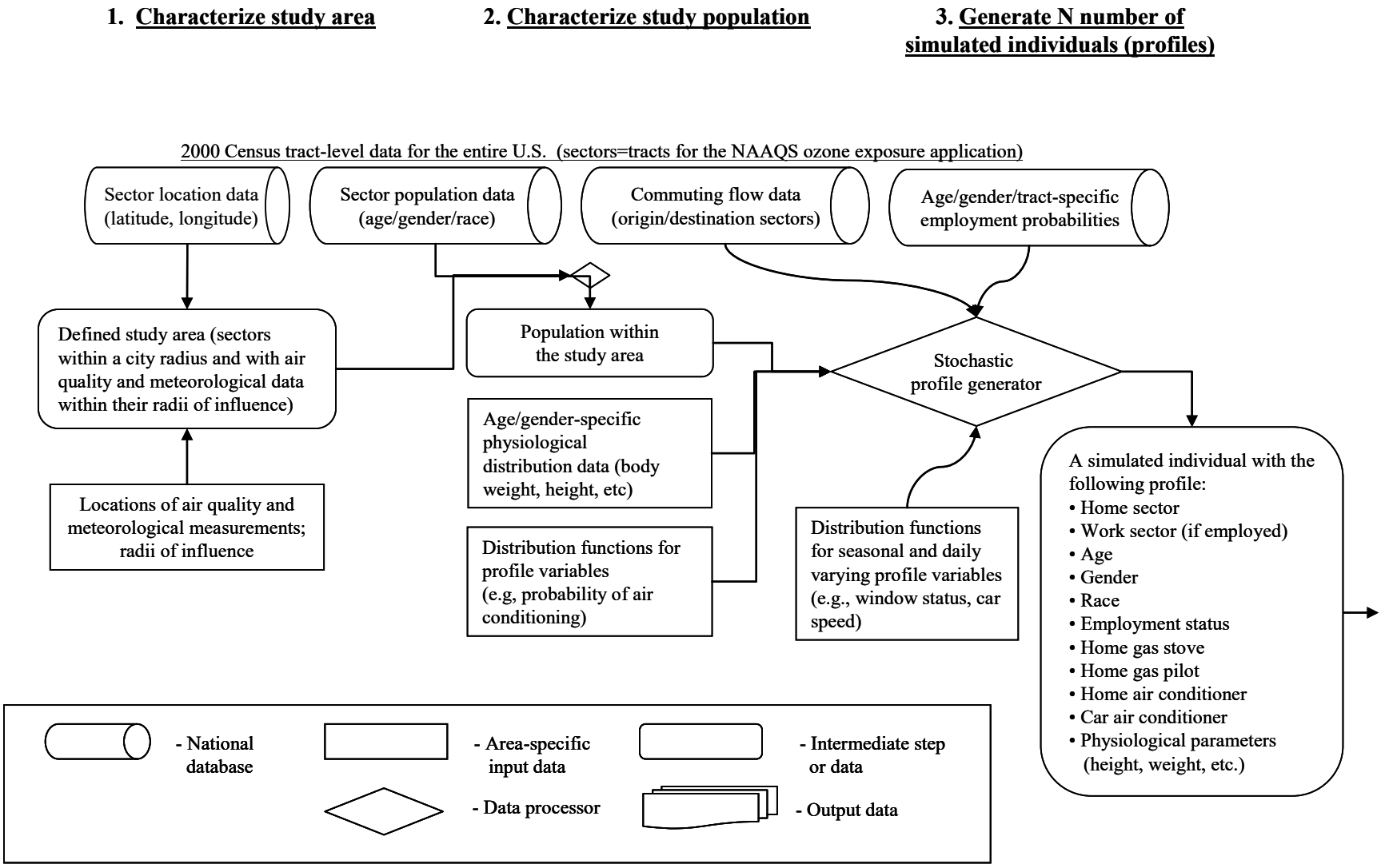

Figure 2.1. Overview of APEX, Part 1

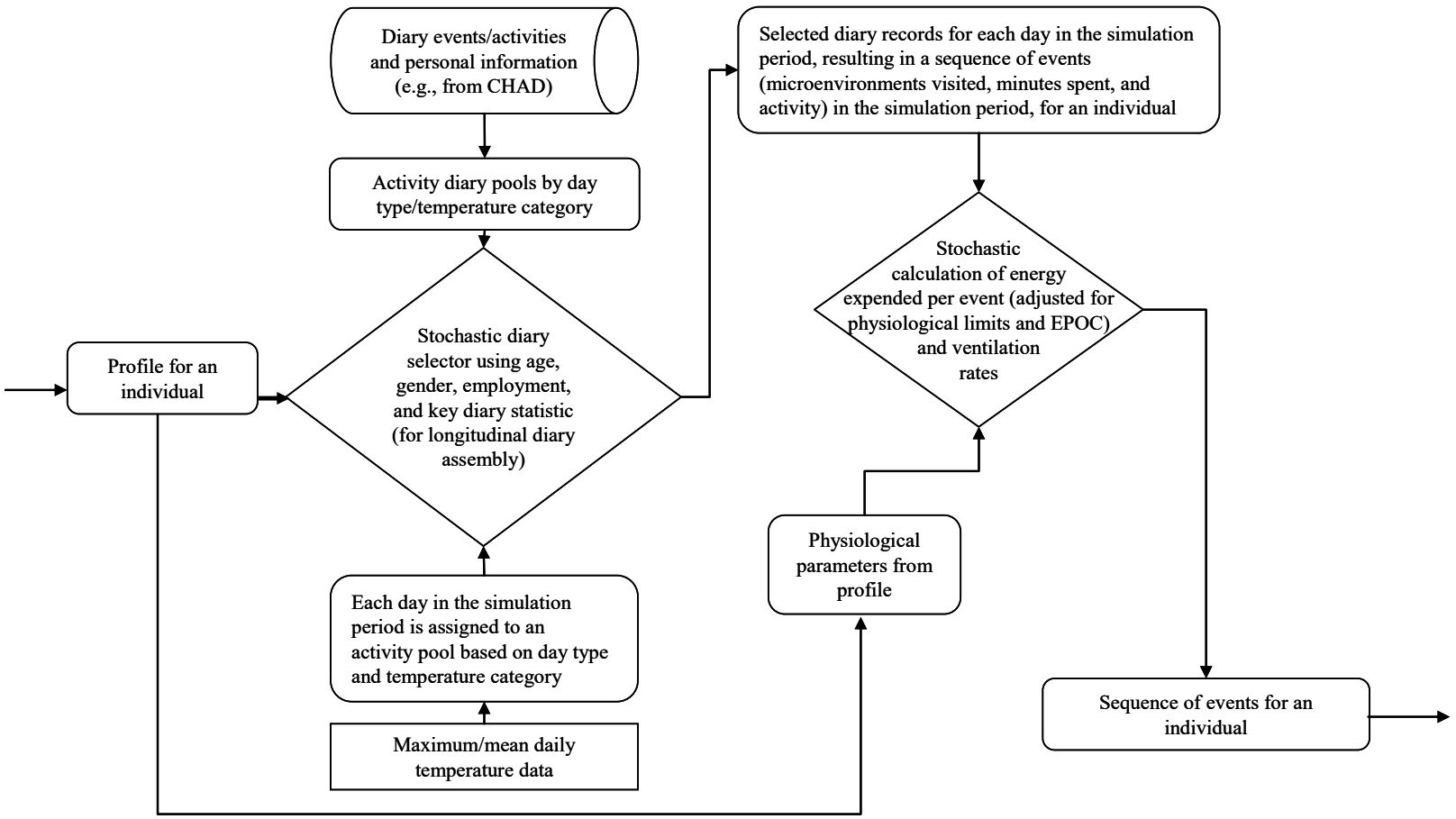

Figure 2.2. Overview of APEX, Part 2

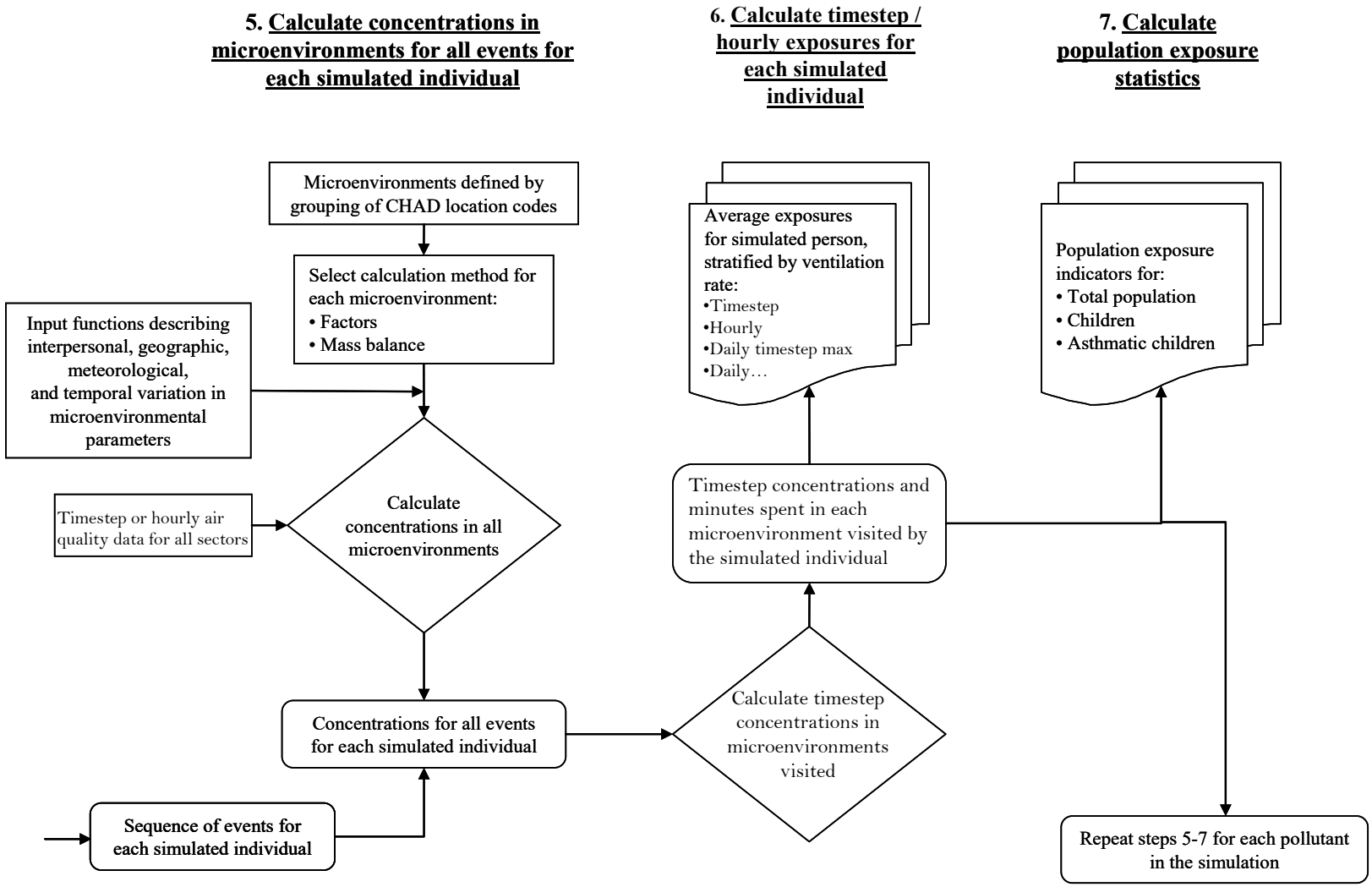

Figure 2.3. Overview of APEX, Part 3

CHAPTER 3. USING PROBABILITY DISTRIBUTIONS IN APEX

APEX is a stochastic model. It makes use of random sampling from probability distributions to model variability in a number of input model parameters. Specifically, the distributions detailed below are used.

1. Model variability in MET (energy expenditures) for different activities: Input MET distributions for each activity are defined in the *MET Distribution* file.
2. Model inter-person variability in physiological parameters: Physiological parameter distributions are defined for each different age-gender cohort in the *Physiology* file.
3. Model timestep, hourly, daily, geographic, or air quality variability in microenvironment parameters: Distributions for microenvironmental parameters are defined in the *Microenvironment Descriptions* file.
4. Model person-to-person variation in hourly air quality data within an air district: Distributions for hourly air quality values can be defined in the *Air Quality Data* file. This is an optional feature of APEX.

This chapter gives direction on how to define distributions in these input files. In addition, each distribution available in APEX and its parameters are defined in detail.

3.1 The APEX Input Distribution Format

In all APEX input files, distributions are defined in the same manner, via a standard APEX format. This format consists of the items shown below.

- ***Shape***: This variable gives the type of the distribution
- ***Par1***: Parameter 1 of the MP distribution. Depends on type.
- ***Par2***: Parameter 2 of the MP distribution. Depends on type.
- ***Par3***: Parameter 3 of the MP distribution. Depends on type.
- ***Par4***: Parameter 4 of the MP distribution. Depends on type.
- ***LTrunc***: Lower truncation point of the distribution
- ***UTrunc***: Upper truncation point of the distribution
- ***ResampOut***: Distribution resampling flag

The distribution shape is a text keyword that defines the type of distribution to be used. The next four items (***Par1***–***Par4***) are numerical values defining the parameters of the distribution. Subsequently, the next two items (***LTrunc*** and ***UTrunc***) are the optional truncation limits for the distribution and the last item is an optional character flag (***ResampOut***, set to either "Y" or "N") indicating how sampled values outside of the truncation limits are handled. All the information is entered on a single line in the appropriate input file.

The probability distributions allowed in APEX are listed in Table 3.1. Equations for each of the distributions in the table are given in Section 3.2.

Table 3.1. Available Probability Distributions in APEX

Distribution	*Shape*	*Par1*	*Par2*	*Par3*	*Par4*	*LTrunc* (optional)	*UTrunc* (optional)	*ResampOut* (optional)
Beta	***Beta***	Minimum	Maximum	Shape1 (s1) > 0	Shape2 (s2) > 0	Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Burr	***Burr***	Scale(b) > 0	Shape1(s1) > 0	Shape2 (s2) > 0	Shift(a)	Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Cauchy	***Cauchy***	Median	Scale (b) > 0			Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Discrete	***Discrete***	This type of distribution has no parameters, rather the keyword is simply followed by a list of up to 100 specific values. One of these values is selected at random, with equal probability for each. Duplicate values are acceptable.						
Exponential	***Exponential***	Decay constant, k > 0	Shift (a)			Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Extreme Value	***Evalue***	Scale (b) > 0	Shift (a)			Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Gamma	***Gamma***	Shape (s) > 0	Scale (b) > 0	Shift (a)		Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Logistic	***Lgt***	Mean	Scale (b) > 0			Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Lognormal	***Lognormal***	Geometric mean (gm) of unshifted distribution	Geometric standard deviation (gsd) > 1	Shift (a)		Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Loguniform	***LUniform***	Minimum > 0	Maximum > 0			Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Normal	***Normal***	Mean	Standard deviation			Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
OffOn	***OffOn***	Probability of being 0 (0-1)						
Pareto	***Pareto***	Shape (s) > 0	Scale (b) > 0	Shift (a)		Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Point	***Point***	Point Value						
Triangle	***Triangle***	Minimum	Maximum	Peak		Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Uniform	***Uniform***	Minimum	Maximum			Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)
Weibull	***Weibull***	Shape (s) > 0	Scale (b) > 0	Shift		Lower truncation limit	Upper truncation limit	Resample outside truncation? (Y/N)

Cells that are grayed out in the table correspond to items not needed for a particular distribution; thus, data entered in these locations will be ignored by APEX. In addition, the ***LTrunc***, ***UTrunc***, and ***ResampOut*** items are in general optional. That is, if ***LTrunc*** and ***UTrunc*** are defined but ***ResampOut*** is not, then the default value of ***ResampOut*** = Y is used. Note that a placeholder period (".") must be used in the distribution definition for each item that is not used.

Consider each of the examples below (one from each input file using distributions):

From the *MET Distribution* file:

```
Row  Act     Age  Occ.      Shape        Par1  Par2  Par3  Par4  LTrunc UTrunc  ResampOut
  1  10000   0    ADMIN     LogNormal    1.7   1.45  0     .     1.4    2.7     Y
```

From the *Physiology* file:

```
!Variable  AgeMin AgeMax    Gen   Shape    Par1   Par2  Par3  Par4    LTrunc UTrunc  ResampOut
NVO2MAX    0      0         M     Normal   48.3   1.7   .     .       44.3   52.2    Y
```

From the *Microenvironment Descriptions* file:

```
Block DType Season Area  C1  C2  C3   Shape      Par1   Par2  Par3 Par4  LTrunc UTrunc  ResampOut
1     1     1      1     1   1   1    Normal     2      0.5   0    .     0.111  10.111  Y
```

Note that in each case the distribution definitions follow the exact same format (starting with the ***Shape*** keyword). The only distribution type that does not follow this format is the Discrete distribution.

Distributions are read from the various input files and stored in **DistributionModule:ReadDist**.

3.2 Details of Distribution Sampling, Truncation, and Resampling

Probability density functions (PDFs) for each of the APEX distributions, parameterized in terms of their input APEX parameters, are given in this section (with the exception of the OffOn and Point distributions, which are trivial). In addition, real examples (of 10000 samples each) from APEX are shown for untruncated distributions and for truncated distributions using both ***ResampOut*** = Y and ***ResampOut*** = N.

When needed, stored distributions (which are read from the input files) are sampled in **DistributionModule:SampleDist.**

3.2.1 Resampling Options

ResampOut determines how truncated distributions are handed by the APEX sampling routines. If ***ResampOut*** = N, then any generated sample outside the truncation points is set to the truncation limit; in this case, samples "stack up" at the truncation points, and the probability associated with the area under the PDF outside the truncation bounds is associated with the truncation limit. If ***ResampOut*** = Y, then (effectively) another random value is selected, and repeated until it is inside the valid range. In this case, the probability outside the limits is spread over the valid values, and thus the probabilities inside the truncation limits will be higher than the theoretical untruncated PDF. In practice, APEX achieves this without the need for potentially open-ended iteration (see Section 3.3 for details).

For each of the distributions defined in this section, the theoretical PDF is shown plotted against APEX results for each truncation case. Some of the "untruncated" cases shown here actually have truncation points set to the 0.1^{st} and 99.9^{th} percentiles, where noted. These distributions had very long tails, and they were truncated so they would fit in the illustration.

3.2.2 Beta Distribution

The PDF for the beta distribution in terms of the APEX input parameters is:

$$p(x) = \frac{(x - min)^{s1-1}(max - x)^{s2-1}\Gamma(s1 + s2)}{\Gamma(s1)\,\Gamma(s2)\,(max - min)^{(s1+s2-1)}}, \quad min \le x \le max \tag{3-1}$$

where Γ indicates the gamma function and s1 and s2 are shape parameters. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the beta distribution is illustrated in Figure 3.1 along with actual examples obtained from APEX using the different sampling options.

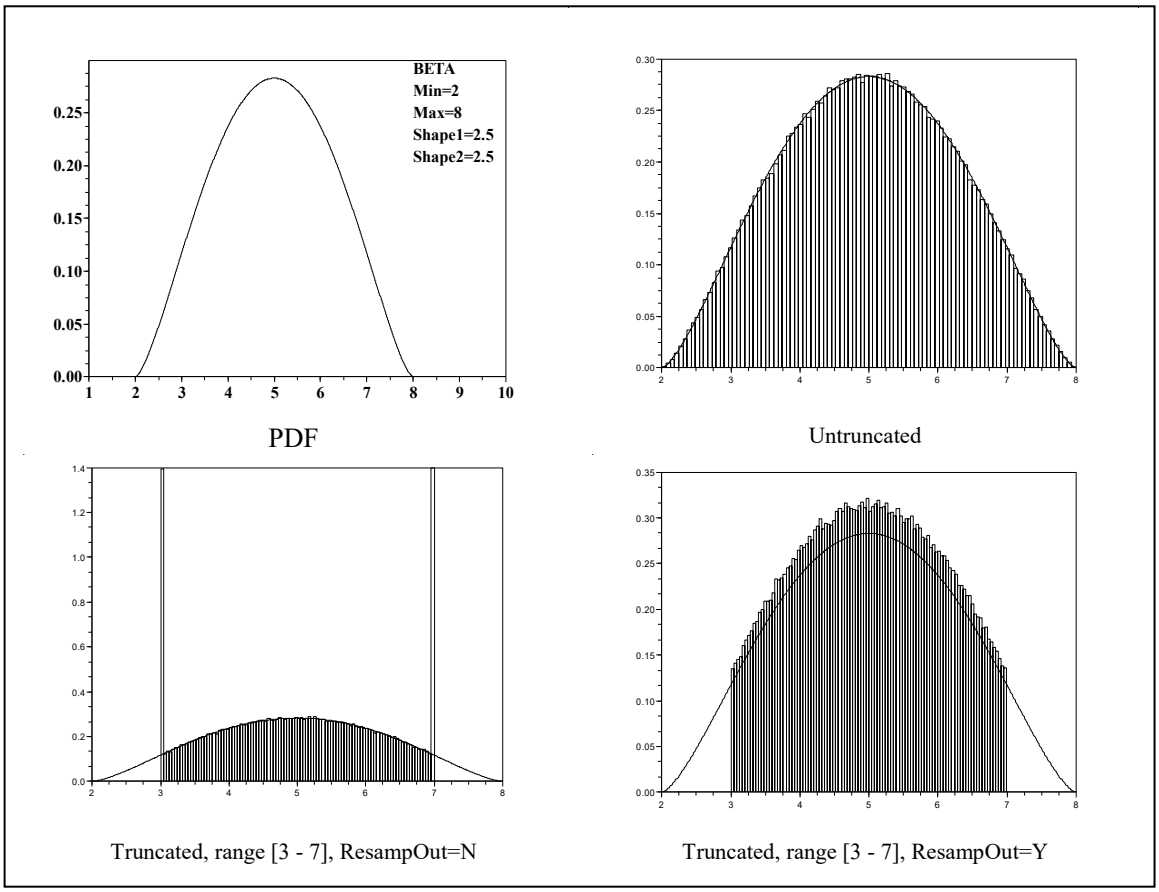

Figure 3.1. The Beta Distribution in APEX

3.2.3 Burr Distribution

The Burr distribution is very flexible because it has two shape parameters along with scale and shift parameters. The shift parameter is ***Par4*** in APEX, and represents the minimum possible

value that may be returned (it is called 'min' in the formulas). It may be omitted, in which case it is assumed to be zero. The other three parameters are required and must be positive. The scale parameter is shown as 'b' in the PDF and is the ***Par1*** parameter in APEX. The two shape parameters, s1 and s2, are ***Par2*** and ***Par3***, respectively.

The PDF for the Burr distribution in terms of the APEX input parameters is:

$$p(x) = s1\ s2\ (x - min)^{s2-1} b^{-s2} (1 + (x - min)^{s2} b^{-s2})^{-(s1+1)}, \qquad for\ x > a \qquad (3\text{-}2)$$

3.2.4 Cauchy Distribution

The PDF for the Cauchy distribution in terms of the APEX input parameters is:

$$p(x) = \frac{1}{b\,\pi\left(1 + \frac{(x - median)^2}{b^2}\right)} \tag{3-3}$$

where b is a scale parameter. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the Cauchy distribution is illustrated in Figure 3.2 along with actual examples obtained from APEX using the different sampling options. While the Cauchy distribution is symmetric, the peak value is labeled as the median because the mean of an untruncated Cauchy is technically not defined due to the heavy tails.

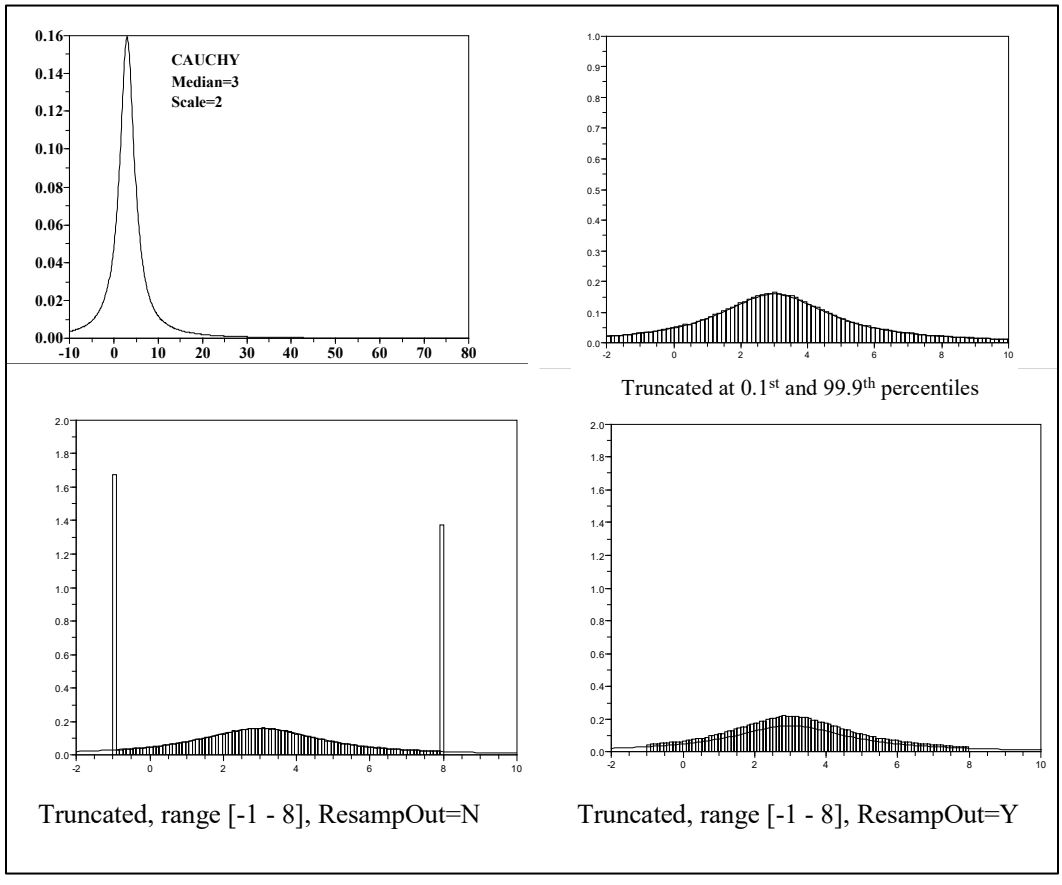

Figure 3.2. The Cauchy Distribution in APEX

3.2.5 Discrete Distribution

The discrete distribution is a custom form of APEX distribution. Rather than being defined by the regular 7 parameters, the discrete distribution is simply given as a space-separated list of up to 100 values. APEX will return all values with equal probability.

An example of a discrete distribution having 6 values is shown in Figure 3.3.

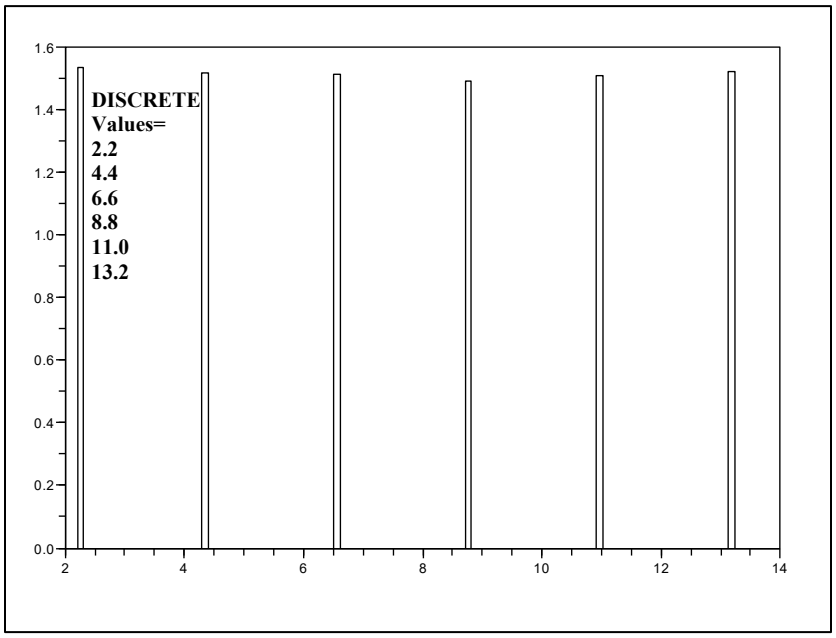

Figure 3.3. The Discrete Distribution in APEX

3.2.6 Exponential Distribution

The PDF for the exponential distribution in terms of the APEX input parameters is:

$$\begin{aligned} p(x) &= ke^{k(a-x)} \qquad \text{for } x > a \\ p(x) &= 0 \qquad\qquad\quad \text{for } x < a \end{aligned} \tag{3-4}$$

where a is a shift parameter and k is the decay constant. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the exponential distribution is illustrated in Figure 3.4, along with actual examples obtained from APEX using the different sampling options.

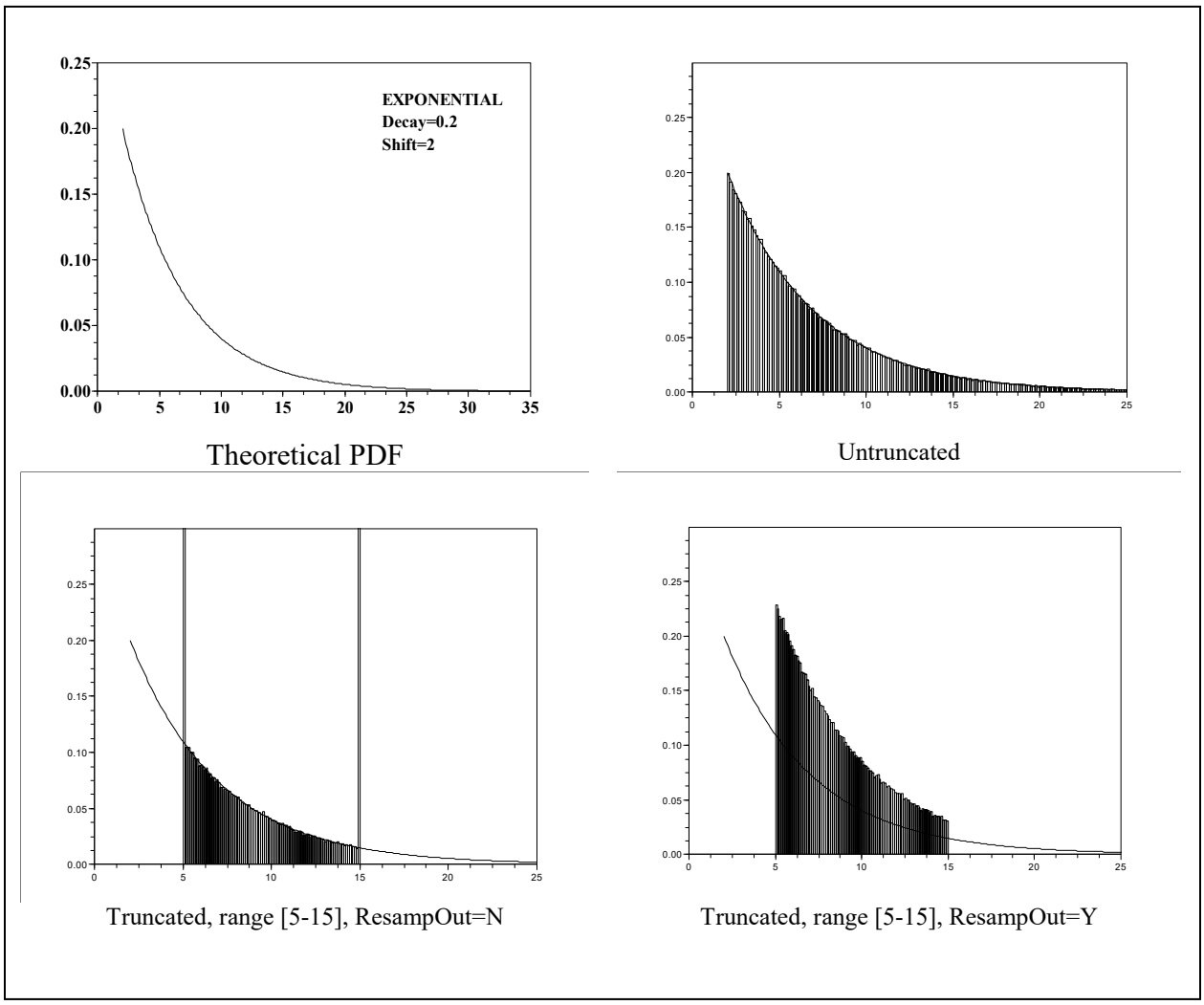

Figure 3.4. The Exponential Distribution in APEX

3.2.7 Extreme Value Distribution

The PDF for the extreme value distribution in terms of the APEX input parameters is:

$$p(x) = \frac{1}{b}\exp\left(\frac{a-x}{b} - \exp\left(\frac{a-x}{b}\right)\right) \tag{3-5}$$

where a is a shift parameter and b is a scale parameter. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the extreme value distribution is illustrated in Figure 3.5, along with physical examples obtained from APEX using the different sampling options.

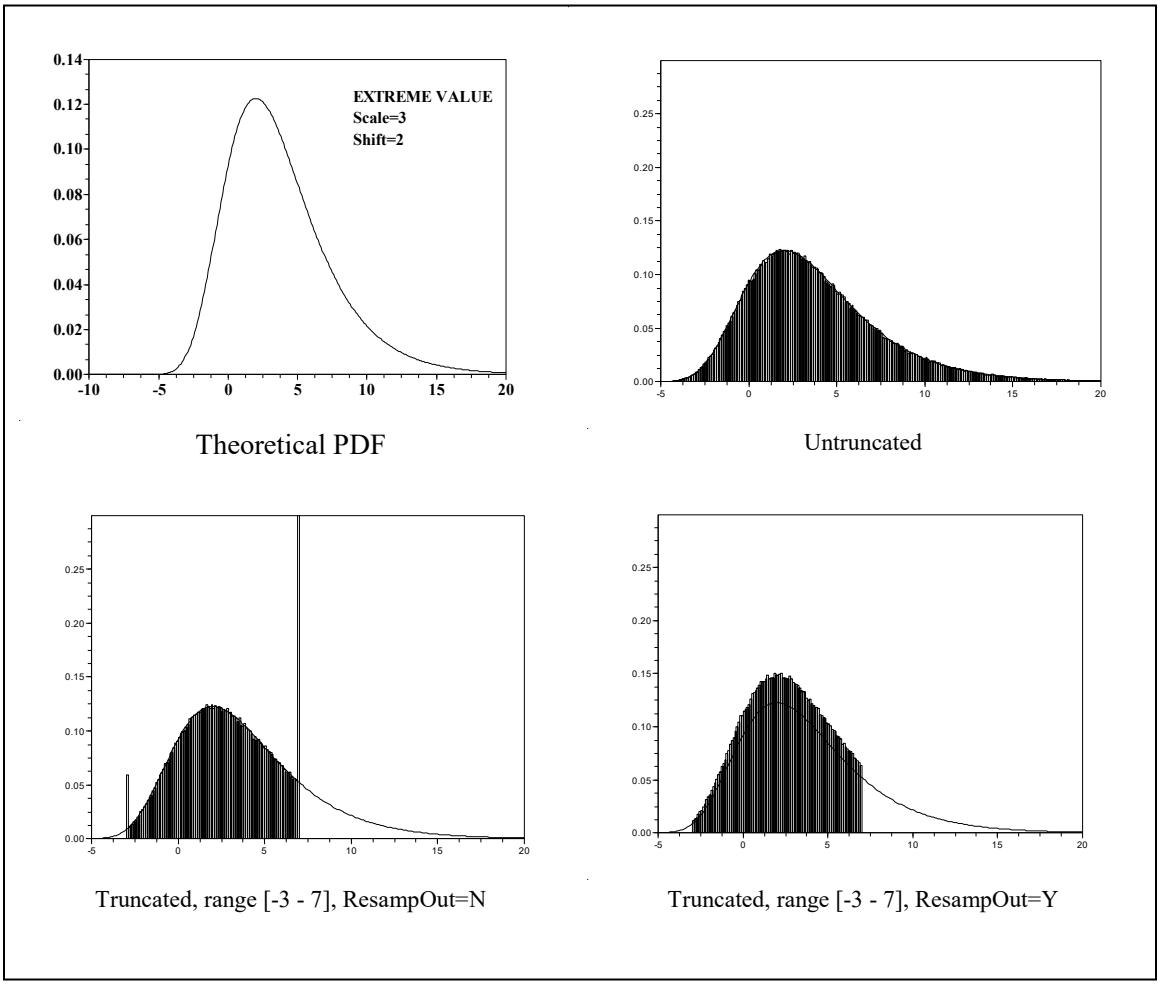

Figure 3.5. The Extreme Value Distribution in APEX

3.2.8 Gamma Distribution

The PDF for the gamma distribution in terms of the APEX input parameters is:

$$p(x) = \frac{b^{-s}}{\Gamma(s)} (x-a)^{s-1} \exp\left(\frac{a-x}{b}\right), \quad for\ x > a \tag{3-6}$$

where a is a shift parameter and b is a scale parameter. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the gamma distribution is illustrated in Figure 3.6, along with tangible examples obtained from APEX using the different sampling options.

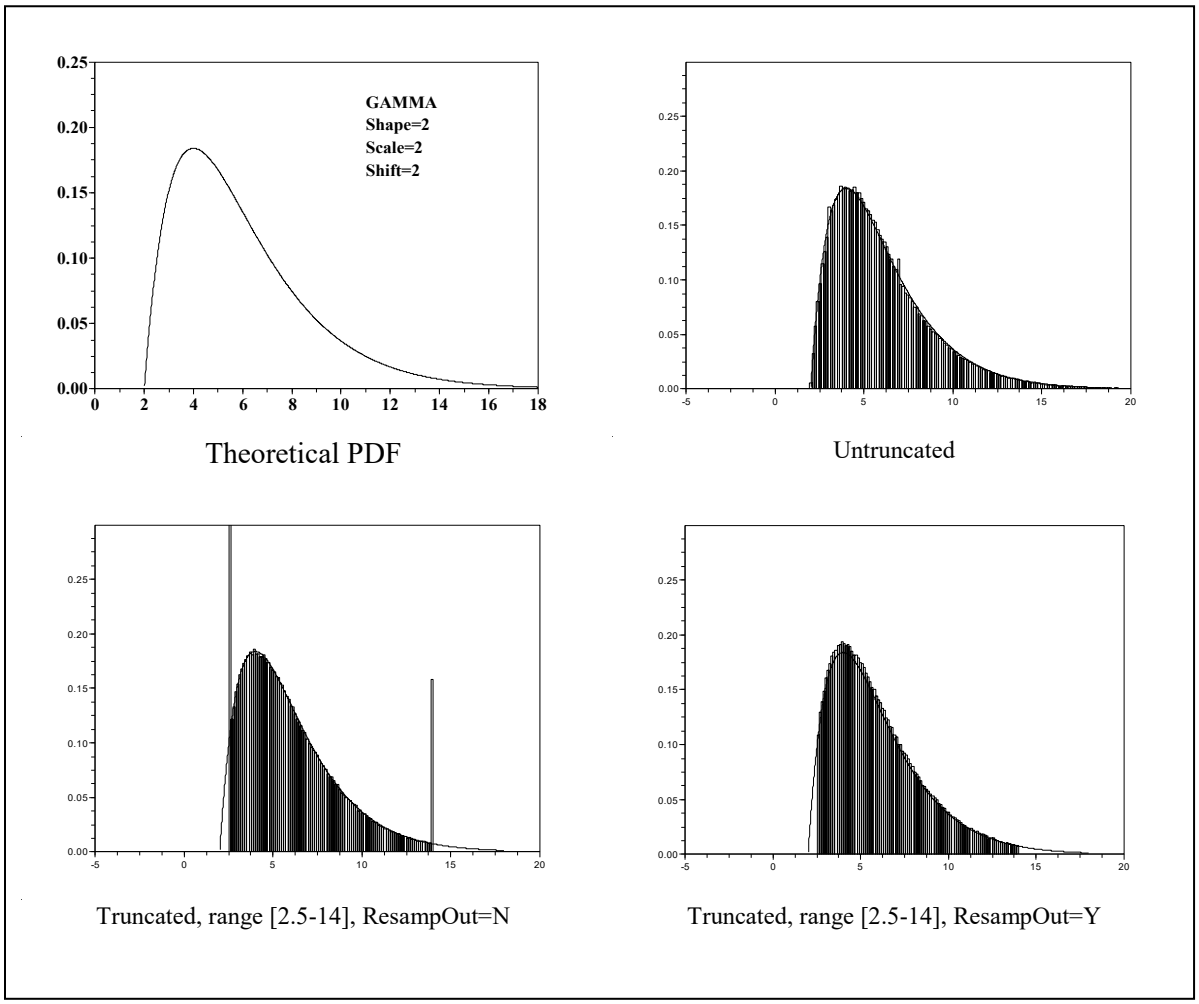

Figure 3.6. The Gamma Distribution in APEX

3.2.9 Logistic Distribution

The PDF for the logistic distribution in terms of the APEX input parameters is:

$$p(x) = \exp\left(\frac{a-x}{b}\right) \Big/ \left(1+\exp\left(\frac{a-x}{b}\right)\right)^2 \tag{3-7}$$

where a is a shift parameter and b is a scale parameter. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the logistic distribution is illustrated in Figure 3.7, along with actual examples obtained from APEX using the various sampling options.

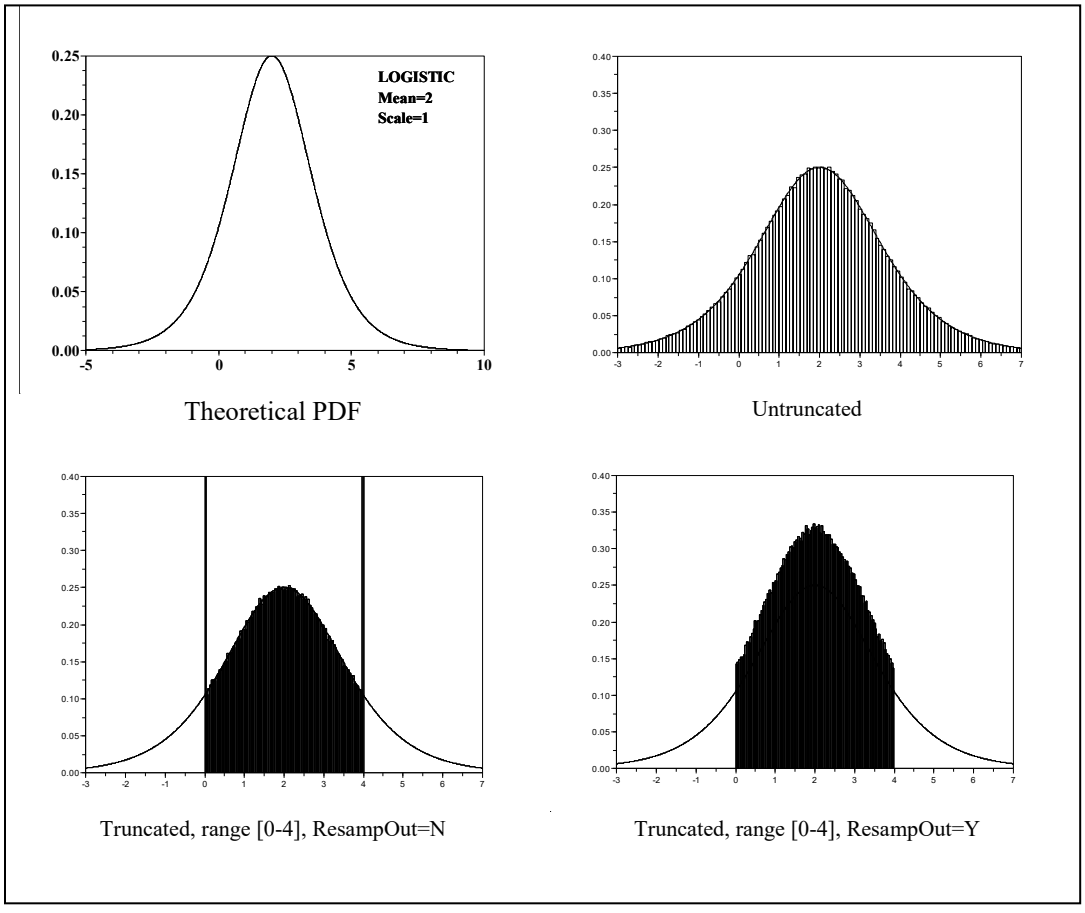

Figure 3.7. The Logistic Distribution in APEX

3.2.10 Lognormal Distribution

The PDF for the lognormal distribution in terms of the APEX input parameters is:

$$p(x) = \frac{1}{\sqrt{2\pi}\ (x-a)\ Log(GSD)} \exp\left(-\frac{1}{2}\left(\frac{Log\left(\frac{x-a}{GM}\right)}{Log(GSD)}\right)^2\right) \qquad (3\text{-}8)$$

where a is a shift parameter, GM is the geometric mean, and GSD is the geometric standard deviation. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***-***Par4***. The theoretical PDF for the lognormal distribution is illustrated in Figure 3.8, along with real examples obtained from APEX using the different sampling options.

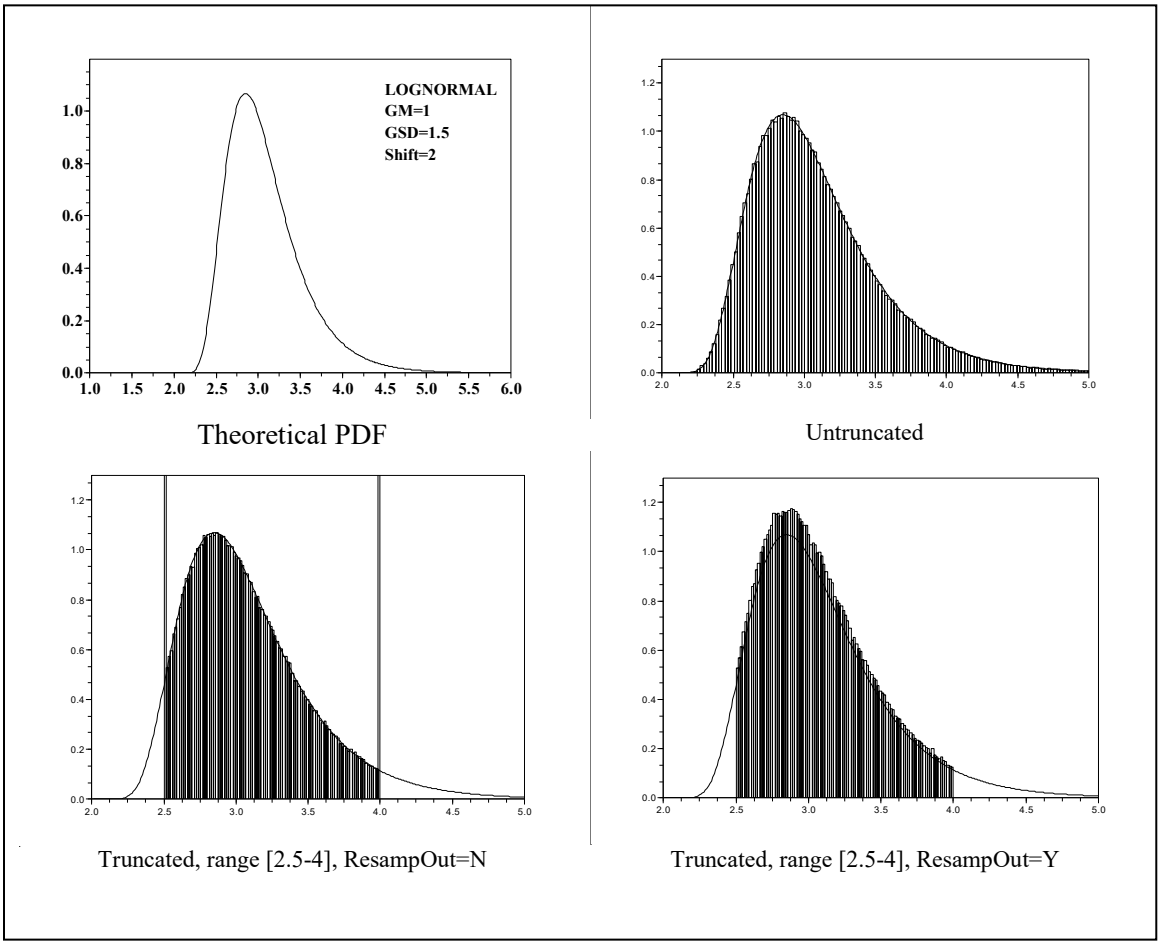

Figure 3.8. The Lognormal Distribution in APEX

3.2.11 Loguniform Distribution

The PDF for the loguniform distribution in terms of the APEX input parameters is:

$$p(x) = \frac{1}{x\ Log(\frac{max}{min})} \tag{3-9}$$

where min and max are the minimum and maximum values of the untruncated distribution, respectively. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the loguniform distribution is illustrated in Figure 3.9, along with tangible examples obtained from APEX using the assorted sampling options.

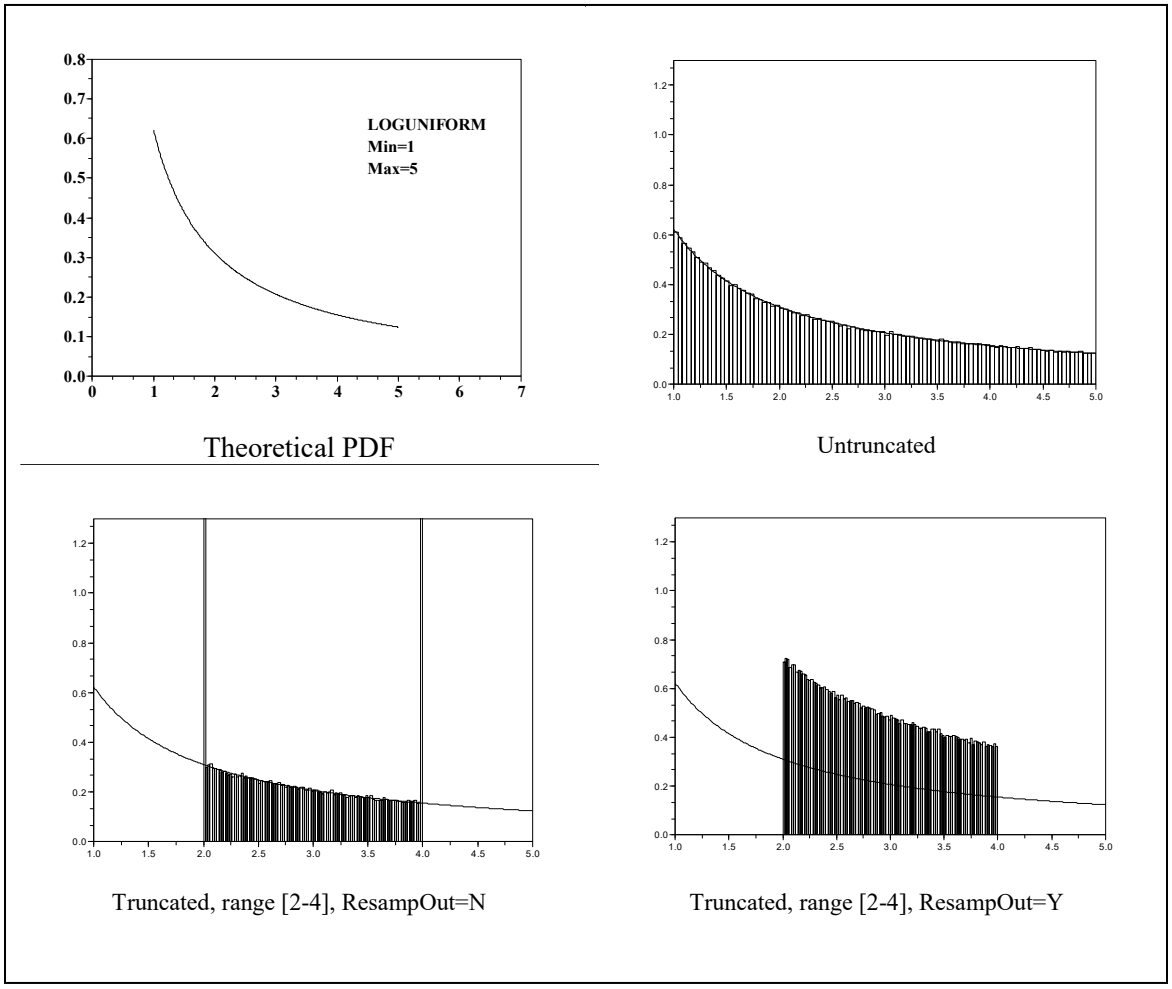

Figure 3.9. The Loguniform Distribution in APEX

3.2.12 Normal Distribution

The PDF for the normal distribution in terms of the APEX input parameters is:

$$p(x) = \frac{1}{SD\sqrt{2\pi}} Exp\left(-\frac{1}{2}\left(\frac{x - mean}{SD}\right)^2\right) \quad (3\text{-}10)$$

where mean is the mean of the untruncated distribution and SD is the standard deviation. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1–Par4***. The theoretical PDF for the normal distribution is illustrated in Figure 3.10, along with real-life examples obtained from APEX using the various sampling options.

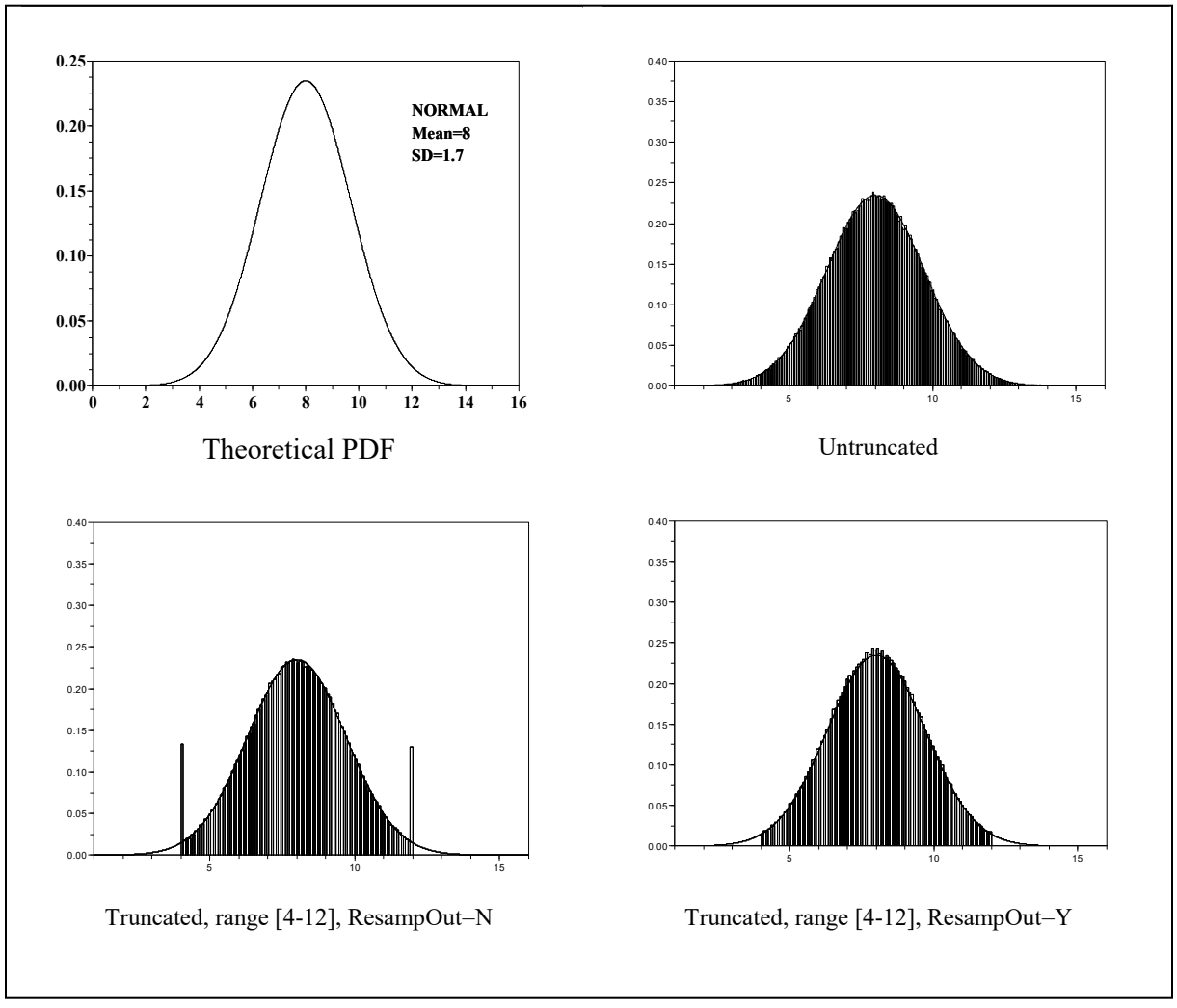

Figure 3.10. The Normal Distribution in APEX

3.2.13 Pareto Distribution

The PDF for the Pareto distribution in terms of the APEX input parameters is:

$$p(x) = \frac{s\ b^s}{(x-a)^{s+1}} \tag{3-11}$$

where a is a shift parameter and b is a scale parameter. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the Pareto distribution is illustrated in Figure 3.11, along with authentic examples obtained from APEX using the assorted sampling options.

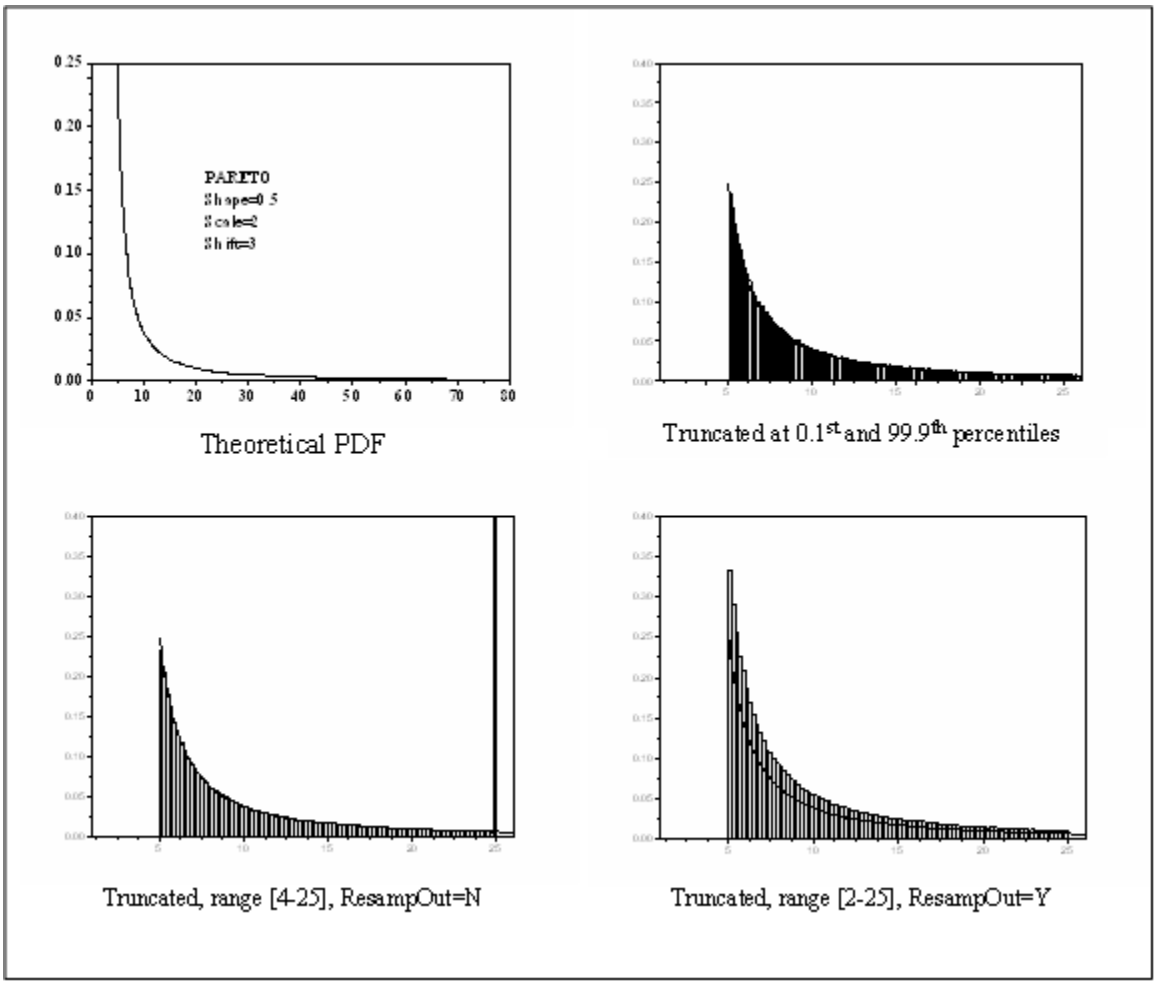

Figure 3.11. The Pareto Distribution in APEX

3.2.14 Triangle Distribution

The PDF for the triangle distribution in terms of the APEX input parameters is:

$$p(x) = \frac{2(x - min)}{(peak - min)(max - min)}, \qquad for\ min \le x \le peak$$
$$p(x) = \frac{2(max - x)}{(max - peak)(max - min)}, \qquad for\ peak \le x \le max \tag{3-12}$$

where min, max, and peak are the minimum, maximum, and peak of the untruncated distribution, respectively. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the triangle distribution is illustrated in Figure 3.12, along with real examples obtained from APEX using the different sampling options.

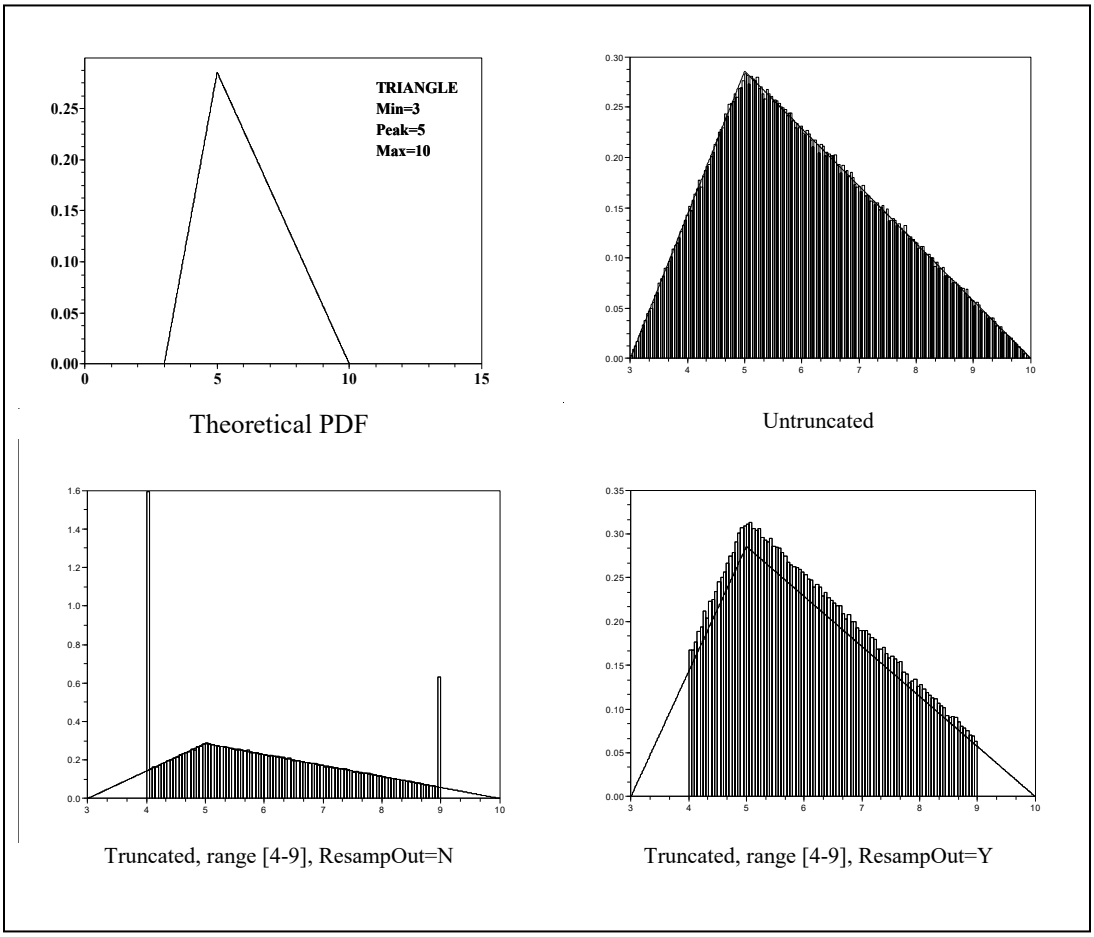

Figure 3.12. The Triangle Distribution in APEX

3.2.15 Uniform Distribution

The PDF for the uniform distribution in terms of the APEX input parameters is:

$$p(x) = \frac{1}{max - min}, \quad for\ min < x < max \tag{3-13}$$

where min and max are the minimum and maximum values of the untruncated distribution, respectively. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1***–***Par4***. The theoretical PDF for the uniform distribution is illustrated in Figure 3.13, along with real-world examples obtained from APEX using the numerous sampling options.

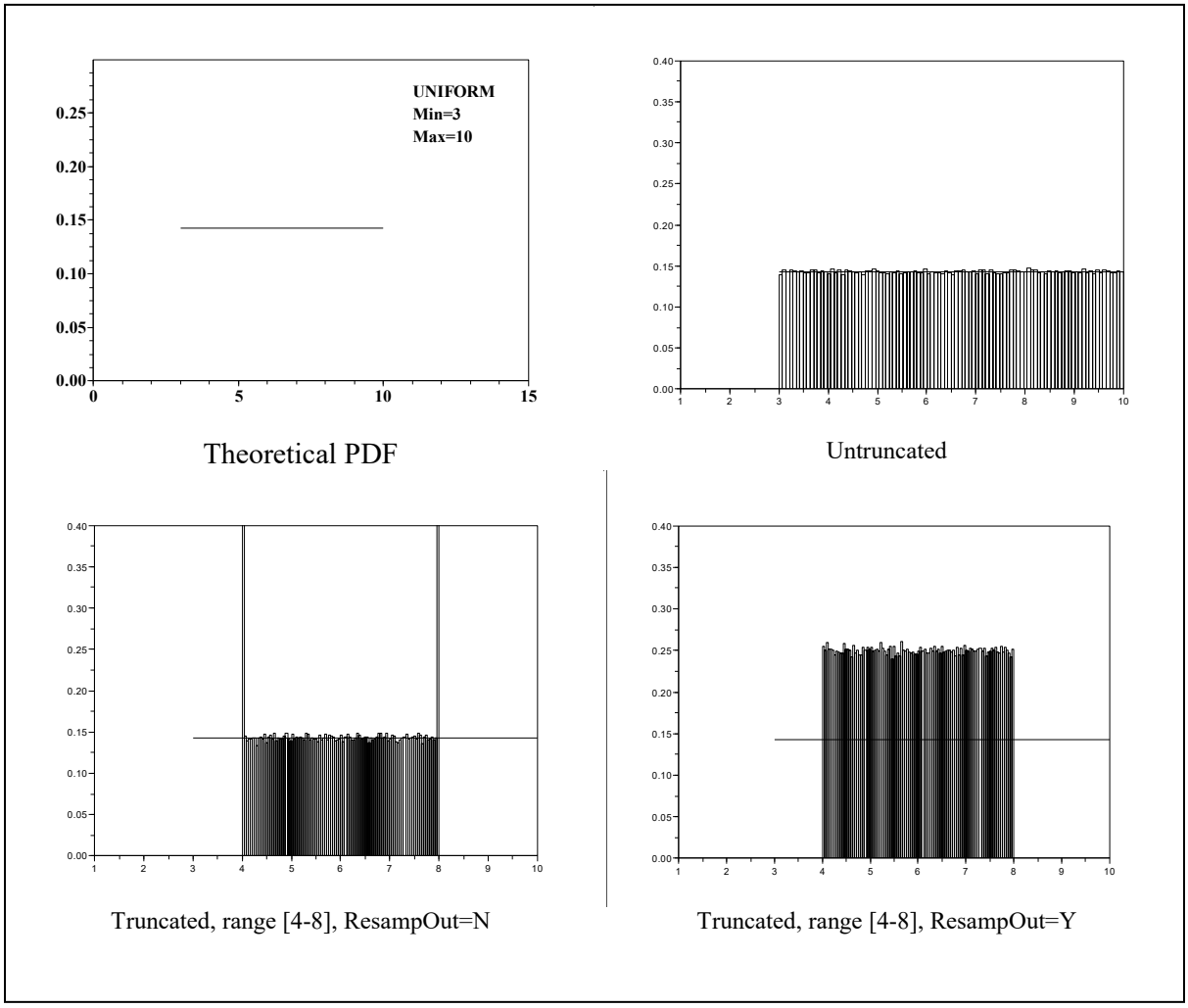

Figure 3.13. The Uniform Distribution in APEX

3.2.16 Weibull Distribution

The PDF for the Weibull distribution in terms of the APEX input parameters is:

$$p(x) = s\, b^{-s}(x-a)^{s}\ \exp\left(-\left[\frac{x-a}{b}\right]^{s}\right), \quad for\ x > a \tag{3-14}$$

where a is a shift parameter and b is a scale parameter. See Table 3.1 for assignment of the parameters in this equation to the APEX parameters ***Par1–Par4***. The theoretical PDF for the Weibull distribution is illustrated in Figure 3.14, along with actual examples obtained from APEX using the different sampling options.

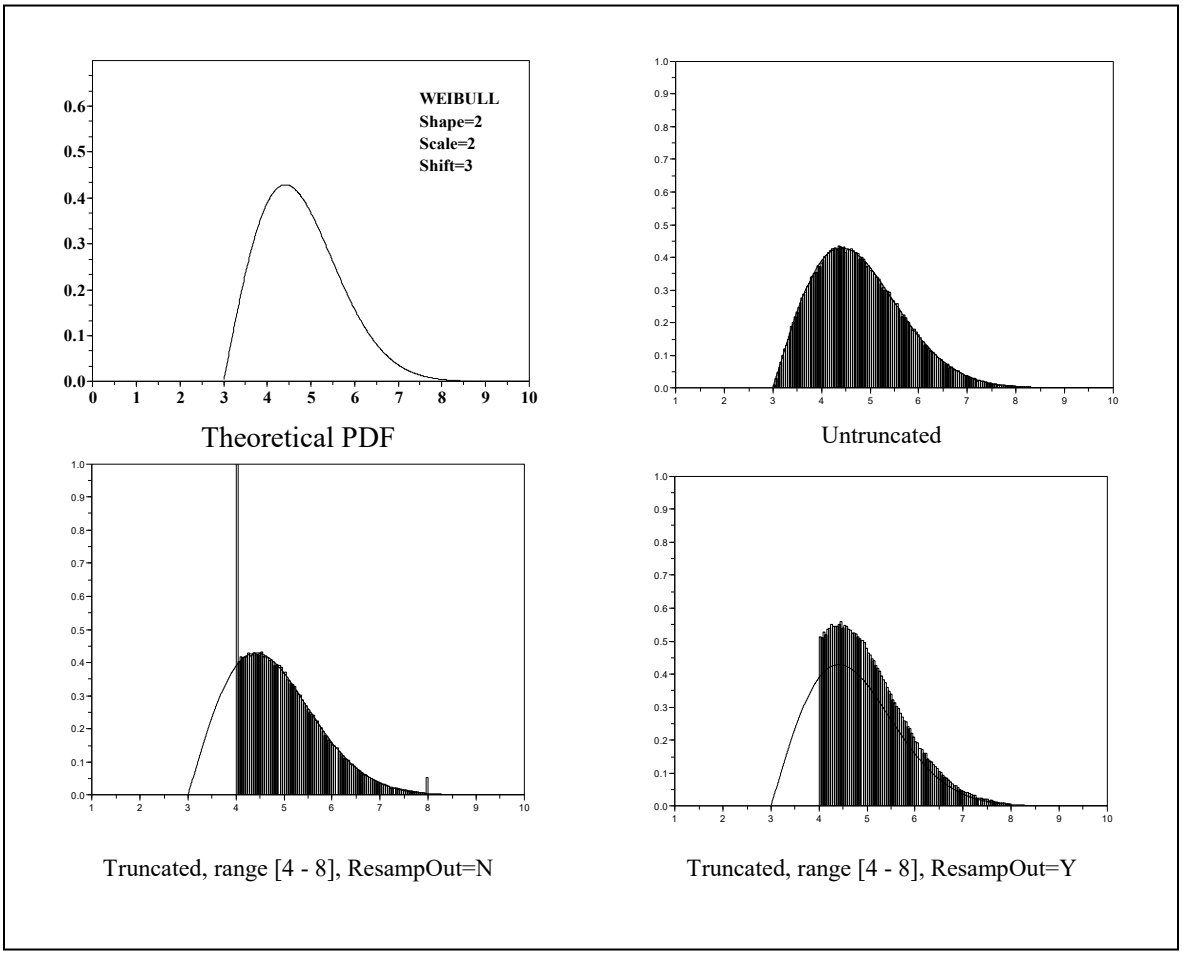

Figure 3.14. The Weibull Distribution in APEX

3.3 Random Number Generation in APEX

Beginning with APEX version 4.5, the system for generating random numbers in APEX was altered substantially, for several reasons. First, the implementation of Sobol sensitivity analysis requires that each random variable may be assigned either of two independent streams of values, without regard to the choices made for other variables. These values must be reproducible from one run to another. Second, it has been noticed that known, simple shifts in the random seeds (see below) produces predictable changes in the returned values from the "Random_Number" function in Fortran. The returned values will exhibit correlation if the seeds have a simple ratio (like 2:1, or 3:1, or 3:2). Finally, for Sobol analysis especially, but also for other purposes such as induced correlation between variables, it is useful to separate the random number generation process into two steps: 1) producing random values uniformly distributed between zero and one, and 2) transforming these using the inverse cumulative distribution function (CDF) to the desired distribution.

APEX is coded in Fortran, and "Random_Number" is the only standard built-in random number generator. This function takes as input an 8-byte integer "seed" (or equivalently, two 4-byte integer seeds), and a real vector to hold the output. The output vector becomes filled with independent random values uniformly distributed between zero and one. Random_Number resets its seed value and retains it internally for the next call, but the user may choose to reset this seed to a specific value at any time by calling the Random_Seed function.

All versions of APEX have allowed seed control. The user supplies the starting seed value for the APEX run. If two runs have identical input data including the same seed, then the entire run will be identical. This allows earlier runs to be replicated in full. Another desirable feature in APEX is the ability to generate the random values assigned to a given simulated person in a later run, without the need to reproduce the entire run. For example, suppose one runs 100,000 profiles and finds that profile number 98,765 has a particularly high exposure. It would be prohibitive to save detailed information on all 100,000 profiles. However, one can request a run with just profile number 98,765 of 100,000 to be output in detail. If the random seeds used for this profile are known, then there is no need to run the other profiles at all.

The method used in earlier versions of APEX for the above feature was to simply offset the initial seed by a known amount for each subsequent profile. This worked as long as the initial seed was large enough that the subsequent ones were unlikely to form simple ratios with it. While problems could be avoided, the potential for trouble was worrying, especially since users might be unaware of the issue. As Sobol analysis was being added in APEX version 4.5, and the random number generation system needed modification to permit this, the opportunity was taken for a larger overhaul.

Under the old system, problems could arise because the difference in seed values between persons was predictable, which in turn, could potentially lead to correlation in output between persons. This was not common, and an accidental correlation between two persons in a run would not matter much if the other persons were not correlated. However, the essence of Sobol analysis is to measure the amount of induced correlation between outputs when some random values (but not all) are reproduced. The size of this induced correlation measures the importance of the reproduced variables. Any accidental correlation that arises without reproduced variables

will act as either bias or noise, and partially mask this effect. Even when not running Sobol analysis, it would not hurt to remove the possibility of accidental correlation.

3.3.1 Random Seeds and Generation of Uniform Samples

To achieve this, a second stage of randomness has been introduced to APEX. A new function called "GenerateSeeds" has been added which internally reproduces the logic used by the Ranuni function in SAS. This has been checked by exhaustively comparing output from the SAS and Fortran versions. Actually, the returned random numbers from Ranuni are not directly used. Instead, the Ranuni logic for updating its seeds has been reproduced in APEX. These Ranuni seeds cover all integers from 1 to 2,147,483,646 inclusive, in a pseudo-random order. Every number is used once, and then the list repeats. Each Ranuni seed is a 4-byte integer.

Suppose the APEX run consists of V random variables and P persons. To model individuals without running the whole set, and to separately be able to reproduce or alter each variable, a minimum of P*V seeds are needed. APEX actually uses 2*P*V seeds, because each seed from Ranuni is 4 bytes, but Random_Number needs an 8 byte seed, so two 4 byte seeds are used. The same pair of seeds are used regardless of the number of times that variable is evaluated (for that one person). For example, if an air exchange rate is evaluated daily, then a vector of the correct length (the number of days in the simulation) is provided for Random_Number to fill, along with the seed for that combination of person and variable. The seed is determined as shown below.

1. Get the overall seed for the run from the *Control Options* file. If it is not defined, set it automatically using the clock. In either case, the initial seed is written to the log so it could be used again in a future run.
2. Once per run, generate and store a list of (2*V*P) 4-byte seeds, generated using the Ranuni logic, using the overall run seed to start the process.
3. Assign each random variable an index ranging from 1 to V. To find the seed for variable "v" for person "p", first define Base = 2*V*(p-1)+2*(v-1).
4. From the list of (2*V*P) seeds, take the two at positions (Base+1) and (Base+2).
5. Construct an 8-byte seed by concatenating these two. For a Sobol run, two possible seeds are needed for each combination of person "p" and variable "v". One of these is constructed by concatenating the two 4-byte seeds in order, and the other by using reverse order. Since all the 4-byte seeds are unique, these 8-byte seeds are always different.
6. Call Random_Number using this 8-byte seed, and return as many uniform random values as are needed for this person and variable.

In Step 6, there are three possibilities: a variable is sampled once per person, once per day, or once per hour. For an APEX simulation covering D days, that means returning one, D, or 24*D values. For

3.3.2 Transformation to Final Distributions

The second stage in producing random numbers is to transform the vector of uniform values to the appropriate final distributions. While the uniform random samples may be produced at any point before they are used, the transformations are not made until the variables are needed because the choice of transformation might depend on other variables. For example, body

weight is a random variable, but it depends on age and gender. The uniform random samples generated for "body weight" actually represents quantiles of body weight for the profiles. For example, a uniform sample of 0.5 represents the median body weight for any age and gender group. The same quantiles may be generated in multiple runs, but if the age and gender assignments have changed, so will the transformations to the final body weights. This transformation is not a random process at all, but is a deterministic function of the known variables.

APEX uses this two-step method for both Sobol runs and standard runs. Even a standard run requires (V*P) different 8-byte seeds (one for each modeling variable, for each person), which requires (2*V*P) calls to the Ranuni function, since each call returns one 4-byte seed. While a Sobol run requires twice as many 8-byte seeds, the "trick" is to combine the same two 4-byte seeds in reverse order, generating the extra 8-byte seeds without needing any more 4-byte seeds. The APEX code generates and stores a list of (2*V*P) seeds exactly the same way in all runs.

Versions of APEX before 4.5 allowed the drawing of new random samples on each diary event. This is no longer possible because the uniform random samples are effectively drawn before the activity diaries are assigned (in fact, before any properties of the simulated person are assigned), so the number of diary events each hour is not known at that point. New uniform random samples may now be drawn using one of three intervals: once per person, once per day (per person), or once per hour (per person). However, the transforms that apply to these random values are not evaluated until the variables are needed, and may still vary on each diary event. For example, the MET distribution which determines the inhalation rate may change with each diary event, as the type of activity changes. All diary events in the same clock hour will now share the same uniform random sample, and hence the same quantile of their respective MET distributions, but the final MET values will differ if the distributions differ.

3.3.3 Truncation of Distributions

The two-step process for generating random samples from distributions has another benefit. It allows truncated distributions to be evaluated without the need for new samples, even when the original samples would be beyond the truncation bounds. For the ***ResampOut*** = No option, obviously one sample is enough; it just needs to be compared to the bound(s) and shifted if necessary. The case of ***ResampOut*** = Yes is more of a concern, because the process for generating seeds and streams of random numbers (as described above) does not have any allowance for generating replacement samples.

First, evaluate the truncation bounds as quantiles of the untruncated distribution. For example, if a truncation bound chops off the upper 2% tail of the distribution, then a uniform sample (interpreted as a quantile) will remain within the bounds after transformation if it is 0.98 or less, and it will exceed the bound (and hence be unacceptable) if it exceeds 0.98. Similar logic applies to the lower truncation bound, if one is defined. The key is to realize that if the two-step generation process were repeated until the final value was in bounds, the uniform random value that produced the acceptable result would be equally likely to have any value between the two cutoff points (which are the quantiles corresponding to the truncation bounds). Hence, the original (0,1) uniform may be first mapped onto this reduced range and then transformed, and the result has the same distribution as if the resampling had actually occurred. The following steps describe this:

1. Use the CDF of the untruncated distribution to locate the quantiles corresponding to the truncation points. Call them 'qlo' (lower bound) and 'qhi' (upper bound). Set qlo=0 if no lower truncation point is defined, and set qhi=1 if no upper truncation point is defined.
2. Let 'u' be the original (untruncated) sample from the U(0, 1) distribution. Let u* = qlo + (qhi-qlo)*u . Then u* will be uniformly distributed between qlo and qhi.
3. Transform to the final value 'x' using $x = CDF^{-1}(u^*)$, where CDF^{-1} is the inverse Cumulative Distribution Function for the untruncated case.

The variable 'x' will then be distributed as if it had been repeatedly resampled from the untruncated distribution until it was within the truncation bounds. To implement this method, APEX requires the CDF and CDF^{-1} functions for the supported types of untruncated distributions, but does not need any for truncated distributions. For instance, no "CDF for Truncated Normal" is required.

The fact that the CDF and CDF^{-1} always apply to the untruncated distributions means that the parametrization (Par1-Par4) and statistical properties such as mean and variance always apply to the untruncated distributions. For example, an untruncated normal distribution with mean 5 and variance 2 may be truncated at 3 and 10. The truncated distribution will have a mean over 5 and a variance under 2. APEX does not use (or calculate) the parameters of the distribution after truncation.

There are other benefits from this procedure. The transformation from u to u* above is rank-preserving. This means that any quantile 'q' of the untruncated distribution is mapped to the same quantile 'q' of the truncated distribution. The preservation of ranks means that certain types of sensitivity analysis (specifically, in which variables are set to preselected percentiles) can be performed without being invalidated by truncation. Also, random variables could be (in principle) rank-correlated with each other, although APEX does not currently have this option.

CHAPTER 4. CHARACTERIZING THE STUDY AREA

An initial study area in an APEX analysis consists of a set of basic geographic units called sectors, typically defined as census tracts (see nomenclature definitions in Section 1.2). The user provides the geographic center (latitude/longitude) and radius of the study area. APEX contains a database with the nominal locations of every census tract in the U.S. Each tract is assigned a single point as its location, as determined by the U.S. Census Bureau. APEX calculates the distances to the center of the study area of all the sectors included in the sector location database, and then selects the sectors within the radius of the study area. One can also provide a list of counties or census tracts as part of the specification of the initial study area. APEX then maps the user-provided *Air District Location* and *Meteorology Zone Location* data to the selected sectors. The sectors identified as having acceptable air and meteorological data within the radius of the study area are selected to comprise a final study area for the APEX simulation analysis. This final study area determines the population make-up of the simulated persons (profiles) to be modeled.

See CHAPTER 3 in *Volume I* for a description of how a final study area is determined in an APEX simulation analysis.

4.1 APEX Spatial Units

4.1.1 Sectors

The demographic data used by the model to create personal profiles is provided at the sector level. For each sector the user must provide demographic information allowing the determination of age, gender, race, and work status.

The sectors are read in **SpaceTimeModule:ReadSiteLists**.

4.1.2 Air Quality Districts

The spatial units for ambient air quality data are called air quality districts. Ambient air quality data are provided as time series at specific locations.

The air quality districts are read in from the input file in **SpaceTimeModule:ReadSiteLists**. The actual air quality is read from the *Air Quality Data* files in **SpaceTimeModule:ReadAirQuality.** If the *Air Quality Data* file uses hourly distributions for each district (an optional feature of APEX, see *Volume I*), then the air data for each simulated individual is sampled in **SpaceTimeModule: GenerateAirQuality.**

4.1.3 Meteorological Zones

Another spatial unit in APEX is the meteorological zone, which is the equivalent to air quality districts but for meteorological data.

The meteorological zone locations are read in from the *Meteorology Zone Location* input file in **SpaceTimeModule:ReadSiteLists**. The actual meteorological data is read from the

Meteorology Data file in **SpaceTimeModule:ReadMeteorologyData.** When the data are read in, daily maximum and average temperatures are calculated for later use.

4.2 Determining the Final Study Area

4.2.1 Matching Sectors, Air Quality Districts, and Meteorological Zones

The APEX code for reading the locations of sectors, air quality districts, and meteorological zones as well as for pairing the sectors to the nearest air quality districts and meteorological zones, is found in **SpaceTimeModule: SelectSites**.

The final study area consists of all the sectors within ***CITYRADIUS*** of the study area central location, restricted to the listed counties or tracts (if provided), that have both an air quality district and a meteorological zone within range. If both tracts and counties are listed, then the resulting study area is the union of the two lists. Sectors for which a valid air quality district (one within ***AIRRADIUS***) or a valid meteorological zone (one within ***ZONERADIUS***) cannot be found are discarded from the final study area. The study area population is the total population in the input *Population Data* files that reside in these sectors.

The **SelectSites** subroutine makes use of the function **SpaceTimeModule: Distance** which calculates the distance between two points given their latitudes and longitudes. It is discussed below.

4.2.2 The Distance Algorithm

APEX uses a computational algorithm (implemented in **SpaceTimeModule:Distance**) to calculate the distances (e.g., study center to sector center, or sector center to air quality district center) required to determine the final study area. The method is accurate, simple in terms of program code, and works satisfactorily worldwide. The algorithm calculates the distance between two points based on their latitudes and longitudes and is based on projecting the points onto a sphere, rather than onto a plane, using the steps below.

- The latitude and longitude of the two points (locations) in question are identified
- The two sets of angles, (θ_1, ϕ_1) and (θ_2, ϕ_2), subtended at the Earth's center by the radial vectors to the two points are calculated
- The net angle between these two radial vectors is found
- The Earth's radius at the average latitude of the two points is calculated
- The results from the previous 2 steps are multiplied to give the distance between the points, D

All angles are measured in radians. In step 1, the angles ϕ subtended by the radial vectors at the Earth's center are the same as the longitudes of the points in question. However, calculating θ from latitude is more complicated since it is affected by the flattening of the poles. The intersection of the Earth's surface with a plane through the poles is an ellipse. Let "e" be the eccentricity of the ellipse. Latitude measures the angle between the polar axis and a locally horizontal surface such as sea level (a tangent surface to the Earth). For a point on an ellipse at an angle θ from the semi-major axis, the tangent line has slope s:

$$s = -(1 - e^2)\cot(\theta) \quad (4\text{-}1)$$

where the square of the eccentricity of the earth is $e^2 = 0.00672265$. This slope is equal to the negative cotangent of the latitude. Hence,

$$\theta = tan^{-1}[(1 - e^2)\ \tan(lat)] \quad (4\text{-}2)$$

When applied to both points, this gives angles θ_1 and θ_2.

The radius of the Earth varies with latitude but not with longitude. At an angle theta relative to the semi-major axis, the distance from the center to a point on the ellipse is given by:

$$R = A\sqrt{\frac{1 - e^2}{1 - e^2\ cos^2(\theta)}} \quad (4\text{-}3)$$

where A is the semi-major axis length, A = 6378.388 km. For purposes of the distance calculation, θ is set to the average of θ_1 and θ_2.

For two points at $(\theta_1, \emptyset_1)$ and $(\theta_2, \emptyset_2)$ on a sphere of radius R, the arc length of the great circle connecting them, the distance D, is given by the product of the radius and the net angle between them:

$$D = R\ cos^{-1}[\cos(\theta_1)\cos(\theta_2)\cos(\emptyset_1 - \emptyset_2) + \sin(\theta_1)\sin(\theta_2)] \quad (4\text{-}4)$$

which follows from the formula for the dot product of the radial vectors, expressed in spherical coordinates. In the code, the argument in square brackets is rounded to one when the points are coincident, to avoid numerical problems.

All of the above formulas are exact for distances on an ellipsoid, except for assuming that a single value of R applies to both points (that is, assuming that the arc of the Earth's surface between the two points lies on a sphere). In most practical cases, the error resulting from this approximation is in the range of one to ten parts per million.

4.3 Modeling Commuting

APEX models commuting by assigning a work sector to each employed individual based on commuting data for that individual's home sector. Two commuting data files are required. These are files consisting of: 1) commuting flow data (the *Commuting Flow* file), and 2) commuting time data (the *Commuting Time* file). Nationwide files are supplied with APEX. The nationwide *Population Data* and both *Commuting* input files use census tracts as the sectors. The *Population Data* files and the two commuting files must refer to the same census. As part of step 2 (Figure 2.1), APEX can extract the flows for the selected home sectors from the *Commuting Flow* file and derive profile level commuting times from the *Commuting Time* file. The development of the commuting files is described below.

4.3.1 Nationwide Commuting Flow Database

A national commuting database was generated from a set of U.S. Census files listing the number of persons living in one tract and working in another. They contain counts on individuals commuting from home to work locations at a number of geographic scales. The 2010 data come from the American Community Survey (ACS). The 5-year aggregate data is required to obtain tract-level information. The ACS also releases 1-year and 3-year data sets, but at a coarser geographical level, such as at the county level. Tract-to-tract data were used for the APEX commuting databases.

One concern identified from these data was that some home-work pairs are very widely separated geographically. For example, some people live in Alabama but work in Alaska. This may be the case for either military personnel or others who are stationed in remote workplaces for weeks or months at a time; APEX, however, requires data regarding daily commuting flows and there is no way to determine from the data which workers commute on a daily basis. A preliminary analysis of the home-work counts from the 2000 version of the data showed that a graph of Log(flows) versus Log(distance) had a near-constant slope up to somewhere around 100 kilometers (km). Beyond 150 km, the slope is also fairly constant, but flatter, meaning that flows were not as sensitive to distance. Between these two distances, there is a smooth transition from one slope to the other. A simple interpretation of this result is that at distances below 100 km, the vast majority of the flow was due to persons traveling back and forth daily, while the numbers of such persons decrease fairly rapidly with increasing distance. At large distances, the majority of the flow are persons who stay at the workplace for extended times, in which case the separation distance is not as crucial in determining the size of the flow.

To apply the home-work data to commuting patterns in APEX, a simple rule was chosen. It was assumed that all persons in home-work flows up to 120 km are daily commuters and that no persons in more widely-separated pairs commute these greater distances on a daily basis. This meant that the list of destinations for each home tract can be restricted to only those work tracts that are within 120 km of the home. In practice, this cutoff has little impact in an APEX model run since persons who live and work more than 120 km apart are not likely to have both their home and work sectors inside the same study area.

The resulting 2010 database contained a total of 73,057 distinct home sectors (roughly a 12% increase from the 2000 data) and 71,566 work sectors. The home-work tract pairs were sorted by the home tract ID, and for each home tract, the possible destinations were listed at one per line. The flows were converted into fractions of the home tract total, and then expressed as cumulative fractions since this is the form that will be needed by APEX. The work sectors were sorted by a decreasing fraction of the home tract total, and the information was written to an ASCII text file. Every record in this file contains three variables: a tract ID, a flow fraction, and a distance. A convention was used that the flow fraction and distance were set to -1.0 to indicate a home tract. The first three records on the nationwide ASCII file for 2010 data are as follows:

```
01001020100 -1.00000 -1.0
01001020200 0.19075 1.8
01001020400 0.27168 4.4
```

The first record contains the ID of the first home tract by sort order, followed by two placeholder values. The second record contains ID for the first destination tract for the current home tract. The destination tracts are sorted in order of decreasing commuting fraction. The number 0.19075 indicates that the largest percentage of commuters of the commuters who live in this home tract (around 19.1%) commute to tract 01001020200. The third number on each line is the separation distance in kilometers between the home and work tracts. The third record shows the second destination tract for this same home tract, followed by a cumulative fraction of 0.27167. This means that roughly 8.1% of the workers in that home tract go to the second work tract more than 4 km away. Further examination of this file reveals that this particular home tract has 29 destination tracts to which workers commute. Note that fractions down to 0.00001 (one in one hundred thousand) are reported. Fractions that round to zero (at this level of precision) are deleted from the database.

The number of destinations per home tract is not constant. There are two ways to know when the last destination for a given home tract has been reached. First, the cumulative flow fraction for the last destination is always 1.00000. Second, the next record in the file has negative flow and distance indicators as placeholders, indicating that the record contains the ID of the next home tract. With the 2010 data, the mean number of associated work tracts per home tract is 54, with a minimum of 1 and a maximum (within 120 km) of 294. Overall, this file has nearly 4 million records and occupies 100 megabytes of disk space.

Note that the tract list on the commuting and population files must match; consequently, an APEX run must consistently use files all based on the same census.

4.3.2 Nationwide Commuting Time Database

The ACS surveys travel time to work from work-age people (https://www.census.gov/topics/employment/commuting.html). The travel time is an estimate of the usual amount of time in minutes it took for the respondent to get from home to work each day. Travel time includes not only time spent in motor vehicles, but also time devoted to waiting for public transportation, carpooling, walking to the bus, etc.

Previously, APEX had hard-coded definitions for age bins and home workers, but with the 2010 data it must read them from the input file. Therefore, the input file requires two header lines to define the bins. For example, for the 2010 file these are:

```
number of bins = 9
boundaries = 0 5 15 20 30 45 60 75 90
```

The boundaries are the lower bounds on the one-way commuting time that goes into that particular bin. The bounding value is not actually included in that bin, so the first bin is more than zero and up to (and including) five minutes. The second bin is more than five minutes, up to (and including) 15 minutes, and so on. The final bin is open-ended in principle, but in APEX, the last bin is assumed to be 30 minutes wide. There are three more counts per line on the input file than the number of bins. The first two counts (after the tract ID) are the total number of workers and the number of non-home workers. The bin counts follow, and the last number on each line is the number of stay-at-home workers.

APEX processes the raw data into cumulative probability distributions for later use. At 6 megabytes, the file is not large; it contains one record for each census tract in the U.S. (i.e., 73,057 tracts for year 2010).

4.3.3 Implementation of Commuting in APEX

The APEX model only extracts the commuting data that pertains to the selected study area. Starting with APEX5, the commuting data may be at a coarser level than the population data. For each home sector in the study area, APEX reads in a list of possible work places, the cumulative probabilities associated with each, and the distances to each. The cumulative probabilities range from 0 to 1; the work places having higher probabilities are associated with a larger "range" of this interval. For each worker (each profile with ***Employed*** = YES), a uniform random number R from zero to one is generated. This random number is used to select a corresponding work place (the one that is associated with smallest cumulative probability that is ≥ R) for the profile. With the selection of the work place, the distance is also recorded. If the commuting data is coarser than the population sectors, then APEX randomly assigns on of the sectors in the destination place as the work sector.

When a profile's activity diary events are read from the *Diary Events* file, they are characterized as occurring in one of three places: "Work," "Home," or "Other" based on the CHAD location and activity codes. For profiles who are commuters, all "Work" events will use the work sector's air quality data when calculating microenvironmental concentrations. "Home" events will use the home sector data; "Other" events will use an average of data from different sectors (either all sectors in the area or a random selection, based on the APEX input settings). See Section 8.2.1 for more information on home, work, and other locations.

It is likely that some profiles will commute to sectors outside of the final study area. The user may choose to include these profiles in the analysis or discard them by setting the *Control Options* file variable ***KeepLeavers***. If ***KeepLeavers*** = YES, the air quality data for the work sector is not available, and it is assumed to be related to the average concentration over all of the study area air quality districts at the same point in time. Calling this average C_{avg}, the ambient concentration C for the person is:

$$C = LeaverMult * C_{avg} + LeaverAdd \quad (4\text{-}5)$$

where ***LeaverMult*** and ***LeaverAdd*** are also set in the *Control Options* file. If ***KeepLeavers*** = NO, then these individuals are still modeled but are excluded from the output tables.

Diaries are weighted by commuting time, so that people with long commute distances also have long commute durations on their diaries. For each tract included in the simulation, APEX uses the list of all possible destinations to calculate a cumulative probability distribution of commuting distances. By comparing a person's home-work distance with the tract-specific probability distribution, the person's commute distance is ranked relative to all other possible destinations in the tract.

For weighting to occur, commuting time must be estimated for each individual. Tract-level commuting time is assigned based on the commuting distance cumulative distribution function: if a person is in the 75th percentile of commuting distance, that person is placed in the 75th

percentile of commuting time. The counts for the commuting time bins are converted to cumulative probabilities. The random number used to select the commuting distance is compared to these probabilities to identify the correct time bin, and then linear interpolation is used between the boundaries of that bin to obtain a specific commuting time.

Additionally, if modeling commuting, then commuting time is a required input on the CHAD *Diary Questionnaire* file. By modifying the *Diary Questionnaire* file, users have the option to use any method of assigning commute time to diaries they wish. In the version of the *Diary Questionnaire* file included with APEX, we defined commuting time as the total time spent in travel locations or activities before and after work activity. Commuting time is calculated on all weekdays for all employed people in CHAD. Commuting activities are defined by either the activity codes 18200—18240, or the location codes 31000—31910. To match the census bins, a cap of 120 minutes (240 minutes total for the day) was placed on the computed daily commute time.

Since commuting time both is a personal variable and a parameter on the *Diary Questionnaire* file, diaries can now be weighted for selection based on commuting time. Previous versions of APEX weighted on three categories: gender, employment, and age. Commuting time is the fourth category. Weights, however, are only applied for employed individuals on weekdays, and diary selection for unemployed people is not affected by the change.

The user has the option to set two commuting time windows. All diaries within the first time window are weighted at 100%. Diaries within the second window, diaries outside both windows are weighted at a level determined by the user using keywords in the *Control Options* file. These variables may be modified to match the desired strength of the correlation to commuting time. For no correlation, all weights can be set to 1.0. These windows are in units of minutes (as opposed to a percentage of the target time), so that matching of diaries applies equally well for both low and high targets.

CHAPTER 5. GENERATING SIMULATED INDIVIDUALS (PROFILES)

APEX stochastically generates a user-specified number of simulated persons to represent the population in the study area. Each simulated person is represented by a "personal profile." The personal profile (see step 2 in Figure 2.1) is a set of parameters that describe the person being simulated. The simulated person has a specific age, a specific home sector, a specific work sector (or does not work), specific housing characteristics, specific physiological parameters, and so on.

The profile does not belong to any particular real person, instead it is a fictional or simulated person. A single profile does not have much meaning in isolation, but a collection of profiles represents a random sample drawn from the study area population. This means that statistical properties of the collection of profiles should reflect statistical properties of the real population in the study area.

APEX generates the simulated person or profile by probabilistically selecting values for a set of profile variables. The profile variables fall into the five categories below.

- Demographic variables, which are generated based on the census data;
- Residential variables, which are generated based on sets of user-defined distribution data;
- Physiological variables, which are generated based on age- and gender-specific distribution data; and
- Daily varying variables, which (in most cases) are based on distribution data that change daily during the simulation period.
- Modeling variables, which can have a variety of meanings, and are usually defined by the user. The individual profile variables are given in Table 5.1.

The profile variables do not depend on the results for any other profile, and in general do not depend on any other results (diary selection, microenvironment concentrations, exposure, or dose) for the current profile. The exception is the person's physical activity index (PAI), which is calculated using the activity diary.

At present, APEX does not allow the user to select which variables are part of the personal profile. If such an enhancement were included, the user would be able to assign (say) a family size to each profile using input files rather than by reprogramming the source code.

The process by which APEX generates the personal profile is illustrated in Figure 5.1. The demographic and residential variables are set in **ProfileModule:GenerateProfiles,** the physiological variables are generated in **ProfileModule:GeneratePhysiology**, and the daily-varying variables are set in **ProfileModule:GenerateDailyVars**.

The following subsections describe the different categories of profile variables in more detail.

Table 5.1. Profile Variables in APEX

Variable Type	Profile Variable	Description
Demographic variables	Index	Internal APEX profile index number
	Gender	Male/Female
	Race	White, Black, Native American, Asian, or Other
	Age	Age (years)
	Profile group	A user-generated profile factor that can vary by age, gender, and sector.
	Home sector	Sector in which a simulated person lives
	Work sector	Sector in which a simulated person works
	Home district	Air quality district assigned to home sector
	Work district	Air quality district assigned to work sector
	Meteorological zone	Meteorological zone assigned to home sector
	Employment status	Indicates employment outside home
	Commuting distance	Distance (km) from home to work
	Commuting time	One-way home-work commuting time (min)
Residential variables	Gas stove	Indicates presence of gas stove
	Gas pilot	Indicates presence of gas pilot light
	Home air conditioning	Indicates type of home ventilation/air conditioning system
	Car air conditioner	Indicates presence of air conditioning in the simulated person's car
Physiological variables	Blood volume	Blood volume of simulated person (ml)
	Body mass	Mass of simulated person (kg)
	Body mass index	Body Mass Index (lbs/in^2)
	Weight	Body weight of a simulated person (lbs)
	Height	Height of a simulated person (in)
	Resting metabolic rate	Resting metabolic activity rate (kcal/min)
	Body surface area	Surface area of the body (m^2)
	Maximum permitted MET value	Maximum multiple of resting metabolic rate that can obtained by the individual (dimensionless)
	Energy conversion factor	Oxygen uptake per unit of energy expended (liters/kcal)

Variable Type	Profile Variable	Description
	Lung CO diffusivity	Lung CO diffusivity parameter used in COHb calculation (ml/min/torr)
	Endogenous CO production rate #1	Endogenous (internally produced) CO production rate #1 (ml/min)
	Endogenous CO production rate #2	Endogenous CO production rate #2 (ml/min) used only for women between ages of 12 and 50 for half the menstrual cycle
	Hemoglobin altitude factor	Correction for blood hemoglobin density at high altitudes
	Recovery time	Time to recover a maximum oxygen deficit (hours)
	Ventilation slope	Slope term for MET to ventilation conversion
	Ventilation intercept	Intercept term for MET to ventilation conversion
	Ventilation residual	Residual term for MET to ventilation conversion
	Hemoglobin density in the blood	Amount of hemoglobin in the blood (g/ml)
	Normalized maximum oxygen consumption	Maximum rate of oxygen consumption by the individual, normalized to body mass (ml of oxygen/min/kg)
	Starting day of menstrual cycle	The day during the first 28 days of the simulation period that menstruation begins; used for calculating endogenous CO
	Disease status	Whether or not the person has the disease. Probability is determined by the input *Prevalence* file.
	β1 – β9	Model parameters for the %ΔFEV1 ozone model
	Variance of U	Model parameter for the %ΔFEV1 ozone model, describes the variability in the effect on an individual
	%ΔFEV1 – age slope	The slope of the age - %ΔFEV1 relationship
	%ΔFEV1 – age y-intercept	The y-intercept of the age - %ΔFEV1 relationship
	%dFEV1 – body mass average	Constant subtracted from body mass in regression equation
	Maximum oxygen deficit	Maximum oxygen deficit obtainable by profile (ml/kg)

Variable Type	Profile Variable	Description
Daily varying variables	Daily endogenous CO production rate	Daily endogenous CO production rate in the simulation period (ml/min)
	Diary number	Index for the selected diary for the day
	First event #	Index of the first event for the day in the composite activity diary
	Diary ID	CHAD (or other database) ID for the diary selected for the day
	Diary age	Age of the CHAD diary selected for the day
	Diary employment	Employment status of the CHAD diary selected for the day
	Last event #	Index of the last event for the day in the composite activity diary
	Residence window position	Daily residence window position (open or closed) during the simulation period
	Car window position	Daily car window position (open or closed) during the simulation period
	Daily average car speed category	Daily average car speed category during the simulation period
	Daily conditional variable # 1	Generic user-defined daily conditional variable # 1
	Daily conditional variable # 2	Generic user-defined daily conditional variable # 2
	Daily conditional variable # 3	Generic user-defined daily conditional variable # 3
	PAI	Daily physical activity index (MET, dimensionless)
	Diary key statistic	Key diary statistic (such as outdoor time of vehicle time) for the day. Used for constructing longitudinal activity diary. Equal to 0 if not using longitudinal assembly.
Modeling Variables	Number of diaries	Number of different activity diaries used to construct the composite activity diary for the person.
	Profile conditional variable #1	Generic user-defined conditional variable # 1.
	Profile conditional variable #2	Generic user-defined conditional variable # 2.
	Profile conditional variable #3	Generic user-defined conditional variable # 3.
	Profile conditional variable #4	Generic user-defined conditional variable # 4.

Variable Type	Profile Variable	Description
	Profile conditional variable #5	Generic user-defined conditional variable # 5.
	Regional conditional variable #1	Regional user-defined conditional variable # 1.
	Regional conditional variable #2	Regional user-defined conditional variable # 2.
	Regional conditional variable #3	Regional user-defined conditional variable # 3.
	Regional conditional variable #4	Regional user-defined conditional variable # 4.
	Regional conditional variable #5	Regional user-defined conditional variable # 5.
	AQ conditional variable #1	User-defined conditional variable #1 that depends on AQ during event
	AQ conditional variable #2	User-defined conditional variable #2 that depends on AQ during event
	AQ conditional variable #3	User-defined conditional variable #3 that depends on AQ during event
	AQ conditional variable #4	User-defined conditional variable #4 that depends on AQ during event
	AQ conditional variable #5	User-defined conditional variable #5 that depends on AQ during event

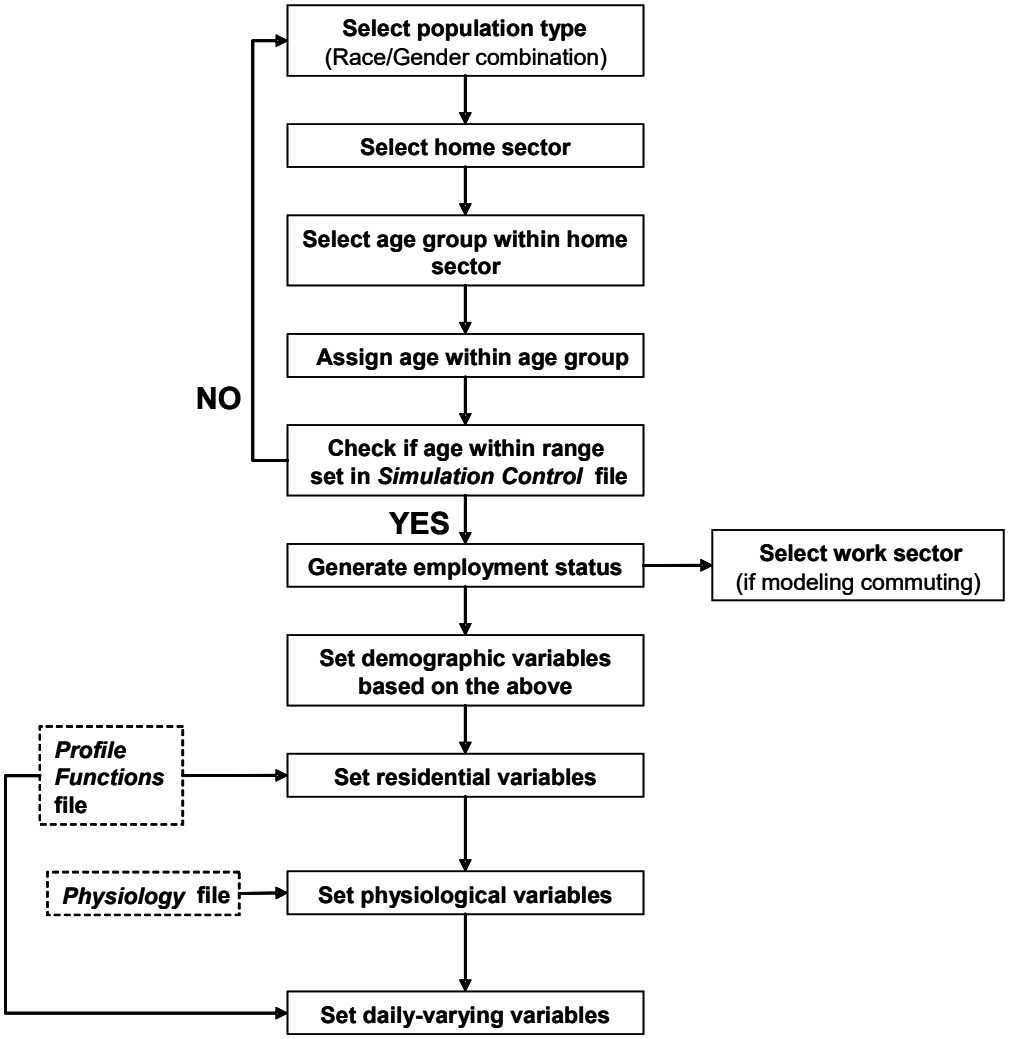

Figure 5.1. Generating a Simulated Profile

5.1 Demographic variables

The profile variables of a demographic nature are listed below.

- ***Age***: Age in years (integer)
- ***Employed***: YES if employed outside home, NO otherwise
- ***Gender***: Male or Female
- ***Profile Factor Group***: A user defined group that can be flexibly defined to be any factor that varies by location, age and gender, and which may include occupation.
- ***HomeD***: Number of the air quality district assigned to home sector
- ***HomeSec***: Number of the home sector for this profile
- ***Race***: White, Black, Native American, Asian, or Other
- ***WorkD***: Number of the air quality district assigned to work sector
- ***WorkSec***: Number of the work sector (if any); same as ***HomeSec*** for non-workers
- ***Index***: Sequential number indexing the set of profiles

- ***Zone***: Number of the meteorological zone for the home sector
- ***CommDist***: The distance from the home to the work tract
- ***CommTime***: The one-way commute time between home and work

The values of demographic variables (gender, age, etc.) for a simulated profile are selected probabilistically according to the steps listed below (see Figure 5.1).

- The fractions of people in each of the gender/race combinations ("population types") in the final study area are calculated and then used as probabilities to randomly select a gender/race type for a simulated individual.
- The fractions of the selected population type in each sector within a study area are found and then used as probabilities to randomly select a home sector for the simulated person.
- The fractions of people in each age group in the selected population type in the selected sector are calculated and then used as probabilities to randomly select an age group for the simulated person.
- A specific age within the selected age group is randomly selected, assuming a uniform distribution.
- The employment probability for the selected age group (for the home sector) is used to randomly determine whether a simulated person will work.
- If the user specifies a profile factor, then the profile factor probability for the selected age group will be used to randomly determine the group of the individual.
- If the commuting option is used and a simulated person works, then use the fractions of people commuting to each of the work sectors for the selected home sector to randomly select a work sector to which the simulated person will commute. If the commuting option is not used, then assume a person who works does so in their home sector.
- When the commuting option is used, APEX estimates the commuting time in addition to the commuting distance (i.e., the distance from the home tract to the work tract). The commuting time is based on the length of the individual's commute compared to all commuters in their home tract. Longer commute lengths lead to longer commute durations.

5.2 Residential Variables

The residential variables are set after the demographic variables described above. The residential variables are categorical variables that are used to indicate whether a residence or a car associated with a simulated person has a specified appliance or component that may affect exposure. The rules for assigning these four variables are specified by the user in the *Profile Functions* input file (see *Volume I*). The residential variables are listed below.

- ***AC_Home***: An integer indicating the type of home ventilation systems (e.g., central air, attic fan, window unit, no A/C).
- ***AC_Car***: YES, if person has air conditioning in car; NO otherwise
- ***HasGasStove***: YES, if person has gas stove in home; NO otherwise
- ***HasGasPilot***: YES, if person has gas pilot for gas stove in home; NO otherwise

APEX randomly determines the result for the variable based on the probabilities specified in the input files. For example, suppose a user specifies probabilities of 0.3 for not having an air-

conditioned car (***AC_Car*** = NO) and 0.7 for having a car with air conditioning (***AC_Car*** = YES). APEX randomly generates a value in the range of 0 to 1, assuming a uniform distribution. If this value is larger than 0.3, the simulated person will own an air conditioner, otherwise not.

These variables may be used as conditional variables for calculating concentrations in microenvironments (see CHAPTER 8).

5.3 Physiological Profile Variables

The physiological variables are used for estimating ventilation, calculating dose (as described in CHAPTER 7 and CHAPTER 10) and classifying profiles in reporting results. This section covers the calculation of these variables. The profile variables relating to physiology are listed below.

- ***Blood***: Volume of blood (milliliters)
- ***BM***: Body mass (kilograms)
- ***BSA***: Body surface area (m^2)
- ***Diff***: Lung CO diffusivity(ml/min/torr)
- ***ECF***: Energy conversion factor (liters of oxygen per kcal)
- ***Endgn1***: Endogenous CO production rate #1 (ml/min)
- ***Endgn2***: Endogenous CO production rate #2 (ml/min)
- ***Height***: Body height (inches)
- ***Hemfac***: Hemoglobin altitude adjustment
- ***Hmglb***: Hemoglobin density (grams per milliliter of blood)
- ***METmax***: Maximum obtainable MET value (unitless)
- ***MOXD***: Maximum obtainable oxygen debt (ml/kg body mass)
- ***NVO2max***: Maximum normalized oxygen consumption (milliliter of oxygen per minute per unit body mass)
- ***PAI***: Median daily physical activity index (MET, dimensionless)
- ***RMR***: Resting metabolic rate (kcal/min)
- ***RecTime***: Oxygen debt recovery time (hours)
- ***Start***: Starting day (phase) for menstrual cycle (if applicable)
- ***VeSlope***: Slope for MET to ventilation rate conversion
- ***VeInter***: Intercept for MET to ventilation rate conversion
- ***VeResid***: Residual for MET to ventilation rate conversion
- ***Weight***: Body weight (pounds)
- ***Ill***: Flag indicating whether or not a person has the disease that is modeled in the *Prevalence* input file.
- ***BMI***: Body Mass Index (kg/m^2)
- ***B1 – B9:*** Model parameters used in the calculation of %ΔFEV1
- ***FEVU***: A variance term describing an individual's responsiveness to ozone (used for %ΔFEV1)
- ***FEVE1 & FEVE2***: Terms that describe error involved in %ΔFEV1 measurements
- ***FEVBMI***: The constant subtracted from body mass index in the regression equation
- ***FEVSLP***: The slope of the linear regression relating %ΔFEV1 and age
- ***FEVINT***: The y-intercept of the linear regression relating %ΔFEV1 and age

The physiological profile variables do not affect the exposure calculations for individuals in any way, as none of them affect either the selection of diaries or the concentrations in the various microenvironments. However, the physiology of a profile affects the ventilation calculations and thus influences (1) the calculation of dose (see CHAPTER 10) and (2) the calculation of exertion level, which is used to group exposure results in the output *Tables* file (see *Volume I*).

The physiological variables are calculated from input data. The *Physiology* input file contains distributions that are used to set the physiological profile variables and parameters for each simulated person. Specifically, the *Physiology* file contains distributions for the following variables and parameters for each age and gender cohort (see *Volume I* for more information):

- ***NVO2max***: normal distributions for ***NVO2MAX*** in ml O_2/min/kg (Isaacs and Smith, 2005)
- ***BM***: lognormal distributions for BM in kg, for ***HTWTMETHOD*** = 1 (Isaacs and Smith, 2005)
- ***RMRSlp***, ***RMRInt***, and ***RMRErr***: Point distributions for regression coefficients for RMR (slope, intercept, and error) as a function of BM in MJ/Day, for ***RMRMETHOD*** = 1 (Johnson et al., 2000, adapted from Schofield, 1985)
- ***RMR_BM***, ***RMR_LBM***, ***RMR_Age***, ***RMR_LAge***, ***RMR_HT***, ***RMR_LHt***: Point distributions for regression coefficients for RMR when using ***RMRMETHOD*** = 2. These must be used as a set (meaning all must be specified, when this method is used), but they may still be present on the *Physiology* file even when using the other method. The six coefficients refer to body mass (BM), log(BM), age, log(1+age), height(in meters), and log(height). If the user chooses to use a regression without one or more of these, define the coefficient to have a point value of zero for all cases (ICF report to EPA, 2017).
- ***BldF1*** and ***BldF2***: Point distributions for blood volume factors for calculation of volume (ml) from height and weight (Johnson et al., 2002)
- ***Hmg***: Normal distributions for blood hemoglobin density (g/lm of blood) (Johnson et al., 2002)
- ***HtTSlp***, ***HtInt***, and ***HtErr***: Point distributions for regression coefficients for height in inches (slope, intercept, and error) as a function of age (children 0–17) or body mass (adults), for ***HTWTMETHOD*** = 1 (Johnson et al., 2000)
- ***Height***, ***LogBM***, and ***HtWtCorr***: Normal distributions for height and natural logarithm of body weight, and point distribution for ***HtWtCorr***, which is the correlation between the two for each age-gender combination, for ***HTWTMETHOD*** = 2 (ICF report to EPA using NHANES data, 2016)
- ***BSAExp1*** and ***BSAExp2***: Point distributions for exponential parameters for calculating body surface area (m^2) as a function of body mass (Burmaster, 1998)
- ***MOxD***: Normal distributions for maximum obtainable oxygen debt in ml per kg body weight (Isaacs et al., 2007) **.** ***MOxD*** is used in the adjustment of MET values; see Section 7.2.
- ***ECF***: Uniform distributions for the energy conversion factor for the person. Liters of oxygen per kcal. (Johnson et al., 2000, adapted from Esmail et al., 1995)
- ***RecTime***: Uniform distributions for the time required to recover a maximum oxygen deficit (hours). (Isaacs et al., 2007) ***RecTime*** is used in the adjustment of MET values; see Section 7.2.
- ***Endgn1*** and ***Endgn2***: Point distributions for endogenous CO production rates in ml/min.

(***Endgn2*** is used for women in 2nd half of menstrual cycle).

- **β1 – β9**, **FEVu**, **FEVE1**, **FEVE2**: Distributions for model parameters used by the %ΔFEV1 ozone calculations.
- ***FEVBMI***: Point value to be subtracted from the body mass index.
- ***FEVSlp*** and ***FEVInt***: Point estimates describing the linear regression relationship between %ΔFEV1 and age.

The *Ventilation* input file contains data for a set of regression parameters used by APEX to estimate the ventilation rate VE. Two methods are now supported, which require different inputs on this file. The first method (the only one available until 2017) allows estimation of the VE-related profile variables ***VeSlope***, ***VeInter***, and ***VeResid***. See *Volume I* for an example of the *Ventilation* file; see Graham and McCurdy (2005) for the derivation of these parameters. The file contains the following parameters for 5 age groups:

- ***VEb0*** and ***VEb0SE***: Mean and standard error for regression parameter b0
- ***VEb1*** and ***VEb1SE***: Mean and standard error for regression parameter b1
- ***VEb2*** and ***VEb2SE***: Mean and standard error for regression parameter b2
- ***VEb3*** and ***VEb3SE***: Mean and standard error for regression parameter b3
- ***VEeb***: Interpersonal variance
- ***VEew***: Intrapersonal variance

Note: the standard error terms defined in the *Ventilation* file are not currently used by APEX, but could be used for future uncertainty analyses.

The second option for VE calculations (***VEMETHOD*** = 2, which is the default starting in 2017) requires point values for the intercept, ln(VO2), (VO2/VO2max)^4, eb (interpersonal variance), and ew (intrapersonal variance), for each age group. The analysis supporting the default values for this method found that all age groups (that is, 0–100) may be fit by the same set of coefficients.

Once the above parameters are read/set by APEX, the physiological profile variables are formulated as follows in **ProfileModule:GeneratePhysiology** using the appropriate parameter values for the profile age and gender. First, ***NVO2max***, ***BM***, ***RMR***, ***Hmg***, ***BldF1, BldF2***, ***HtSlp, HtInt, HtErr, MOxD, ECF, RecTime , BSAExp1, BSAExp2, Endgn1, Endgn2, β1-β9, FEVu, FEVSlp,*** and ***FEVInt*** are sampled from the appropriate input distribution (using the APEX sampling methods, Section 3.2) for the profile's age and gender.

There are two methods for generating height and weight in APEX. In ***HTWTMETHOD*** = 1, the body mass BM is calculated from age and gender-specific distributions, and then height is calculated later. This option is discussed first. For ***HTWTMETHOD*** = 2, weight and weight are sampled from bivariate distributions, as discussed further below.

Once the basic physiological variables have been set, the other profiles variables are calculated as follows:

$$Weight = 2.2046\,(BM) \tag{5-1}$$

where ***BM*** is in kg and weight is given in pounds (lbs.). ***BMI*** is calculated as:

$$BMI(kg\ m^{-2}) = \frac{703\ Weight\ (lbs)}{Height^2\ (in^2)} \tag{5-2}$$

BSA (in m^2) (Burmaster, 1998) is calculated as:

$$BSA = e^{BSAEXP1} BM^{BSAEXP2} \tag{5-3}$$

RMR is given in units of kcal/min. ***RMRMETHOD*** = 1 uses regressions from (Johnson et al., 2000). All of them have the following form:

$$RMR = 0.166(BM \times RMRSlp + RMRInt + RMRErr) \tag{5-4}$$

where 0.166 is the conversion factor for converting MJ/day to kcal/min. The parameters ***RMRSlp***, ***RMRInt***, and ***RMRErr*** are different for each of several age-gender groupings. The first two parameters (***RMRSlp*** and ***RMRInt***) are point values, and ***RMRErr*** is a normal distribution with mean zero and a specific standard deviation for each group.

Under ***RMRMETHOD*** = 2, the population is still divided into several age-gender groupings, but now three independent variables are used (***BM***, ***Age***, and ***Height***), along with three logarithmic terms. One important point is that the new regression was developed using different units: ***RMR*** is in kcal/day and height (HT) is in meters. The new form is

$$\begin{aligned} BMterm &= RMR\_BM * BM + RMR\_LBM * \ln(BM) \\ AgeTerm &= RMR_{AGE} * Age + RMR_{LAGE} * \ln(1 + age) \\ HTterm &= RMR\_HT * HT + RMR\_LHt * \ln(HT) \\ RMR &= BMterm + AgeTerm + HTterm + RMRInt + RMRErr \end{aligned} \tag{5-5}$$

BM is body mass in (kg), ***Age*** is in full years, and HT = ***Height***/39.37 because "height" in APEX is reported in inches. There are six coefficients, for ***BM***, ln(BM), age, ln(1+***Age***), HT, and ln(HT). The logarithm of age uses a shifted age to prevent taking a logarithm of zero for children under one year. In effect, age is rounded up to the nearest integer instead of rounded down, so a newborn starts life at 1. No difference would result by similarly using ***RMR_Age****(1+Age), because that would simply add a constant amount in all cases, which would then be removed by altering the intercept term ***RMRInt*** by the same amount in the other direction. Hence, the regression line would not change. As usual, ***RMRErr*** is sampled from a normal distribution with mean zero for each group.

The user may also choose another regression form, for example without HT or ln(HT). An example was provided in the same memo detailing the above regression. In that case, set the coefficients of the missing variables to have point values of zero.

After producing ***RMR*** in (kcal/day) using the above equation, it must be converted to the APEX units of (kcal/min) by dividing by 1440 minutes per day.

The maximum MET value an individual can obtain is given by their personal maximum energy expenditure divided by their ***RMR***. Personal maximum energy expenditure is simply maximum oxygen consumption converted to kcals via ECF, and thus ***METmax*** can expressed as:

$$METmax = \frac{(0.001)(BM)(NVO2max)}{ECF * RMR} \quad (5\text{-}6)$$

where 0.001 is the conversion factor for ml to liters O_2 such that ***METmax*** is a unitless number. The lower bound for ***METmax*** is 5, and the higher bound is 20; values outside of this range are set to the corresponding bound.

Under ***HTWTMETHOD*** = 1, for children under the age of 18, height (in inches) is a function of age (in years):

$$Height = HTINT + age \times HTSLP + HTERR \quad \text{(children under 18)} \quad (5\text{-}7)$$

for adults, height (in inches) is a function of body weight (in pounds):

$$Height = HTINT + ln(weight) \times HTSLP + HTERR \quad \text{(adults 18 and older)} \quad (5\text{-}8)$$

Alion staff fit these equations for height, and derived the accompanying coefficients.

HTWTMETHOD = 2 uses the empirically derived relationship that for all age and gender combinations, the height and the logarithm of body weight form a bivariate normal distribution. To use this method, set ***HTWTMETHOD*** = 2 on the *Control Options* file, and ensure that distributions are specified for ***LogBM***, ***Height***, and ***HtWtCorr*** on the *Physiology* input file. ***LogBM*** is the natural logarithm of body weight in kilograms, height is in centimeters (unfortunately, this differs from both standard APEX usage, which is inches, and from the variable in the ***RMR*** regressions, which is in meters). Both ***LogBM*** and ***Height*** should be normal distributions. ***HtWtCorr*** must be a point value, but in general is age- and gender-specific. APEX now has equations for generating samples from correlated bivariate normal distributions.

Several physiological variables are calculated differently for males and females, including lung CO diffusivity in ml/min/torr (Johnson et al., 2000). Lung CO diffusivity is only used in the APEX CO dose algorithm:

$$Diff = 0.361(Height) - 0.232(Age) + 16.3 \quad \text{(males)} \quad (5\text{-}9)$$

$$Diff = 0.556(Height) - 0.115(Age) - 5.97 \quad \text{(females)} \quad (5\text{-}10)$$

where height is in inches and age in years. Blood volume in milliliters (Johnson et al., 2000) is calculated as:

$$Blood = BLDF1(Weight) + BLDF2(Height)^3 - 30 \quad (5\text{-}11)$$

where weight is in pounds and height in inches.

For males (when using ***VEMETHOD*** = 1), ***VeInter*** is calculated as:

$$VeInter = Intterm - [VEB3] \qquad \text{(males)} \qquad (5\text{-}12)$$

and for females

$$VeInter = Intterm + [VEB3] \qquad \text{(females)} \qquad (5\text{-}13)$$

where

$$Intterm = VEB0 + Z(VEEB) + [VEB2][ln(1 + Age)] \qquad (5\text{-}14)$$

where Z is a number drawn from a unit normal distribution ranging from -4 to 4.

For both genders:

$$VeSlope = VEB1 \qquad (5\text{-}15)$$

$$VeResid = VEEW \qquad (5\text{-}16)$$

The variables ***VeSlope***, ***VeInter***, and ***VeResid*** are used in the APEX ventilation algorithm ***VEMETHOD*** = 1 (see CHAPTER 7 and Graham and McCurdy, 2005). ***VEMETHOD*** = 2 does not require the use of these three variables.

5.4 Daily-Varying Variables

The daily varying variables are generated for each day in the model simulation for each profile in **ProfileModule:GenerateDailyVars**. There are eight daily-varying profile variables:

- ***DailyConditional1***: User-defined daily conditional variable #1
- ***DailyConditional2***: User-defined daily conditional variable #2
- ***DailyConditional3***: User-defined daily conditional variable #3
- ***Endgn***: Endogenous CO production rate on given day (ml/min)
- ***PAI***: Daily physical activity index (MET, dimensionless)
- ***SpeedCat***: Category for average vehicle speed on given day
- ***WindowRes***: Residence window status (open or closed) for given day
- ***WindowCar***: Car window status (open or closed) for given day
- ***DiaryID***: ID of the activity diary selected for the day (as it appears in the *Diary Questionnaire* and *Diary Events* files)
- ***DiaryEmp***: Employment status of the activity diary selected for the day
- ***DiaryAge***: Age of the activity diary selected for the day
- ***Stat***: Value of the key diary statistic for the activity diary selected for the day

The rules for defining WindowRes, WindowCar, SpeedCat, DailyConditional1, DailyConditional2, and DailyConditional3 are provided by the user in the *Profile Functions* input file. These variables may be used as conditional variables for use in calculating the concentrations in microenvironments. The *Profile Functions* file and conditional variables are

covered in detail in *Volume I.* Conditional variables are also explained further in the Microenvironments chapter (Chapter 8) of this Volume.

The variable, Endgn, is used in the evaluation of the blood COHb level. Even though it is stored in an array of daily values, for males, it always has the same value from day to day; that is, it is always set to the physiological profile variable Endgn1. For females, it may have one of two values (the physiological profile variables Endgn1 and Endgn2), depending on the phase of the menstrual cycle.

The variable PAI is calculated as the time-averaged MET value for the entire simulation day. As it depends on the daily activity diaries, it is not set in the profile module, but rather later, in the ventilation subroutine **ExposureDoseModule:Ventilation**.

The variables DiaryID, DiaryAge, and DiaryEmp are primarily used for QA of the diary selection and diary assembly methods. The variable Stat contains the value for the key diary variable, which is used to construct the longitudinal diary, and thus is usually a variable important to exposure (such as time spent outdoors or time spent in vehicles). See Section 6.3. These variables are all set in **DiaryModule:CompositeDiary**.

5.5 Modeling Variables

The remaining profile variables are general modeling variables. They are:

- ***Number of diaries***: Number of different activity diaries used to construct the composite activity diary for the person.
- ***Profile Conditional Variable #1-5***: Generic user-defined profile conditional variables. Can be defined by the user model any property of the simulated person (for example, behavior or residential properties) that may affect microenvironmental parameters.
- ***Regional Conditional Variable #1-5***: Regional user-defined conditional variables. Can be defined by the user model any property of the simulated person (for example, behavior or residential properties) that vary by county or sector.
- ***Air Quality Variable #1-5***: User-defined conditional variables that depend on the ambient air quality during the event, and are recalculated with each timestep.

The number of diaries is a convenience variable that is set during the diary assembly process, in **DiaryModule:CompositeDiary**. It depends on the number of appropriate diaries available in CHAD for the person, the diary pools, and other factors. The quantity of diaries affords the modeler an idea of the heterogeneity in the diary selection for the person.

The profile and regional conditional variables are defined by the user in the *Profile Functions* file (see *Volume I*). They are set in ***ProfileModule:GenerateProfiles.***

CHAPTER 6. CONSTRUCTING A SEQUENCE OF DIARY EVENTS

APEX probabilistically creates a composite diary for each of the simulated persons by selecting a 24-hour diary record—or diary day—from an activity database for each day of the simulation period. The Consolidated Human Activity Database (CHAD) has been supplied with APEX for this purpose. A composite diary is a sequence of events that simulate the movement of a modeled person through geographical locations and microenvironments during the simulation period. Each event is defined by geographic location, start time, duration, microenvironment visited, and activity performed. Events crossing a time step boundary will be separated into two distinct events.

The APEX model generates sets of exposure time series—one for each simulated individual—and both mean exposures over time and variation in exposures are important. The ability to realistically reproduce these exposure metrics depends on the method used to construct the composite activity diaries for the simulated population. APEX provides three methods of assembling composite diaries. The first (basic) method, which is adequate for estimating mean exposures, simply involves randomly selecting an appropriate activity diary for the simulated individual from the available diary pool. The second method is a more complex algorithm for assembling longitudinal diaries that realistically simulates day-to-day (within-person) and between-person variation in activity patterns (and thus exposures). The third method uses a Markov-chain clustering algorithm to recreate realistic patterns of day-to-day variability. All methods are covered in this chapter.

As all methods of composite diary assembly require the creation of diary pools, their construction is covered first.

6.1 Constructing the Diary Pools

6.1.1 Diary Data

The composite diary is created by concatenating individual one-day activity diaries from the CHAD database. APEX currently provides this database as two files: a file containing personal information of the studied individual, and one containing the actual diary events (see *Volume I* for more information). The *Diary Questionnaire* file contains the following variables, in this order, as comma-separated values:

- CHAD ID
- Day of the week (e.g., Monday)
- Gender (M/F)
- Race
- Status of employment (Y/N)
- Age (years)
- Maximum hourly temperature on day of study (°F)
- Average temperature on day of study (°F)
- Occupation
- Count of missing time in minutes (when activity and/or location codes are missing from

the *Diary Events* file)
- Number of events
- Commuting time (in minutes; only required for simulations modeling commuting)

The *Diary Events* file contains the following variables as comma-separated values:

- CHAD ID
- Start time (of the event)
- Duration (minutes)
- Activity code
- Location code

See *Volume I* for a description of the CHAD activity and location codes. The *Diary Events* file contains a record for each of the events indicated by the number of events in the *Diary Questionnaire* file. Note that while CHAD data are provided with APEX, other activity data could be used instead as long as the input file formats are followed and the CHAD coding conventions are used.

Diary Occupation is an optional file in which users can specify new occupations or occupation groups for each individual. These occupations overwrite those found in the *Diary Questionnaire* file. This file consists of two columns: CHAD ID, and a string containing the occupation of the individual. This file may be useful for matching via occupation group.

The APEX diary input files are read by **DiaryModule:ReadDiaries**.

6.1.2 Grouping the Available Diaries into the Diary Pools

A diary pool is a group of CHAD diaries, appropriate for a given simulation day, from which a daily diary may be drawn. The criteria for creation of the Diary Pools are defined using the DiaryPool variable in the *Profile Functions* input file (see *Volume I*). Briefly, the user can define different pools for different combinations of ranges of maximum temperature, average temperature, and day of the week. Thus, the definition of a diary pool for a single (hypothetical) simulated day may be something like "all CHAD activity diaries for a weekend day for which the maximum temperature was between 70 and 80 degrees, and the average temperature was between 50 and 80 degrees." The idea behind this logic is that temperature and day of the week affect the type of activities people perform, and thus it is important to match these real properties of the activity diaries to corresponding simulated days. It should be noted that in APEX, CHAD diaries that are missing temperature data are thrown out (that is, not assigned to any diary pool.)

The number of diary pools that are defined affects the number of diaries that available for selection on a given day. Therefore, it is important not to define the pools too narrowly, as it could result in the same activity diaries being selected over and over again, or could result in pools with no diaries in which case APEX will fail with a fatal error. Additionally, pools with very few diaries may be poorly characterized because those diaries might contain activities that are quite rare in the general population. For example, if a pool contained only two diaries and one of them played golf most of the day, one would conclude from an APEX run that people in this diary pool average many hours outdoors per day. This would not change simply by running more profiles. However, this finding would probably not hold up with larger diary pools.

The diary pools are created in **DiaryModule:ReadDiaries**.

6.2 Basic (Random) Composite Diary Construction

Basic (random) composite diary construction is implemented in **DiaryModule:CompositeDiary**. In basic composite diary construction, APEX develops a composite diary for each of the simulated individuals (profiles) in the following manner:

Once the diary pools have been created (based on temperature and day of the week as described above), a selection probability for each diary within the pool is calculated based on age/gender/employment similarity of the simulated person to the "real" diaries. If desired, the user can use the Profile Factor Groups to match diaries by occupation as well. The probabilities are calculated in **DiaryModule:DiaryProbabilities**. The selection probability is a product of several probabilities between 0 and 1—one each for age, gender, and employment (and occupation, if selected). The gender and employment probabilities are straightforward: if the gender or the employment status of the diary matches that of the profile, then the probability for these factors is set to 1. If they do not match, then the probability is set to 0. The exception is for the employment probability for children under 16. Since APEX does not model employment status for children under sixteen (the employment profile variable will always be 0), then the employment probability is always set to 1 for this age group. This prevents APEX from discarding the CHAD activity diaries for children under age 16 that had an employment status=YES. Matching by occupation is similar to matching via employment and gender. However, any occupation listed on a diary that is not listed as one of the input occupation groups is automatically set to missing (*X*).

The age selection probability is a bit more complicated. APEX provides the user with the option of using activity diaries that have an age close to that of the simulated profile, although it may not match exactly. This range is controlled by the *Control Options* file setting ***AgeCutPct*** (see *Volume I*). For example, if the simulated person is age 30 and ***AgeCutPct***=10, then the diaries from persons within 3 years (which is 10% of 30) of the target age are within range and will be given a probability of 1. In addition, APEX allows for the use of "shoulder ages," which are the age ranges (of width ***AgeCutPct***) above and below the main age window. These ages are given reduced probability equal to the value of the *Control Options* file setting the variable ***Age2Prob***. In this example, diaries from persons between the ages of 27-33 have a full probability of being selected for an age 30 target, while diaries from ages 24-26 and 34-36 have a probability of Age2Prob of being selected. If the employment status, occupation, gender, or age for an activity diary is missing, then the selection probability for that variable is determined directly from the *Control Options* file variables ***MissEmpl***, ***MissOcc***, ***MissGender***, and ***MissAge***, respectively.

The final selection probability for each diary is the product of the age, gender, employment, and occupation selection probabilities. Then, on each day of the simulation period, APEX randomly selects a diary day from the appropriate diary pool, based on selection probability value. This method does not use information on prior selections when making the selection for the next day.

6.3 D&A Longitudinal Activity Diary Assembly

The second method of multi-day diary construction in APEX, which is required for characterizing within-person and between-person exposure variability, is a longitudinal diary

assembly algorithm that constructs multi-day diaries based on reproducing realistic variation in a user-selected key diary variable. The key variable must be numeric and may reflect any diary property. Values for the key variable areprovided to APEX in the *Diary Statistics* input file. If the user prefers to use a key variable that is not on the default file, they may prepare a replacement file with the variable of their choice. It is assumed that the key variable has a dominant influence on exposure. Otherwise, it would not matter to the exposure results whether or not consecutive days exhibited consistency in this variable.

The APEX release provides *Diary Statistics* files for outdoor time and vehicle time, which were constructed by summing the total time associated with "outdoor" and "vehicle" CHAD location codes for each diary (see *Volume I*). For some scenarios, the key variable might be travel time or time performing a particular activity. The key variable could also be a composite formed from several different variables, for example, a weighted average of diary variables. The necessary condition for implementing the method is that every single-day diary be assigned a numeric value for this key variable. This allows the set of available diaries in every diary pool to be ranked in terms of this key variable, from lowest to highest. The method uses this key variable to preferentially select appropriate diaries from the available pool on the different days in order to produce a final longitudinal activity diary that has specific statistical properties. The method is primarily contained within **DiaryModule:DiaryRanks**, which is called from **DiaryModule:DiaryProbabilities**.

The longitudinal diary construction method targets two statistics, ***D*** and ***A***. The ***D*** statistic reflects the relative importance of within-person variance and between-person variance in the key variable. The ***A*** statistic quantifies the lag-one (day-to-day) variable autocorrelation, which characterizes the similarity in diaries from day to day. Desired ***D*** and ***A*** values for the key variable are selected by the user and set in the *Control Options* file, and the algorithm constructs a longitudinal diary that preserves these parameters. See Section 6.3.2 for guidance on selecting appropriate ***D*** and ***A*** values for a particular simulation.

6.3.1 The D&A Longitudinal Diary Assembly Algorithm

The longitudinal diary selection method is based on the scaled rank, or "x-score" for the key variable for each diary. The x-scores are calculated within diary pools. First, a pool is sorted from lowest to highest on the key variable and given a corresponding rank, R. If there are K diaries in a pool and each diary has equal statistical weight, then the x-score for the diary at rank R is:

$$x = \frac{\left(R - \frac{1}{2}\right)}{K} \tag{6-1}$$

If the diaries have unequal statistical weight, then the x-scores will not be evenly spaced from 0 to 1 (as some diaries will correspond to a greater "interval" on the 0 to 1 scale). In APEX, eq. 6-1 is not used, but rather the x-scores are assigned to each diary by sorting the diary pool on the key statistic, and then applying the age, gender, and employment probabilities as described in Section 5.2. The diaries are then assigned to days in the simulation based on their x-scores by the algorithm described below.

An overview of the longitudinal diary method is shown in Figure 6.1. For each simulated person, the following steps are performed:

1. An individual target x-score T_i is selected from a beta distribution (β_1) that depends on the value of D.
2. For each day in the simulation, a daily scaled x-score (scaled rank) is generated. It is picked from a different beta distribution (β_2) having a peak near T_i.
3. An individual target correlation A_i is sampled from a beta distribution (β_3) having a peak near the population autocorrelation A.
4. The independently sampled daily x-score values are re-ordered to induce the target autocorrelation A_i.
5. Diaries are assigned from the diary pool according to the final time-series of daily x-score values.

All of the random number generation in the new method involves drawing numbers from beta distributions, with bounds of min=0, max=1. All of the random number distributions are bounded both above and below, which is a natural property of the beta distribution. Given these fixed end points, the beta distribution has two shape parameters which allow a great variety of forms. Both shape parameters (for all the distributions) are positive. The means and shape parameters for each of the beta distributions were formulated to 1) properly reproduce D for the population, while 2) producing unbiased x-scores across the population. The resulting equations for the parameters are given below. See Glen et al. (2008) for the derivation of these equations.

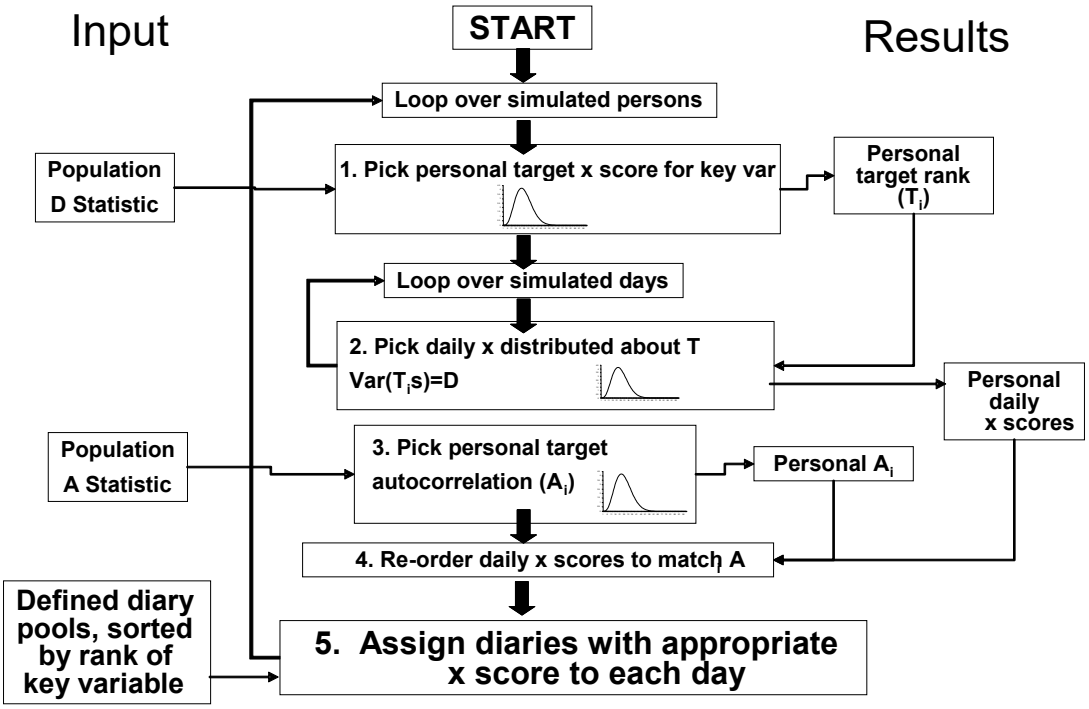

Figure 6.1. Overview of the Longitudinal Diary Assembly Algorithm

Each of the above steps of the algorithm is described in detail below:

1. An individual target x-score T_i is selected from a uniform distribution that depends on the value of *D*. The distribution of T_i depends on the value of ***D*,** and is chosen from a uniform distribution described as:

$$T_i = Uniform((1-\sqrt{D})/2, (1+\sqrt{D})/2) \tag{6-2}$$

2. For each day in the simulation, a daily x-score is generated. To generate daily scores, a beta distribution (β_2) dependent on ***D*** with a peak near the target value T_i is constructed. Since each individual has a personal value of T_i, each will have a unique β_2. As ***D*** approaches zero, β_2 flattens into a uniform distribution. As ***D*** approaches one, β_2 narrows to a spike at T_i. One x value is selected randomly from the PDF for each day in the simulation, plus a few extra (15 extra is sufficient for a one-year simulation). Since the β_2 for two individuals differ, the between-person variance does not go to zero as the number of draws becomes large. The shape parameters a and b for β_2 are a function of T_i and ***D***:

$$a = \frac{2T_i}{(1-D)} \tag{6-3}$$

$$b = \frac{2(1-T_i)}{(1-D)} \tag{6-4}$$

Thus for each day in the simulation, an x-score is generated as

$$x=\beta_2 =\beta(a,b) \tag{6-5}$$

3. An individual target correlation A_i is sampled from a different beta distribution (β_3) having a peak near the population autocorrelation *A*. The width of this distribution is w_2:

$$w_2 = min(2-2|A|,1) \tag{6-6}$$

The personal target for ***A***, A_i, is then generated as:

$$A_i = A + w_2(\beta(2.625,2.625)-0.5) \tag{6-7}$$

4. The independently sampled daily x-scores are re-ordered to induce the target autocorrelation A_i. This is done with a single pass, from first simulation day to the last. The reordering process involves selecting each x-score in the time series from another beta distribution (β_4) that is a function of both the individual autocorrelation target A_i and the previous x-score. The "extra" diaries come into play here; they allow the avoidance of undesirable forced selections. The shape parameters c and d for β_4 for day j are:

$$c = s\left(\frac{1-A_i}{2} + A_i x_{j-1}\right) \tag{6-8}$$

$$d = s\left(\frac{1+A_i}{2} - A_i x_{j-1}\right) \tag{6-9}$$

where

$$s = \frac{2}{(1 - A_i^2)} \tag{6-10}$$

and x_{j-1} is the previous day's x-score. The x-score for the first day is picked at random from a uniform distribution ranging from 0 to 1. Then, for day j, the x-score is selected as:

$$x_j = \beta(c, d) \tag{6-11}$$

The APEX code for performing this process keeps track of which x-scores are picked. If the same score is selected more than once, then the algorithm finds the closest unused score. The process laid out by equations 6-8 through 6-11 continues until x-scores for all simulation days are selected.

On very short time series (less than 30 days), the autocorrelation step has a slight effect on the resulting ***D*** for the population (a few percent increase).

5. Assign the selected x-score (scaled rank) values to each day in the simulation, and assign corresponding diaries from the diary pool. The result of steps 1-4 is a time series of x-score values mapped to simulation days, for example,

Jan 1	Jan 2	Jan 3
0.148	0.372	0.324

The diary pool has already been defined by APEX (typically by factors including age, gender, employment, day type, and season, and perhaps others, see Section 6.1). The pool is then sorted into rank order from lowest to highest score for the key variable, and sort order is mapped onto a corresponding x-score. The x-score reflects the behavior of an individual relative to their peer group (for example, a person with score 0.75 is above 75% of the people in the same cohort and pool, in terms of the key variable). Scores can be moved across diary pools, whereas absolute values for the key variable might not. Thus, there might be a diary with six hours of outdoor time in the Sunday pool, but no such diary in the Monday pool. However, a score of 0.75 exists on all days. The use of scores also helps in ensuring that all the available diaries are collectively sampled with the correct frequency. Note that the use of scores does not destroy information. All that matters in terms of diary assembly is the ability to specify which diary should be used on a given simulation day. For this purpose, requesting the available diary nearest to score 0.38 is no different than requesting the available diary nearest to (for example) 73 minutes of outdoor time.

APEX assigns the diary whose x-score is closest to the daily x-score value to the day. No distinction is made or needed between day types, seasons, etc., because that is already taken into account. Note that any diary matching criteria such as day type, season, temperature, rainfall, workday, holiday, etc., affect the list of diaries that belong to the pool for a given day, but have no effect on the x-scores.

6.3.2 Selecting Appropriate D and A Values For a Simulated Population

The statistic ***D*** for a population of individuals is given by:

$$\boldsymbol{D} = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2) \tag{6-12}$$

where σ_b^2 and σ_w^2 are the between- and within- person variances in the key variable. Note that σ_b^2 is the variance between persons in their long-term (not daily) means for the key variable. Since both variances are non-negative, it is clear that ***D*** is in the interval [0,1]. ***D***=0 means that σ_b^2 is zero, or that each person has the same mean score. A small ***D*** means that σ_b^2 is substantially smaller than σ_w^2, indicating that the overall variability between people in the key diary statistic is smaller than the variability observed over days within the same person. A ***D*** near one means that σ_b^2 is much larger than σ_w^2, or that each person shows little variation over time relative to the variability between persons.

The lag-one autocorrelation ***A*** is simpler to calculate than ***D***, because each time series can be examined independently. The first step is to determine the x-score for each day, relative to the entire time series. If there are J days in the time series, and a given day is at rank R in terms of the rank for the key variable among the J days, then the x-score for that day is (R-1/2) / J. The overall mean and variance in these scores for the time series is then calculated. Due to the properties of the discrete uniform distribution of the scores (neglecting tied scores), the mean must be 1/2 and the variance is:

$$\sigma^2 = \frac{1}{12}\left(1 - \frac{1}{J^2}\right) \tag{6-13}$$

which is very close to 1/12 for large J. The lag-one covariance is calculated by:

$$COV = \frac{1}{J}\sum\left(x(J) - \frac{1}{2}\right)\left(x(J+1) - \frac{1}{2}\right) \tag{6-14}$$

where x(*j*) is the x-score on day *j* (see for example, Box et al., 1994). The lag-one autocorrelation for the individual time series is given by the ratio of the covariance to the variance:

$$A_i = \frac{\mathrm{COV}}{\sigma^2} \tag{6-15}$$

This calculation is repeated for each time series, and the statistic ***A*** is the mean of these individual autocorrelations. The statistic ***A*** has a range from -1 to +1, with positive values indicating that each day has a tendency to resemble the day before. Random selection of diaries from day to day produces ***A*** values near zero. Negative ***A*** values imply dissimilarity between consecutive days.

A study of children conducted in Southern California (Xue et al., 2004) provided about 60 days of data on each of 163 children. The time series are not continuous as the monitoring consisted of twelve six-day periods; one per month over a year. Furthermore, only roughly 40 children were measured simultaneously as the other children were sampled in different weeks. However,

a sample size of 40 is sufficient to calculate reliable rankings across persons. The number of consecutive day pairs was substantially less than the number of days due to the gaps in the time series. However, ***D*** and ***A*** statistics were calculated for three variables directly recorded on the activity diaries (outdoor time, travel time, and indoor time), and also for a fourth variable, the physical activity index or PAI (McCurdy, 2000). The analyses were performed for all children together and for two gender cohorts. The separation into two cohorts reduces the number of children measured simultaneously to fewer than 20. Further division into more cohorts is therefore not practical, as the reliability of the scores would become very uncertain. The results for these analyses are given in Table 6.1.

Table 6.1. D and A Statistics Derived from the Southern California Children's Study

Variable	Group	*D*	*A*
Outdoor time	all	0.19	0.22
Outdoor time	boys	0.21	0.21
Outdoor time	girls	0.15	0.24
Travel time	all	0.18	0.07
Travel time	boys	0.18	0.05
Travel time	girls	0.18	0.08
Indoor time	all	0.17	0.22
Indoor time	boys	0.21	0.2
Indoor time	girls	0.17	0.24
PAI	all	0.16	0.23
PAI	boys	0.16	0.2
PAI	girls	0.16	0.25

For all variables and each group, the standard deviation between persons for autocorrelation was about 0.20, and the standard error in the mean A was about 0.02. The values in Table 6.1 indicate that gender differences for both ***D*** and ***A*** are small, if present at all. The variance in ***A*** over the population was also examined for the four diary variables. While the absolute values of ***A*** were different across variables, it was found the variance in individual autocorrelations was very similar (approximately 0.2) in all variables. This variance was used to derive the parameters for the target ***A*** distribution (eq. 6-7) in the APEX algorithm, and thus the method returns diaries that reproduce a variance of 0.2 in the daily autocorrelation, no matter what diary variable is modeled or what absolute ***A*** value is used.

6.4 Cluster-Markov Chain Diary Assembly

APEX has had a diary clustering method since 2009. In 2018 it was extensively revised, and makes use of new files and parameters on the *Control Options* file. The older method had limitations, particularly in the number of diaries it could handle, and it has been discontinued. The new method is faster and more memory efficient than the old, as well as being more flexible. For specificity, the diary database is called CHAD, although the APEX user can use their own database, provided that it is coded in a manner consistent with the default CHAD database.

This method clusters diaries along user-defined axes, based on the time spent in various CHAD locations. The user may customize this mapping as desired. For each simulated person, one

suitable activity diary is selected from each cluster, from each diary pool. The pool for each simulation day is determined by the day of the week and the temperature. The cluster for each day is randomly chosen, based on the pool and cluster for the previous day. Empirical transition probabilities are derived for the next day's cluster selection, based on examples in CHAD. This is a first-order Markov chain approach to diary selection.

The steps in the algorithm are as follows:

- Calculate cluster axis scores for every CHAD diary. Determine the ratio of each axis score to the mean score for that axis over all of CHAD. Assign the CHAD diary to the axis with the highest score (that is, highest ratio to the mean).
- Identify all the examples of transitions from one diary day to the next (for the same person) that are found in CHAD. Assign the pool and cluster for each day. Stratify CHAD into age groups for the determination of cluster weights. Three groups is typical. More groups can be used, but would create cases with very few examples, making empirical estimates more uncertain. Determine the per-diary statistical weight for each cluster pair.
- For each simulated individual, a single time-activity diary day is randomly selected from each pool-cluster combination, subject to the usual APEX diary selection probabilities on age, gender, employment, and (optionally) commute time and occupation.
- Select a cluster for each simulation day, based on the correct pool and the cluster weights, given the pool and cluster for the prior day. Append the previously selected diary day for the appropriate pool-cluster combination, to form a longitudinal diary over the simulation period.

The algorithm is selected by setting ClusterDiary=YES on the *Control Options* file. One cannot have both LongitDiary=YES and ClusterDiary=YES. If both are NO, then diaries are selected at random from the ones that are demographically matched.

The diary clustering algorithm above can be broken down into the following steps:

6.4.1 Clustering of CHAD

1. Read the mapping of CHAD location codes to cluster axes

This is similar to the first step in the old clustering method, but now all CHAD locations must be mapped to one of the axes, including codes such as "U" or "X". The keyword "diaryclus file" is used to name the file. Currently, APEX limits the number of axes to 5 or fewer, but this may be altered in the source code (and then recompiled) if desired.

2. Score every diary day

APEX totals the time spent in locations mapped to each axis, for each diary. Since all diaries have 1440 minutes of time, and all time is mapped to an axis, the raw axis scores total 1440 minutes for any given diary. For each axis, find the mean score over all diaries. These means also total to 1440 minutes.

3. For each diary and axis, find the ratio of the raw score to the mean

Divide each axis score by the mean score for that axis over CHAD. Every CHAD diary must have at least one axis at or above the mean (that is, at least one ratio of 1.0 or above). There is always one cluster for each axis, so the terms "cluster" and "axis" become interchangeable. Assign each diary to the cluster corresponding to the axis with the highest ratio. If there are ties, assign the diary to the lowest numbered axis (among the ties).

4. Write out the cluster assignments (optional)

This is activated using the keyword ClusterOut=YES on the COF. If so, then a filename is needed as well, using the keyword "Cluster file = ", followed by the filename. This produces a text file listing every CHADID, the axis scores, and the cluster assignment.

6.4.2 Evaluation of transitions

1. Find the examples of transitions in CHAD

This requires determining which CHAD diaries belong to the same person. This can no longer be determined directly from the CHADID. Instead, a new input file is read, containing each CHADID and a chronological sequence number, which resets to one for the first diary from each person. This file is named using the "DiaryTrans file" keyword. This file needs to be updated only when new diaries are added to CHAD.

2. Renumber the diaries

APEX rejects diaries for various reasons, such as missing age, gender, date, or temperature. Also, diaries are rejected if too much time is spent in location or activities "U" or "X". Because of this, the sequential numbering on the DiaryTrans file must be recalculated. However, the person identifier on that file is still crucial. Transitions are always between calendar days with no other diary day in between. If a person has N diaries (after filtering out the rejects), there are (N-1) examples of transitions for that person. The N diaries are sorted chronologically, when their diary numbers become consecutive. Transitions between two consecutive calendar dates are called "adjacent". Two lists are made for later use, one for adjacent transitions, and the other for all transitions. The latter group includes gaps, that is, calendar days on which the person does not have a valid diary in the database. Every transition consists of two diary days, called the "from" day and the "to" day.

3. Stratify persons with transitions into age groups

This is based on the list of ages supplied with the COF keyword "ClusterAges". For example, if the user has the line "ClusterAges = 16, 50 " on the *Control Options* file, then ages 0-15 are in the first age bin, 16-49 are in the second, and ages 50 and over are in the third. If the user creates too many age bins, then there are likely to be few examples of certain transitions. The current hard-coded maximum is 9 age cutpoints, although this number is not recommended unless the diary database becomes much larger. The default is that all ages are in the same bin.

4. Read the diarypools definitions and assign pools

There are two pool assignments for each transition, namely, the "from" pool (for the earlier diary day) and the "to" pool (for the later day). Many APEX runs use 6 diary pools, so there are 36

combinations of "from" and "to" pools in that case. These are also stratified by age bin, so there are 108 combinations if there were 6 pools and 3 age bins. There are over 20,000 transitions in CHAD, but that does not translate to 200 examples for each combination, because some are much more common than others.

5. Count the number of diaries in each cluster

This is applied to every combination of pool and age bin. If N clusters have been defined, then there are N^2 possibilities for "from" cluster and "to" cluster. For each "from" cluster, the "to" clusters are likely to be unevenly populated, because there is a general tendency to remain in the same cluster (that is, at a greater than random chance). For example, if the database has 30 transition for a given "from" cluster (and pool and age bin combination), then random chance would indicate that each of the three "to" clusters should have 10 examples among these 30 examples. But in practice, the "to" cluster than matches the "from" cluster will likely have more than 10, and the others will have fewer than 10. With very few diaries, this tendency may be obscured by sampling variation, so for reliable transition probabilities it is necessary to have higher numbers of examples. The keyword UseAdjacent on the COF gives the smallest number of diaries for a given pool and agebin combination for which the cluster counts are restricted to adjacent transitions only. If there are fewer transitions in that combination, then all transitions (including non-adjacent) are counted.

Even with the use of all transitions instead of just adjacent transitions, there will be some pool combinations that have very few examples. This frequently occurs between two pools belonging to non-adjacent temperature bins. For example, the first day is in a cold bin (maximum temperature below 54 degrees F, say), and the next day is warm (over 84 degrees, say). Since the pools and these temperature bins are user-defined on each APEX run, one cannot predict before the run which combinations will be extremely rare. In such cases, the user may also define a "PoolTrans" function on the "functions" input file (the same input file that has "DiaryPools"), which tells APEX which pool combinations to join together for purposes of defining probabilities. In the above example, it is rare for a below 54 day to be followed by an above 84 day, so in that case the above 84 days should be combined with the 54 to 83 days by the PoolTrans function, to allow the calculation of cluster transition probabilities when going from a below 54 day to any warmer day. Any count of zero is replaced by a count of 0.5, to avoid a transition probability of zero.

6. Convert cluster counts to cluster weights

The formula for the cluster weights was explained in the report "The new diary clustering algorithm in APEX" by ICF in August 2018. Arrange all the counts for a given pool-agebin combination in a matrix with a row for each "from" cluster and column for each "to" cluster. Then the weights are given by

Cell weight = (cell count * total count) / (row total * column total)

If every cell of the matrix had the same count, then all the weights would be 1.00. Note that the replacement of any cell that has zero by a value of 0.5 means that none of the row totals or column totals can ever be zero. This prevents problems with division by zero.

If the COF has the keyword CWeightout = YES, then the weights for all pool-age-cluster combinations are written to an output file.

6.4.3 Diary Assembly using Clustering

The other diary assembly methods can conceivably select a different CHAD diary on every day of the simulation (if there are enough suitable diaries). However, the longitudinal assembly method gives greater weight to diaries that are similar (or identical) to ones previously used, so that is not likely. The clustering method is different in that re-use of the same diaries is guaranteed. For every pool-cluster combination, one diary is selected, using the standard APEX selection weights. These weights are based on matching the gender, employment status, age (with a window), and optionally the occupation and commute time. All these factors are constant over the simulation period for one person.

For the first day of the simulation, there is no prior day pool or cluster. The pool for the first day is known. The cluster is chosen at random, using the combined probability for all the diaries in each cluster, stored previously in the CProbs array. For example, if cluster 1 has 40% of the total diary weight (for the correct pool for day 1), then it has a 40% chance of being chosen as the cluster for the first day.

For subsequent simulation days, the pool and cluster for the previous day affect the cluster selection. Each cluster has its raw probability (measured by CProbs) multiplied by the weight for that cluster (from the CWeight array), based on the information on the previous day. It is unlikely that these adjusted probabilities still sum to exactly one, so each one is divided by the total (a process called "standardization"), to ensure that the sum of the probabilities over clusters equals one. A new uniform random value from zero to one is generated for each simulation day, and this is compared to the weighted cluster probabilities to determine the cluster for each day. Once the cluster is selected, the previously chosen diary for that cluster and pool is added to the composite activity diary.

For example, with 6 diary pools and 3 clusters, a total of 18 diaries that match age, gender, and the other selection variables are chosen. As one moves through the simulation, these 18 diaries are re-used. If the pool changes from one day to the next, then a different one of the 18 diaries must be used. If the pool remains the same on the next day, there is a chance (based on the cluster weights) for the same diary to be used again.

CHAPTER 7. ESTIMATING ENERGY EXPENDITURES AND VENTILATION

Ventilation rates are used in APEX for:

- Calculating exertion level for use in tabulating exposure summaries for the population
- Estimating dose

Ventilation does not influence the exposures for a simulated person.

Ventilation is a general term for the movement of air into and out of the lungs. Minute or total ventilation is the amount of air moved in or out of the lungs per minute. Quantitatively, the amount of air breathed in per minute (V_i) is slightly greater than the amount expired per minute (V_e). Clinically, however, this difference is not important, and by convention minute ventilation is always measured on an expired sample, V_e. Alveolar ventilation (V_a) is the volume of air breathed in per minute that (1) reaches the alveoli and (2) takes part in gas exchange. The ventilation rate needed for the %COHb calculation is this ventilation rate, V_a, and is derived for use in APEX based on work by Adams (1998), Astrand and Rodahl (1977), Burmaster and Crouch (1997), Esmail et al. (1995), Galetti (1959), Johnson (1998), Joumard et al. (1981), McCurdy (2000), McCurdy et al. (2000), Schofield (1985), and many others. Only a brief description of V_a is described below; for the complete derivation, see Johnson (2002).

Ventilation is calculated on an activity event-by-activity event basis. Ventilation is derived from the energy expenditure rate (MET, given as a multiple of resting metabolic rate), associated with the diary activities. The general steps in the estimating ventilation are:

- Generate the MET event time-series based on the diary activities
- Adjust the resulting MET series for fatigue and excess post-exercise oxygen consumption
- Convert the MET time-series into a ventilation rate time series

These steps are covered in the following sections.

7.1 Generating the MET Time Series

MET—which comes from "metabolic equivalents of task"—is a dimensionless ratio of the activity-specific energy expenditure rate to the basal or resting energy expenditure rate. While different people have very different basal metabolic rates, it is generally found that the MET ratios do not exhibit as much variability. Thus, standing still might require two times the basal energy expenditure, or two MET, for most people, with relatively little variation. The basal rate is constant (it only has to be determined once per profile), while the activity-specific MET ratio is calculated for each of the activities reported on the composite activity diary.

Each possible diary activity code (see *Volume I*), is mapped to a corresponding APEX MET distribution number the *MET Mapping* input file. Each of these distributions is then defined in the *MET Distributions* file. This file (see *Volume I*) gives the properties of the MET distribution for each type of activity, in some cases as a function of age or occupation. The distributions are defined in the standard APEX format (see Section 3.1). These distributions are based on many

available data on energy expenditure, and in general should not be changed. This file is read into APEX in the **DiaryModule:ReadMETS**. In the subroutine **DiaryModule:METSEval**, APEX steps through the activity diary and assigns a MET value to each event by selecting a value from the appropriate distribution as defined by the activity code and the profile, using a random quantile that is resampled hourly. The result is consider the "raw" MET time-series, which is then adjusted to be physiologically realistic. The adjustments are covered in the next section.

7.2 Adjusting the MET Time Series for Fatigue and Excess Post-Exercise Oxygen Consumption

As discussed in the previous section, APEX assigns distributions for MET level to each diary event, based on the reported event activity (and in some cases, age and occupation). However, these raw MET time-series do not consider the sequence of the events (i.e., the order in which they occur). It is well known that a person's capacity for work will diminish as they get tired, and in practice, this means that the upper bound on MET is lowered if events in the recent past have been at unusually high MET levels. Furthermore, once high activity levels have ended, people tend to breathe heavily even while resting as they recover their accumulated oxygen deficit. This effect is called excess post-exercise oxygen consumption (EPOC), and results in raising the MET levels above the 'raw' values pulled from the activity-based distributions. APEX contains an algorithm for adjusting the MET time series for both of these effects. The algorithm is implemented in **ExposureDoseModule:Ventilation**.

The adjustment method is based on keeping a running total of the oxygen deficit as a simulated individual proceeds chronologically through his or her activity diary. The oxygen deficit is the amount of energy supplied to the muscles by non-aerobic systems during exercise. It reflects a need for increased post-exercise ventilation to "pay back" this energy. The oxygen deficit calculations were derived from a synthesis of numerous published studies (see below).

Oxygen deficit is measured as a percentage of the maximum oxygen deficit an individual can attain prior to deterioration of exercise performance. Limitations on MET levels corresponding to post-exercise diary events were based on maintaining an oxygen deficit below this maximum value. In addition, adjustments to MET were simultaneously made for EPOC. The EPOC adjustments are based in part on the modeled oxygen deficit and in part on data from published studies on EPOC, oxygen deficit, and oxygen consumption.

The methods are constructed in terms of reserve MET, which is the amount over the basal rate (MET=1). Furthermore, we defined M as the normalized reserve, so that M=0 at MET=1, and M=1 at maximum MET:

$$M = \frac{MET - 1}{MET_{max} - 1} \tag{7-1}$$

Recall that MET_{max} is a profile variable assigned for each simulated profile (see Section 5.3). Using a normalized reserve assures that the method can be applied identically to the entire population of profiles, each having a unique MET_{max} value.

A number of terms will be used in the description of the algorithm. They are defined below:

- MET Metabolic equivalent of task (unitless)
- MET_{max} Maximum achievable metabolic equivalent for an individual (unitless)
- M Normalized MET reserve (unitless, M, bounded between 0 and 1)
- ΔM Change in M from one diary event to the next (M)
- D_{max} Absolute maximum oxygen deficit that can be obtained (M-hr)
- F Fractional oxygen deficit (percent of individual maximum, unitless)
- t_e Duration of activity diary event (hours)
- t_r Time required to recover from an F of 1 to an F of 0 at rest (recovery time, hours)
- dF_{inc} Rate of change of F due to deficit increase (F/hr, will be positive)
- dF_{rec} Rate of change of F due to deficit recovery (F/hr, will be negative)
- dF_{tot} Total rate of change of F, $dF_{inc+} dF_{rec}$ (F/hr)
- ΔF_{inc} Increase in F due to anaerobic energy expenditure (F)
- ΔF_{rec} Decrease in F due to recovery of oxygen deficit (F)
- ΔF_{tot} Change in F due to simultaneous anaerobic work and oxygen recovery, $\Delta F_{inc}+\Delta F_{rec}$ (F)
- ΔF_{fast} Total change in F during the fast recovery phase (F)
- S_{fast} Magnitude of the rate of change in M during fast component (M/hr)
- $EPOC_{fast}$ Change in M due to fast-component EPOC (M)
- $EPOC_{slow}$ Change in M due to slow-component EPOC (M)

See Isaacs et al. (2007) for a complete derivation of the method.

7.2.1 Simulation of Oxygen Deficit

This section presents the theoretical development of the equations describing the accumulation of oxygen deficit. The method was developed using a large number of studies on oxygen consumption, oxygen deficit, and EPOC. Individual studies will be referenced below. The first two subsections below describe the equations themselves, while the last section describes the determination of the values for the model parameters.

7.2.1.1 Fast Processes

There exists a component of the accumulated oxygen deficit that is due to transition from one M level to another (McArdle et al., 2001). This component derives from the anaerobic work that is required by sudden muscular motion. There is also a corresponding fast component of oxygen recovery which occurs very quickly after a change from a high M level to a lower one. In the absence of any data to the contrary, it is assumed that these fast deficit accumulation and fast recovery processes occur at the same rate. These processes are illustrated in Figure 7.1. The adjustment to F is equal to the area of the triangle associated with either a positive or negative change in M, normalized by the maximum obtainable accumulated oxygen deficit (D_{max}). The normalized area can thus be calculated as:

$$\Delta F_{fast} = 0.5 \frac{\Delta M |\Delta M|}{S_{fast} D_{max}} \tag{7-2}$$

where $\Delta M = M_i - M_{i-1}$ and S_{fast} is the slope of the change in M (in M/hr). Note that this change in F will be positive if ΔM is positive, and negative otherwise.

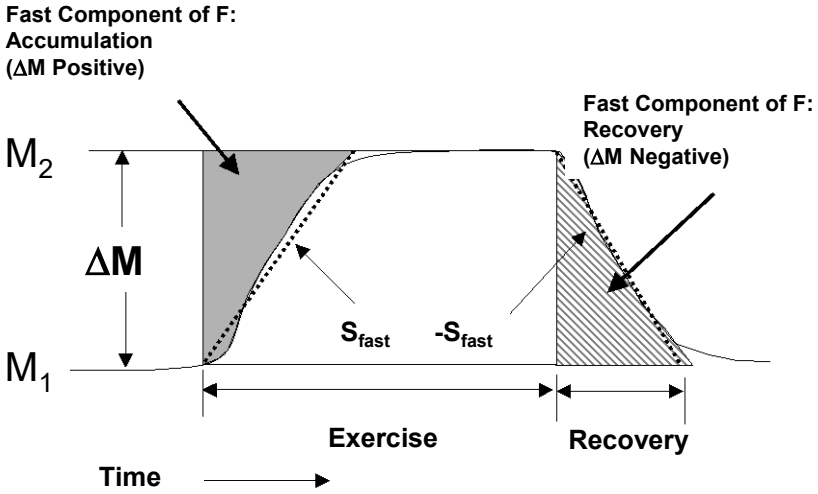

Figure 7.1. Fast Components of Oxygen Deficit and Recovery

7.2.1.2 Slow Processes

The slow component of the increase in oxygen deficit corresponds to the accumulation of deficit over a period of heavier exercise (rather than that associated with an increase in activity level). The method was derived from the analysis of a number of studies on exercise and EPOC including: Bahr (1992), Bahr et al. (1987), Bielinski et al. (1985), Brockman et al. (1993), Gillette et al. (1994), Gore and Withers (1990), Hagberg et al. (1980), Harris et al. (1962), Kaminsky and Whaley (1993), Katch et al. (1972), Maehlum et al. (1986), and Sedlock (1991a,b). The following data were considered: the time it took for subjects to reach exhaustion, their accumulated oxygen deficit, their MET_{max}, the MET value at which they exercised, and the corresponding normalized reserve MET (M). Note that the MET and MET_{max} quantities were derived from the published VO_2 and VO_{2max} measurements. The data indicated that oxygen deficit accumulates at a much faster rate when M is high. For example, an M value near 0.5 requires about 5 times longer to reach exhaustion than an M value near 0.75 (on average), indicating that F is nonlinear in M.

Let the rate of increase in F be given by dF_{inc}. The relationship between dF_{inc} and M is a power law:

$$dF_{inc} = aM^b \tag{7-3}$$

where a and b were estimated from available data. The slow recovery of oxygen deficit must also be accounted for, as it occurs simultaneously with debt accumulation. A slow, but continual, process for recovering oxygen deficit is modeled, independent of the MET level. EPOC recovery is modeled as constant over time until the oxygen deficit is erased. Assuming this takes t_r hours, the slow recovery of oxygen deficit occurs at a rate:

$$dF_{rec} = -\frac{1}{t_r} \tag{7-4}$$

The total net rate of change in F from slow processes during an event with duration t_e is given by:

$$dF_{slow} = dF_{inc} + dF_{rec} \tag{7-5}$$

and the associated change in F is:

$$\Delta F_{slow} = \left(aM_i^b - \frac{1}{t_r} \right) t_e \tag{7-6}$$

For an individual starting with an F of 0 and exercising to exhaustion (neglecting the transitory effects), the change in ΔF is 1.0. In this case, rearranging and taking the logarithm gives:

$$log\left(\frac{1}{t} + \frac{1}{t_r} \right) = log(a) + b\, log(M) \tag{7-7}$$

Equation 7-7can be used to fit data to estimate the parameters a and b. This will be discussed in the next subsection.

The starting normalized oxygen deficit for the next event (i +1), taking into account both the fast and slow changes in F, is then:

$$F_{i+1} = F_i + \Delta F_{slow} + \Delta F_{Fast} \tag{7-8}$$

7.2.1.3 Derivation of Appropriate Values for the Model Parameters

The values of the model parameters t_r, a, and b, were derived from summaries of published data on EPOC and oxygen debt (see references listed in Section 7.2). Several of these studies reported t_r values; however, due to variability in measurement and protocol differences, these recovery times varied from 0.5 hours to 24 hours. From a modeling viewpoint, it would be unacceptable to allow recovery to significantly carry over from one day to the next. To do so could lead to a perpetual delay in recovering an oxygen deficit, e.g., by repeatedly encountering new exercise events before recovery is complete. In APEX, t_r, which is a profile variable, is selected from a uniform distribution having a minimum of 8 and a maximum of 16 hours.

In APEX, a and b are modeled as a function of t_r:

$$a = 5.20 - \left(\frac{1.54}{t_r} \right) + \left(\frac{3.92}{t_r^2} \right) \tag{7-9}$$

$$b = 3.93 - \left(\frac{3.57}{t_r} \right) + \left(\frac{3.66}{t_r^2} \right) \tag{7-10}$$

These expressions were derived from the experimental data.

Appropriate distributions for maximum oxygen debt (MOD) in ml/kg were derived from data from a number of studies in adults (Bickham et al., 2002, Billat et al., 1996, Buck and McNaughton, 1999, Demarle et al., 2001, Doherty et al., 2000, Faina et al., 1997, Gastin and Lawson, 1994, Gastin et al., 1995, Hill et al., 1998, Maxwell and Nimmo, 1996, Olesen,1992, Renoux et al., 1999, Roberts et al., 2003, Weber and Schneider, 2000); adolescents (Naughton et al., 1998); and children (Berthoin et al., 1996, Carlson and Naughton, 1993). The studies covered multiple types of exercise protocols, some having more than one protocol per study. Normal distributions for MOD were defined for all three age groups, based on average mean and standard deviation values from the studies:

- adults (>17 yrs): 54.95±14.46 (ml/kg)
- adolescents (12-17 yrs): 63.95±21.12 (ml/kg)
- children (<12 yrs): 34.74±13.10 (ml/kg)

These mean and standard deviations are read in from the *Physiology* file (see Section 5.3). Values are selected from normal distributions with these characteristics. These values are constant for an individual over the simulation period. The bounds of these distributions are selected as two standard deviations from the mean; these ranges were found to be reasonable when compared to reported ranges (Olesen 1992). These values are transformed to $D_{max,}$ via a units conversion factor and the normalization needed for use with reserve MET:

$$D_{max}\ (M\text{-}hr) = \left(\frac{MOD}{60\ METtoO_2}\right)(MET_{max} - 1)^{-1} \tag{7-11}$$

where $METtoO_2$ is the conversion factor for ml O_2 to MET-min, 3.5 [(ml O2/min)/kg]/MET. Note that the variability in this factor is not addressed here.

A number of studies on EPOC (Almuzaini et al., 1998, Dawson et al., 1996, Frey et al., 1993, Harms et al., 1995, Kaminsky et al., 1990, Knuttgen, 1970, Maresh et al., 1992, Pivarnik and Wilkerson, 1988, Short and Sedlock, 1997, and Trost et al., 1997) were used to derive S_{fast}. These were all studies in which oxygen consumption was measured relatively soon (within a few minutes) after the end of exercise and at a frequency high enough to capture the kinetics of the change in oxygen consumption. The data were found to be relatively uniform from the minimum (0.6 MET/min) to the maximum (3.7 MET/min) slope values, and so values were selected from a uniform distribution having these bounds. Converting units and normalizing to M, one obtains:

$$S_{fast}\ (M/hr) = \frac{60\ Uniform\ (0.6, 3.7)}{(MET_{max} - 1)} \tag{7-12}$$

7.2.2 Adjustments to M for Fatigue

The equations provided in the previous section describe a method for keeping a running total of the fractional oxygen deficit (F) for each diary event for an individual. These event F values are used to limit M for each event to appropriate values. Basically, the maximum M value that can be maintained for an entire event is the value that would result in an F_{i+1} (eq. 7-8) equal to 1 (i.e., the maximum value) at the end of the diary event. The approach used in APEX is to set M for each event equal to the raw MET value, and test if $F_{i+1}>1$. If it is, then the M_i value is reduced by a predetermined amount (currently 0.01) and F_{i+1} is recalculated. The process continues until

an appropriate value of M_i, called $M_{max,i}$ is found. As the exposure model marches through the events of the activity diary, the M values associated with each event are adjusted if necessary:

$$M_i = min(M_i, M_{max,i}) \tag{7-13}$$

7.2.3 Adjustments to M for EPOC

As noted above, it has been observed in many studies that EPOC is characterized by both slow and fast components. The fast increase in oxygen consumption occurs within minutes of exercise, while the slow component may persist for many hours. Both fast and slow EPOC components were modeled.

7.2.3.1 Fast Processes

The fast EPOC component, which takes place in the first few minutes after exercise, is also characterized by the slope $S_{fast.}$ The energy recovered during those first few minutes corresponds to the recovery triangle in Figure 7.1, and this increase in the rate of energy expenditure for a post-exercise event is modeled as the area of the triangle divided by the event duration:

$$EPOC_{fast} = 0.5\frac{(\Delta M)^2}{S_{fast} t_e} \tag{7-14}$$

$EPOC_{fast}$ will thus have units of M (normalized reserve MET). The M level for the post-exercise events will be incremented by $EPOC_{fast.}$

7.2.3.2 Slow Processes

The increase in M associated with the slow EPOC component is estimated as the amount required to maintain the slow recovery of F. Since the deficit D_{max} is recovered in full in the recovery time t_r, the time-averaged adjustment to MET for the slow recovery process must be:

$$EPOC_{slow} = \frac{D_{max}}{t_r} \tag{7-15}$$

Every diary event with the full rate of slow recovery will have its M value adjusted upward by $EPOC_{slow}$. An appropriate fraction of $EPOC_{slow}$ is used if only partial recovery is needed to eliminate the deficit (i.e., return F to 0). The final adjusted M value for the diary event is thus:

$$M_{adj} = M + EPOC_{fast} + EPOC_{slow} \tag{7-16}$$

and the new MET value for the event is:

$$MET_{adj} = M_{adj}(MET_{max} - 1) + 1 \tag{7-17}$$

7.3 Calculating PAI and the Ventilation Rates

APEX calculates three different ventilation rates from the adjusted MET time series. These are the expired ventilation rate (V_e), the alveolar ventilation rate (V_a), and the effective ventilation rate (EVR). All three are reported for each hour in the simulation in the APEX output files. V_a is used in the CO dose calculations, and EVR is used in compiling summary exposure tables for different populations during different levels of exertion.

In addition, the final MET time-series is used to calculate a physical activity index (PAI) for each individual. Finally, an intermediate rate, oxygen consumption (VO_2), is also calculated. The equations for calculating these rates from the MET time series are given below. All of these calculations are implemented in **ExposureDoseModule:Ventilation**.

While there is still just one method for calculating V_a and VO_2, a second option has been added for calculating V_e. See Section 7.3.2 for details.

7.3.1 Calculating PAI and Energy Expenditure

Once the final MET time series is calculated, the timestep and hourly physical activity index (PAI) for each hour in the simulation for the simulated individual is calculated as the time-weighted average of MET:

$$PAI = \frac{1}{t}\sum_{i=1}^{N} MET_i \times t_i \qquad (7\text{-}18)$$

where MET_i is the MET value for event i, t_i is the event duration in minutes, and t is the length of the timestep in minutes (or 60 minutes, if calculating hourly values). N_{events} is the number of diary events in the considered timestep or hour. These PAI values can be written to the *Timestep* or *Hourly* files. The daily PAI value is simply the average of the 24 hourly values. Finally, a median daily PAI value is calculated for each profile. The daily and median daily PAI values are saved as profile variables (see Section 5.3). The median daily PAI value is used in the characterization of persons as "active" when creating the output exposure summary tables (see *Volume I* and Section 9.2).

The energy expenditure (kcal/min) is:

$$EE = PAI \times RMR \qquad (7\text{-}19)$$

where RMR is the profile resting metabolic rate in kcal per minute. EE is calculated for both timesteps and hours, and can be written to the corresponding output files.

7.3.2 Calculating Oxygen Consumption and Ventilation Rates

The oxygen consumption rate in ml/min/kg (McCurdy, 2000), normalized to body mass, is given by:

$$\frac{VO_2}{BM} = \frac{MET \times ECF \times RMR}{BM} \quad (7\text{-}20)$$

where ECF is the profile energy conversion factor in liters of oxygen per kcal and BM is the profile body mass. The alveolar ventilation rate is also calculated from MET using:

$$V_a = MET \times 19630 \times ECF \times RMR \quad (7\text{-}21)$$

where the constant, 19630, is the oxygen to air conversion factor (19,630 ml of air/l of O_2).

Method 1 for the calculation of V_e is based on the VeSlope, VeResid, and VeInter profile variables (see Section 5.3). The calculation of those variables and the following equations for V_e comprise the V_e regression equations derived by Graham and McCurdy (2005). V_e is calculated as:

$$V_e = BM(e^X) \quad (7\text{-}22)$$

where BM is the profile body mass and the exponent term X is given by:

$$X = VeInter + VeSlope \times \ln(VO_2 / BM) + e_b + e_w \quad (7\text{-}23)$$

where e_b and e_w are random numbers pulled from the specified distributions, sampled hourly. EVR is then:

$$EVR = \frac{V_e}{BSA} \quad (7\text{-}24)$$

where BSA is the profile body surface area.

Method 2 for calculating Ve is based on the idea that as one approaches one's personal limit VO2max, the efficiency of extracting oxygen from air decreases. It therefore incorporates the fraction F = VO2/VO2max into the regression equation. The report on the development of this method is a February 2017 memorandum from ICF to EPA, which includes new algorithms for both RMR and VE. A brief summary is presented here.

Every simulated person in APEX is assigned both a resting metabolic rate (RMR) and a personal maximum for VO2 (called VO2max). The algorithm for assigning VO2max has not changed, and is based on sampling distributions for NVO2max (which is VO2max per kilogram body weight), specific to each age-gender combination, which are on the *Physiology* input file. Each activity diary event is assigned a MET value, as described earlier in this chapter. The oxygen requirement for the diary event is given by equation (7-20) above. Thus, both VO2 and F=VO2/VO2max are available in APEX for every diary event.

The regression equation is

$$\log(VE) = 3.300 + 0.8128 \log(VO2) + 0.5126\, F^4 + e_b + e_w \quad (7\text{-}25)$$

The left hand side is the natural logarithm of VE, in units of (L/min). Log(VO2) is also the natural logarithm, in units of (L/min). F is the unitless ratio of VO2 to VO2max. The two residuals are normal distributions with mean zero. The first is e_b, the between-person term, which has a standard deviation of 0.09866 and is sampled once per person. The second is e_w, the within-person variation, which has a standard deviation of 0.07852 and is sampled daily.

Using the above equations, APEX generates an event time-series for Ve and Va and EVR. Ve and Va are output on the *Events* output file. Timestep and hourly values (time-weighted timestep and hourly averages of the event values) for EVR, Va, Ve, are calculated and can be written to the *Timestep* and *Hourly* output files (see *Volume I*).

7.4 Calculating Ozone-Induced Changes to Forced Expiratory Volume

Studies of exposure to ozone have shown that there is a significant relationship between ozone exposure and reversible decrements in the forced expiratory volume in 1s (FEV1). Activity levels, duration of exposure, age, height, and weight have also been shown to be significant factors in determining ΔFEV1. APEX uses a model developed by McDonnell, Stewart and Smith (2010), later revised in another paper by McDonnell, Stewart and Smith (2013). The current implementation in APEX matches Model 3 from the 2013 paper, with extensions to other age groups. The 2013 paper is referred to as MSS, below.

First, define an age-dependent term y_{age} as:

$$y_{age} = FEVSlp \times age + FEVInt \tag{7-26}$$

The values of FEVSlp and FEVInt are input from the *Physiology* file, and are specific to certain age ranges. That is, different regression fits have been made for several different age ranges. Next, construct a centered body mass term z_{bmi} as:

$$z_{bmi} = BMI - FEVBMI \tag{7-27}$$

BMI (Eq. 5-2) is the body-mass index, which is an APEX profile variable that depends on the individual's height and weight. The parameter FEVBMI is fitted from the experimental data that was used for the beta parameters, and is input from the *Physiology* file.

Define a time-dependent variable X(t) which is a measure of the external "stress" on the lungs due to ozone. The differential equation from MSS is:

$$d\,X(t)/dt = C\left(\frac{VE}{BSA}\right)^{\beta_6} - \beta_5\,X \tag{7-28}$$

For a single diary event in APEX, the ozone concentration C and the ventilation rate VE are both assumed to remain constant. This equation can then be integrated as:

$$X(t) = X_o\, e^{-\beta_5 t} + \frac{C}{\beta_5}\left(\frac{VE}{BSA}\right)^{\beta_6}(1 - e^{-\beta_5 t}) \tag{7-29}$$

Here X_0 is the value of X at the start of the diary event, or equivalently, the value at the end of the previous diary event. For a diary event of duration D minutes, this becomes:

$$X(D) = X_o\, e^{-\beta_5 D} + \frac{C}{\beta_5}\left(\frac{VE}{BSA}\right)^{\beta_6}(1 - e^{-\beta_5 D}) \tag{7-30}$$

Here, *C* is the concentration in the microenvironment (ppm), *VE* is the instantaneous expired minute volume (L min^{-2}), *BSA* is the body surface area (m^2), and *D* is the duration of the event.

For use below, define X_{TH} as the amount by which $X(D)$exceeds a threshold:

$$X_{TH}(D) = Max(0, X(D) - \beta_9) \tag{7-31}$$

The value of $X(D)$ must exceed the threshold β_9, or else there is no effect (which means that X_{TH} is zero, it cannot be negative).

Define a response function M as:

$$M = \frac{\beta_1 + \beta_2 y_{age} + \beta_8 z_{bmi}}{1 + \beta_4\, e^{-\beta 3\, X_{TH}}} - \frac{\beta_1 + \beta_2 y_{age} + \beta_8 z_{bmi}}{1 + \beta_4} \tag{7-32}$$

Here, $\beta_1 - \beta_9$ are unitless fitted model parameters (see paper for details of fit). They are chosen once for each individual and remain constant throughout the simulation. By construction, when $X_{TH} = 0$, then M=0. Since β_4 is positive, when $X_{TH} > 0$ then M>0 (as the denominator of the first term is less than that of the second). Since X_{TH} is never negative, neither is M.

While the variable X_{TH} can never be negative, it has been found that the numerator in equation (7-32) could be, for certain values of age and body mass index. These occur due to extrapolation beyond the range of the original study data, and are likely not correct. Therefore, a minimum value of zero is now set for the numerators in equation (7-32).

The decrement in lung function is expressed as %ΔFEV1. Note that positive values of this variable mean a decrease in effective lung volume. The model is:

$$\%\Delta FEV1 = e^U M + E1 + e^U M\, E2 \tag{7-33}$$

E1 and E2 are residual variability terms, both of which are normally distributed with mean zero. One of the main changes from the earlier FEV calculations is the introduction of a variability

term which is proportional to M. Distributions for E1 and E2 are specified on the Physiology input file using the keywords "FEVE1" and "FEVE2." These should be specified as normal distributions with mean zero. Note that par2 is the standard deviation, which is the square root of the variance (the MSS paper reports the variance of E1 and E2). The sampling rates of E1 and E2 are controlled by the parameters *HourlyFEVE1* and *HourlyFEVE2* in the *Control Options* file.

CHAPTER 8. CALCULATING POLLUTANT CONCENTRATIONS IN MICROENVIRONMENTS

APEX calculates concentrations of all modeled air pollutants in all microenvironments at each timestep of the simulation period separately for each of the simulated individuals. The default APEX timestep is 1 hour (See *Volume I*). The timestep must match the data on the air quality input files, and is fixed throughout the APEX simulation. For example, for a 3-hour timestep, the air quality files must have 8 values per day. The air quality input files define the ambient air concentration for each district, for each pollutant. The microenvironmental concentrations are based on these ambient concentrations and are set in **MicroEnvModule:MicroConcs**. The input files and algorithms for these calculations are described in the following sections.

8.1 Defining Microenvironments

APEX gives the user great flexibility in defining the number and properties of the microenvironments (see step 4 in Figure 2.2). (Note that the term microenvironment is generally shortened to micro in the computer code and files.) Along with this flexibility, however, is the need for the user to specify a substantial amount of information about the microenvironments in the input files.

There are three input files that relate to microenvironments. The first is the *Microenvironment Mapping* file, which contains the mapping from the location categories used in the activity diaries to the APEX microenvironments. The second is the *Microenvironment Descriptions* file, which contains rules for calculating pollutant concentrations in each microenvironment. The third file is the *Profile Functions* file, in which profile variables influencing the microenvironmental concentrations can be defined. Examples are the presence or absence of air conditioning or a gas stove in the home. See *Volume I* for a discussion of the *Profile Functions* file.

The *Microenvironment Mapping* file gives the user control over how many microenvironments will be modeled and what CHAD (or other activity database) locations should be grouped into each microenvironment. This file contains one row for each CHAD location code, indicating which APEX microenvironment is to be used whenever that CHAD code is encountered. Thus, the 100+ location codes defined in the activity (CHAD) database are mapped into a smaller subset of user-defined microenvironments amenable to modeling. In addition, location codes are also mapped to concentration locations (e.g., Home, Work, Other, Road, Road Work, Near Home, Near Work, Last, H/W/O/R/RW/NH/NW/L), which tells APEX which set of ambient data are to be used. Road and Road Work are optional locations and are used in conjunction with a set of specific roadway air quality data. See Section 8.2.1 for details.

Table 8.1 lists the 115 location codes currently in CHAD and examples of microenvironments which can be assigned in the *Microenvironment Mapping* file.

Table 8.1. Example Mapping of CHAD Location Codes to APEX Microenvironments

CHAD Location Code	CHAD Location Description	APEX Microenv. Code	APEX Microenviroment Description	Location (see Section 8.2.1)
U	Uncertain of correct code	−1	Use previous microenvironment	U
X	No data	−1	Use previous microenvironment	U
30000	Residence, general	1	Indoors − Residence	H
30010	Your residence	1	Indoors − Residence	H
30020	Other residence	1	Indoors − Residence	H
30100	Residence, indoor	1	Indoors − Residence	H
30120	Your residence, indoor	1	Indoors − Residence	H
30121	..., kitchen	1	Indoors − Residence	H
30122	..., living room or family room	1	Indoors − Residence	H
30123	..., dining room	1	Indoors − Residence	H
30124	..., bathroom	1	Indoors − Residence	H
30125	..., bedroom	1	Indoors − Residence	H
30126	..., study or office	1	Indoors − Residence	H
30127	..., basement	1	Indoors − Residence	H
30128	..., utility or laundry room	1	Indoors − Residence	H
30129	..., other indoor	1	Indoors − Residence	H
30130	Other residence, indoor	1	Indoors − Residence	H
30131	..., kitchen	1	Indoors − Residence	H
30132	..., living room or family room	1	Indoors − Residence	H
30133	..., dining room	1	Indoors − Residence	H
30134	..., bathroom	1	Indoors − Residence	H

CHAD Location Code	CHAD Location Description	APEX Microenv. Code	APEX Microenviroment Description	Location (see Section 8.2.1)
30135	..., bedroom	1	Indoors − Residence	H
30136	..., study or office	1	Indoors − Residence	H
30137	..., basement	1	Indoors − Residence	H
30138	..., utility or laundry room	1	Indoors − Residence	H
30139	..., other indoor	1	Indoors − Residence	H
30200	Residence, outdoor	10	Outdoors − Other	H
30210	Your residence, outdoor	10	Outdoors − Other	H
30211	..., pool or spa	10	Outdoors − Other	H
30219	..., other outdoor	10	Outdoors − Other	H
30220	Other residence, outdoor	10	Outdoors − Other	H
30221	..., pool or spa	10	Outdoors − Other	H
30229	..., other outdoor	10	Outdoors − Other	H
30300	Residential garage or carport	7	Indoors − Other	H
30310	..., indoor	7	Indoors − Other	H
30320	..., outdoor	10	Outdoors − Other	H
30330	Your garage or carport	1	Indoors − Residence	H
30331	..., indoor	1	Indoors − Residence	H
30332	..., outdoor	10	Outdoors − Other	H
30340	Other residential garage or carport	1	Indoors − Residence	H
30341	..., indoor	1	Indoors − Residence	H
30342	..., outdoor	10	Outdoors − Other	H
30400	Residence, none of the above	1	Indoors − Residence	H
31000	Travel, general	11	InVehicle − Cars and Trucks	O/R
31100	Motorized travel	11	InVehicle − Cars and Trucks	O/R
31110	Car	11	InVehicle − Cars and Trucks	O/R
31120	Truck	11	InVehicle − Cars and Trucks	O/R
31121	Truck (pickup or van)	11	InVehicle − Cars and Trucks	O/R

CHAD Location Code	CHAD Location Description	APEX Microenv. Code	APEX Microenviroment Description	Location (see Section 8.2.1)
31122	Truck (not pickup or van)	11	InVehicle − Cars and Trucks	O/R
31130	Motorcycle or moped	8	Outdoors − Near Road	O/R
31140	Bus	12	InVehicle − Mass Transit	O/R
31150	Train or subway	12	InVehicle − Mass Transit	O
31160	Airplane	0	Zero concentration	O
31170	Boat	10	Outdoors − Other	O
31171	Boat, motorized	10	Outdoors − Other	O
31172	Boat, other	10	Outdoors − Other	O
31200	Non-motorized travel	10	Outdoors − Other	O/R
31210	Walk	10	Outdoors − Other	O/R
31220	Bicycle or inline skates/skateboard	10	Outdoors − Other	O/R
31230	In stroller or carried by adult	10	Outdoors − Other	O
31300	Waiting for travel	10	Outdoors − Other	O/R
31310	..., bus or train stop	8	Outdoors − Near Road	O/R
31320	..., indoors	7	Indoors − Other	O
31900	Travel, other	11	InVehicle − Cars and Trucks	O/R
31910	..., other vehicle	11	InVehicle − Cars and Trucks	O/R
32000	Non-residence indoor, general	7	Indoors − Other	O
32100	Office building/ bank/ post office	5	Indoors − Office	O
32200	Industrial/ factory/ warehouse	5	Indoors − Office	O
32300	Grocery store/ convenience store	6	Indoors − Shopping	H
32400	Shopping mall/ non-grocery store	6	Indoors − Shopping	O
32500	Bar/ night club/ bowling alley	2	Indoors − Bars and Restaurants	O
32510	Bar or night club	2	Indoors − Bars and Restaurants	O
32520	Bowling alley	2	Indoors − Bars and Restaurants	O
32600	Repair shop	7	Indoors − Other	O
32610	Auto repair shop/ gas station	7	Indoors − Other	O

CHAD Location Code	CHAD Location Description	APEX Microenv. Code	APEX Microenviroment Description	Location (see Section 8.2.1)
32620	Other repair shop	7	Indoors − Other	O
32700	Indoor gym /health club	7	Indoors − Other	O
32800	Childcare facility	4	Indoors − Day Care Centers	O
32810	..., house	1	Indoors − Residence	O
32820	..., commercial	4	Indoors − Day Care Centers	O
32900	Large public building	7	Indoors − Other	O
32910	Auditorium/ arena/ concert hall	7	Indoors − Other	O
32920	Library/ courtroom/ museum/ theater	7	Indoors − Other	O
33100	Laundromat	7	Indoors − Other	H
33200	Hospital/ medical care facility	7	Indoors − Other	O
33300	Barber/ hair dresser/ beauty parlor	7	Indoors − Other	H
33400	Indoors, moving among locations	7	Indoors − Other	O
33500	School	3	Indoors − Schools	O
33600	Restaurant	2	Indoors − Bars and Restaurants	O
33700	Church	7	Indoors − Other	H
33800	Hotel/ motel	7	Indoors − Other	O
33900	Dry cleaners	7	Indoors − Other	H
34100	Indoor parking garage	7	Indoors − Other	O
34200	Laboratory	7	Indoors − Other	O
34300	Indoor, none of the above	7	Indoors − Other	O
35000	Non-residence outdoor, general	10	Outdoors − Other	O
35100	Sidewalk, street	8	Outdoors − Near Road	O/R
35110	Within 10 yards of street	8	Outdoors − Near Road	O/R
35200	Outdoor public parking lot /garage	9	Outdoors − Public Garage / Parking	O/R
35210	..., public garage	9	Outdoors − Public Garage / Parking	O/R
35220	..., parking lot	9	Outdoors − Public Garage / Parking	O/R
35300	Service station/ gas station	10	Outdoors − Other	O
35400	Construction site	10	Outdoors − Other	O

CHAD Location Code	CHAD Location Description	APEX Microenv. Code	APEX Microenviroment Description	Location (see Section 8.2.1)
35500	Amusement park	10	Outdoors − Other	O
35600	Playground	10	Outdoors − Other	H
35610	..., school grounds	10	Outdoors − Other	O
35620	..., public or park	10	Outdoors − Other	H
35700	Stadium or amphitheater	10	Outdoors − Other	O
35800	Park/ golf course	10	Outdoors − Other	O
35810	Park	10	Outdoors − Other	O
35820	Golf course	10	Outdoors − Other	O
35900	Pool/ river/ lake	10	Outdoors − Other	O
36100	Outdoor restaurant/ picnic	10	Outdoors − Other	O
36200	Farm	10	Outdoors − Other	O
36300	Outdoor, none of the above	10	Outdoors − Other	O

All of the other properties of the microenvironments are provided in the *Microenvironment Descriptions* file. All microenvironments assigned locations in the *Microenvironment Mapping* file must be defined in the *Microenvironment Descriptions* file. More microenvironments may be described than are assigned locations, but they will not be used by APEX (since in this case simulated people will never enter the microenvironment). The total number of microenvironments described in the *Microenvironment Descriptions* file must also be indicated in the *Control Options* input file. (These input files are also covered in *Volume I*.)

Definition of the microenvironments using the *Microenvironment Descriptions* file is covered in Section 8.2.

8.2 Calculating Concentrations in Microenvironments

APEX calculates concentrations of all the air pollutants in all the microenvironments at each timestep of the simulation period for each of the simulated individuals, based on the ambient air quality data specific to the geographic locations visited by the individual. APEX provides two methods for calculating microenvironmental concentrations: the mass balance (MASSBAL) method and the simpler factors (FACTORS) method. The MASSBAL method starts with the previous timestep's concentration in each microenvironment, which is modified over time by exchange with the ambient air. The FACTORS method uses a simple equation to relate the concentration in each microenvironment to the current ambient concentration. Both methods require that a number of parameters including proximity, penetration and pollutant sources be specified over time; however, the MASSBAL method uses additional parameters such as air exchange, volume, and decay rates. All pollutants use the same method for a given microenvironment. These methods are described in the next two subsections. The user is required to specify the calculation methods for each of the microenvironments in the simulation in the *Microenvironment Descriptions* file (see *Volume I*). Microenvironments within a single simulation may use either method; mixing the methods (across micros, not pollutants) is allowed with no restrictions.

8.2.1 Microenvironmental Concentrations in Locations

In APEX, *locations* determine the source of ambient pollutant concentration data. In general, these represent the geographical locations a person moves through in a day. The locations are "Home" (H), "Work" (W), "Other" (O), "Roadway" (R), "Road near Work" (RW), "Near Home" (NH), and "Near Work" (NW). An ambiguous location, "Last" (L), is set to either Near Home or Near Work, depending on the last location the individual was in. A person who is not employed has identical work and home locations. The H and W concentrations are calculated from the air quality data in a person's home and work sectors, respectively. The concentrations in the O location are calculated from a composite of set of air districts. By default, APEX uses the city-average air concentration to calculate O concentrations. However, the user can customize this average concentration using the *Control Options* file settings ***SampleOtherLocs***, ***#OtherDistricts***, and ***HomeProbab*** (see *Volume I*). If the user specifies roadway air quality districts, then APEX will use these AQ data to determine microenvironmental concentrations for R and RW locations. R is drawn from air concentrations near the Home location, while RW is drawn from concentrations near the work location. NH is randomly sampled from a tract within a given distance from the H location, while W is sampled near the work location. L is set to either NH or NW, depending on the last location.

Location is determined event-by-event for each simulated person. However, APEX calculates concentrations in all microenvironments, for all times, for all locations, and repeats this process for each simulated person. Most of these are not encountered by the simulated person. For MASSBAL micros, prior concentrations are always needed, even when the person was not there before the current event. FACTORS micros do not have this requirement, but it is simpler (and faster) to perform extra calculations rather than perform logic tests to determine which calculations are necessary.

APEX uses the CHAD codes to determine location. The *Microenvironment Mapping* file assign location definitions to each activity database location code (see Table 8.1). APEX also assigns diary events with a "work" activity code to the W location. This assignment overrides the location assignment based on location code. By default, APEX assigns CHAD activity codes 10000-10300 to the work location (see Table 4-5 in *Volume I*). However, the user also has the ability to customize this setting using the *Control Options* file variable ***CustomWork***.

APEX defines one set of micro parameter distributions for each micro (i.e., there are no unique distributions for the H, W, O, R, RW, NH, NW, or L locations). However, the values of the parameters themselves may differ between locations. By using the ***ResampleWork*** keyword in a micro parameter description, APEX will select a different value from the distribution to use for the W location (see Section 8.3.3 for details). If ***ResampleWork*** is used, then the micro parameters for the O location will be the average of the parameters for the H and W locations. If not, the H, W, and O locations will all use the same values of the parameters for the micro. If the user specifies roadway AQ data, then APEX will use these data for all CHAD locations indicated by the R, or RW value. If the user chooses ***RoadLast*** =Y, then roadway concentrations will be chosen similar to the Last location; R or RW will be selected based on the last event to occur in either the home or work location.

8.2.2 Mass Balance Method

The mass balance method assumes that an enclosed microenvironment (e.g., a room in a residence) is a single, well-mixed volume in which the air concentration is approximately spatially uniform. The concentration of an air pollutant in such a microenvironment is estimated using the following four processes (as illustrated in Figure 8.1):

- Inflow of air into the microenvironment;
- Outflow of air from the microenvironment;
- Removal of a pollutant from the microenvironment due to deposition, filtration, and chemical degradation; and
- Emissions from sources of a pollutant inside the microenvironment.

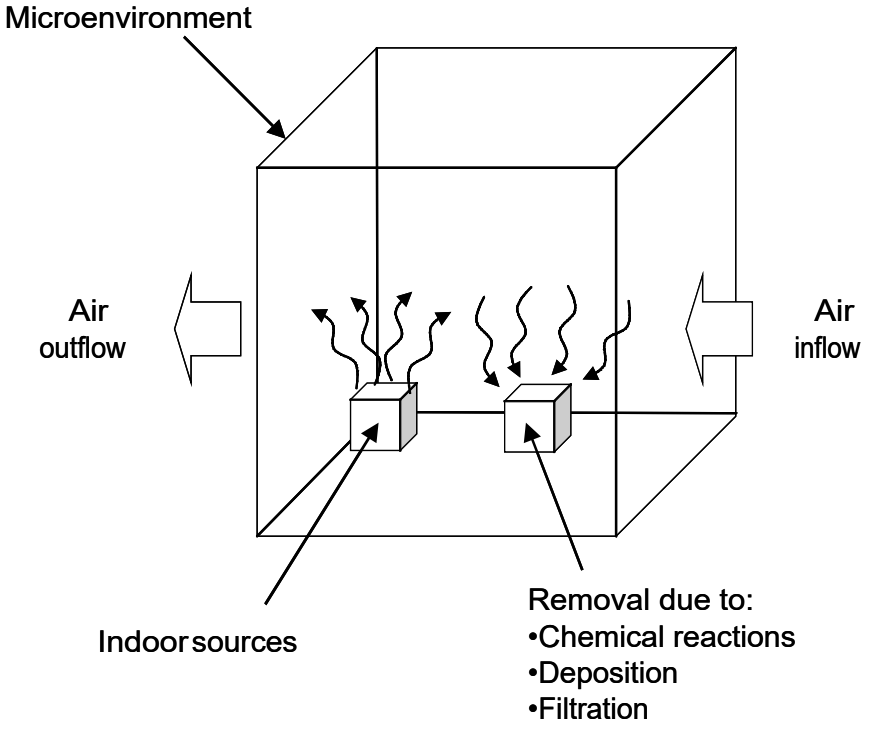

Figure 8.1. The Mass Balance (MASSBAL) Model

It is assumed that the amount of outside air flowing into the microenvironment is equal to that flowing out of the microenvironment. This rate is given by the air exchange rate, $R_{air\ exchange}$, with units of [1/hr]. The air exchange rate can be interpreted as the number of times per hour the entire volume of air in the microenvironment is replaced.

Considering the microenvironment as a distinct, well-mixed volume of air, the mass balance equation for a pollutant can be described by:

$$\frac{dC(t)}{dt} = \dot{C}_{in} - \dot{C}_{out} - \dot{C}_{removal} + \dot{C}_{source} \qquad (8\text{-}1)$$

where:

$C(t)$	=	Concentration in the microenvironment at time t ($\mu g/m^3$)
$\dot{C}_{in}$	=	Rate of change in C(t) due to air entering the micro
$\dot{C}_{out}$	=	Rate of change in C(t) due to air leaving the micro
$\dot{C}_{removal}$	=	Rate of change in C(t) due to all removal processes
$\dot{C}_{source}$	=	Rate of change in C(t) due to all source terms

Note that concentration must be in the same units as the ambient air quality data, i.e., either ppm, ppb, or $\mu g/m^3$, although throughout these equations concentration is shown only in $\mu g/m^3$ for brevity.

The change in microenvironmental concentration due to influx of air, $\dot{C}_{in}$, is:

$$\dot{C}_{in} = C_{ambient} \times f_{proximity} \times f_{penetration} \times R_{air\,exchange} \tag{8-2}$$

where:

$C_{ambient}$	=	Ambient timestep concentration ($\mu g/m^3$)
$f_{proximity}$	=	Proximity factor (unitless)
$f_{penetration}$	=	Penetration factor (unitless)
$R_{airexchange}$	=	Air exchange rate between micro and outdoors (1/hour)

The proximity factor $f_{proximity}$ is used to account for differences in ambient concentrations between the geographic location represented by the ambient air quality data (e.g., a regional fixed-site monitor) and the geographic location of the microenvironment. That is, the outdoor air at a particular location may differ systematically from the outdoor air at the center of the air quality district. For example, a house might be located next to a busy road in which case the air outside the house would have elevated levels for mobile source pollutants such as carbon monoxide. The concentration $C_{outdoor}$ in the air directly outside the microenvironment is given by the product of the ambient concentration and $f_{proximity}$:

$$C_{outdoor} = f_{proximity} C_{ambient} \tag{8-3}$$

For some pollutants (especially particulate matter), the process of infiltration may remove a fraction of the pollutant from the air. The fraction that is retained in the air is given by the penetration factor $f_{penetration}$. During exploratory analyses, the user may examine how a microenvironment affects overall exposure by setting the microenvironment's proximity or penetration factor to zero, thus effectively eliminating the microenvironment.

Change in microenvironmental concentration due to outflux of air is calculated as the concentration in the microenvironment C(t) multiplied by the air exchange rate:

$$\dot{C}_{out} = R_{air\,exchange} \times C(t) \tag{8-4}$$

The third term in the MASSBAL calculation represents removal processes within the microenvironment. There are three such processes in general: chemical reactions, deposition, and filtration. Chemical reactions are significant for ozone, for example, but not for carbon monoxide. The amount lost to chemical reactions will generally be proportional to the amount present, which in the absence of any other factors would result in an exponential decay in the concentration with time. Similarly, deposition rates are usually given by the product of a (constant) deposition velocity and a (time-varying) concentration, also resulting in an exponential decay. The third removal process is filtration, usually as part of a forced air circulation or HVAC system. Filtration will normally remove particles but not gases. In any case, filtration rates are also proportional to concentration. Changes in concentration due to deposition, filtration, and chemical degradation in a microenvironment are simulated based on the first-order equation:

$$\dot{C}_{removal} = \left(R_{deposition} + R_{filtration} + R_{chemical}\right) C(t) = R_{removal} \times C(t) \qquad (8\text{-}5)$$

where:

$\dot{C}_{removal}$	=	Change in microenvironmental concentration due to removal processes (µg/m3/hour)
$R_{deposition}$	=	Removal rate of a pollutant from a microenvironment due to deposition (1/hour)
$R_{filtration}$	=	Removal rate of a pollutant from a microenvironment due to filtration (1/hour)
$R_{chemical}$	=	Removal rate of a pollutant from a microenvironment due to chemical degradation (1/hour)
$R_{removal}$	=	Removal rate of a pollutant from a microenvironment due to the combined effects of deposition, filtration, and chemical degradation (1/hour)

For unreactive gases like carbon monoxide, all three removal terms could be zero, in which case $R_{removal}$ =0.

The fourth term in the MASSBAL calculation represents pollutant sources within the microenvironment. This is the most complex term primarily due to the fact that several sources may be present. APEX allows two methods of specifying source strengths: emission sources (***ESource*** or ***ES***) or concentration sources (***CSource*** or ***CS***). Either may be used for MASSBAL microenvironments, and both can be used within the same microenvironment. The source strength values are used to calculate the source term, $\dot{C}_{source}$.

Emission sources are expressed as emission rates in units of µg/hr. To determine the source term associated with an emission source, ES must be divided by the volume V of the microenvironment in m^3:

$$\dot{C}_{source,ES} = \frac{ES}{V} \qquad (8\text{-}6)$$

Concentration sources, however, are expressed in units of concentration. These must be the same units as used for the ambient concentration (e.g., μg/m^3, ppm or ppb). Concentration sources are normally used as additive terms for microenvironments using the FACTORS method. Strictly speaking, they are somewhat inconsistent with the MASSBAL method, since concentrations should not be inputs but should be consequences of the dynamics of the system. Nevertheless, a suitable meaning can be found by determining the source strength $\dot{C}_{source}$ that would result in a mean increase of CS in the concentration, given constant parameters and equilibrium conditions, in this way:

Assume that a microenvironment is always in contact with clean air (ambient = zero) and it contains one concentration source. Then the mean concentration over time in this microenvironment from this source should be numerically equal to CS. The mean source strength, expressed in ppm/hr, ppb/hr or μg/m^3/hr, is the rate of change in concentration $\dot{C}_{source,CS}$. In equilibrium,

$$CS = \frac{\dot{C}_{source,CS}}{R_{air\ exchange} + R_{removal}} \tag{8-7}$$

$\dot{C}_{source,CS}$ can be written as:

$$\dot{C}_{source,CS} = CS \times R_{mean} \tag{8-8}$$

where R_{mean} is the chemical removal rate. From eq. 8-7, R_{mean} is equal to the sum of the air exchange rate and the removal rate ($R_{air\ exchange}$+$R_{removal}$) under equilibrium conditions. In general, however, the microenvironment will not be in equilibrium, but in such conditions there is no clear meaning to attach to $\dot{C}_{source,CS}$ since there is no fixed emission rate that will lead to a fixed increase in concentration. The simplest solution is to use $R_{mean} = R_{air\ exchange}+R_{removal}$. However, the user is given the option of specifically specifying R_{mean} (see discussion of parameters below). This may be used to generate a truly constant source strength $\dot{C}_{source,CS}$ by making CS and R_{mean} both constant in time. If this is not done, then R_{mean} is simply set to the sum of ($R_{air\ exchange}$ + $R_{removal}$). If these parameters change over time, then $\dot{C}_{source,CS}$ also changes. Physically, the reason for this is that in order to maintain a fixed elevation of concentration over the base conditions, then the source emission rate would have to rise if the air exchange rate were to rise.

Multiple emission and concentration sources within a single microenvironment are combined into the final total source term by combining equations 8-6 and 8-8:

$$\dot{C}_{source} = \dot{C}_{source,ES} + \dot{C}_{source,CS} = \frac{1}{V}\sum_{i=1}^{n_e} ES_i + R_{mean}\sum_{i=1}^{n_c} CS_i \tag{8-9}$$

where:

ES_i	=	Emission source strength for emission source *i* (µg/hour)
CS_i	=	Emission source strength for concentration source *i* (µg/m^3 or ppm or ppb, same as InputUnits)
n_e	=	Number of emission sources in the microenvironment
n_c	=	Number of concentration sources in the microenvironment

A note on units: The above equation is modified if the units of air quality are ppm or ppb rather than µg/m^3. For ppm, *1/V* is replaced by 1/(*V*ppmFact*); or for ppb, *1/V* is replaced by 1000/(*V*ppmFact*). The value of ppmFact is user-supplied in the *Control Options* input file; it expresses the number of µg/m^3 that equate to 1 ppm. For the pollutant CO, past runs used ppmFact=1145, but this is not hard-coded and needs to be specified on the *Control Options* file. For example, if a CO source had ES=2290 µg/hr in a room of volume 20 m^3, then $\dot{C}_{source}$ would be 2290/(20*1145) = 0.10 ppm/hr. That is, in the absence of losses, the CO concentration would increase at a rate of 0.10 ppm each hour due to the CO source.

Equations 8-2, 8-4, 8-5,and 8-9 can now be combined with 8-1 to form the differential equation for the microenvironmental concentration C(t). Within the time period of a timestep, $\dot{C}_{source}$ and $\dot{C}_{in}$ are assumed to be constant. Using $\dot{C}_{combined} = \dot{C}_{source} + \dot{C}_{in}$ leads to:

$$\frac{dC(t)}{dt} = \dot{C}_{combined} - R_{air\ exchange}C(t) - R_{removal}C(t)$$

$$= \dot{C}_{combined} - R_{mean}C(t) \qquad (8\text{-}10)$$

Solving this differential equation leads to:

$$C(t) = \frac{\dot{C}_{combined}}{R_{mean}} + \left(C(0) - \frac{\dot{C}_{combined}}{R_{mean}} \right) e^{-R_{mean}t} \qquad (8\text{-}11)$$

where:

C(0)	=	Concentration of a pollutant in a microenvironment at the beginning of a timestep (µg/m^3)
C(t)	=	Concentration of a pollutant in a microenvironment at time *t* within the time period of a timestep (µg/m^3).

Based on eq. 8-11, the following three concentrations in a microenvironment are calculated:

$$C_{equil} = C(t \to \infty) = \frac{\dot{C}_{combined}}{R_{mean}} = \frac{\dot{C}_{source} + \dot{C}_{in}}{R_{air\ exchange} + R_{removal}} \qquad (8\text{-}12)$$

$$C_{end} = C_{equil} + \left(C(0) - C_{equil} \right) e^{-R_{mean}t} \qquad (8\text{-}13)$$

$$C_{mean} = \frac{\int_0^1 C(t)dt}{\int_0^1 dt} = C_{equil} + \left(C(0) - C_{equil}\right)\frac{1 - e^{-R_{mean}t}}{R_{mean}t} \quad (8\text{-}14)$$

where:

t = length of the APEX timestep (hours)
C_{equil} = Concentration in a microenvironment (µg/m^3) if t $\rightarrow \infty$ (equilibrium state).
$C(0)$ = Concentration in a microenvironment at the beginning of the timestep (µg/m^3)
C_{end} = Concentration in a microenvironment at the end of the timestep (µg/m^3)
C_{mean} = Mean concentration in a microenvironment for timestep (µg/m^3)
R_{mean} = $R_{air\ exchange} + R_{removal}$ (1/hour)

At each timestep of the simulation period, APEX uses Eqs. 8-12, 8-13, and 8-14 to calculate the equilibrium, ending, and mean concentrations. APEX reports mean concentration as the concentration for a specific timestep. The calculation continues to the next timestep by using C_{end} for the previous timestep as $C(0)$.

The microenvironmental parameters for the MASSBAL method that can be defined by the user in the *Microenvironment Descriptions* file are summarized in the Table 8.2, with their valid ranges and their corresponding names in the file.

Table 8.2. Microenvironmental Parameters

Parameter	Definition	Units	Range	Default Value	Name[a]
$f_{proximity}$	Proximity factor	unitless	$f_{proximity} \geq 0$	1	PR
$f_{penetration}$	Penetration factor	unitless	$0 \leq f_{penetration} \leq 1$	1	PE
CS	Concentration source	µg/m^3, ppm, or ppb	$CS \geq 0$	0	CS
ES	Emission source	µg/hr	$ES \geq 0$	0	ES
$R_{removal}$	Removal rate due to deposition, filtration, and chemical reaction	1/hr	$R_{removal} \geq 0$	0	DE
$R_{air\ exchange}$	Air exchange rate	1/hr	$R_{air\ exchange} \geq 0$	none	AE
R_{mean}	Mean removal rate:	1/hr	$R_{mean} \geq 0$	$R_{removal} + R_{air\ exchange}$	MR
V	Volume of microenvironment	m^3	$V > 0$	none	V

Parameter	Definition	Units	Range	Default Value	Name[a]

[a] Designation in *Microenvironment Descriptions* file

Not all of the possible parameters are always needed and several of them have natural default values. Based on the above equations, the following generalizations can be made about the definition of the MASSBAL parameters in the *Microenvironment Descriptions* file:

- Air exchange rate is a critical parameter that is always needed in a MASSBAL calculation. It must always be defined in the file as it has no default value.
- Air exchange rate and volume are not pollutant-specific, and therefore are defined only once for each micro. All other parameters must be defined for each pollutant.
- Removal rate must also be user-defined in the file if not assumed to be zero. For some pollutants, it can be assumed to have a natural default value of zero.
- The proximity and penetration factors must be defined in the file unless assumed to be unity, which is the natural default value for both factors that should be used in the absence of data to the contrary.
- If any emission source terms are present, then volume must be defined. Volume has no default value.
- If any concentration source terms are present, then the mean removal rate may be user-defined, but if appropriate, it may assume a default value of ($R_{air\ exchange} + R_{removal}$).

The details for specifying these input parameters in the *Microenvironment Descriptions* file are provided in the *Volume I* of this User's Guide. Further details on the options for designating these parameters are given in Section 8.3.

In APEX, it is assumed that the outdoor concentration and the other modeling parameters for the MASSBAL method remain constant during any timestep. Of course, recalling that the APEX default timestep is one hour, in many cases the MASSBAL parameters may not remain constant for one timestep at a time. For example, a person may enter a microenvironment and smoke a cigarette for five or ten minutes and then leave. Or, someone might enter a kitchen and cook for a few minutes using a gas stove. Or one might alter an air exchange rate by opening or closing a window. There are two reasons why it is difficult to model such events in APEX. First, there is already a large computational burden in calculating concentrations in every microenvironment for every timestep for every simulated person. This burden is substantially large if very fine time resolution were demanded. Second, most examples of fine-scale parameter variation are driven by human actions. The CHAD activity diaries generally do not contain sufficient detail to determine when each cigarette is lit, each time a stove is used, or a window is opened or closed. Furthermore, the diaries only follow the activities of a single person. It is quite possible for these actions to be performed by other people. For example, if the activity diary follows a child, then the child's parents may be doing these things that affect the properties of the microenvironments that the child is in. Since the diaries do not reliably report such information, it was decided that a very fine time resolution could not reliably be used for the calculation of concentrations.

In a MASSBAL microenvironment, the concentration during any timestep depends on the concentration for the previous timestep. Ultimately, all timesteps depend on some method for establishing initial conditions. To avoid the problem of establishing new initial conditions every

time the activity diary indicates that a MASSBAL microenvironment is entered, the time series is evaluated for all timesteps in the simulation period. An extra 24-hour period is added prior to the start of the APEX simulation period by duplicating the properties from the first day of the simulation period. It is assumed that 24 hours is sufficient so that the initial concentration becomes irrelevant. The entire simulation period is then evaluated timestep-by-timestep, without gaps, with each timestep being used to determine the next.

8.2.3 Factors Method

The FACTORS method is simpler than the mass balance method. In this method, the value of the concentration in a microenvironment is not dependent on the concentration during the previous timestep. Rather, the method uses the following equation to calculate timestep concentration in a microenvironment from the user-provided air quality data:

$$C_{timestep} = C_{ambient}\, f_{proximity}\, f_{penetration} + \sum_{i=1}^{n_C} CS_i \qquad (8\text{-}15)$$

where:

$C_{timestep}$	=	Timestep concentration in a microenvironment ($\mu g/m^3$)
$C_{ambient}$	=	Timestep concentration in ambient environment ($\mu g/m^3$)
$f_{proximity}$	=	Proximity factor (unitless)
$f_{penetration}$	=	Penetration factor (unitless)
CS_i	=	Mean air concentration resulting from source i ($\mu g/m^3$)
n_c	=	number of concentration sources in the microenvironment

The user may provide values for proximity, penetration, and any concentration source terms, or may allow them to assume default values (see Table 8.2); however, it is not mandatory that the user supply any values if the default values are suitable. An undefined proximity or penetration is assumed to be unity at all times. Missing (i.e., undefined) sources are assumed to be zero. Parameters are left undefined by simply omitting them from the *Microenvironment Descriptions* input file. If all parameters are missing, then the concentration in the microenvironment is always the same as the ambient concentration. All of the parameters in the above equation are evaluated for each timestep, although these values might remain constant for several hours, entire days, or even for the entire simulation. For the ambient concentration, the timestep values come from the input *Air Quality Data* file, and may be either measurements or modeled results, or may be sampled for each hour from a distribution (see *Volume I* for available formats of this file). For the other parameters, the timestep values are the result of the calculations based on the information specified in the *Microenvironment Descriptions* input file.

The proximity and penetration factors operate similarly, so the user can choose how the split their effects, or whether to combine them and leave the other factors at a point value of one. Proximity is intended to represent the relationship between the specific location of the micro (for example, the person's house) and the ambient monitoring site for the district. A house beside a park may have cleaner air than the district as a whole, and therefore have a proximity factor below one. Another house may be on a busy road, and for vehicular-related pollutants would have a proximity factor over one. The penetration factor represents the fraction of pollutant that

manages to enter the micro during air exchange. For gaseous pollutants the penetration factor should generally be one. For particulate pollutants, some may be removed when air passes through the small cracks and openings to enter the micro.

8.3 Microenvironment Parameter Definitions

The second section of the *Microenvironment Descriptions* input file contains the rules for determining the values of the parameters used in the MASSBAL and FACTORS methods. Instructions for specifying microenvironmental input parameters (those in Table 8.2) in the *Microenvironment Descriptions* file are provided in *Volume I* of this User's Guide, but further details on the options for using resampling rates; conditional variables; periodic (daily, weekly, and monthly) groupings; and random seeds are provided in this section. This includes an explanation of ways to specify CS terms as products of distributions.

Both of the concentration calculation methods require multiple user-defined input parameters. These microenvironmental parameters are defined by probability distributions defined in the *Microenvironment Descriptions* input file, with values being assigned to each timestep of the simulation. The file is read in **MicroEnvModule:ReadMicroData**. Any of the following distribution types in Table 3.1 can be used for microparameters.

The user may provide different distribution data for the parameter for any combination of the following temporal and spatial variables:

- Hour in a day
- Day in a week
- Month in a year (i.e., season)
- Air quality district

For example, the user can define probability distributions for a parameter that vary depending on the time of day and whether the simulated timestep is on a weekday or a weekend, or the user can define a distribution that changes with the season of the year and with the air quality district associated with the microenvironment being considered (recall that home and work sectors for each profile will be associated with a unique air quality district).

The distributions for the microenvironmental parameters may also depend on conditional variables that are a subset of the profile variables for an individual. A single microenvironmental parameter may depend on up to three of the above conditional variables. Conditional variables may change on an interpersonal, geographic, meteorological, or temporal basis, influencing the microparameters accordingly.

The rules for determining any microenvironmental parameter (MP) are unique to each microenvironment. For example, the rules for proximity for a house may differ completely from the rules for proximity for a car. Every combination of parameter, pollutant, and microenvironment is distinct (with the exception of air exchange rate and volume, which can only be defined once per micro). The order in which the MP definitions are presented is not significant to APEX, although it might help the user to group them by microenvironment, pollutant, or MP. Note that the entire definition of an MP can be omitted if its default value is acceptable, so for example if proximity is always to be unity in some microenvironment, then no

MP definition is needed for proximity in that microenvironment. The example in Exhibit 8-1 shows one possibility for the proximity parameter for microenvironment #1. The data are only for illustrative purposes and are not intended to properly represent this MP in any real scenario.

```
Micro number      = 1
Pollutant = 3
Parameter Type    = Proximity
Hours - Block     = 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1
Weekday-DayType   = 1 1 1 1 1 1 1
Month-Season      = 1 1 2 2 2 3 3 3 4 4 4 1
District-Area     = 1 1 1 1 1 1
Condition #1 = 0
Condition #2 = 0
Condition #3 = 0
ResampHours = NO
ResampDays  = YES
ResampWork  = YES
RandomSeed  = 0
Block DType Season Area  C1  C2  C3   Shape      Par1   Par2  Par3 Par4  LTrunc UTrunc  ResampOut
1     1     1      1     1   1   1    Normal      1.5    1.2     .    .    0     4.0      Y
2     1     1      1     1   1   1    Point       2.0      .     .    .    .     .        .
1     1     2      1     1   1   1    Lognormal   1.2    1.5     .    .    0     10       Y
2     1     2      1     1   1   1    Lognormal   0.4    1.2     .    .    0     10       Y
1     1     3      1     1   1   1    Triangle    0      3       2    .    .     .        Y
2     1     3      1     1   1   1    Normal      2.5    1.5     .    .    .     .        Y
1     1     4      1     1   1   1    Uniform     0      3       .    .    .     .        Y
2     1     4      1     1   1   1    Lognormal   3      2       .    .    0     10       Y
```

Exhibit 8-1. Example of a Microenvironmental Parameter Description

The general format is the same for all MPs. The first three lines are mandatory and specify the microenvironment number, the pollutant (indicated by its order in the *Control Options* file), and the MP, or parameter type (the **Pollutant** line may be absent for AER and Vol). This combination should be unique for every MP in the input file, with the possible exception of enumerated sources (discussed later).

The parameter types are indicated using standard keywords (given in Table 8.2). Note that only the first two characters of the parameter type are checked by APEX, and the keyword is not case sensitive, so the example above could use "PR." The user may spell out the parameter types if desired, providing greater clarity.

After the microenvironment number and parameter type, any or all of the remaining lines containing an equal sign may be omitted. These indicate the settings for various options, all of which have default values. These settings may appear in any order within the description; they are recognized via the keyword that precedes the equal sign. One option, missing in the above example since it only applies to parameter types ES and CS, is the source number. See section 8.3.5 for an example using this option. The other seven options in this example are the mappings that determine the values for the seven indices that label each distribution. This is followed by three resampling options and a random seed initialization option. These options are covered in the following subsections.

After all the options are specified, the next line (starting with "Block") indicates that the following lines contain descriptions of distributions. At least one distribution is always required; the exact number needed depends on the settings of the seven indexing options. The shortest possible MP description (other than one completely missing) consists of five lines. Such a description would have the first four mandatory lines, the header line indicating that distributions follow, and a single distribution that applies to all timesteps of the simulation, as in Exhibit 8-2.

```
Micro number = 1
Pollutant = 1
Parameter Type = Proximity
Block DType Season Area  C1  C2  C3  Shape  Par1  Par2  Par3 Par4  LTrunc UTrunc  ResampOut
1     1     1      1     1   1   1   Normal 1.5   1.2   .    .     0      4        Y
```

Exhibit 8-2. Example of the Shortest Possible MP Description

These rules state that the proximity in microenvironment #1 is to be drawn from a normal distribution with mean 1.5 and standard deviation 1.2. If the value drawn is below zero or above 4.0, then another value is drawn until one is found that is within bounds. This single value is then applied to all timesteps of the simulation—for that particular simulated individual and for their home air quality district. A separate single value is applied to that individual for all timesteps in their work district. Actually, a third value, the average of the first two, is applied to all timesteps in the "other" (non-home, non-work) air quality district. In effect, all the microenvironments are modeled in triplicate to account for the three different places. In the above example, the three values would all be the same if an extra line "`ResampleWork=NO`" were added after the second line. Each simulated individual is modeled independently so new values are drawn for all parameters when starting another profile.

It is necessary that all the parameter distribution data be given in the correct units (i.e., those that are compatible with the data in the *Air Quality Data* input files.)

The motivation for the rather complex programming that defines and evaluates the MP is that the various parameters that enter the MASSBAL and FACTORS equations may have widely divergent properties. For example, a parameter like house volume should have a single value that does not change over time. Another parameter such as an air exchange rate may change every hour. Some parameters like source strengths from cooking or from traffic may show strong diurnal patterns that may repeat on a daily or a weekly basis. Parameters that relate to temperature may show seasonal variation.

There are a number of possibilities for each optional rule in the MP definitions. The options fall into the following categories:

- Time and area mappings;
- Conditional variables;
- Correlation settings;
- Resampling options;
- Random number seeds; and
- Source number specification.

Each of these are discussed in the following subsections, after which comes a brief subsection on the specification of distributions.

8.3.1 Time and Area Mappings

Each MP is evaluated for every timestep of the simulation period. Normally, this is accomplished by drawing a value at random from a distribution. The user may specify that different distributions apply at timesteps within different hours of the day. Furthermore, the

frequency of sampling can be controlled by the user. It is not the case that a new value must be drawn every timestep; instead, values drawn for other times may be reused. The primary purpose of the time and area mappings is to specify which distribution applies to each timestep. The secondary purpose, in conjunction with the resampling options, is to establish periodic reuse of values on a daily, weekly, monthly, or geographical basis.

As an example, suppose some parameter should be sampled from one distribution during typical working hours and from another distribution at other times. This can be accomplished by defining two "blocks" and then assigning each hour of the day to either block #1 or block #2. Perhaps hours 1-7 (midnight to 7 a.m.) and hours 19-24 (6 p.m. to midnight) belong to block #1, while hours 8-18 (7 a.m. to 6 p.m.) belong to block #2. The distributions must then be defined for block #1 and block #2. Hours that fall into block #1 will have their parameter values drawn from the distribution that applies to block #1, and so on. If, for example, the distribution for block #2 has the higher mean, then the daytime values for the parameter will generally be higher than the nighttime values, generating a diurnal pattern. Similarly, weekly and seasonal patterns may be generated.

The Hour-Block (HB) mapping indicates the block to which each hour of the day belongs. The mapping must contain 24 numbers (even if the time steps do not equal an hour). The first is the block number for hour 1 (midnight to 1 a.m.), and so on. All timesteps belonging to the same block use the same distribution. The block numbers range from 1 upwards; there can be anything from 1 to 24 blocks. The number of blocks (#blocks) is determined from this mapping itself. If the Hour-Block mapping is missing, it is assumed that there is only one block, which implies that the parameter values for all 24 hours of the day (that is, all timesteps) are taken from the same distribution. The term "block" might suggest that the hours belonging to a given block should be adjacent chronologically, but this is not necessary in APEX. It is also not necessary for each block to contain the same number of hours.

The Weekday-Daytype (WT) mapping is similar to the Hour-Block mapping, except that it contains seven values instead of 24. The first value is the day type for Sunday, the second is for Monday, and so on until the last which is the day type for Saturday. The seven days of the week are in the same order as on a standard calendar. Thus if day type 1 is weekday and 2 is weekend, the vector should be (2 1 1 1 1 1 2), but without the parentheses. The mapping (1 2 2 2 2 2 1) would be equivalent if the distributions presented further down were appropriately renumbered. If the WT mapping is missing in the *Microenvironment Descriptions* input file, then only one day type is assumed, that is, the default mapping is (1 1 1 1 1 1 1).

The Month-Season (MS) mapping is similar to the previous two, except that 12 numbers are needed. The first number indicates the season for January, and so on through December. Again, if this mapping is missing then a single season is assumed to apply to all months.

The District-Area (DA) mapping is similar to the others, except that the number of air quality districts is not a universal constant, but may vary from one simulation to another. If the mapping is present, the user must ensure that it contains the correct number of terms (one area assignment for each air quality district in the study area). The district indices represent the APEX air district numbers, as enumerated in the *Sites* output file.

If the user defines 2 blocks, 2 day types, 4 seasons, and 3 areas, then a total of 48 distributions (that is, 2 x 2 x 4 x 3) must be specified—one for each possible combination. The number would be even larger if any conditional variables were used. To ease the burden of data requirements, the number of cases should be kept to a minimum. For example, if the seasonal dependence of this parameter were weak, one could eliminate it and reduce the number of distributions from 48 to just 12. If this is too extreme, perhaps two seasons would suffice to capture the variation. The number of seasons (or blocks, day types, or areas) can be defined differently for each MP, even ones that belong to the same microenvironment. Each MP is evaluated for all timesteps in the simulation period, independently of other MP, so there is no reason why the rules for one MP should match or correspond with the rules for any other MP.

8.3.2 Conditional Variables

Selected profile variables may be used to influence the parameter values in the MASSBAL or FACTORS equations. These profile variables are known as *Conditional Variables*. Conditional variables can be used to vary parameters on a profile, daily, hourly, or timestep basis. The list of variables to use for the current simulation are set in **ProfileModule: SetCVlist**. The allowable conditional profile variables are:

- ***Gender***
- ***Population category (Race/gender combination)***
- ***Employed***
- ***HasGasStove***
- ***HasGasPilot***
- ***AC_Home***
- ***AC_Car***
- ***Window_Res***
- ***Window_Car***
- ***SpeedCat***
- ***ProfileConditional1***
- ***ProfileConditional2***
- ***ProfileConditional3***
- ***RegionalConditional1***
- ***RegionalConditional2***
- ***RegionalConditional3***
- ***RegionalConditional4***
- ***RegionalConditional5***
- ***FactorGroup***

These variables influence the parameters on a profile basis. See Chapter 5 for definition of these variables. With the exception of the first three, rules for setting these variables for each profile are defined in the *Profile Functions* input file. In addition to these variables, the MPs can also depend on seven meteorological variables:

- ***TempCat*** Hourly temperature, binned into categories.
- ***HumidCat*** Hourly humidity, binned into categories.
- ***PrecipCat*** Hourly precipitation category.

- ***WindCat*** Hourly wind speed, binned into categories.
- ***DirCat*** Hourly wind direction, binned into categories.
- ***MaxTempCat*** Daily maximum temperature, binned into categories.
- ***AvgTempCat*** Daily average temperature, binned into categories

The first five can influence parameters on an hourly basis, while the last 2 are daily-varying parameters.

The MPs can depend daily on 5 user-defined daily varying functions:

- ***DailyConditional1***
- ***DailyConditional2***
- ***DailyConditional3***
- ***DailyConditional4***
- ***DailyConditional5***

Finally, the MPs can depend daily on 5 user-defined functions that depend on the ambient air quality and vary with each timestep:

- ***AQConditional1***
- ***AQConditional2***
- ***AQConditional3***
- ***AQConditional4***
- ***AQConditional5***

These hourly, timestep, and daily varying variables are not profile variables. However, rules for defining them are also designated in the *Profile Functions* input file (see *Volume I*).

Note that Conditional Variables must be integer, since their values are used as indices to select the distribution to be sampled. In practice, the number of categories must be fairly small (generally 2 or 3), otherwise defining distributions for every case becomes burdensome. It should be noted that while the concentrations in various microenvironments may depend on the profile through the conditional variables, these concentrations do not depend on the activity diaries or the event structure.

The user may select up to three Conditional Variables from the list for each MP. If one is used, it does not matter whether conditional variable #1, #2, or #3 is used. If more than one is used then the order they are designated does not matter.

The conditional variable to be applied is identified by its name. For example, if a parameter were to depend on gender then it would be indicated as follows in the input file:

```
Condition # 1     = Gender
```

Note that the word "gender" must be spelled out in full as in the above list, but it is not case sensitive. There are two ways to indicate that a conditional variable is not used. Either the line for it can be omitted, or else the variable name can be set to anything that is not on the list. The standard practice (if the line is not simply omitted) is to set the right hand side to zero:

```
Condition # 1  = 0
```

There is a complication for the user when it comes to specifying the distributions that are applicable to conditional variables. Each distribution has seven indices (four for time and area mappings and three for Conditional Variables). The values that any index may have must be integers. Thus, if gender is used as a Conditional Variable, then the user must specify distributions for gender=1 and for gender=2. The user cannot use mnemonic devices such as "M" or "F" instead. The numerical codes for the Conditional Variable are set as constants in **GlobalModule** and are as follows:

- ***Gender:*** 1=Male, 2=Female
- ***Employed:*** 1=YES, 2=NO
- ***HasGasStove:*** 1=YES, 2=NO
- ***HasGasPilot:*** 1=YES, 2=NO
- ***AC_Car:*** 1=YES, 2=NO
- ***Window_Res:*** 1=OPEN, 2=CLOSED
- ***Window_Car:*** 1=OPEN, 2=CLOSED

For the population category conditional variable, the conditional variable value codes are integers that represent the order in which the population categories are defined within the "Population file" definitions in the *Control Options* file. For example, if "Native American females" is the third population file defined, the code for that population category would then be 3. For the other Conditional Variables, the numerical values are provided by the user in the *Profile Functions* input file, so there are no pre-assigned ranges or interpretations. The list of conditional variables to be used in defining MPs is set in **DistributionModule:SetCVList**.

8.3.3 Resampling Options

Random sampling from distributions in APEX is a two-step process. First, a uniform random value ranging from zero to one is drawn. This is later transformed to a sample from the appropriate distribution using the inverse CDF method. "Resampling" in APEX refers to the drawing of a new uniform random value. Each of the "Resamp" settings (namely, ResampTS, ResampHours, ResampDays, and ResampWork) indicates when new uniform random samples are produced, and each may be set to YES or NO independently.

Hour and timestep resampling determine whether the various hours or timesteps within a day share the same uniform random samples, or not. ResampHours=YES works only if the timestep is one hour or less. If the timestep is one hour, then ResampTS and ResampHours have the same effect, and will be implemented if either is set to YES. The defaults are ResampTS=NO and ResampHours=NO, which means that all timesteps within the same simulation day will share the same uniform random sample.

The third resampling option is ResampDays. If ResampDays=NO, then the same daily profile of uniform random samples is used on all days in the same category. If ResampDays=YES, then all days have new sets of values drawn for them. The last resampling option is ResampWork. APEX normally generates different parameter values for home and work locations. However, there are cases when this is not logical. For example, if a car is defined to be a

microenvironment and it uses the mass balance method, the volume of the car should be the same whether the car is at home or at work. If ResampWork=YES, then home and work always draw parameter values independently. If ResampWork=NO, then the same values are used for work as for home. Since such cases are rare, ResampWork=YES is the default, meaning that the workplace will have its values sampled independently of the home.

To summarize, the default values for ResampTS, ResampHours and ResampDays are NO and for ResampWork it is YES. This means that the default is to draw only two values (one for home and one for work) from each distribution listed for that MP, for each person. If the default is to be used for any of the Resamp options, then the line may be omitted from the input file, but it does not hurt to show the lines anyway for purposes of clarity.

8.3.4 Random Number Seeds

One of the features of the APEX model is the ability to conduct paired runs by controlling the random number seeds. In normal use, the model stochastically generates random profiles, selects diaries, and generates MP values from distributions. If multiple model runs are to be independent, then the main ***RandomSeed*** value in the *Control Options* input file should be set to zero or to different values. Setting this seed to zero means that the code uses the internal clock as the seed, so every run will be different (unless APEX was run at exactly the same time on the same day on different machines). If model runs are to be paired, then identical streams of random numbers are desired, in which case RandomSeed should be set to the same positive number in both runs.

Creating paired runs in which everything is identical will create results that are identical. The usual mode of operation is for sensitivity analysis, setting one specific difference between paired runs and then see how much effect it has on the results. In APEX version 4.5 and later, the seeds used for random number generation depend on three quantities: the number of random variables, the number of profiles simulated, and the initial random seed. To conduct paired runs, it is necessary that these three remain the same in both runs. While it is easy to keep the same number of persons and the same initial seed, sometimes one run will have extra terms in the micro-parameter definitions. This must be balanced out by creating "dummy" definitions to equalize the numbers, as discussed below.

Each combination of microenvironment, pollutant, and modeling variable is called a "micro-parameter," or MP for short. "The penetration factor for pollutant 2 in micro 1" is a single MP, regardless of the number of different distributions that it may have. Starting with APEX version 4.5, each MP is assigned a unique MP#. Gaps in MP# are permitted, although they result in extra seeds being generated that are never used. To avoid duplication of MP#, it is simplest to order them sequentially on the input file. For paired runs, the MP# should match for variables that are the same in both runs. It is important that the largest MP# be the same in both runs, as that determines the number of random seeds needed per person. Thus, if one run contains a variable that the other does not, then skip that MP# in the run without that variable, making sure that the variables common to both runs have the same MP#.

8.3.5 Source Strength Specification

As described in the sections on the MASSBAL and FACTORS methods (Sections 8.2.1 and 0), APEX allows two types of sources to be defined. ***ESource*** terms are emission sources expressed in units of micrograms per hour. ***CSource*** terms are sources expressed in concentration units of $\mu g/m^3$ (or ppm, ppb). ESource and CSource strengths can optionally be specified as the product of two or more values drawn from different distributions.

As an example, an ESource term representing emissions from gas stoves can be constructed as the product of three terms: a binary switch for Use/Non-use for each timestep, a duration of usage term, and an emission rate per minute of usage. One advantage of subdividing this term into three parts is that different rules for the time and area mappings and the resampling rate can be defined for each of the three terms.

Each source in APEX, meaning each MP of type ES or CS, may be assigned an optional source number. This is done by adding a line to the MP definition as illustrated in Exhibit 8-3:

```
Micro number   = 4
Parameter Type = ES
Source number = 2
Pollutant = 1
Block DType Season Area  C1  C2  C3   Shape      Par1   Par2  Par3 Par4  LTrunc UTrunc
ResampOut
1     1     1      1     1   1   1    Lognormal  1000   2     .     .    100    10000   Y
```

Exhibit 8-3. Use of Source Number in MP Definition

For clarity, the source number should appear right after the parameter type. It cannot appear earlier, nor can it appear after the header line starting with "Block." It only applies to types ES or CS and is not relevant for other parameter types. If the source number is omitted, then it is assumed to be zero, which is the catch-all category for additive source terms.

All MP of type ES or CS that share the same microenvironment number and source number are evaluated separately for all timesteps according to their own rules; then the results are multiplied together on a timestep basis. For example, if another MP has the description shown in Exhibit 8-4, then both this MP and the previous one are evaluated for each timestep, and then the results are multiplied together, timestep by timestep.

```
Micro number  =  4
Parameter Type=  ES
Source number = 2
Pollutant = 1
Hours - Block = 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 1 1 1 1
Block DType Season Area  C1  C2  C3   Shape      Par1   Par2  Par3 Par4  LTrunc UTrunc
ResampOut
1     1     1      1     1   1   1    Normal     10     10    .    .     0      60      Y
2     1     1      1     1   1   1    Point      45     .     .    .     .      .       .
```

Exhibit 8-4. Second MP Definition with Source Number 2

Since the first MP was not resampled, only one value is generated per person and that value remains constant over the simulation. The second MP is resampled every hour (although the timesteps using the Point distribution will all return the same value).

It is possible to have a large number of terms sharing the same microenvironment and source numbers, in which case all terms are evaluated separately and the timestep results are multiplied. Note that all terms sharing the same microenvironment and nonzero source numbers must also share the same parameter type; one cannot mix ES and CS types since it does not make sense to multiply them together. APEX will generate an error message if this is attempted. It is possible for some sources to be ES while others are CS, even in the same microenvironment, as long as they are not assigned the same nonzero source numbers.

Once the product is evaluated, the result is treated exactly the same as any additive sources for that microenvironment. All ESource terms are added together timestep by timestep, whether each resulted from a product or was a separate term by itself. The same applies to all CSource terms in the same microenvironment. In effect, there is no change at all to either the MASSBAL or the FACTORS equations; there is simply a change in how the ESource and CSource terms are determined.

When defining product terms, a line must be added to the *Control Options* input file indicating the largest source number used in the run for each pollutant. This line has the keyword ***#SOURCES*** and it might appear as in the pollutant parameters section of the *Control Options* file as shown in Exhibit 8-5.

```
Pollutant   = CO

DoDose      = YES
InputUnits  = ug/m3
OutputUnits = ug/m3
#Sources    = 3
```

Exhibit 8-5. Use of #sources Setting in the Pollutant Parameters section of the *Control Options* File

This value reported here (3 in this case) is echoed to the *Log* file after "#microenvironments" and it is called "#Enumerated sources." Actually, like #microenvironments, this value is only used to allocate array space, and APEX will run correctly as long as #sources is large enough to accommodate the source numbers used in the *Microenvironment Descriptions* input file. If #sources is larger than is needed, no error occurs although job execution will be slightly less efficient. This line can only be omitted from the *Control Options* file if no source numbers are assigned to any MP.

Not all sources need be expressed as products. These sources can either be left without source numbers, or can explicitly have source number of zero, or may even have a positive source number. If there is only one MP with a given microenvironment number and source number, then it essentially constitutes a product with only one term. If only source numbers = 0 are used (i.e., only additive terms are defined), then #Sources may be set to 0 in the *Control Options* file or omitted.

8.3.6 Specification of Distribution Data

The number of distributions that the user must supply for each MP is the product of these numbers:

$$N_{Distributions} = N_{blocks} \times N_{daytypes} \times N_{seasons} \times N_{areas} \times N_{C1} \times N_{C2} \times N_{C3} \quad (8\text{-}16)$$

Here N_{C1}-N_{C3} are the number of possible responses for conditional variables #1-#3. If a conditional variable is not used, then it has one possible value (its index is always equal to 1). The number of possible responses varies from one conditional variable to another. All preset Conditional Variables have two values. Population Category has as many values as population files are defined in the *Control Options* file. For other Conditional Variables, the number is determined from the definition supplied by the user in the *Profile Functions* input file.

Each distribution occupies one line in the *Microenvironment Descriptions* input file. Every combination of the seven indices must have a distribution defined, even if that specific combination never occurs during the simulation. For example, if a winter season exists then it must have a set of distributions defined, even if the simulation period only covers the summer. The reason is that the distributions are stored in an array with room for all possible combinations of index values, and this array is checked once for gaps (missing data) before the first profile is begun. This is more efficient than checking to see if a distribution exists every time it is called, as a model run can contain millions or even billions of calls to distributions. The price is that distributions that ultimately are never called must still exist in the array.

The user does not need to number the distributions; this is done internally by the program. The distributions are assigned index numbers in standard Fortran order (the block index changes fastest, and Condition #3 changes slowest). Thus, distribution#1 is for (1,1,1,1,1,1,1), and if there is more than one block then distribution#2 is for (2,1,1,1,1,1,1), etc. In the input file, the distributions can appear in any order; the standard order, however, is preferred for consistency.

Each line describing a distribution contains the following information. First, the seven indices are listed—block, day type, season, area, c1, c2, and c3. The seven indices must all appear explicitly in the set order. Any superfluous indices must be given a value of one. Thus, if a conditional variable such as c3 is not used, then the c3 index number is 1 for all distributions for that MP.

After the seven indices, the distribution definition is given in standard APEX distribution format (Section 3.1). Any of the APEX distributions can be used; the available shapes and their parameters are given in Table 3.1.

CHAPTER 9. CALCULATING EXPOSURES

A description of how the microenvironment concentrations and other information are used by APEX to estimate exposure (see step 6 in Figure 2.3), and how the model then presents and summarizes the exposure results is illustrated below. Exposures are calculated in **ExposureDoseModule:Exposure**.

9.1 Estimating Exposure

For inhaled pollutants such as the ones APEX models, exposure is defined as the time-integrated concentration of the pollutant in the breathing zone of a person. We refer to the concentration to which a person is exposed as the *exposure concentration*. This concentration is assumed to be approximately spatially uniform within each microenvironment and also approximately temporally uniform for the duration of any one activity event (at most one hour). A time series of exposure concentrations can be constructed by following the sequence of microenvironments and locations (e.g. "Home," "Work," or "Other") visited according to the composite activity diary assembled for the target profile.

As an example, assume the activity diary indicates that the first 40 minutes of the simulation are spent in microenvironment #3 ("Home"), the next 20 minutes in microenvironment #2 ("Other"), and the next 60 minutes in microenvironment #5 ("Work"), and so on. Then the exposure time series has its first 40 minutes at the concentration for hour 1 in microenvironment #3 in location "Home," the next 20 minutes at the concentration for hour 1 in microenvironment #2 for location "Other," and the next 60 minutes at the concentration for hour 2 in microenvironment #5 for location "Work," and so on. The exposure itself depends on location, but depends neither on the activities that the person is performing nor on any personal physiological properties.

The user may select the units for reporting the exposure output—either ppm, ppb, or $\mu g/m^3$ may be chosen. This applies to both concentration and to exposure. It also applies to the levels used as cutpoints in the various exposure tables. The parameter ***OutputUnits*** in the *Control Options* file controls this; if set to PPM, then parts per million (ppm) are used; if set to PPB, then parts per billion (ppb) are used; otherwise, micrograms per cubic meter ($\mu g/m^3$) are used.

APEX calculates exposure as a time series of exposure concentrations that a simulated individual experiences during the simulation period. APEX calculates the exposure by identifying the concentrations in the microenvironments visited by the person according to the composite activity diary. In this manner, a time-series of event exposures are found. The timestep exposure concentration at any clock hour during the simulation period is then calculated using the following equation:

$$C_i = \frac{\sum_{j=1}^{N} \left(C_{timestep,j} t_j \right)}{T} \qquad (9\text{-}1)$$

where:

C_i	=	Exposure concentration at hour i of the simulation period (µg/m^3, ppm, or ppb)
N	=	Number of events (i.e., microenvironments visited) in timestep i of the simulation period.
$C_{timestep\ (j)}$	=	Timestep concentration in microenvironment j (µg/m^3, ppm, or ppb)
$t_{(j)}$	=	Time spent in microenvironment j (minutes)
T	=	Length of timestep (minutes)

From the timestep exposures, APEX calculates time series of 1-hour, 8-hour and daily average exposures that a simulated individual would experience during the simulation period. APEX then statistically summarizes and tabulates the timestep, hourly, 8-hour, and daily exposures. Note that if the APEX timestep is greater than an hour, the 1-hour and 8-hour exposures are not calculated and the corresponding tables are not produced. Exposures are calculated independently for all pollutants in the simulation.

9.2 Exposure Summary Statistics

The exposure time series provides a wealth of detail on the exposure experienced by each profile. However, this is difficult to analyze since the number of events differs for each profile, and even the number of events on any given calendar day is unpredictable. Furthermore, a model run may consist of thousands of profiles, so it is not practical to retain the exposure time series for all profiles in memory. For output tables and for analysis, summaries of the exposure time series are required.

Exposure summary statistics are calculated for each pollutant in the simulation and are written to pollutant-specific output files: the *Microenvironment Summary* file, the *Microenvironment Results* file, and the *Tables* file. Each exposure metric in this section is thus calculated for each pollutant, and the *Control Options* file keywords listed below are pollutant-specific (see *Volume I*).

The first step in summarizing exposure is to calculate the time series of event-level, timestep-level, and hourly average exposures for the profile. The event exposures for all pollutants are written to the *Events* output file (if the *Control Options* file setting ***EventsOut*** = YES), while the timestep values are written to the *Timestep* output file (if the *Control Options* file setting ***TimestepOut*** = YES). Daily averages and maxima for each pollutant can be written to the *Daily* output file as well. The timestep values are used to derive most of the other exposure summary statistics.

There are two exposure time series reported on an hourly basis, namely the series of 1-hour averages and running 8-hour averages. For the first seven hours of the simulation, the nominal 8-hour average is actually taken over fewer than eight values; otherwise it is always the current hour and the previous seven hours. These hours may cross day, month, or even yearly boundaries. The 1-hour and 8-hour averages are not calculated for timesteps greater than 1 hour (i.e., when the *Control Options* file variable TimestepsPerDay is less than 24).

There are three exposure statistics calculated on a daily basis. The first is the daily average exposure or ***DAvgExp***, which is the arithmetic mean of all the timestep exposures that fall on a

given calendar day. All days in APEX contain 24 hours (or an equivalent number of timesteps); the effects of Daylight Savings Time are removed from the model. The daily average exposure is then binned into levels according to the cutpoints provided by the user in the *Control Options* file. For example, the input line:

```
DAvgExp  = 2, 5, 8, 12, 20
```

indicates that the first bin for daily average exposure extends from 0.0 to 2.0 (ppm, ppb, or $\mu g/m^3$), and the second bin from 2.0 to 5.0, etc. The final or sixth bin in this example contains all values over 20.0. The number of days at each level is recorded for each profile.

The second daily exposure statistic is the maximum timestep for each calendar day, or ***DMTSExp*** in the code (for Daily Maximum Timestep Exposure). It is the highest of the timestep values on each day. Like DAvgExp, these are also converted to bins or levels using cutpoints from the *Control Options* input file. The number of days at each level is recorded for each profile.

The third daily exposure statistic is the maximum 1-hour average for each calendar day, or ***DM1HExp*** in the code (for Daily Maximum 1-Hour Exposure). It is the highest of the 24 hourly values on each day. Like DAvgExp, these are also converted to bins or levels using cutpoints from the *Control Options* input file. The number of days at each level is recorded for each profile.

The final daily summary statistic for exposure is ***DM8HExp*** which is the Daily Maximum 8-Hour Exposure. It is the largest of the 24 8-hour running averages for each calendar day. Like the other daily statistics, this is also binned into levels and recorded for each profile.

The average exposure over the entire simulation period is also calculated. As for the daily summary statistics, the average exposure is binned using the cutpoints designated in the *Control Options* input file by ***SAvgExp***.

In addition to the exposure statistics described above, there are a few others that are derived directly from the event-based exposure time series. The variable ***TimeExp*** represents the number of minutes spent in each bin in each micro, based not on hourly averages but on the original event exposures. Again, the cutpoints for the TimeExp bins are provided by the user in the *Control Options* input file.

The user can also set a threshold exposure level in the *Control Options* file. This variable is called ***AlertThresh***. If the exposure time series exceeds AlertThresh for any event, then the following three things occur. First, the count of high exposure events for this profile is incremented by one. Second, the total duration over the threshold exposure is incremented by the duration of this event. Third, the exposure is checked to see if it exceeds the maximum exposure previously experienced by this profile. These results are reported in the *Log* file for the model run for each profile that exceeds the threshold.

9.3 Exposure Summary Tables

APEX can write out over a hundred different exposure summary tables for the statistics described above. The content and interpretation of these tables (including examples) are covered in *Volume I*. There are 11 basic types of tables. All tables are written to the *Tables* files (one file for each pollutant) and optionally the *Log* file in **ExposureDoseModule:Output:**

1. Minutes in each Exposure Interval by Microenvironment (***TimeExp***)
2. Minutes at or above each Exposure Level by Microenvironment (***TimeExp***)
3. Person-Days at or above each Daily Maximum 1-Hour Exposure Level (***DM1HExp***)
4. Person-Days at or above each Daily Maximum 8-Hour Exposure Level (***DM8HExp***)
5. Person-days at or above each Daily Maximum Timestep Exposure Level (***DMTSExp***)
6. Number of Simulated Persons with Multiple Exposures at or above each Daily Maximum 1-Hour Exposure Level (***DM1HExp***)
7. Number of Simulated Persons with Multiple Exposures at or above each Daily Maximum 8-Hour Exposure Level (***DM8HExp***)
8. Number of Simulated Persons with Multiple Exposures at or above each Daily Maximum Timestep Exposure Level (***DMTSExp***)
9. Number of Simulated Persons with Multiple Exceedances (in the Simulation) of the Threshold Timestep Exposure Levels (***TSExp***).
10. Person-Days at or above each Daily Average Exposure Level (***DAvgExp***)
11. Persons at or above each Overall Average Exposure Level (***SAvgExp***)

For determining whether a variable is "at or above" a bin boundary, the variable is rounded to nine decimal places before the comparison is made. Due to roundoff error, occasionally a value which should algebraically be exactly at a bin boundary would test as being below it, if this rounding were not performed.

The levels written to each table are given by the *Control Options* file keywords in parentheses. The definition of terms in the titles of these tables are as follows:

- **Person-day**: A single simulated day for one simulated individual. A 100-day simulation of 10 persons contains 1000 person-days, as does a single-day simulation of 1000 persons. APEX in general counts exposures in Person-days. Multiple event-level exposures at the same exposure cutpoint level during a single day are counted as one Person-day of exposure.
- **Multiple Exposures**: Multiple exposures for the same person on different simulation days. This term refers to multiple person-days of exposure in a simulation corresponding to the same profile. Multiple exposures (at a single level) during a single day are counted as one Person-day of exposure.
- **Multiple Exceedances**: Multiple exposures for the same person on different timesteps of the simulation. Multiple exposures during the same day (at a single level) do count as different exceedances.

Table types 1, 2, 10, and 11 are generated only once for the entire population. Table types 3 to 9 are generated for seven population subgroups, under three exertion levels. Tables may be omitted if the subgroup contains no simulated persons.

The seven population subgroups are listed below.

1. All Persons. The table statistics are based on the entire population.
2. Children. The table statistics are based on the population of children, as defined by the age range given by the *Control Options* file settings ***CHILDMIN*** and ***CHILDMAX***.
3. Active Persons. The table statistics are based on the population of people having a median ***PAI*** over the whole simulation period that exceeds the value designated by the *Control Options* file setting ***ACTIVEPAI***.
4. Active Children. The table statistics are based on the population of active children, as determined by the *Control Options* file settings ***CHILDMIN***, ***CHILDMAX***, and ***ACTIVEPAI***.
5. Ill Persons. The table statistics are based on the population of ill people. The population is determined by the probabilities given in the *Prevalence* file. This population is only considered if the input variable ***DISEASE*** is set in the *Control Options* file.
6. Ill Children. The table statistics are based on the population of ill people. The population is determined by the probabilities given in the *Prevalence* file and the *Control Options* file settings ***CHILDMIN*** and ***CHILDMAX***. This population is only considered if the input setting ***DISEASE*** is set in the *Control Options* file.
7. Employed Persons. The table statistics are based on the population of all employed people.

The three exertion levels are listed below.

1. All Exertion Conditions. The table statistics are based on exposures experienced by the population subgroup under any ventilation conditions.
2. Moderate Exertion. The table statistics are based on exposures experienced by the population subgroup only during periods in which their average ***EVR*** is in the "moderate" range. The period of time during which ***EVR*** is averaged is either 1 hour or 8 hours, based on the table being generated. The "moderate" ***EVR*** ranges are defined by the *Control Options* file settings ***MODEVR1*** and ***HEAVYEVR1*** (for 1-hour exposures) and ***MODEVR8*** and ***HEAVYEVR8*** (for 8-hour exposures). An individual's ***EVR*** is in the moderate range if it is greater than or equal to the ***MODEVR#*** setting and less than the ***HEAVYEVR#*** setting for the exposure period.
3. Heavy Exertion. The table statistics are based on exposures experienced by the population subgroup only during periods in which their average ***EVR*** is in the "heavy" range. The period of time during which ***EVR*** is averaged is either 1 hour or 8 hours, based on the table being generated. The "heavy" ***EVR*** ranges are defined by the *Control Options* file settings ***HEAVYEVR1*** (for 1-hour exposures) and ***HEAVYEVR8*** (for 8-hour exposures). An individual's ***EVR*** is in the heavy range if it is greater than or equal to the ***HEAVYEVR#*** setting for the exposure period.

The exertion level statistics are calculated in the following manner: For each day in the simulation, the mean EVR level during every timestep, 1-hour, and running 8-hour time period is calculated. If the mean timestep, 1-hour, or 8-hour EVR is higher, then the EVR levels indicated (in the *Control Options* file) for moderate or heavy exertion and the exposure during the same time period is compared with the levels given in the *Control Options* file; the corresponding statistics and tables are updated if the exposure exceeds the level.

NOTE: Many of the tables can include statistics at the 0.0 exposure level if it is indicated on the *Control Options* file. This can be used to obtain some useful statistics (such as total final study

area of the subpopulation population), since all persons will have exposures equal to or exceeding 0.0 level exposures on all person-days. However, use caution in examining these "0.0" level statistics in the case of the exertion-level tables. If a simulated person has no timestep, 1-hour, or 8-hour periods at Moderate or Heavy EVR, then they will NOT have an exposure in that table for the 0.0 level, and the 0.0 level statistics will not correspond to the entire subpopulation.

CHAPTER 10. CALCULATING DOSE

APEX contains algorithms for estimating pollutant doses. The term "dose" refers to some measure of the amount of pollutant in the body of the target person. The situation is not as clear as for exposure since there are numerous specific definitions of dose that are used in various contexts. For an airborne pollutant, dose could refer to the amount inhaled, the amount currently in the lungs, the amount crossing from the lungs to the body, the total amount in the body, the total in some specific target organ, or a number of other things. Dose may be more useful or accurate than exposure for evaluating the effects of air pollutants because it accounts better for differences in pollutant uptake resulting from: (1) the variation in physiology and activities across populations, and (2) the variation in physiological responses to activities within an individual. In APEX, dose is generally defined at the amount inhaled. APEX contains a special algorithm for the pollutant CO, in which the model calculates the percent of carboxyhemoglobin (%COHb) in the blood. Carboxyhemoglobin is hemoglobin that has carbon monoxide instead of the normal oxygen bound to it. In addition, APEX contains algorithms to estimate the deposited lung dose in the case of particulate matter (PM). When the APEX model is extended to other pollutants, it will likely require the development of specific dose algorithms for each new pollutant, or at least for each class of pollutant.

The simple algorithm for calculating inhaled dose is discussed in the next section, followed by the algorithms for the %COHb calculation and for PM. The final topic relating to dose is the explanation of the summary statistics for reporting dose. Doses are calculated in **ExposureDoseModule:Dose.**

10.1 Inhaled Dose Calculation

Currently, for all pollutants other than CO, APEX calculates inhaled dose. Inhaled dose is simply the amount of pollutant inhaled over the course of a specified time period. Inhaled dose for each timestep is calculated as:

$$D_i = V_e C_i \qquad (10\text{-}1)$$

Where:

C_i = Timestep exposure concentration at timestep hour i of the simulation period ($\mu g/m^3$, ppm, or ppb)
V_e = Expired ventilation rate (ml/min)

The calculation of the expired ventilation rate V_e is discussed in Section 7.3. The exposure C_i is discussed in Section 9.1. Although there may be a very small difference in inspired and expired ventilation rates, this is not considered in APEX. (The expired rate is the one usually measured, so APEX uses that, even for inhalation estimates.) From the timestep doses, APEX calculates time series of 1-hour, 8-hour, and daily average dose that a simulated individual would experience during the simulation period. APEX then statistically summarizes and tabulates the hourly, 8-hour, and daily doses. Note that if the APEX timestep is greater than an hour, the 1-hour and 8-hour dose are not calculated, and the corresponding tables are not produced. Doses are calculated independently for all pollutants in the simulation.

10.2 Carboxyhemoglobin (COHb) Calculation

The calculation of CO dose is complex. It starts with the exposure time series, which indicates the pollutant concentration in the inhaled air at each moment in time. It also requires the ventilation rate which is activity dependent; knowledge of a number of physiological parameters; and the current %COHb level, which depends on recent history. The discussion of physiological parameters is in the chapter on personal profiles. In APEX, dose is the delivered dose of CO as measured by the biotransformation product carboxyhemoglobin (COHb)—specifically %COHb in the blood. For example, a %COHb of 2% means that 2% of the hemoglobin molecules in the blood are bound to CO and therefore cannot bind to and transport oxygen.

The %COHb calculation in APEX uses the time series for exposure to CO and the time series for alveolar ventilation rate, V_a as inputs (among other factors). The dose calculation is based on the solution to the non-linear Coburn, Forster, Kane (CFK) equation, as detailed in Johnson (2002). As pointed out by that report, the CFK equation does not have an explicit solution, so an iterative solution or approximation is needed to calculate the %COHb. An iterative solution, however, was determined to be unsuitable because of the model execution time necessary (i.e., a typical model run of one calendar year represents roughly 14,000 events per person and several thousand people, or tens of millions of diary events). Therefore, the CFK equation is solved using a modified Taylor's series method in which the event duration is restricted in time (if necessary) to ensure convergence with only a few terms. This method avoids the dangers of non-convergence that arise in some other methods.

As the mathematical derivation in the above report is very detailed, only the main results are presented here. First, it should be noted that the literature discusses two forms of the CFK equation, namely a linear and a non-linear form. The linear form itself is an approximation that allows an explicit solution, but is not accurate under all conditions. The non-linear form is considered to be more correct and is the one being discussed here.

Restricting %COHb(t) to between 0 and 100 (percent), the CFK equation takes the form of the following differential equation:

$$\%COHb'(t) = C_0 - C_1 \times \frac{\%COHb(t)}{100 - \%COHb(t)} \tag{10-2}$$

where C_0 and C_1 are constants over the duration of one event that depend on physical and physiological parameters, including V_a and the CO exposure. C_0 is given by:

$$C_0 = \frac{Endgn + P_{CO}/B}{RHB_0 + Blood} \tag{10-3}$$

where Endgn and Blood are physiological profile variables (Section 5.3). P_{CO} is the partial pressure of CO (torr):

$$P_{CO} = P_{gases} \times Exposure \times 10^{-6} \tag{10-4}$$

where Exposure is the event CO exposure concentration and P_{gases} is the partial pressure (torr) of dry gases at the study altitude (EPA, 1978):

$$P_{gases} = 760e^{-0.0000386\ Altitude} - 47 \tag{10-5}$$

where altitude is in feet. The variable RHB_0, is the total reduced blood hemoglobin level (ml O_2 or ml CO per ml blood), adjusted for altitude (as in EPA, 1978):

$$RHB_0 = 1.39 \times 0.01 \times 0.995 \times Hmglb\left(1 + \frac{2.76e^{0.0001429\ \ Altitude}}{100}\right) \tag{10-6}$$

where Hmglb is the profile variable for hemoglobin density. The factor, B, is given by:

$$B = \frac{1}{DIFF} + \frac{P_{gases}}{V_a} \tag{10-7}$$

Va is the event alveolar ventilation (see Section 7.3), and DIFF is lung CO diffusivity (ml/min/torr) adjusted for ventilation rate :

$$DIFF = Diff + 0.000845V_a - 5.7 \tag{10-8}$$

where Diff is the CO diffusivity profile variable (which corresponds to a baseline ventilation).

The constant C_1 is given by:

$$C_1 = \frac{1 + 0.32 \times PO_2}{69.76 \times RHB_0 \times Blood} \tag{10-9}$$

where $P0_2$ is the partial pressure of oxygen in the lungs (torr):

$$PO_2 = 0.209P_{gases} - 49 \tag{10-10}$$

See Johnson (2002) for a detailed derivation of the above equations for C_1 and C_2.

Time zero represents the start of the current event. The concentration *%COHb(0)* (at time zero) is assumed to be known. The first derivative, *%COHb'(0)*, can easily be found from the above equation. The solution *%COHb(t)* is a smoothly-varying function of time, without sudden discontinuities or changes in slope. It therefore can be expanded in a Taylor's series about $t = 0$, which should converge fairly rapidly. One simplification is to rescale the time variable to the unitless parameter *z*:

$$z = \frac{(C_0 + C_1) \times t}{(100 \times D_0 \times D_0)} \tag{10-11}$$

where

$$D_0 = \frac{1 - \%COHb(0)}{100} \tag{10-12}$$

The Taylor's series up to the fourth order term is:

$$T_4(z) = \%COHb(0) + 100 \times D_0 \times Dz - \frac{100 \times A_1 \times D_0 \times Dz^2}{2} +$$

$$\frac{100 \times A_1 \times D_0 \times D \times (A_1 - 2D) \times z^3}{6} +$$

$$\frac{100 \times A_1 \times D_0 \times D \times \left(A_1^{\,2} - 8DA_1 + 6D^2\right) \times z^4}{24} \qquad (10\text{-}13)$$

where

$$A_1 = \frac{C_1}{C_0 + C_1} \qquad (10\text{-}14)$$

$$D = D_o - A_1 \qquad (10\text{-}15)$$

For typical values for the constants C_0, C_1, and %COHb(0), convergence occurs for z<1. For z values below this limit but still close to one, the convergence is slow, so the terms beyond fourth order would be needed if high accuracy were desired. It is found that z<1 generally corresponds to %COHb(0) values below 40 to 50% for 1-hour events.

The z value is proportional to t. In APEX, an event can have a duration of no more than one hour. For 1-hour events, there are conditions where z is close to one and convergence of the Taylor's series may require more than four terms. However, it is not necessary to evaluate the entire hour in one step. By dividing the event into shorter subevents, each will have a smaller z value. For example, if a 1-hour activity diary event has initial conditions that correspond to z=0.9, then by dividing this into three 20-minute subevents, each will have a z value around 0.3. Actually, the first subevent will have exactly z=0.3. The others will be slightly different as the initial conditions for those subevents will have changed slightly. In practice, a dimensionless subevent duration is used, and all the subevents have an equivalent length. An accumulated dose is calculated as the running sum of the average %COHb level over the subevents. At the end of all subevents, the average dose is this accumulated dose divided by the number of subevents while the final value of the dose is simply the final value of COHb itself. These values are saved and the next diary event is processed.

10.3 Calculating PM Dose

In APEX, PM dose is modeled as the mass of PM depositing in the entire respiratory system, including the extrathoracic regions (mouth, nose, and oropharynx) and the lungs. The PM dose algorithm was developed from the empirical lung deposition equations of the International Commission on Radiological Protection's Human Respiratory Tract Model for Radiological Protection (ICRP 1994).

The algorithm calculates the deposition fraction (f) in each of nine filters (see Figure 10.1), and determines the PM mass deposited in each for each event. The nine filters correspond to the following regions of the respiratory system: 1. Nose (inhalation, f_N), 2. Oropharynx (inhalation, f_O), 3. Tracheobronchi (inhalation, $f_{TB,i}$), 4. Bronchioles (inhalation, f_B), 5. Alveoli (f_A),

6. Bronchioles (exhalation, f_B), 7. Tracheobronchi (exhalation, $f_{TB,e}$), 8. Oropharynx (exhalation, f_O), and 9. Nose (exhalation, f_N). The f-values determine the fraction of the particle mass entering the region that deposits. Only the fraction in the Tracheobronchial filters differs for inhalation versus exhalation; for the other filters, the value of the fraction is the same in both cases. Deposition via both aerodynamic and thermodynamic (diffusive) mechanisms are estimated for each filter.

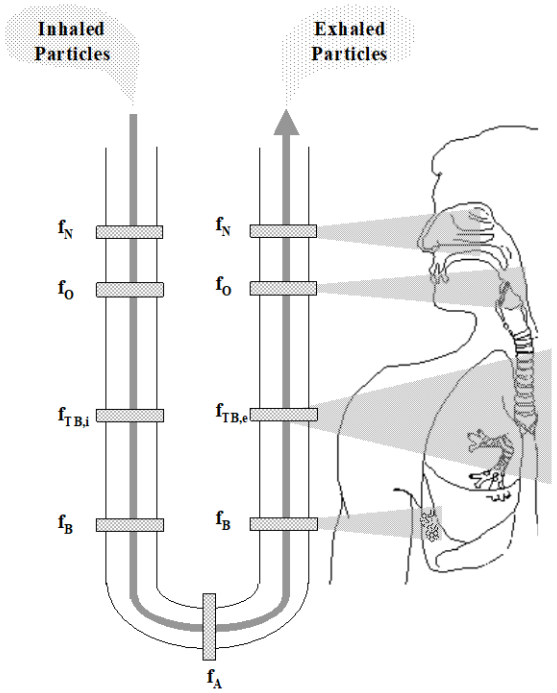

Figure 10.1. Structure of the ICRP Deposition Model

The fractions, f , and the resulting deposited doses are calculated for each particle size considered. The size-specific mass are summed to get the mass deposited in each filter. The deposited filter masses are then combined into four deposited masses corresponding to total PM dose in the entire respiratory system and in the extrathoracic (ET, nose + oropharynx), tracheobronchial (TB), and pulmonary (Pulm, bronchiole + alveoli) regions.

10.3.1 Particle Sizes, Inhalability, and Diffusion Coefficient

The particle size used by the ICRP algorithm is the aerodynamic diameter. Aerodynamic diameter (d_{ae}) is the property of particles measured by the majority of particle samplers, and the EPA designations of $PM_{2.5}$ and PM_{10} are based on aerodynamic diameter. The aerodynamic diameter is provided to the model in the *Control Options* file using the pollutant-specific keyword *Size* (see *Volume I*).

Particle inhalability is defined as the fraction of the ambient particles that are inhaled. This fraction is calculated as:

$$f_{inh,i} = 1 - 0.5(1 - [7.6 \times 10^{-4} d_{ae,i}^2 + 1]^{-1}) \quad (10.16)$$

The particle diffusion coefficient D_i is also required to calculate the deposition fractions. D_i is calculated using Equation 10.17:

$$D_i = \frac{C_C(d_{th_i})kT}{3\pi\mu d_{th,i}} \qquad (\text{cm}^2\ \text{s}^{-1}) \qquad (10.17)$$

where

k is the Boltzmann's constant, 1.38 x 10^{-16} g cm2 s^{-1} K^{-1}
$d_{th,i}$ is an equivalent thermodynamic diameter of particle size i
T is absolute temperature (Kelvin)
μ is the absolute viscosity of air (1.82 X 10^{-4} g cm^{-1} s^{-1})
and C_C is the Cunningham correction factor, given in Equation 10.18:

$$C_C(d_{th,i}) = 1 + \frac{2\lambda}{d_{th,i}}[1.257 + 0.4\ \ exp(-\frac{1.1 * d_{th,i}}{2\lambda})] \qquad (10.18)$$

where λ is the mean free path of air, 6.5 X 10^{-6} cm.

The thermodynamic diameter is calculated from the aerodynamic diameter:

$$d_{th,i} = \sqrt{\frac{1.5}{\rho}\frac{C_c(d_{ae,i})}{C_c(d_{th,i})}} \qquad (\text{microns}) \qquad (10.19)$$

where ρ is the particle density in g/cm^3. The value of $d_{th,i}$ is found by recursively solving this equation using an initial guess of $d_{ae,i}(\rho)^{-1/2}$.

10.3.1.1 The ICRP Deposition Equations

As described above, the ICRP model considers the respiratory system as a system of nine filters corresponding to the particles deposited in the nose (N), oropharynx (O), trachea/bronchi (TB), bronchioles (B), and alveoli (A) during inhalation and exhalation. The ICRP equations considers fractions for both aerodynamic (ae) and thermodynamic (th) deposition processes in each region. These fractions in each filter j are exponential functions of 3 empirical coefficients a, R, and p, as shown in Equations 10.20 and 10.21.

Aerodynamic Deposition Fraction:

$$f_{ae,j} = 1 - exp(-a_{ae,j}R_{ae,j}^{p_{ae,j}}) \qquad (10.20)$$

Thermodynamic Deposition Fraction:

$$f_{th,j} = 1 - exp(-a_{th,j}R_{th,j}^{p_{th,j}}) \qquad (10.21)$$

There are coefficients a, R, and p for each filter for both aerodynamic and thermodynamic mechanisms. The coefficients may be constant or may be functions of parameters describing the respiratory system physiology and activity level of the individual being modeled. The coefficients, which may also be a function of the mode of breathing (nasal or oral), are given in Table 10.1.

Table 10.1 The Values of a, R, and P for Each Filter for Oral and Nasal Breathing

	Nose (N)	Oropharynx (O)	Tracheobronchial (TB)	Bronchioles (B)	Alveoli (A)
Aerodynamic Deposition					
a	0.0003	nose breathing: 0.000055 oral breathing: 0.00011	inhalation: 0.00000408 exhalation: 0.00000204	0.1147	$0.146S_3^{0.98}$
R	$d^2\dot{V}_n(S_1)^3$	nose breathing: $d^2\dot{V}_n(S_1)^3$ oral breathing: $d^2(\dot{V})^{0.6}(V_T)^{-0.2}(S_1)^{1.4}$	$d^2\dot{V}(S_1)^{2.3}$	$0.056 + t_B^{1.5}d^{t_B^{-0.25}}$	d^2t_A
P	1	nose breathing: 1.17 oral breathing: 1.4	1.152	1.173	0.6495
Thermodynamic Deposition					
a	18	nose breathing: 15.1 oral breathing: 9	$22.02(S_1)^{1.24} \times [1-100e^{-[\log_{10}(100+\frac{10}{d^{0.9}})]^2}]$	$-76.8+167S_2^{0.65}$	$170+103S_3^{2.13}$
R	$D\dot{V}_nS_1^{-0.25}$	nose breathing: $D\dot{V}_nS_1^{-0.25}$ oral breathing: $d^2(\dot{V})^{0.6}(V_T)^{-0.2}(S_1)^{1.4}$	Dt_{TB}	Dt_B	Dt_A
p	0.5	nose breathing: 0.538 oral breathing: 0.5	0.6391	0.5676	0.6101

As indicated in the table, the coefficients a, R, and P are functions of a number of physiological variables, including lung volumes, inspiratory flow rates, and residence times. These are described below.

10.3.1.2 Lung Volumes and Age Scaling Factors

The ICRP publication provides reference values for the required lung volumes as a function of subject age and gender. The volumes are: (1) the dead spaces for the entire lung (V_d), and for the ET, TB, and B regions ($V_{d,ET}$, $V_{d,TB}$, $V_{d,B}$), and (2) the functional residual capacity (FRC). The ICRP model also makes use of three scaling factors to adjust some of the a and R model coefficients as a function of age. Physiologically, these factors represent ratios of specific airways of a modeled person to those of a reference adult male. S_1 gives the ratio for the trachea, S_2 for the ninth airway generation, and S_3 for the 16th airway generation. Both the volumes and

the scaling factors are modeled in APEX based on the reference values, using an equation of the form:

$$P = Ah^2+Bh+C \quad (10.22)$$

The variable, P, is the volume or scaling factor of interest and h is the individual's height (in cm). The values of the A, B, and C coefficients for the different parameters for males and females were calculated by fitting eq 10.22 to the ICRP parameters; they are given in Table 10.2.

Table 10.2 Coefficients for the Lung Volumes and Scaling Factors

	A	B	C
Volumes			
Male			
FRC	0.0002	-0.0279	1.0353
V_d	0.0078	-0.7135	29.316
$V_{d,ET}$	0.0031	-0.3175	10.907
$V_{d,TB}$	0.0026	-0.2392	9.7143
$V_{d,B}$	0.0023	-0.2119	11.375
Female			
FRC	0.0002	-0.0265	0.96
V_d	0.0079	-0.7403	30.381
$V_{d,ET}$	0.0029	-0.2861	9.5306
$V_{d,TB}$	0.0022	-0.1523	5.9635
$V_{d,B}$	0.0029	-0.3225	16.098
Scaling Factors			
Male			
S_1	1.1354E-04	-4.0700E-02	4.6711
S_2	2.3020E-05	-1.1168E-02	2.2567
S_3	7.8360E-05	-3.2100E-02	4.2381
Female			
S_1	1.2800E-04	-4.3542E-02	4.7975
S_2	2.3200E-05	1.1225E-02	2.2597
S_3	8.0500E-05	-3.2543E-02	4.2585

10.3.1.3 Tidal Volume and Activity Level

Tidal volumes (V_t) were calculated as a function of age, gender, and activity level. The starting point for the V_t calculations are the ICRP reference values for men and women of reference ages for four different activity levels: sleep, sitting, light exercise, and heavy exercise. Other ages are interpolated from the values at the reference ages. After age 30, the values are assumed to be constant with increasing age. The correct V_t for each event is determined as a function of age, gender, and activity level. Activity level is determined by the event MET values for the individual being studied. Activity level was based on normalized MET (M), defined in Equation 10.23:

$$M = \frac{MET - 1}{MET\ max - 1} \quad (10.23)$$

METmax is the maximum obtainable MET value for the person. If M was less than or equal to 0, then activity level was assumed to be "sleep." If M was greater than 0, then the activity level was assigned as follows:

$M \leq 0.333$: activity level = sitting
$0.333 < M \leq 0.667$: activity level = light exercise
$M > 0.667$: activity level = heavy exercise

10.3.1.4 Inspiratory Ventilation

Several of the model parameters are a function of the inspiratory ventilation. This flow rate was calculated as:

$$\dot{V} = 2V_e \quad \text{(ml/s)} \tag{10.24}$$

where V_e is the exhaled ventilation rate.

10.3.1.5 Residence Times

The residence times in the lungs are required to calculate deposition via diffusive mechanisms. Residence times are a function of the flow rate, the dead spaces, FRC, and V_t. The residence times (in seconds) in the tracheobronchial, bronchial, and alveolar regions are given by:

$$t_{TB} = \frac{V_{d,tTB}}{\dot{V}}\left(1 + \frac{0.5V_t}{FRC}\right) \quad \text{(s)} \tag{10.25}$$

$$t_B = \frac{V_{d,B}}{\dot{V}}\left(1 + \frac{0.5V_t}{FRC}\right) \tag{10.26}$$

$$t_A = \frac{V_t - V_{d,ET} - \left[V_{d,TB} + V_{d,B}\right]\left(1 + \frac{V_t}{FRC}\right)}{\dot{V}} \tag{10.27}$$

10.3.1.6 Final Deposition Fractions and Deposited Masses

The total deposition efficiency for filter j is:

$$f_j = \sqrt{f_{ae,j}^2 + f_{th,j}^2} \tag{10.28}$$

The total mass deposited in each filter j is:

$$m_j = (m_{inh} - \sum_{k=1}^{j-1} m_k) f_j \tag{10.29}$$

The variable m_{inh} is the total inhaled PM mass that is calculated for each event, Equation 10.30.

$$m_{inh} = f_{inh} * VE * minutes * C_{PM} \quad (10.30)$$

C_{PM} is the microenvironmental PM concentration for the event, VE is the exhaled ventilation, and minutes is the event duration.

The mass deposited in each region of the respiratory tract can be calculated by summing the mass deposited by the inhalation and exhalation filters associated with that region. The total deposited mass is given by:

$$m_{tot} = \sum m_j \quad (10.31)$$

10.4 Definition of Dose Summary Statistics

A flag called ***DODOSE*** is available in the *Control Options* file. The default is DODOSE = YES. If DODOSE = NO then the dose calculation in APEX is skipped.

If the dose calculation is performed, the initial result is the dose level for each event in the simulation. Unlike exposure, there are actually three doses calculated per event. The first is the time-average dose over the event. The second is the running 8-hour average of event doses. The third is the final instantaneous dose at the end of the event. For CO, the final dose is found directly using the series $T_4(z)$ given in the previous section. The average dose is easily found since $T_4(z)$ is a polynomial in time, and therefore can be integrated without difficulty. For all other pollutants, the instantaneous end-of-event dose has no meaning, and is simply equal to the value calculated in Eq. 10-1.

Most of the summary statistics for dose are analogous to those for exposure. The average dose values over an event and the instantaneous dose at the end of the event are written to *Events* output file (if it exists). Timestep and hourly-average dose time series are created by taking appropriate duration-weighted averages of dose over the events in each clock hour. These doses may be written to the *Timestep* and *Hourly* output files. A vector of instantaneous dose at the end of each hour (FDose) is also saved for the calculation of a daily statistic for maximum end-of-hour dose (see next section). This is not an average, but simply a subset of the values of the event-end dose corresponding to the events that end on a clock hour. For pollutants other than CO, this end-of-hour dose is simply the dose on the last event of the hour.

There are five daily summary statistics for dose. The first is ***DAvgDose*** (Daily Average Dose) which is the average of the 24 hourly average dose values that fall on the same calendar day. The result for each day is binned according to the levels defined by the cutpoints set in the *Control Options* file. The second summary statistic is ***DMTSDose*** (Daily Maximum Timestep Dose), the largest dose over all the timestep values for the day. The third summary statistic is ***DM1HDose*** (Daily Maximum 1-Hour Dose) which is the largest of the 24 hourly dose values for a day. The fourth is ***DM8HDose*** (Daily Maximum 8-Hour Dose), which is the largest of the 24 8-hour running dose averages. The fifth is ***DMEHDose*** (Daily Maximum End-of-Hour Dose), which is the largest of the 24 values of the instantaneous dose level at the end of each clock hour on a day. All four daily summaries are binned according to the appropriate set of cutpoints in the

Control Options file. Finally, there is the average dose over the entire simulation period, ***SAvgDose***. It too is binned and tabulated according to cutpoints set in the *Control Options* file.

CHAPTER 11. SOBOL SENSITIVITY ANALYSIS

APEX is a stochastic model which may have roughly 100 or so random variables, each of which has some influence on the exposures of the simulated persons. All the simulated persons share the same study area and input data. Therefore, all of the differences in exposure and dose between simulated persons are due to differences in the random variables assigned to them. Sobol analysis attributes these differences to differences in the random variables (see Glen and Isaacs (2012), Saltelli et al. (2004) for details on this process).

Sobol analysis is based on variance decomposition. The variance (across the simulated persons) in the selected output variable is partitioned among all the randomly sampled variables in the model. Note that if all randomly sampled variables happened to be assigned the same values for two (or more) simulated persons, any output variable would have the same value for each of them. In that sense, all variation in output is due to differences in the assignment of the random input variables. Sobol analysis is based conducting pairs of model runs, with selected inputs being held at the same values in both runs, while other inputs are "resampled" with new stochastic sample values. By varying the set of inputs that are held common in both runs, the contribution of each set of input variables to the variance of the output may be quantified.

11.1 Introduction and Background

First, an output variable is selected for analysis. It must have a single, definite numeric value for each simulated person. The Sobol runs examine the variance across persons in this variable, and apportion this variance into a series of terms that reflect single variables or combinations of variables. The single variable terms are called "main effects" and the multiple-variable terms are called "interactions." A "total effect" may be calculated for each variable, which consists of the sum of its main effect plus all of its interactions. Each partial variance term is standardized by dividing it by the total variance in order that the sum of all the main effects plus all unique interactions always equals one. No terms may be negative, because no variances may be negative.

If there are no interaction terms, then the main effects sum to one (when summed over the full set of random variables), and the total effect equal the main effect for every variable. With interactions, the main effects sum to less than one, while the total effects sum to more than one (since each interaction term is counted multiple times in that sum). For example, suppose a model has only three random variables which fully determine the output value. Number these variables from 1 to 3. Label each Sobol sensitivity index as "s," with subscripts indicating which variables it represents. Then:

$$s_1 + s_2 + s_3 + s_{12} + s_{13} + s_{23} + s_{123} = 1 \qquad (11.1)$$

Every unique combination of subscripts appears once in such an equation, regardless of the number of random variables. The order of the subscripts does not matter, so $s_{12} = s_{21}$, and by convention the subscripts are placed in numerical order. All models with exactly three random variables are represented by equation (11.1), although the values of the indices may be different in each case. The main effects and total effects are shown in Table 11.1.

Table 11.1 Main and Total Effects for a Three-variable Model

Variable	Main effect	Total Effect
1	s_1	$s_1 + s_{12} + s_{13} + s_{123}$
2	s_2	$s_2 + s_{12} + s_{23} + s_{123}$
3	s_3	$s_3 + s_{13} + s_{23} + s_{123}$

While the above example is quite simple, the problem quickly becomes more complex with additional random variables. With N random variables, there are (2^N-1) indices in the equation corresponding to (11.1). Even for N=10, this is over 1000 indices. For a model like APEX with N around 100, it is not possible to even write out such an equation, or to evaluate all its terms.

Fortunately, there are two practical ways to handle the case of large N. The first is *grouping*, in which variables may be combined. Each group then has its own main effect and interactions. If there were three groups, then equation (11.1) and Table 11.1 would apply again, only this time the subscripts would represent group numbers. The choice of which variables to group together is entirely up to the modeler. It is possible to have a single variable in a group, or nearly all the variables in the model in a single group. Evidently, there must be at least two groups to learn anything from the analysis. A good, practical way to start Sobol analysis is to define a small number of groups that can be interpreted fairly easily. An example would be to have groups for demographic variables, physiological variables, diary variables, microenvironmental variables, and a catch-all group for all other variables. Once the indices are obtained, the groups may be redefined to learn more from another round of analysis. For example, if the physiological variables are unimportant in determining exposure, they could be lumped with the "other" variables into a single group. If the microenvironmental variables are accounting for most of the variance, it may be useful to split them into two or more groups to obtain separate indices for each.

Actually, it is possible to obtain the main and total effects for all variables without grouping. With just (2N+2) model runs, one can evaluate all main and total effects for N variables. It is not necessary to evaluate all the interaction terms separately to do this. Hence, for N=100, only 202 model runs are needed to obtain the 100 main effects and 100 total effects. The other (2^{100}-101) interaction terms cannot be evaluated without a prohibitive number of runs, although in principle, a few selected interaction terms could be evaluated using a small number of additional runs. This is not currently programmed into APEX.

Each of the "model runs" mentioned above would consist of P profiles, with P selected by the user. With finite P, the indices are only estimates which are subject to statistical error. This error (that is, the standard deviation) decreases as the square root of P, so four times as many profiles are needed per run to cut the size of the error bars in half. As a practical matter, with 1000 profiles per run, each index is accurate to at least one significant digit, with error possible in the second digit.

APEX has been programmed to automatically perform all (2N+2) model runs in a single job submission, and to calculate and save the main and total effect Sobol indices (but not the interactions). The instructions needed on the input files are discussed in the next section.

Warning: Any one of the (2N+2) passes may be used as a random sample representative of the target population. However, the passes should not be combined, because they are not independent. If there are non-influential variables, then some of these passes will have identical output to others. Suppose there are 24 Sobol groups and P=5000 persons. Then APEX performs 50 runs of 5000 persons = 250,000 simulated persons to generate the Sobol indices. But not more than 5000 of these can be used to populate the standard output tables. Sobol runs are not an efficient way to produce regular output tables, so those tables are suppressed.

11.2 Submitting a Sobol Analysis Run

Changes are needed on three input files to submit a Sobol run: the seed file, the *Microenvironment Descriptions* file, and the *Control Options* file. Additional details on these settings are found in *Volume I*. For the *Seed offsets and Sobol grouping* file and *Microenvironment Descriptions* file, the user must specify the Sobol group number for each random variable. The groups may be numbered from zero upwards. Group zero is typically used as the "other" group which represents all the unimportant or uninteresting variables. The group numbering may contain gaps, so when regrouping it is possible to combine groups 1 and 2 (for example) without having to re-number all the subsequent groups.

Table 11.2 Stochastic Input Variables Available for Sobol Analysis

Random Variable	Description
Gender	Male or female
Age	Age in full years (range 0-99)
HomeSec	Home sector
WorkSec	Work Sector
Race	Race
Employ	Employment status (yes or no)
ProfFactor	Sector dependent profile factor
GasStove	Has gas stove (yes or no)
GasPilot	Has gas pilot on stove (yes or no)
AC_Home	Has air conditioning at home (yes or no)
AC_Car	Has air conditioning in car (yes or no)
ProfCond1	Profile conditional variable #1
ProfCond2	Profile conditional variable #2
ProfCond3	Profile conditional variable #3
ProfCond4	Profile conditional variable #4
ProfCond5	Profile conditional variable #5
RegCond1	Regional conditional variable #1
RegCond2	Regional conditional variable #2
RegCond3	Regional conditional variable #3
RegCond4	Regional conditional variable #4
RegCond5	Regional conditional variable #5
OtherD	District other than home or work
NearHome	District when in Near Home location

Random Variable	Description
NearWork	District when in Near Work location
WindowRes	Daily window status at residence
WindowCar	Daily window status in car
SpeedCat	Daily speed category for travel
DayCond1	Daily conditional variable #1
DayCond2	Daily conditional variable #2
DayCond3	Daily conditional variable #3
BodyMass	Body mass in kg (weight is equivalent in lbs)
Height	Height in inches
BSA	Body surface area in m2
RMR	Resting metabolic rate (initially MJ/day)
VEAge	Ventilation-age regression terms
Disease	Presence of disease
VO2max	Max. normalized O2 consumption (ml O2/min/kg)
ECF	Energy conversion factor (liters O2 / kcal)
RecoveryT	Recovery time for oxygen debt (hours)
Hemog	Hemoglobin density in blood
MaxOxD	Maximum possible oxygen debt (ml/kg)
EndogCO	Endogenous CO production rate (ml/min)
BloodVol	Blood volume regression factors
SFast	Slope of fast oxygen debt recovery
FEVB1_9	dFEV (lung function loss) parameters B1-B9
FEVreg	FEVSLP and FEVINT parameters
FEVU	dFEV interperson variation parameter U
FEVE1	dFEV intraperson variation parameter E1
FEVE2	dFEV intraperson variation parameter E2
VEBM	Ventilation/body mass regression parameters
METS	Activity-specific MET values
AQData	Air quality data drawn from distributions
DiarySel	Daily activity diary selection
DAutoCor	Autocorrelation of diaries (D&A method only)
Clus1	First diary clustering parameter
Clus2	Second diary clustering parameter

Each row on the *Seed offsets and Sobol grouping* file (see Table 11.2) represents one random variable. APEX runs will not use the full set, but all should be listed on the *Seed offsets and Sobol grouping* file anyway. (For example, depending on the diary selection method, either DAutoCor is used, or Clus1 and Clus2, but never all three.) The user assigns a Sobol group number to each row. In a non-Sobol APEX run, the group numbers are ignored, so it is simplest (but not required) to set all of them to zero. On the *Microenvironment Descriptions* file, a Sobol run requires that each MP# also have a line with the keyword "SobolGroup," followed by an equal sign and a group number. The MP are also random variables, but are not listed on the *Seed offsets and Sobol grouping* file in order to allow all the MP to be defined in one place, and to avoid having to force consistency across two input files.

Once the group assignments are made (on both the *Seed offsets and Sobol grouping* file and the *Microenvironment Descriptions* file), the remaining changes needed for a Sobol run are on the *Control Options* file. There are three necessary changes. First, an output file must be named on which the Sobol indices will be written. The keyword for this is "Sobol file." Second, a switch with the keyword "SobolRun" must be set to "Y" or "YES" (not case sensitive) to enable a Sobol run. The default is NO, although one can also set "SobolRun=No," for clarity. The final step uses the keyword "SobolVarList," followed by the equal sign and a list of the output variables to be analyzed. Fortunately, the same set of (2N+2) runs may be used to simultaneously generate indices for all of them. The possible variables are shown in Table 11.3.

Table 11.3 Output Variables Available for Sobol Analysis

Variable	Description
AVGEXP	Average daily exposure
MAX1EXP	Maximum daily 1-hour exposure
MAX8EXP	Maximum daily 8-hour exposure
MAXTSEXP	Maximum timestep exposure
MAX8EC	Exposure/ambient ratio at time of max. 8-hour exposure
AVGDOSE	Average daily inhaled dose
MAX1DOSE	Maximum daily 1-hour inhaled dose
MAX8DOSE	Maximum daily 8-hour inhaled dose
MAXTSDOSE	Maximum timestep inhaled dose
MAX1FDOSE	Maximum daily end-of-hour dose
INTAKE	Daily particulate matter intake dose (for PM only)
DEP	Daily particulate deposited mass in lungs (PM only)

The units of exposure are the same as for air concentration. Note that the user may set these units on the *Control Options* file. The units of dose are those of (exposure x ventilation rate). Dose in APEX is actually a "dose rate," since it measures the rate at which chemical enters the body. But the Sobol indices would not depend on the choice of units anyway.

The variables INTAKE and DEP are calculated by APEX only if the pollutant is particulate matter. The variables ending in DOSE are calculated only if the keyword DoDose is set to YES on the *Control Options* file. The exposure variables are always calculated in an APEX run. Sobol analysis may select any or all of the calculated variables that are in Table 11.3 to be analyzed.

For each variable selected from Table 11.3, two tables of Sobol indices are generated. One set is for the average of the selected variable over all days in the simulation period, and the other is for the worst (maximum) day for each person. Each such table contains both main effect and total effect indices.

11.3 Code Implementation of Sobol Analysis

In APEX, a Sobol run is not very different from a standard run. The main APEX module consists of some initialization steps and two nested loops—one over profiles and the other over pollutants. To allow Sobol analysis, a third loop has been added outside the other two. In a standard run, this outer loop is executed just once. In a Sobol run, it is executed (2N+2) times, where N is the number of Sobol groups, which is the number of different group numbers on the *Seed offsets and Sobol grouping* file and the *Microenvironment Descriptions* file taken together.

At the start of each of the (2N+2) loops, each random variable is flagged as to whether the two 4-byte seeds (called A and B) assigned to it should be combined into an 8-byte seed as AB or as BA. Since both AB and BA are equally valid as seeds, there is no reason to prefer any one pass through the loop over another. Either convention would be equally valid as a standard, single pass run. However, one cannot save all the passes and combine them into one large run, since they are not independent of each other.

In a Sobol run, the usual Output function calls a new SobolOutput function that stores information on the selected Sobol output variables. A new global array, SobolOut, is used for this purpose. All of the information that goes into SobolOut is calculated during a standard run, so that Sobol runs are not slower on each pass through the outer loop. The only difference is that there are (2N+2) such passes instead of just one. At the end of the last pass, a Sobol run calls the SobolSummary function, which calculates the sensitivity indices using the SobolOut data, and writes the indices to an output file.

11.4 Interpreting the Tables of Sobol Indices

Each table covers one pollutant, one output variable, and one type, which are listed in the table header. The output variable is always a quantity that can be assigned a specific value on each day of the simulation period. For example, MAX1EXP is the largest of the 24 1-hour exposures that occur on each day. The type is either “average day,” or “maximum day.” The former means the average of the daily values, over all days in the simulation period. The latter means the highest of all the daily values for that person. Essentially, the Sobol analysis for the “average day” indicates which random variables are most important for chronic, long-term effects, while the “maximum day” analysis indicates the same for acute or episodic effects.

Each table consists of four columns. The first is the rank, the second is the group number, the third is the main effect index, and the last is the total effect index. Group numbers are used rather than naming specific random variables because there may be several in the same group. The output file has a section which lists the Sobol group assignments for all the random variables, and these same assignments apply to every Sobol table in that run. The rows in each table are ordered by decreasing values of the total effects index.

In theory, both the main and total effect indices should always be between zero and one. In practice, it is possible to obtain small negative numbers as estimates of indices that are in reality small but positive. Small indices are generally less interesting than large ones. Often, one places all the variables with small effects into one larger group, to reduce the computing time.

The tables on the output file report the indices for all groups, rounded to four decimal places. It is not practical to run enough profiles to require more digits than that. A variable with a large main effect is one that has a clear impact on the exposure or dose, regardless of the values assigned to other variables. In an APEX run, the home sector is often such a variable because some have worse air quality than others. For some other variables, the main effect may be negligible, but the total effect is not. An example might be air exchange rate. A high air exchange rate will tend to increase exposure when the outdoor air has a higher concentration than the indoor air, or decreased exposure otherwise. Thus, the effect of the air exchange rate will not be strong in isolation, but be tied to the effect of the microenvironmental parameters that determine the indoor/outdoor ratio, like source strengths. With just (2N+2) passes, no details are obtained regarding which other variables are interacting with the one of interest; the method can only determine the collective effect of all such interactions, which is the difference between the total effect and main effect. In principle, more runs could be made to identify specific interaction terms, but this option has not been coded into APEX.

Although there may be 100 random variables or more, the majority have little or no influence on the result. For the predominant effects, typically about 5 random variables have a substantial effect. For total effects, the number is usually higher, but seldom more than 15. In general, the "average day" results depend on more variables than the "maximum day" results, which is logical since much more data (including more uncommon events) are used in determining the "average day" output.

REFERENCES

Adams WC. 1998. Letter to Tom McCurdy, National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina. August 21.

Almuzaini KS, Potteiger JA, and Green SB. 1998. Effects of split exercise sessions on excess postexercise oxygen consumption and resting metabolic rate. Can J Appl Physiol. 23(5):433-43.

Astrand PO and Rodahl K. 1977. Textbook of Work Physiology. 2nd ed. McGraw-Hill, New York, New York.

Avol EL, Linn WS, Shamoo DA, Valencia LM, Anzar UT, Venet TG, Hackney JD. 1985. Respiratory effects of photochemical oxidant air pollution in exercising adolescents. Am Rev Respir Dis. 132(3):619-22.

Avol EL, Linn WS, Shamoo DA, Spier CE, Valencia LM, Venet TG, Trim SC, Hackney JD. 1987. Short-term respiratory effects of photochemical oxidant exposure in exercising children. JAPCA. 37(2):158-62.

Bahr R, Ingnes I, Vaage O, Sejersted OM, and Newsholme EA. 1987. Effect of duration of exercise on excess postexercise O2 consumption. J Appl Physiol. 62(2):485-90.

Bahr R. 1992. Excess postexercise oxygen consumption--magnitude, mechanisms and practical implications. Acta Physiol. Scand Suppl. 605:1-70.

Berthoin S, Baquet G, Dupont G, Blondel N, and Mucci P. 1996. Critical velocity and anaerobic distance capacity in prepubertal children. Can J Appl Physiol. 28(4):561-75.

Bickham D, Le Rossignol P, Gibbons C, and Russell AP. 2002. Re-assessing accumulated oxygen deficit in middle-distance runners. J Sci Med Sport. 5(4):372-82.

Bielinski R, Schutz Y, and Jequier E. 1985. Energy metabolism during the postexercise recovery in man. Am J Clin Nutr. 42(1):69-82.

Billat V, Beillot J, Jan J, Rochcongar P, and Carre F. 1996. Gender effect on the relationship of time limit at 100% VO2max with other bioenergetic characteristics. Med Sci Sports Exerc. 28(8):1049-55.

Biller WF, Feagans TB, Johnson TR, Duggan GM, ,Paul RA, McCurdy T, and Thomas HC. 1981. A general model for estimating exposure associated with alternative NAAQS. Paper No. 81-18.4 in Proceedings of the 74th Annual Meeting of the Air Pollution Control Association, Philadelphia, Pa.

Box G, Jenkins G, and Reinsel G. 1994. Time Series Analysis: Forecasting and Control, Prentice Hall, Englewood Cliffs, NJ.

Brockman L, Berg K, and Latin R. 1993. Oxygen uptake during recovery from intense intermittent running and prolonged walking. J Sports Med Phys Fitness. 33(4):330-6.

Buck D and McNaughton L. 1999. Maximum accumulated oxygen debt must be calculated using 10 min time periods. Med Sci Sports Exerc. 31(9):1346-1349.

Burmaster DE and Crouch EAC. 1997. Lognormal distributions of body weight as a function of age for males and females in the United States, 1976 – 1980. Risk Analysis 17(4).

Burmaster DE. 1998. LogNormal distributions for skin area as a function of body weight. Risk Analysis. 18(1):27-32.

Carlson JS and Naughton GA. 1993. An examination of the anaerobic capacity of children using maximum accumulated oxygen debt. Pediatr Exerc Sci. 5:60-71.

Dawson B, Straton S, and Randall N. 1996. Oxygen consumption during recovery from prolonged submaximum cycling below the anaerobic threshold. J Sports Med Phys Fitness. 36:77-84.

Demarle AP, Slawinski JJ, Laffite LP, Bocquet VG, Koralsztein JP, and Billat VL. 2001. Decrease of O(2) deficit is a potential factor in increased time to exhaustion after specific endurance training. J Appl Physiol. 90(3):947-53.

Doherty M, Smith PM, and Schroder K. 2000. Reproducibility of the maximum accumulated oxygen deficit and run time to exhaustion during short-distance running. J Sports Sci. 18(5):331-8.

Drechsler-Parks DM, Bedi JF, Horvath SM. 1987. Pulmonary function responses of older men and women to ozone exposure. Exp Gerontol. 22(2):91-101.

Drechsler-Parks DM, Bedi JF, Horvath SM. 1989. Pulmonary function responses of young and older adults to mixtures of O3, NO2 and PAN. Toxicol Ind Health. 5(3):505-17.

DuBois D, DuBois EF. 1916. A formula to estimate the approximate surface area if height and weight be known. Arch Intern Medicine. 17:863-71.

Esmail S, Bhambhani Y, and Brintnell S. 1995. Gender differences in work performance on the Baltimore therapeutic equipment work simulator. Amer. J. Occup. Therapy. 49: 405 - 411.

Faina M, Billat V, Squadrone R, De Angelis M, Koralsztein JP, and Dal Monte A. 1997. Anaerobic contribution to the time to exhaustion at the minimal exercise intensity at which maximum oxygen uptake occurs in elite cyclists, kayakists and swimmers. Eur J Appl Physiol. Occup Physiol. 76(1):13-20.

Frey GC, Byrnes WC, and Mazzeo RS. 1993. Factors influencing excess postexercise oxygen consumption in trained and untrained women. Metabolism. 42(7):822-828.

Galetti, P. M. 1959. Respiratory exchanges during muscular effort. Helv. Physiol. Acta. 17: 34 - 61.

Gastin PB and Lawson DL. 1994. Variable resistance all-out test to generate accumulated oxygen deficit and predict anaerobic capacity. Eur J Appl Physiol. Occup Physiol. 69(4):331-6.

Gastin PB, Costill DL, Lawson DL, Krzeminski K, and McConell GK. 1995. Accumulated oxygen deficit during supramaximum all-out and constant intensity exercise. Med Sci Sports Exerc. 27(2):255-63.

Geyh, AS, Xue, J, Ozkaynak, H, and Spengler, JD. 2000. The Harvard Southern California chronic ozone exposure study: Assessing ozone exposure of grade-school-age children in two Southern California communities. *Environ Health Persp.* 108:265-270.

Gillette CA, Bullough RC and Melby CL. 1994. Postexercise energy expenditure in response to acute aerobic or resistive exercise. Int J Sport Nutr. 4(4):347-60.

Glen, G, and Isaacs K. 2012. Estimating Sobol Indices using Correlations. Environmental Modelling and Software 37: 157-166.

Glen G, Smith L, Isaacs K., McCurdy T., and Langstaff J. 2008. A new method of longitudinal diary assembly for human exposure modeling. J. Expos. Sci. Environ. Epidemiol. 18:299-311.

Gong H Jr, Shamoo DA, Anderson KR, Linn WS. 1997. Responses of older men with and without chronic obstructive pulmonary disease to prolonged ozone exposure. Arch Environ Health. 52(1):18-25.

Gore CJ and Withers RT. 1990. Effect of exercise intensity and duration on postexercise metabolism. J Appl Physiol. 68(6):2362-8.

Graham S and McCurdy T. 2005. Revised Ventilation Rate (Ve) Equations for Use in Inhalation-Oriented Exposure Models, A NERL Internal Research Report.

Hagberg JM, Hickson RC, Ehsani AA, and Holloszy JO. 1980. Faster adjustment to and recovery from submaximum exercise in the trained state. J Appl Physiol. 48(2):218-24.

Harms CA, Cordain L, Stager JM, Sockler JM, and Harris M. 1995. Body fat mass affects postexercise oxygen metabolism in males of similar lean body mass. Med Exer Nutr Health. 4:33-39.

Harris JM, Hobson EA, and Hollingsworth DF. 1962. Individual variations in energy expenditure and intake. Proc Nutr Soc. 21: 157-169.

Hazucha MJ, Folinsbee LJ, Bromberg PA. 2003. Distribution and reproducibility of spirometric response to ozone by gender and age. J Appl Physiol. 95(5):1917-25.

Hill DW, Ferguson CS, and Ehler KL. 1998. An alternative method to determine maximum accumulated O2 deficit in runners. Eur J Appl Physiol. Occup Physiol. 79(1):114-7.

ICRP Publication 66. 1994. Human Respiratory Tract Model for Radiological Protection. *Annals of the ICRP*. International Commission on Radiological Protection.

Isaacs K and Smith L. 2005. New Values for Physiological Parameters for the Exposure Model Input File Physiology.txt. Memorandum submitted to the U.S. Environmental Protection Agency under EPA Contract EP-D-05-065. NERL WA 10. Alion Science and Technology.

Isaacs K, Glen G, McCurdy T., and Smith L. 2007. Modeling energy expenditure and oxygen consumption in human exposure models: Accounting for fatigue and EPOC. J. Expos. Sci. Environ. Epidemiol. 18: 289-298.

Johnson TR and Paul RA. 1983. The NAAQS Exposure Model (NEM) Applied to Carbon Monoxide. EPA-450/5-83-003. Prepared for the U.S. Environmental Agency by PEDCo Environmental Inc., Durham, N.C. under Contract No. 68-02-3390. U.S. Environmental Protection Agency, Research Triangle Park, N.C.

Johnson T, Capel J, Olaguer E, Wijnberg L. 1992. Estimation of Ozone Exposures Experienced by Residents of ROMNET Domain Using a Probabilistic Version of NEM. Report prepared by IT Air Quality Services for the Office of Air Quality Planning and Standards, U. S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Johnson T, Capel J, and McCoy M. 1996a. Estimation of Ozone Exposures Experienced by Urban Residents Using a Probabilistic Version of NEM and 1990 Population Data. Report prepared by IT Air Quality Services for the Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Johnson T, Capel J, Mozier J, and McCoy M. 1996b. Estimation of Ozone Exposures Experienced by Outdoor Children in Nine Urban Areas Using a Probabilistic Version of NEM. Report prepared for the Air Quality Management Division under Contract No. 68-DO-30094, April.

Johnson T., Capel J, McCoy M and Mozier J. 1996c. Estimation of Ozone Exposures Experienced by Outdoor Workers in Nine Urban Areas Using a Probabilistic Version of NEM. Report prepared for the Air Quality Management Division under Contract No. 68-DO-30094, April.

Johnson T. 1998. Analysis of Clinical Data Provided by Dr. William Adams and Revisions to Proposed Probabilistic Algorithm for Estimating Ventilation Rate in the 1998 Version of pNEM/CO. Memorandum submitted to the U.S. Environmental Protection Agency under EPA Contract No. 68-D6-0064. TRJ Environmental, Inc.

Johnson T, Mihlan G, LaPointe J, Fletcher K, Capel J, Rosenbaum A, Cohen J, Stiefer P. 2000. Estimation of carbon monoxide exposures and associated carboxyhemoglobin levels for residents of Denver and Los Angeles using pNEM/CO. Appendices. EPA constract 68-D6-0064.

Johnson T. 2002. A Guide to Selected Algorithms, Distributions, and Databases Used in Exposure Models Developed By the Office of Air Quality Planning and Standards. Revised Draft. Prepared for U.S. Environmental Protection Agency under EPA Grant No. CR827033.

Joumard R, Chiron M, Vidon R, Maurin M, and Rouzioux J-M. 1981. Mathematical models of the uptake of carbon monoxide on hemoglobin at low carbon monoxide levels. Environmental Health Perspectives. 41: 277 - 289.

Kaminsky LA, Padjen S, and LaHam-Saeger. 1990. J Effect of split exercise sessions on excess post-exercise oxygen consumption. Br J Sports Med. 24(2):95-8.

Kaminsky LA, and Whaley MH. 1993. Effect of interval-type exercise on excess post-exercise oxygen consumption in obese and normal-weight women. Med Exer Nutr Health. 2:106-111.

Katch FI, Girandola RN, and Henry FM. 1972. The influence of the estimated oxygen cost of ventilation on oxygen deficit and recovery oxygen intake for moderately heavy bicycle ergometer exercise. Med Sci Sports. 4:71-76.

Knuttgen HG. 1970. Oxygen debt after submaximum physical exercise. J Appl Physiol. 29(5):651-657.

Langstaff, J.E. 2007. Analysis Of Uncertainty In Ozone Population Exposure Modeling. Office of Air Quality Planning and Standards,U.S. Environmental Protection Agency.

Maehlum S, Grandmontagne M, Newsholme EA, and Sejersted OM. 1986. Magnitude and duration of excess postexercise oxygen consumption in healthy young subjects. Metabolism. 35(5):425-9.

Maresh CM, Abraham A, De Souza MJ, Deschenes MR, Kraemer WJ, Armstrong LE, Maguire MS, Gabaree CL, and Hoffman JR. 1992. Oxygen consumption following exercise of moderate intensity and duration. Eur J Appl Physiol. Occup Physiol. 65(5):421-6.

Maxwell NS and Nimmo MA. 1996. Anaerobic capacity: a maximum anaerobic running test versus the maximum accumulated oxygen deficit. Can J Appl Physiol. 21(1):35-47.

McArdle WD, Katch FI, and Katch VL. 2001. Exercise Physiology: Energy, Nutrition, and Human Performance, Fifth Edition. Lippincott, Williams, and Wilkins, Philadephia.

McCurdy T. 2000. Conceptual Basis for Multi-Route Intake Dose Modeling Using an Energy Expenditure Approach. Journal of Exposure Analysis and Environmental Epidemiology. 10:1 - 12.

McCurdy T, Glen G, Smith L, and Lakkadi Y. 2000. The National Exposure Research Laboratory's Consolidated Human Activity Database, Journal of Exposure Analysis and Environmental Epidemiology 10: 566-578.

McDonnell WF, Chapman RS, Leigh MW, Strope GL, Collier AM. 1985. Respiratory responses of vigorously exercising children to 0.12 ppm ozone exposure. Am Rev Respir Dis. 132(4):875-9.

McDonnell WF, Stewart PW, Smith MV. 2007. The temporal dynamics of ozone-induced FEV_1 Changes in Humans: An Exposure-Response Model. Inhal Toxicol. 19:483-94.

McDonnell WF, Stewart PW, Smith MV. 2010. Prediction of ozone-induced lung function responses in humans. Inhal Toxicol. 22(2):160-8.

Naughton GA, Carlson JS, Buttifant DC, Selig SE, Meldrum K, McKenna MJ, and Snow RJ. 1998. Accumulated oxygen deficit measurements during and after high-intensity exercise in trained male and female adolescents. Eur J Appl Physiol. Occup Physiol. 76(6):525-31.

Olesen HL. 1992. Accumulated oxygen deficit increases with inclination of uphill running. J Appl Physiol. 73(3):1130-4.

Pivarnik JM and Wilkerson JE. 1988. Recovery metabolism and thermoregulation of endurance trained and heat acclimatized men. Sports Med Phys Fitness 28(4):375-80.

Renoux JC, Petit B, Billat V, and Koralsztein JP. 1999. Oxygen deficit is related to the exercise time to exhaustion at maximum aerobic speed in middle distance runners. 1: Arch Physiol. Biochem. 107(4):280-5.

Roberts AD, Clark SA, Townsend NE, Anderson ME, Gore CJ, and Hahn AG. 2003. Changes in performance, maximum oxygen uptake and maximum accumulated oxygen deficit after 5, 10 and 15 days of live high:train low altitude exposure. Eur J Appl Physiol. 88(4-5):390-395.

Roddin, MF, Ellis HT, and Siddiqee WM. 1979. Background Data for Human Activity Patterns, Vols. 1, 2. Draft Final Report prepared for Strategies and Air Standards Division, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, N.C.

Rosenbaum, AS. 2008. *The Cluster-Markov algorithm in APEX.* Memorandum prepared for Stephen Graham, John Langstaff. USEPA OAQPS by ICF International.

Rosenbaum, AS, and Cohen JP. 2004. *Evaluation of a multi-day activity pattern algorithm for creating longitudinal activity patterns.* Memorandum prepared for Ted Palma. USEPA OAQPS by ICF International.

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto (2004). Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. John Wiley & Sons, Ltd, Chichester, England.

Schofield, WN. 1985. Predicting basal metabolic rate, new standards, and review of previous work. Hum Nutr Clin Nutr, 39C(Supplement 1):5 - 41.

Sedlock DA. 1991a. Effect of exercise intensity on postexercise energy expenditure in women. Br J Sports Med. 25(1):38-40.

Sedlock DA. 1991b. Postexercise energy expenditure following upper body exercise. Res Q Exerc Sport. 62(2):213-6.

Short KR and Sedlock DA. 1997. Excess postexercise oxygen consumption and recovery rate in trained and untrained subjects. J Appl Physiol, 83(1):153-159.

Trost S, Wilcox A, and Gillis D. 1997. The effect of substrate utilization, manipulated by nicotinic acid, on excess postexercise oxygen consumption. Int J Sports Med 18(2):83-88.

U.S. EPA. 1978. Altitude as a Factor in Air Pollution. Environmental Criteria and Assessment Office. EPA-600/9-78-015.

U.S. EPA. 1999. Total Risk Integrated Methodology. [On-line]. Available: https://www.epa.gov/fera/total-risk-integrated-methodology-trim-overview

U.S. EPA. 2007. *Ozone Population Exposure Analysis for Selected Urban Areas*. EPA-452/R-07-010 http://www.epa.gov/ttn/naaqs/standards/ozone/data/2007-01_o3_exposure_tsd.pdf

U.S. Environmental Protection Agency. 2019. The Consolidated Human Activity Database (CHAD) Documentation and User's Guide, EPA-452/B-19-001. Available at: https://www.epa.gov/fera/human-exposure-modeling-databases-support-exposure-modeling

Weber CL and Schneider DA. 2000. Maximum accumulated oxygen deficit expressed relative to the active muscle mass for cycling in untrained male and female subjects. Eur J Appl Physiol. 82(4):255-61.

Xue J, McCurdy T, Spengler O, Özkaynak, H. 2004. Understanding variability in the time spent in selected locations for 7-12 year old children. J Exposure Anal Environ Epidem 14(3) : 222-233.

Xue J, Liu SV, Ozkaynak H, Spengler J. 2005. Parameter evaluation and model validation of ozone exposure assessment using Harvard Southern California Chronic Ozone Exposure Study Data. *J. Air & Waste Manage. Assoc.* **55**:1508–1515.

APPENDIX

Flexible Duration for Running Averages

All APEX versions prior to 5.12 computed and reported eight-hour running averages for several variables, specifically Run8EVR, Run8Exp, Run8Dose, and Run8AmbHome. Starting with version 5.12, a switch has been added to the APEX control file, allowing the user to specify a different number of hours to use for these running averages. This switch is optional, and if not specified it defaults to 8 hours. To set it, enter a new line on the control file with keyword "RunHours". For example,

…
```
RunHours = 7
```
…

The above directs APEX to use seven-hour running averages for Run8EVR, Run8Exp, Run8Dose, Run8AmbHome, and subsequently derived variables such as DM8HExp and DM8HDose. These variables have generally been renamed in the code to remove the '8', but there are many places in the manuals, output tables, and code comments where the '8' may still be found. In particular, the control file still uses the same list of variables as in previous versions. Thus, to obtain tables of running averages of exposure and dose, the 'TablesList' keyword must be followed by 'Exp8H' and 'Dose8H, as in the example

```
TablesList = EXP1H EXP8H DOSE1H DOSE8H ILLNESS MOD
```

If RunHours=7, then the corresponding output tables will be for 7-hour running averages, which should be indicated in the table headers. Note that the word 'MOD' above indicates tables at moderate exertion, which is defined as having a running average of EVR at or above the moderate threshold. The running average for EVR automatically is based on the same number of hours as exposure and dose.

Caution should be used in setting RunHours far from 8, because the experiments that determine the physiological response to prolonged exercise and exposure are typically close to 8 hours in duration, and it is expected that this data may inform any risk estimates based on the APEX results.

As in all versions of APEX, RunHours may be used only in runs with a timestep of one hour or less.

United States
Environmental Protection
Agency

Office of Air Quality Planning and Standards
Health and Environmental Impacts Division
Research Triangle Park, NC

Publication No. EPA-452/R-19-005b
October 2019