# Update and Refinement of an Indoor Exposure Assessment Methodology

## Contract 98-327

## Final Report

**May 2002**

*Prepared for:*
California Air Resources Board
Research Division
1001 I Street
Sacramento, CA 95814

*Prepared by:*
ICF Consulting
60 Broadway
San Francisco, CA 94111
(415) 677-7100

Arlene S. Rosenbaum
Jonathan P. Cohen
Farzad Kavoosi

*This page intentionally left blank.*

## Disclaimer

The statements and conclusions in this Report are those of the contractor and not necessarily those of the California Air Resources Board. The mention of commercial products, their source, or their use in connection with material reported herein is not to construed as actual or implied endorsement of such products.

*This page intentionally left blank.*

# Acknowledgments

The primary author of this report, Arlene Rosenbaum, served as principal investigator, overseeing all aspects of the effort.

Dr. Jonathan Cohen, in consultation with Ms. Rosenbaum designed and directed the implementation of the uncertainty software module, UNC. The design was reviewed by Dr. Christopher Frey. The implementation was performed by Cynthia Van Landingham and Michael Huang. Dr. Cohen also designed the verification procedures of UNC, which were implemented by Dr. Willard Hobbs.

Dr. Cohen designed the procedures for estimating the post-stratification weights for the activity patterns, and managed their implementation. These procedures were reviewed by Johnny Blair.

Farzad Kavoosi ported the original CPIEM software from a DOS platform to a Visual Basic platform, designed and implemented the new user interface, and incorporated all the CPIEM model enhancements.

Overall project management was provided by Ms. Susan Lum of the California Air Resources Board with the assistance of Ms. Peggy Jenkins, Manager of the Indoor Exposure Section at the California Air Resources Board.

This report was submitted in fulfillment of California Air Resources Board Contract No. 98-327 “Update and Refinement of An Indoor Exposure Assessment Methodology” by ICF Consulting under the sponsorship of the California Air Resources Board. Work on this project was completed as of May 2002.

*This page intentionally left blank.*

# Abstract

The primary mission of the ARB Indoor Program is to identify and reduce Californians' exposures to indoor pollutants. To fully consider indoor exposures in assessing risk, the ARB needs estimates of average and peak indoor exposures for the general California population as well as certain subgroups of that population such as individuals who may be highly sensitive to indoor pollutants. The model described in this document--California Population Indoor Exposure Model, Version 2 (CPIEM 2.0)--is a software product that has been designed to expedite the exposure-assessment process by providing a user interface and calculation tools for supplying and integrating all required information. It is an enhanced version of the original CPIEM 1.4F released by ARB in 1998.

Enhancements include greatly improved ease of use through a Windows interface, superior graphic outputs, an updated default database, as well as enhanced and new calculation capabilities, including uncertainty analysis. The CPIEM is a software tool that combines:

- Air pollutant concentration distributions for several microenvironments, including outdoors, and
- Population activity patterns that specify time spent in each microenvironment

in a Monte Carlo framework to predict distributions of exposure concentrations for the California population. The default databases of microenvironment concentration distributions and activity patterns are specific to California, but the model allows the user to easily add his or her own data as well.

For many air pollutants, the indoor concentration data are either sparse or nonexistent. To address this limitation, and to provide a means of evaluating hypothetical exposure reduction activities, the CPIEM also includes a mass-balance algorithm so that the user can estimate indoor concentration distributions based on distributional information for parameters such as indoor source emission rates, building volumes, and air exchange rates.

The Windows platform of this new version of CPIEM greatly improves the software's efficiency and ease of use with standard, easily understood drop-down menus and dialogue boxes. The graphic outputs are presentation quality. Scenarios are easily saved and edited to facilitate sensitivity analysis. The default databases have been updated with more recent data on indoor and outdoor pollutant concentrations, mass-balance parameters, and the demographic composition of California's population.

The exposure distributions predicted by CPIEM reflect the variability of exposure concentrations across population groups, but not our uncertainty about them. A new supplementary software program, designed to be used in conjunction with the new CPIEM, facilitates the estimation of the uncertainty of these exposure distributions. The uncertainty supplement creates alternative distributions for the CPIEM input variables with Monte Carlo sampling to reflect our uncertainty about the parameters of the input distributions. The user provides each alternative to CPIEM for iterative simulations. At the conclusion of the simulations of the alternatives the uncertainty supplement combines the resulting exposure distributions to estimate uncertainty distributions for selected percentile values.

Other enhancements to the CPIEM capabilities include refinement of the pollutant removal process calculation in the mass-balance algorithm, and additional output metrics.

*This page intentionally left blank.*

# Executive Summary

The primary mission of the ARB Indoor Program is to identify and reduce Californians' exposures to indoor pollutants. To fully consider indoor exposures in assessing risk, the ARB needs estimates of average and peak indoor exposures for the general California population as well as certain subgroups of that population such as individuals who may be highly sensitive to indoor pollutants. The model described in this document--California Population Indoor Exposure Model, Version 2 (CPIEM 2.0)--is a software product that has been designed to expedite the exposure-assessment process by providing a user interface and calculation tools for supplying and integrating all required information. It is an enhanced version of the original CPIEM 1.4F released by ARB in 1998.

The primary function of the CPIEM software is to combine indoor-air concentration distributions with Californians' location/activity profiles to produce exposure and dose distributions for different types of indoor environments. This function, referred to as Level 1-2 of the model, is achieved through a Monte Carlo simulation whereby a number of location/activity profiles that were collected in prior ARB-sponsored surveys are combined with airborne concentrations for specific types of environments (e.g., residences, office buildings). For many compounds, the concentration data are either limited or nonexistent. Consequently, a second function of the model (Level 3) is to estimate indoor-air concentration distributions based on distributional information for mass-balance parameters such as indoor source emission rates, building volumes and air exchange rates.

The goals of this project were to enhance the original version of CPIEM, implemented as an MS-DOS program for personal computer, in order to improve the accuracy of estimates of the exposure of Californians to air pollutants, to enhance the characterization of uncertainty and variability or the estimates, and to upgrade the underlying technology to take advantage of Microsoft Windows.

The specific objectives of the project were threefold.

- Identify, review, and incorporate new default data into CPIEM. New data includes indoor concentration distributions, outdoor concentration distributions, building air exchange rates, pollutant penetration factors, pollutant reactivity factors, and pollutant adsorption factors.
- Identify and incorporate improvements to the CPIEM estimation capabilities. New capabilities include an uncertainty module designed to be used in conjunction with CPIEM, various adjustments to activity pattern weights, disaggregation of the removal rate term of the mass-balance equation to represent various processes, and additional exposure statistic (time-weighted average exposure concentration).
- Identify and incorporate improvements to the efficiency and ease of use of the CPIEM by converting CPIEM from QuickBasic to VisualBasic, improving user interfaces, and improving output reports.

### New Input Data

At the direction of ARB new indoor concentration data was limited to studies conducted in California. A total of 24 concentration distributions were added covering residential, office, school, vehicle, and public access building microenvironments. The pollutants addressed with the new distributions include formaldehyde, benzene, trichloroethylene, perchloroethylene, benzo(a)pyrene, carbon monoxide, nitrogen dioxide, PM10, ozone, 1-3, butadiene, and MTBE.

At the direction of ARB new mass balance parameter data on air exchange rates and emission factors for consumer products were limited to studies in California, but new data for penetration factor and removal rates were not. A total of 10 distributions were added to the default database covering the parameters of adsorption, reactive decay, penetration, and air exchange rates.

The outdoor concentration database in the original version of CPIEM includes measurements from the late 1980's to the early 1990's. These were updated with more recent measurements.

Daily averages for selected air pollutants measured in California between 1997 and 1999 were taken from the ARB monitoring network, as well as data from the San Francisco Bay Area Air Quality Management District (BAAQMD) toxics monitoring network, and the South Coast Air Quality Management District (SCAQMD) toxics monitoring network. Normal and lognormal distributions were fitted to the daily averages, and the best fitting distribution selected. A total of 96 distributions were added to the outdoor concentration default database, covering benzene, formaldehyde, benzo(a)pyrene, chloroform, trichloroethylene, and perchloroethylene. A distribution was provided for each pollutant/region/year combination, for each pollutant/region combination for the overall 3-year period, for each pollutant/year combination throughout the state, and for each pollutant for the overall 3-year period throughout the state.

## New Estimation Capabilities

An additional software program, UNC1.0, is also provided to facilitate the estimation of uncertainty of the exposure distributions predicted by CPIEM2.0. UNC1.0 was designed to be used in conjunction with CPIEM2.0. For input variable distributions that are uncertain, UNC1.0 creates sets of input distribution parameters with Monte Carlo sampling. Each set is provided to CPIEM2.0 for iterative simulations. At the conclusion of the simulations UNC1.0 combines the resulting exposure distributions to estimate uncertainty distributions for selected percentile values.

The original CPIEM provided only a single set of weights for the population activity patterns when constructing exposure and dose distributions. CPIEM 2.0 has been enhanced so that the user may select a set of activity pattern weights from the following choices:

- No weights, i.e. all weights are equal.
- TIMEWT, the original model default, which adjusts for deliberate oversampling of certain populations and day-types.
- SAMPWT, which adjusts for deliberate oversampling of certain populations.

The original weights, which reflect the population structure at the time the activity surveys were taken, are provided in the default file, POP.mdb. However, a second file of activity patterns has been provided with post-stratification values for TIMEWT and SAMPWT (POP_NEW.mdb). The age and gender post-stratification adjustments were computed using California Department of Finance year 2000 projected population counts. Age groups were chosen as 0-4, 5-11, 12-17, 18-29, 30-39, 40-49, 50-65, and 66 or greater. The user may substitute this file of more recent data for POP.mdb by re-naming it as POP.mdb. (To avoid overwriting the original POP.mdb, first save it under another name.) In addition, the user is now able to supply his or her own set of weights by providing a properly formatted file in the same subdirectory and re-naming it to POP.mdb.

The original version of provided only a single factor, *k*, to represent pollutant removal in the mass balance algorithm of the Level 3 module. However, there are several removal processes that pertain to various pollutants, including reactive decay, deep adsorption, and deposition. CPEIM 2.0 allows the user to specify a distribution for one or more of these processes to be used in the mass-balance algorithm.

The original version of CPIEM provided only 2 output metrics: integrated exposure, measured in units of $\mu$g-h/m$^3$, and dose, measured in units of $\mu$g. CPIEM 2.0 provides an additional output metric of time-weighted average exposure, measured in units of $\mu$g/m$^3$.

## Efficiency Improvements

In order to take advantage of newer technology to improve efficiency and ease of use the CPIEM computer code from a QuickBasic platform to a Visual Basic\Windows platform. The Windows platform of this new version of CPIEM greatly improves the software's efficiency and ease of use with standard, easily understood drop-down menus and dialogue boxes. The graphic outputs are presentation quality. Scenarios are easily saved and edited to facilitate sensitivity analysis.

Several characteristics of the QuickBasic version of CPIEM presented problems for porting to the new platform, such as the extensive use of global variables, the lack of modularity of the source code, and the lack of documentation within the code. These problems were addressed in re-writing the CPIEM software by the separation of the 3 application tiers, i.e. User Interface, Calculation Engine, and Database. To achieve this objective the software code was systematically disassembled and re-constructed in a modular fashion, retaining calculation and other core algorithms in the model as necessary. The same data structure for the existing dBase input files was maintained to minimize introduction of new changes and additional testing required to verify the new modifications. However, the bulk of input files were consolidated into Access databases where necessary and efficient, and additional new tables were created to link and store scenario runs and maintain application level settings.

A systematic identification and repair all of the potential problems in the original QuickBasic product was not attempted, although a number of problems discovered in the course of translating and documenting the existing code were repaired. The Visual Basic product was verified to demonstrate that the key elements of the original model have been ported effectively by duplicating results obtained by the original code for the same inputs. After the addition of the enhancements, additional testing was conducted to assure that the extensions were implemented correctly.

The new Windows-based user interface closely replicates the original one, but provides improved reporting capabilities. Specific improvements include:

- More legible model reports, suitable for reproduction in color or in black and white.
- Elimination of the check-mark model for "visiting" data screens.
- Implementation of field-level data validation checks to block users from entering values in the wrong formats.
- Creation of an installation program to install the complete model on a target PC running Windows.

## Validation with Uncertainty Estimation

In addition to the replication/verification tests for CPIEM 2.0, additional verification tests were conducted for the new uncertainty module to demonstrate that it is performing correctly. In addition, one of the validation tests conducted for the original version CPIEM was repeated. Using the original version of CPIEM, the significance of any discrepancies in the model predictions and observed values could be judged only subjectively. However, utilization of the new uncertainty module provides uncertainty bounds for the model output statistics. It can then be determined whether the observed values fall within the model uncertainty bounds. For this validation test, uncertainty was characterized by sampling uncertainty only, i.e., the portion of uncertainty related to the limited size of the sample on which the parameter distribution is based. The result showed that most of the observed values did fall within the predicted uncertainty bounds, but there were some exceptions, perhaps due to an underestimation of the total uncertainty.

## Recommendations for Future Refinements

Several further enhancements of CPIEM are recommended for future work. These are development of default distributions for breathing rates by a methodology provided in Appendix A, incorporation of additional California activity pattern data as available from US EPA's Consolidated Human Activity Database (CHAD), expanded options for characterization of data input distributions, and provide several refinements to the uncertainty module, including integration into the CPIEM software to create a seamless package.

# Table of Contents

## Table of Figures



## Table of Tables

# Glossary

| | |
|---|---|
| Absorption - | the process by which a contaminant penetrated the exchange boundaries of an organism after contact. |
| Adsorption - | adhesion of the molecules of a compound to a surface. |
| Activity pattern - | an accounting of how an individual's time over a defined period (e.g., hour or day) is spent in various types of environments at various levels of activity or exertion, i.e., location/activity profile. |
| Air exchange rate - | the rate at which air is exchanged between the airspace in an indoor environment and the surrounding outdoor airspace, expressed as the indoor/outdoor airflow rate ($m^3$/h) divided by the indoor volume ($m^3$), with resultant units of inverse time (1/h). |
| Breathing rate - | the rate at which an individual breathes or inhales air, expressed in units of volume/time (e.g., $m^3$/h); also called inhalation rate. |
| Concentration - | the extent of occurrence of a pollutant in air, expressed in terms of pollutant mass per unit volume of air (e.g., μg/$m^3$), or as parts of the pollutant per billion parts (ppb) or per million parts (ppm) of the air-pollutant mixture, both by volume. |
| Conservation of mass - | a principle stating that decreases in pollutant mass in a defined airspace (e.g., indoors) are equally compensated by corresponding increases in mass in other airspaces (outdoors) or media (indoor sinks) to which the pollutant is transported. |
| Decay rate - | the first-order rate of reduction in the indoor-air concentration of a pollutant, due to physical/chemical reactions in air or with indoor sinks, expressed in units of inverse time (e.g., 1/h). |

Deposition rate - the rate of removal of airborne substances to available surfaces that occurs as a result of gravitational settling and diffusion, as well as electrophoresis and thermophoresis.

Dose (inhaled) - the integral over time of the product of a pollutant concentration times a breathing rate, expressed in units of mass (e.g., μg).

Emission rate - the rate at which an indoor source emits a pollutant into the indoor airspace, expressed in units of mass/time (e.g., μg/h).

Environment - a type of place or building where an individual spends time, such as a residence, school or office building.

Exposure - the contact at one or more boundaries (e.g., mouth or skin) between a human and a pollutant at a specific concentration for a period of time; the three principal routes of exposure are inhalation of air (the subject of this report), ingestion of food or liquids, and dermal contact.

Factor k - a parameter used in the CPIEM level 3 module that describes the rate of removal of airborne pollutants from the atmosphere due to chemical decay, net adsorption (adsorption minus de-adsorption), or net deposition (deposition minus re-entrainment). See indoor sink.

Indoor-air model - an equation, algorithm, or series of equations/algorithms used to calculate the average or time-varying pollutant concentration in an indoor environment for a specific situation.

Indoor sink - a material, furnishing or appliance used or installed indoors that causes the indoor-air concentration of a pollutant to decrease by processes such as deposition, adsorption or chemical reaction.

Indoor source - a product, material, appliance or activity indoors that causes the indoor air concentration of a pollutant to increase.

Integration period - the time period (e.g., hour or day) over which inhalation exposure or inhaled dose is mathematically integrated.

Latin HyperSquare Sampling –

a method of sampling from one or more probability distributions that cuts each distribution into k slices and selects exactly one value from each slice.

Level 1-2 - the primary function of the CPIEM software, which is to combine indoor-air concentration distributions with location/activity profiles to produce exposure and dose distributions for different types of indoor environments.

Level 3 - a secondary function of the CPIEM software, which is to estimate indoor-air concentration distributions based on distributional information for mass-balance parameters such as indoor source emission rates, building volumes and air exchange rates.

Loading - a type of indoor emission source for which the quantity of relevant material present is specified in relation to the building volume, e.g., square feet of carpet per cubic meters of indoor space. (See no loading.)

Location/activity profile - an accounting of how an individual's time over a defined period (e.g., hour or day) is spent in various types of environments at various levels of activity or exertion, i.e., activity pattern).

Mass-balance equation - a differential equation, based on the conservation of mass, stating that changes in pollutant mass indoors over time are related to gains (from indoor sources or from outdoor mass transported indoors) and losses (indoor mass transported to outdoors or to indoor sinks).

Mitigation - an action or series of actions intended to reduce the concentration of a pollutant indoors.

Model evaluation - a series of steps through which a model developer or user assesses a model's performance for selected situations.

Model parameter - a mathematical term in an indoor-air model that must be estimated by a model developer or user before model calculations can be performed.

Model validation - a series of evaluations undertaken by an agency or organization to provide a basis for endorsing a specific model (or models) for a specific application (or applications).

Model verification - a series of checks to ensure that model logic has been correctly programmed and that model calculations are mathematically correct.

Monte Carlo simulation - a stochastic process by which values are repeatedly sampled from distributions for various model parameters and used to estimate an indoor-air or exposure model; each repetition of the process is called a trial.

Net deposition rate - first-order net rate of pollutant loss from the atmosphere due to the combined processes of deposition and re-entrainment, expressed in units of inverse time (e.g., 1/h).

Net surface adsorption rate - first-order net rate of pollutant loss from the atmosphere due to the combined processes of adsorption and de-adsorption, , expressed in units of inverse time (e.g., 1/h).

No loading a type of indoor emission source for which the quantity of relevant material present is specified in absolute terms rather than in relation to the building volume, e.g., fuel burned in a pilot light. (See loading.)

Penetration factor - the fraction of the outdoor pollutant concentration that bypasses the outer envelope of an indoor environment and enters the indoor airspace.

Post-stratification weights- a method of re-weighting survey responses to reflect changes in the population demographics (e.g. age and gender distributions) since the survey was carried out.

Reactive decay rate - the first-order rate of reduction in the indoor-air concentration of a pollutant, due to chemical transformation, expressed in units of inverse time (e.g., 1/h).

Scenario - a complete set of input values and parameters required by CPIEM for a simulation

Seed Number - a number used by a computer program to generate a sequence of pseudo-random numbers; if the same seed number is used then identical sequences will be generated.

Sensitivity analysis - an analysis of a model that examines how the model predictions vary when selected input values vary.

Uncertainty analysis - an analysis of a model such that the model inputs are assigned probability distributions representing the user's knowledge of their true value and the resulting model predictions are used to develop probability distributions representing the user's knowledge of the true output values.

*This page intentionally left blank.*

# 1. Background

The goal of the California Air Resources Board's (ARB) Indoor Air Quality and Personal Exposure Assessment Program is to identify and reduce Californians indoor and personal exposures to air pollutants, including toxic air contaminants. To meet these goals, the ARB needs accurate estimates of Californians indoor exposures to air pollutants. In recent years ARB has sponsored studies of activity patterns of California residents and indoor environmental data, as well as the development of a methodology that combines those data and outdoor air quality data to estimate indoor and total exposures to pollutants. These data and methodologies are embodied in the California Population Indoor Exposure Model (CPIEM). CPIEM, which includes primarily California-specific data, can be used by ARB to both estimate exposure of Californians to several air pollutants and to assess the effectiveness of possible mitigation measures in reducing exposures.

CPIEM consists of two modules, both of which use Monte Carlo simulation to produce frequency distributions of results, rather than point estimates. The first module combines indoor air concentration distributions with population activity data to estimate population indoor exposure and inhaled dose. If additional data on outdoor concentration distributions are used, the module estimates total air pollutant exposure and dose. The second module combines data on indoor source emissions, outdoor air quality, building air exchange rates, pollutant penetration, and pollutant removal factors to estimate indoor concentration distributions, which can be used as input to the first module. The original version of CPIEM, CPIEM 1.4F, was implemented as an MS-DOS program for personal computer (Koontz et al. 1998).

The goals of this project were to enhance the original version of CPIEM in order to improve the accuracy of estimates of the exposure of Californians to air pollutants, to enhance the characterization of uncertainty and variability or the estimates, and to upgrade the underlying technology to take advantage of Microsoft Windows.

The specific objectives of the proposed project were threefold.

- Identify, review, and incorporate new default data into CPIEM. New data includes indoor concentration distributions, outdoor concentration distributions, building air exchange rates, pollutant penetration factors, pollutant reactivity factors, and pollutant adsorption factors.
- Identify and incorporate improvements to the CPIEM estimation capabilities. New capabilities include and an uncertainty module designed to be used in conjunction with CPIEM, various adjustments to activity pattern weights, disaggregation of the removal rate term of the mass-balance equation to represent various processes, and an additional exposure statistic (time-weighted average exposure concentration).
- Identify and incorporate improvements to the efficiency and ease of use of the CPIEM by converting CPIEM from QuickBasic to VisualBasic, improving user interfaces, and improving output reports.

These modifications will make it easier for ARB to address their goals of assessing human exposure to air contaminants in indoor environments, identifying the relative contribution of indoor human exposure to total human exposure, and evaluating potential mitigation measures for indoor microenvironments.

*This page intentionally left blank.*

# 2. Overview of CPIEM Structure and Algorithms

In this section we will briefly summarize the CPIEM structure and algorithms used to calculate the population exposure distributions. Because the structure and algorithms are unchanged from the original version of CPIEM (with a couple of exceptions noted), a more detailed discussion of them may be found in the model development document for the original version (Koontz et al., 1998).

The primary objective of the CPIEM is to produce estimates of exposure and dose distributions by combining indoor-air concentration distributions with location/activity profiles of California residents. In order to accomplish this objective, the model contains the following three levels of capabilities:

- **Level 1**—provide estimates of total indoor exposure distributions by aggregating environment-specific exposure estimates (see Level 2). This capability is combined with the Level 2 capability into a single module, and thus is not separately accessible.
- **Level 2**—provide estimates of indoor exposure and dose distributions for specific indoor environments by combining indoor concentration data with population activity data.
- **Level 3**—provide estimates of indoor-air concentration distributions based on factors such as indoor source emissions, building volumes, and indoor-outdoor air exchange rates.

The CPIEM has two major components—calculation of exposure/dose distributions and calculation of indoor-air concentration distributions. The model provides a user-friendly interface for making choices and providing inputs at each level. This Windows-based interface is described in detail in the User's Guide for the software. It also provides defaults where possible and is equipped with summaries of distributional information for various inputs, in cases where such information is known to exist.

## 2.1. Level 1-2 Module

The exposure/dose module is called the Level 1-2 module because it simultaneously calculates exposure and dose for all selected environments while providing estimates of “total indoor air” exposure and dose across these environments. Thus, Level 1-2 is a single module with all outputs produced at the same time; Level 1 is not a separately accessible function. Level 1-2 of the model uses measured or modeled concentration distributions for one or more environments, together with location/activity patterns (i.e., amount of time spent in each environment at specific activity levels), to calculate exposure and inhaled-dose distributions (Level 2) for the chosen environment(s).

The Level 1-2 model carries out the primary function of the CPIEM, which is to combine indoor-air concentration distributions with Californians' location/activity profiles to produce exposure and dose distributions for different types of indoor environments. This function is achieved through a Monte Carlo simulation whereby a number of location/activity profiles that were collected in prior ARB-sponsored surveys are combined with airborne concentration distributions for specific types of environments such as residences and offices. Concentration values for a given environment are sampled from user-selected distributions, either provided in the CPIEM database or supplied by the user. These distributions are specified for various pollutants/integration time combinations. The model uses integration periods of 1, 8, 12 (specified as 12AM or 12PM) or 24 hours. The exposure/dose calculations are performed for

either the general California population or a subset of that population, and are performed for up to nine different types of environments.

Inputs for the Level 1-2 module include the population subgroup for which exposures are to be estimated, environment-specific concentration distributions, and breathing rates for adult males, adult females and children. The model outputs include (1) graphs of the differential (density) and cumulative probability distributions, (2) summary statistics for the distributions, (3) an output file containing additional environment-specific and total exposure/dose distribution statistics, and (4) an output file containing values for each trial (person) in the simulation.

### 2.1.1. Level 1-2 Algorithms

Sampled concentration values for a given environment are multiplied by time durations in the environment, as sampled from surveyed Californians' location/activity profiles, to simulate exposure as either a time-integrated exposure concentration (e.g., $\mu$g-hr/m$^3$) or a time-weighted average exposure concentrations (e.g., $\mu$g/m$^3$). Multiplication of integrated exposure values by breathing rates determined from the location/activity profiles and pulmonary ventilation data yields an estimate of the distribution of potential-inhaled-dose (e.g., $\mu$g) for each modeled environment. The model then aggregates the environment-specific exposure and dose estimates to develop distributions of "total indoor air" exposures and doses (Level 1); that is, the contribution to total (24-hour) exposure /dose associated with time spent indoors. Because the outdoors is included as one of the environments for the model, it is also possible to simulate total (24-hour) exposure and dose distributions.

The time-integrated exposure concentration encountered by an individual while in an indoor environment (National Academy of Sciences 1991; USEPA 1989) is given by:

$$C_T = \int_0^T C(t)dt \qquad (2\text{-}1)$$

where $C(t)$ is the concentration in the environment at time $t$, $T$ is the amount of time spent in the environment, and $C_T$ is the time-integrated concentration. If the concentration is measured in $\mu$g/m$^3$ and time in hours, then the units for $C_T$ are $\mu$g-hr/m$^3$.

An enhancement to the outputs of the model, described in Section 4 is the time-weighted average exposure concentration.

Potential inhaled dose is defined as "an exposure multiplied by rate and assumes total absorption of the contaminant" (National Academy of Sciences 1991; USEPA 1989). This can be mathematically represented as the time-integrated product of the exposure concentration and the individual's breathing rate (i.e., amount of air inhaled per unit time while in the environment):

$$D_T = \int_0^T B(t)C(t)dt \qquad (2\text{-}2)$$

where $B(t)$ is the breathing rate at time $t$ and $D_T$ is the potential inhaled dose over the time duration $T$. If the breathing rate is in units of m$^3$/h and the units for $C_T$ are as above, then $D_T$ is expressed in $\mu$g (= m$^3$/h $\times$ $\mu$g-h/m$^3$). If the breathing rate is assumed to be constant and this constant rate is expressed as $\underline{B}_T$, then the potential inhaled dose can be expressed as:

$$D_T = \underline{B}_T \int_0^T C(t)dt = \underline{B}_T \times C_T \qquad (2\text{-}3)$$

For the model, the average breathing rate while in the environment is assigned from activity codes contained in each location/activity profile; this assignment is conditional on the individual's age/sex category—adult male, adult female or child (i.e., under age 12).

## 2.2. Level 3 Module

For many compounds, data on measured concentrations are either limited or nonexistent. Consequently, a second function of the model is to estimate indoor-air concentration distributions based on distributional information for mass-balance parameters such as indoor source emission rates, building volumes and air exchange rates.

The concentration module (Level 3 of the model) utilizes a mass-balance equation, based on the principle of conservation of mass, to estimate concentration distributions for specific types of indoor environments such as residences, offices and schools. This module samples values from user-specified distributions for parameters such as emission rates for indoor sources, building volumes, outdoor-air concentrations and indoor-outdoor air exchange rates, which are used as inputs to the mass-balance equation. The output from the concentration module can be used as one of the inputs to the exposure/dose module, and would be particularly useful for pollutant-environment combinations for which concentration data are not available from field monitoring studies.

Inputs for concentration calculations include indoor sources, outdoor concentrations, building penetration factors, indoor sinks, building volumes and indoor-outdoor air exchange rates. The model outputs include (1) graphs of the differential (density) and cumulative probability distributions, (2) summary statistics for the distributions, (3) and output file containing additional distribution statistics, and (4) an output file containing daily and hourly average concentration values for each trial (building such as a residence) in the simulation.

For modeling purposes, indoor sources are classified into three types—long-term, episodic and frequent. Long-term sources, such as interior finishings, furnishings and some appliances, tend to be relatively static features of buildings. Many of these sources, because they contain a finite amount of pollutant material that can be emitted (as in the case of interior finishings that offgas volatile organic compounds), emit at a declining rate over time as the reservoir of available material is gradually depleted. Thus, distributions for the initial emission factor ($\mu$g/hr per quantity), the rate of decline in emission factor ($months^{-1}$), and the duration since installation (months) must be specified by the user so that the model can calculate the emission factor at a given time. Additional parameters specified by the user are the percentage of buildings with sources present and a distribution of the quantity or load present in each building.

Episodic sources typically are used or present on a weekly, monthly or less frequent basis; some examples are carpet cleaning, painting and bringing home dry-cleaned clothes. There are two types. The first type of episodic source (e.g., painting) is typically characterized by an initial emission rate at the time of use that subsequently declines, like long-term sources. Thus, the inputs specified by the user are similar to those for long-term sources, but with a time scale of days rather than months. These are the percentage of buildings with episodic sources present, and distributions for the initial emission factor ($\mu$g/hr per quantity), the rate of decline in emission factor ($days^{-1}$), and the time since use (days) and the quantity used or present in each building. The second type of long-term source is one that occurs in a relatively short time period (e.g., an hour or two), like carpet cleaning. For this type of source the user specifies a rate of decline for the emission factor of zero. In both cases an additional input of duration of use (hr) is specified.

Frequent sources tend to be used on a daily basis, often more than once a day; cooking, showering and tobacco smoking are good examples. The time scale of the parameters for frequent sources is much shorter than for long-term of episodic sources, and the episodes can overlap. The inputs specified by the user are the percentage of buildings with the source present, and distributions for the initial emission factor (μg/hr per quantity), the rate of decline in emission factor ($hrs^{-1}$), rate of use or quantity used (quantity/min), duration of episode (min), and the number of episodes per day. Other inputs are the distribution of start hours per episode (% share for each hr), and whether overlapping episodes are allowed.

### 2.2.1. Level 3 Algorithms

The fundamental relationship used to estimate the microenvironmental concentration is as follows (see section 6 of Koontz et al., 1998 for a detailed development).

$$\overline{C_i} = \frac{-C_{i-1}}{LT}(1-e^{LT}) - \frac{G}{T(L^2)}(1-e^{LT}) - \frac{G}{L} \qquad (2\text{-}4)$$

where

$\overline{C_i}$ = the average concentration for time period *i*;

$C_{i-1}$ = the concentration at the endpoint of the previous time period;

$L$ = -(*k*+*a*), where *k* is the first-order loss rate and *a* is the air exchange rate;

$G$ = $S/V + pa\ C_{out}$, where *S* is the indoor source rate, *V* is the compartment volume, *p* is the penetration factor, and $C_{out}$ is the outdoor concentration; and

$T$ = the averaging time.

The new version of this relationship disaggregates the first-order loss rate, *k*, into three separate terms, as described in Section 4.

# 3. Addition of New Data

## 3.1. Microenvironment Concentration Data

ICF identified, acquired and reviewed publications to determine whether they contain indoor concentration data that could potentially be added to the CPIEM default database. At the direction of ARB new data was limited to studies conducted in California. We considered only data for which the distribution form and defining parameters were explicitly specified in the document.

Table 3-1 lists those studies identified that met the criteria for data inclusion. The last column in the table indicates the format in which the data were incorporated. All concentration units were converted to those currently used in CPIEM, as necessary, before incorporation.

In addition a reference number was provided for each default data set, which is cross-referenced to the literature citations in the User's Guide.

## 3.2. Mass Balance Parameter Data

ICF also identified, acquired, and reviewed publications to determine whether they contain data that could potentially be added to the CPIEM data base pertaining to the mass balance parameters for the level 3 module, i.e., ventilation, filtration, building characteristics, pollutant penetration factors, and pollutant removal rates. New data for penetration factor and removal rates were not limited to California studies, but new data on air exchange rates and emission factors for consumer products were limited to studies in California. We considered only data for which the distribution form and defining parameters were explicitly specified in the document, or could be easily derived.

Table 3-2 lists those studies identified that met the criteria for data inclusion. The last column in the table indicates the format in which the data were incorporated.

Although no particle deposition rate data that met the inclusion criteria were identified in time to incorporate into CPIEM2.0, such data can easily be added to the database by the user when identified. The process is explained in the CPIEM2.0 User's Guide.

In addition a reference number was provided for each default data set, which is cross-referenced to the literature citations in the User's Guide Appendix B.

## 3.3. Outdoor Concentrations

The outdoor concentration database in the original version of CPIEM includes measurements from the late 1980's to the early 1990's. For this version more recent measurements were added.

Daily averages for selected air pollutants measured in California between 1997 and 1999 were taken from the ARB monitoring network, as well as data from the San Francisco Bay Area Air Quality Management District (BAAQMD) toxics monitoring network, and the South Coast Air Quality Management District (SCAQMD) toxics monitoring network. Normal and lognormal distributions were fitted to the daily averages, and the best fitting distribution selected.

### 3.3.1. Details

Daily averages for the following selected air pollutants measured in California between 1997 and 1999 were taken from the ARB monitoring network, as well as data from the San Francisco Bay Area Air Quality Management District (BAAQMD) toxics monitoring network, and the South Coast Air Quality Management District (SCAQMD) toxics monitoring network:

- Benzene
- Benzo(a)pyrene (total PM10 and vapor)
- Chloroform
- Formaldehyde
- Perchloroethylene
- Trichloroethene.

Half the detection limit was used to substitute for values below the detection limit.

The data was subset by region and by year. The years were 1997, 1998, 1999, and all there years combined. The regions were:

- San Francisco Bay Area: All monitors in the Bay Area Air Quality Management District.
- South Coast: All monitors in the South Coast Air Quality Management District or in San Diego County.
- Other: All other monitors.
- All: All monitors in California.

These regions are as defined in the 1991 California activity pattern surveys used in CPIEM (Wiley et al. 1991a, b).

Normal and lognormal distributions were fitted to the sets of daily averages stratified by pollutant, region and year. The better fitting of the two distributions was determined using the Shapiro-Wilk W statistic for 2000 or fewer daily averages, and using the Kolmogorov-Smirnov D statistic for more than 2000 averages. The Shapiro-Wilk W statistic approximately measures the squared correlation between the observed and fitted order statistics on a normal probability plot, so that higher values of W indicate a better fit; a perfect fit is when W =1. SAS software does not provide W for sample sizes above 2000. The Kolmogorov-Smirnov D statistic measures the maximum difference between the observed and fitted cumulative distribution function, so that lower values of D indicate a better fit; a perfect fit is when D = 0.

In all but one case, the lognormal fitted better than the normal. The only exception was for formaldehyde in 1998 in the South Coast region, although in that case the histograms show that both distributions fit the data well. For trichloroethene in the San Francisco Bay Area in 1999, all 114 daily averages were equal (probably because they were all below the detection limit), and so the standard deviation is zero.

In general, the results show that the lognormal distribution fits the benzene and formaldehyde data very well. The lognormal fit is moderately good for perchloroethylene and trichloroethene. Both distributions fit the B(a)P and chloroform data poorly, although the lognormal fit is slightly

better. In general, the fitted distributions are similar from year to year but substantially different from region to region. Therefore, for modeling purposes, it should not matter too much which year or year group is selected, but it is important to select the most appropriate region for the modeling application. The CPIEM can be applied using the four alternative regions to evaluate the sensitivity of the results to the outdoor concentration region.

## Table 3-1
## Summary of Studies Added That Include Microenvironment Concentration Measurements in California.

| Title | Author(s) | Pollutants Addressed | Concentration / Distribution | Averaging Time | Environment/ Comments | Recommendations |
|---|---|---|---|---|---|---|
| Modeling Ozone Levels In and Around Southern California Homes | Avol, E.L., Navidi, W.C., and Colome, S.D., 1998, *Environ. Sci. & Technol.*, 32: 463-468. | ozone | Mean and distribution | 24 hours | Simultaneous measurement of indoor and outdoor ozone in 126 southern California **homes**. | Concentration distribution, as percentiles, specified for southern CA in summer. |
| Benzene and Toluene Concentrations Inside and Outside of Homes in California | Colome, S.D, Fung, K., Behrens, D.W., Billick I.H., Tian, Y., and Wilson A.L., 1994, Presented at the A&WMA's 87th Annual Meeting and Exhibition. Cincinnati, OH. 94-WP90.03. | benzene and toluene | Arith means, std, devs,, and percentiles by utility district | 48 hours | PGE&E, SCE, and SD utility districts. | Concentration distributions, as data sets, for CA residences by region for benzene. |
| Volatile Organic Compounds in 12 California Office Buildings: Classes, Concentrations and Sources | Daisey, J.M., Hodgson, A.T., Fisk, W.J., Mendell, M.J., and Ten Brinke, J., 1994, *Atmos. Environ..,* 28: 3557-3562, 1994. | VOCs including benzene and trichloroethylene *(see attachment A)* | Mean + S.D. | 8 hours | Reports 39 individual VOCs measured in 12 northern California city and county **office buildings** of three different ventilation types. | (Not a probability sample) concentration distribution, as lognormal, for northern CA offices for benzene and trichloroethene. |
| A Pilot Study to Measure Indoor Concentrations and Emission Rates of Polycyclic Aromatic Hydrocarbons | Offermann, F.J., Loiselle, S.A., Hodgson, A.T., Gundel, L.A., and Daisey, J.M., 1990, In: *Indoor Air '90, Proceedings of the International Conference in Indoor Air Quality and Climate, Ottawa, Canada,* 2: 379-384. | PAHs (*see attachment B).* | Average concentration | 12 hours | Simultaneous indoor and outdoor meas of PAHs in **homes and offices** (1) without active combustion sources and (2) with controlled sources of gas/wood stoves and tobacco smoking. | (Not a probability sample) PAH and BaP concentration distributions, as data sets, for CA residences and commercial buildings. |

| Title | Author(s) | Pollutants Addressed | Concentration / Distribution | Averaging Time | Environment/ Comments | Recommendations |
|---|---|---|---|---|---|---|
| Epidemiological Investigation to Identify Chronic Health Effects of Ambient Air Pollutants in Southern California | Peters, J.M. 1997, *Epidemiological Investigation to Identify Chronic Health Effects of Ambient Air Pollutants on Southern California*. USC-LA Final report to CARB NTIS No. PB 98-140833/XAB. | formaldehyde | Mean, SD, Median, Min., and Max | | Measurements are reported for **residences**. | Concentration distributions, as normal or lognormal, to the CPIEM default database for residences. |
| Measuring Concentrations of Selected Air Pollutants Inside California Vehicles | Rodes, C., L.Sheldon, D.Whitaker, A.Clayton, K, Fitzgerald, J. Flanagan, F.DiGenova, S. Hering, and C. Frazier., 1998, Prepared for the California Air Resources Board, Contract No, 95-339. | several VOCs, CO, PM10, PM2.5, several elements *(see attachment C)* | Raw data | 120 minutes (2 hours) | Sacramento and Los Angeles—2 test vehicles under various commute conditions. | *(Not a probability sample.)* Concentration distributions, as data sets, for CA vehicles by region for benzene, CO, formaldehyde, and PM10. |
| Assessing the Indoor Air Impact from a Hazardous Waste Site: A Case Study | Underwood, M.C., 1996, *Toxicology and Industrial Health*, 12: 179-188. | Benzene, tetrachloroethylene, trichloroethylene, vinyl chloride | 30-min integrated samples | 30 minutes | Measurements are reported for **schools**. | *(Not a probability sample.)* Concentration distributions, as data sets, for schools. |
| Developing Baseline Information on Buildings and Indoor Air Quality | Womble, S.E., Ronca, E.L., Girman, J.R., and Brightman, H.S., 1995, Proceedings of Healthy Buildings '95 Milan, Italy, Vol. 3, 1995, pp. 1305-1310. | CO, PM10, PM2.5, formaldehyde, VOC's (only sum of VOC's given in the article) | PM10 and formaldehyde measurements for 2 CA buildings at 3 sites each (6 measurements) | 8 hours | Baseline data to characterize public and commercial **office buildings** with respect to indoor air quality and occupant perceptions. | *(Not a probability sample.)* The 6 CA values for PM10 and formaldehyde as data sets for office buildings. |

## Table 3-2
## Summary of Studies Added That Include Measurements or Estimates of Mass Balance Parameters

| Title | Author(s) | Pollutants Addressed | Mass Blanace Parameters? | Averaging Time | Environment/ Comments | Recommendations |
|---|---|---|---|---|---|---|
| Adsorption of Selected Volatile Organic Compounds on a Carpet, a Wall Coating, and a Gypsum Board in a Test Chamber | Colombo, A., De Bortoli, M., Knoppel, H., Pecchio, E., and Vissers, H., 1993, *Indoor Air* 3:276-282. | Tetrachloroethylene, $\alpha$-pinene, 1,2,4-trimethyl 2-butoxyethanol, benzene 1,4-dichlorobenzene, 2-ethylhexanol, n-dodecane, n-decane | "Deep adsorption" constants for several materials | Variable. But generally 48 hours | Estimation of "deep adsorption" rates by various materials for several VOCs, including tetrachloroethylene. | "Deep adsorption" removal rate distributions, as normal (with 0.0 std dev), for tetrachloroethylene in residences and other indoor Mes. |
| Sources of Air Pollutants Indoor: VOC and Fine Particulate Species | Lewis, C.W., 1991, *Journal of Exposure Analysis and Environmental Epidemiology* 1(1) 31-44. | PM fine + VOCs *(see Attachment D)* | Plots of infiltration vs. air change rate | 12 hours | Average concentrations of fine particle aerosol and VOC species in ten Boise, Idaho, **residences** in winter time have been apportioned according to their contributions from all inside and outside sources. | Penetration rate distribution, as a normal distribution with a mean of 1 and a std dev of 0 to characterize a single datum of unity, for VOC pollutants. |
| Characteristics of Airborne Particles Inside Southern California Museums | Ligocki, M.P., L.G. Salmon, T. Fall, M.C. Jones, W.W. Nazaroff and G.R. Cass, 1993, *Atmos. Environ*. 27A(5):697-711. | PM fine and coarse (sulfate, nitrate, ammonium, soil dust, organic matter, elemental carbon) | Air exchange rate, penetration factor, but no distributional data | 24 hours | PM fine and coarse measured inside and outside five southern California **museums** over summer and winter. | (Not a probability sample) air exchange rate distribution, as lognormal, for public access buildings. |
| Transformations, Lifetimes, and Sources of $NO_2$, HONO, and $HNO_3$ in Indoor Environments | Spicer, C.W., Kenny, D.V., Ward, G.F., and Billick, I.H., 1993, *Journal of the Air & Waste Management Association*, v 43 n 11 , p1479+. | NO2, HONO, and HNO3 | Surface reactivity removal rates | Variable | Experimental measurements in an unoccupied building to estimate removal rates. | Surface reactivity removal rates, as a lognormal distribution, for NO2. |
| California Residential Air Exchange Rates and Residence Volumes | Wilson, A.L., S.D.Colome, Y.Tian, E.W.Becker, P.E.Baker, D.W.Behrens, I.H. Billick, C.A.Garrison., *J Expo Anal Environ Epidemiol*; Vol 6(3):311-326. | N/A | **Mean and distribution** of air exchange rates | 2 days | Air exchange rates from two **residential** indoor air quality studies in southern California are presented, and one for other CA regions. | Air exchange rate distributions, as lognormal, for winter in CA residences by region. |

# 4. Enhancement of Estimation Capabilities

## 4.1. Addition of Uncertainty Distributions to Model

### 4.1.1. Inputs and Outputs

For the enhanced version of the model, we added the capability of estimating the uncertainty of model outputs based on the uncertainties of the model inputs, with a separate companion module named “UNC”. UNC allows the user to create CPIEM input files to take into account the uncertainty of the model inputs, and then performs a summary uncertainty analysis of the CPIEM output files. It also provides default uncertainty distributions that reflect sampling uncertainty of the input data, as explained below. This capability automates a process that previously had to be performed manually by a user wanting to evaluate the uncertainty of the CPIEM results.

In this section, we describe the procedures for estimating uncertainty using UNC in conjunction with CPIEM. In section 7 we discuss possible improvements to these procedures that have been postponed for implementation in a future project; these improvements include directly incorporating the capabilities of UNC into a seamless version of CPIEM that could perform uncertainty analyses without significant user intervention. A Users’ Guide for UNC is available separately.

In addition to defining distributions for the CPIEM model inputs, the user now has the option of defining the uncertainty distribution for each of those input distributions. Options for defining input uncertainty include:

- No uncertainty;
- Default uncertainty distributions;
- User-supplied continuous uncertainty distributions; and
- User-supplied discrete uncertainty distributions.

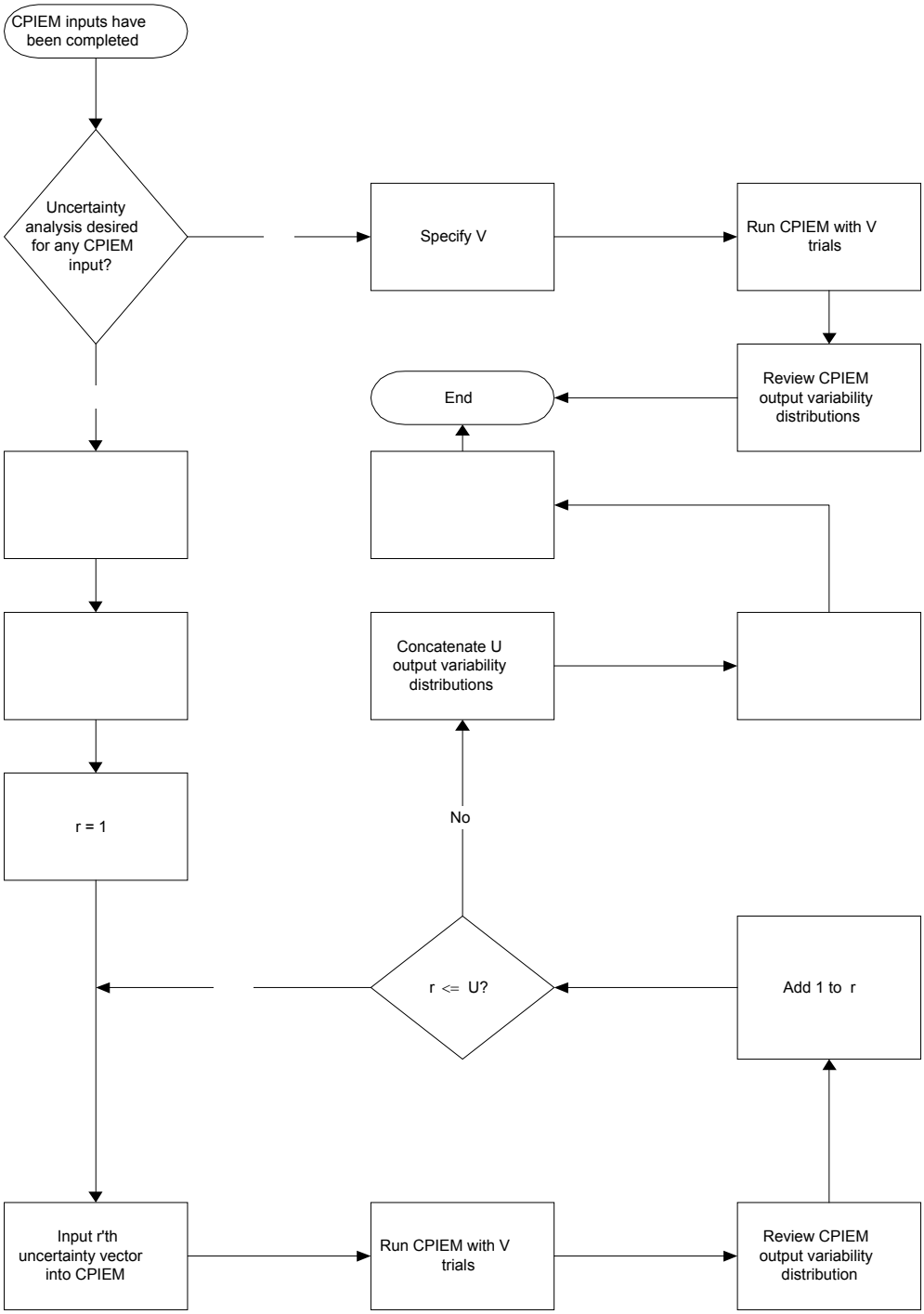
Details for how the UNC software allows the input uncertainty distributions to be specified by the user are given in the following discussion. The UNC software creates the set of CPIEM model input distributions for each of U runs of the CPIEM model. The CPIEM outputs of these U runs are then combined so that the UNC software can provide the user with graphical and statistical summaries of the uncertainty and variability of the exposure, dose, or concentration.

### 4.1.2. Model Structure

A flowchart with the basic model structure for the implementation of the variability/uncertainty analysis is shown in Figure 4-1.

To clarify the terminology, a single run of the CPIEM is made after the user provides: the pollutant; values or distributions for each CPIEM model input; the averaging period (1, 8, 12, or 24 hours) for the level 1-2 module; and the number of trials, V. For the level 3 module, each trial for each run produces 24 hourly concentration values and a daily average. For the level 1-2 module with 12- or 24-hour averaging, each trial for each run selects one individual’s activity pattern and produces an exposure or dose value for the 12- or 24-hour period. For the level 1-2

**Figure 4-1**
**Model Structure for Uncertainty/Variability**



module with 1- or 8-hour averaging, each trial for each run selects one individual's activity pattern and produces an exposure or dose value for 24 1-hour periods or 24 overlapping 8-hour periods, respectively. The number of replicates is the number of observations used to generate the CPIEM output variability distributions. For level 3 daily values and for level 1-2 12- or 24-hour averages, the number of replicates is V. For level 3 hourly values and for level 1-2 1- or 8-hour averages, the number of replicates is 24V.

For each CPIEM model input, the user must choose whether to define a unique data input distribution (e.g., normal, with a mean of 3 and a standard deviation of 1) or to define a set of two or more possible data input distributions, when the user is uncertain about which data input distribution should be applied. There may be a finite number of possible distributions (e.g., concentration distributions from different studies) or an infinite number of possible distributions (e.g., a normal distribution with an uncertain mean between 2 and 4, and a standard deviation of 1). If all the model input distributions are uniquely defined, then the model structure is the same as the previous version of CPIEM: The user will select V, the number of trials. The CPIEM Monte Carlo simulation model will be run once with V trials and the model outputs will tabulate and graph the output variability distribution If any model input is uncertain, then the user will select U, the number of uncertainty vectors, as well as V, the number of trials. The UNC model will be run to randomly generate the U uncertainty vectors. Then, the CPIEM model will be run U times, with V trials for each of the U randomly selected uncertainty vectors. The user can examine the CPIEM tables and graphs for each of the U uncertainty vectors. Then, the UNC model will be run using the compiled summary statistics files to generate a table and graph displaying the uncertainty of the variability distribution, as described below.

With this additional capability of uncertainty distributions, each CPIEM simulation consists of several iterations of the variability analysis. For each iteration, UNC randomly selects a set, or vector, of parameter values from the specified uncertainty distributions, The selected parameters define for that iteration the distributions of the model inputs, which are used by the CPIEM model algorithms to simulate a single variability distribution of concentrations or exposure/doses. The vector of uncertain parameters is then re-selected for the next iteration, and a second concentration or exposure/dose variability distribution simulated. After several iterations the result is a set of concentration or exposure/dose variability distributions, reflecting the uncertainty of the true variability distribution.

Consider the following example using the CPIEM level 3 module for a single microenvironment: The model input distribution for outdoor concentrations is chosen to be lognormal with mean 10 (no uncertainty) and standard deviation uniformly distributed from 2 to 4. The distribution for the penetration factor is chosen to be normal, with the mean and standard deviation both lognormally distributed with mean 6 and standard deviation 1. (These distributions are for illustrative purposes only and do not represent realistic assumptions about these distributions.) For the first uncertainty iteration, the standard deviation of the concentration distribution is selected from its uniform distribution—for example the value might be 3.1—and the mean and standard deviation of the penetration factor are selected from their lognormal distributions. Values for any other uncertain model input parameters are also selected. This gives a vector of model input parameters. Using this vector, the model input distributions are completely defined and can be used to simulate a set of V daily microenvironment concentration values, where V is the selected number of trials. Then, a new vector of model input parameters is generated from the uncertainty distributions and this is used to generate another set of V daily microenvironment concentration values.

More generally, the first stage (outer loop) of the simulation is to select an uncertainty vector, defined as a single random selection from each of the uncertainty distributions for the

parameters of the model input distributions. The second stage (inner loop) is to simulate the variability of indoor concentrations or exposure/doses by selecting values from the model input distributions. For 12- or 24-hour averages, for each of U randomly selected uncertainty vectors, V simulations of concentrations or exposure/doses will be generated using the model input distributions. For 1- or 8-hour averages, for each of U randomly selected uncertainty vectors, 24V simulations of concentrations or exposure/doses will be generated using the model input distributions. This gives a total of UV trials and either UV or 24UV simulated values, where U and V are the number of iterations for the outer and inner loop, respectively. The number of iterations for both stages (i.e., U and V) is selected by the user. This feature allows the user to determine his or her priorities with respect to precision vs. both data entry and run time, which could increase substantially for application of the uncertainty analysis.

### 4.1.3. Number of Simulations

The optimum number of simulations (U and V) that the user should select cannot be determined in advance, since it depends on several considerations. Our recommended approach is presented in Appendix B.

### 4.1.4. Solicitation of model input uncertainty distributions

A flowchart showing the scheme for soliciting model input uncertainty distributions is given in Figure 4-2.

The CPIEM user specifies the distributions for each model input using the various CPIEM screens but does not specify the uncertainty distributions for those model input distributions. Instead, the user needs to consider each model input to be entered into CPIEM and decide a) whether to specify and uncertainty distribution for that model input distribution, and b) the type of uncertainty distribution to be used. The four types of uncertainty distribution to be described below are: continuous, discrete, default, and case name. Having made those decisions, the user will enter into UNC the uncertainty specifications for all the model inputs that will have an uncertainty distribution. These uncertainty specifications include the type of uncertainty distribution and its statistical distribution. UNC will generate the inputs for each of the U CPIEM uncertainty simulations.

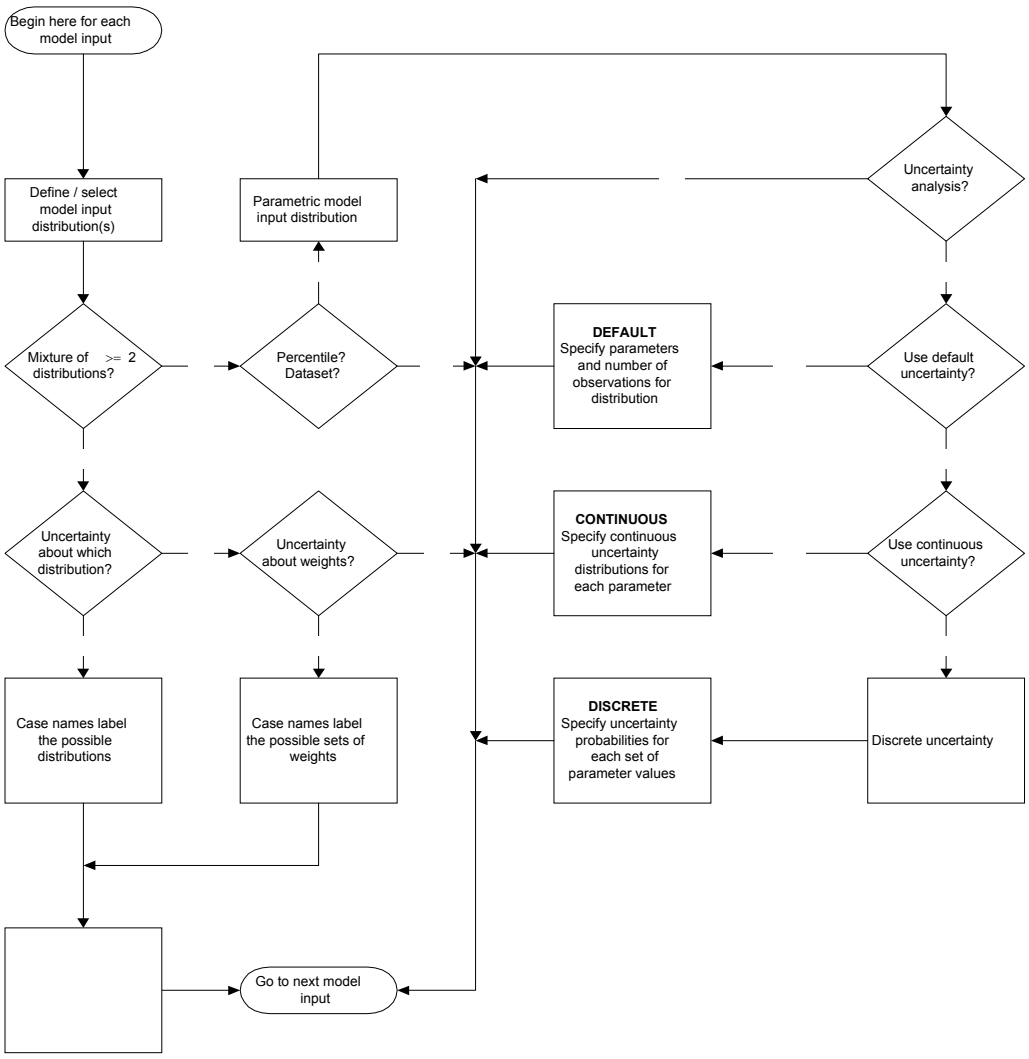
If all the CPIEM model inputs are treated as having no uncertainty, then the following procedures would not be used. The CPIEM model would be run for V trials using the selected model input distributions. The UNC module would not be used.

Each model input has a distribution characterized either in the form of a set of percentiles, a data set of possible values, or a parametric distribution (e.g., lognormal). The user may also specify that the data input is a mixture of two or more distributions, so that for each trial, the distribution itself is randomly selected with a probability given by the user-supplied “weight.” An uncertainty distribution can be specified for each input parametric distribution. (In the case of a mixture, the user may choose to specify uncertainty distributions for none, some, or all of the distributions in the mixture.)

*Mixtures*

If the input is a mixture of two or more distributions, then the user decides whether to: specify uncertainty about which distribution to use; specify uncertainty about the weights; or specify no uncertainty about the mixture. The user may also wish to specify uncertainty about the distributions themselves. For more details, please see the Appendix B.

**Figure 4-2**
**Solicitation of model inputs uncertainty distributions.**




In the first case, the user is uncertain about which distribution to use. For example, this situation would arise if different studies reported different distributions for that input (e.g., the outdoor concentration distribution) and the user was uncertain about which studies to believe. In this case, the mixture weights represent uncertainty rather than variability, so they should NOT be entered as user-supplied weights in CPIEM. Instead, assign a case name label to each distribution and use the case name uncertainty option in UNC to determine which distribution is selected for each CPIEM run of V trials.

In the second case, the user is uncertain about the weights. For example, this situation would arise if the distributions were indoor concentration distributions for kitchens with gas, electric, or wood stoves and different studies reported different fractions of residences with gas, electric, or wood stoves. In this case, the mixture weights represent variability across California residences, so they should be entered as user-supplied weights in CPIEM. Since the weights themselves

are uncertain, assign a case name label to each set of weights and use the case name uncertainty option in UNC to determine which set of weights is selected for each CPIEM run of V trials.

In the third case, the possible distributions and user-supplied weights are treated as having no uncertainty and so this model input mixture would not be entered into UNC.

### *Percentile or Data Set*

If the input distribution is specified as a percentile or data set, then an uncertainty analysis for that input is unavailable. Entry for that model input is complete. If the input distribution is specified as parametric, then the user needs to decide whether to specify an uncertainty analysis for the parameters. If not, entry for that model input is complete. These are the "no uncertainty" options.

### *Default Uncertainty*

As shown in Table 4-1, the possible parametric distributions for each CPIEM model input are all continuous: normal, lognormal, triangular, uniform, and, for the Time Since Use level 3 input, exponential. The user may choose to use the default uncertainty distribution for the given parametric distribution. These default uncertainty distributions reflect the portion of the uncertainty associated with sampling error, i.e., uncertainty of the parameters for the distribution due to the fact that the specified distribution is based on observations of only a subset of the entire population of interest. Other types of uncertainty that would not be reflected in the default uncertainty distributions include uncertainty about the correct distributional form, uncertainty about the representativeness of the population from which the samples were taken (i.e., proper sampling frame), and uncertainty about the randomness of the sampling procedure.

An example of UNC's process for providing the recommended default uncertainty distributions is as follows. Suppose the user specifies to UNC a normal distribution with a mean of 30 and a standard deviation of 6, based on n = 31 observations. For the first CPIEM run, UNC randomly selects 31 values from the specified normal distribution (with mean 30 and standard deviation 6). It then computes the arithmetic mean and arithmetic standard deviation of those 31 values. These two parameter values are provided to the user to enter into CPIEM for the first run. The process is repeated, using another set of 31 randomly selected values from the specified distribution, to obtain the arithmetic mean and standard deviation pair for the second run, and continued for U CPIEM runs.

An alternative approach, detailed in Appendix C, is based on the approximate statistical distribution of the arithmetic mean and standard deviation from n normally distributed values. The default uncertainty distribution for the mean parameter is also normal, and the default uncertainty distribution for the standard deviation parameter is a normally distributed variable raised to the power 1.5. Under the alternative approach, it is not necessary to generate U samples of n values from the normal distribution. The method uses a total of only 2U-simulated values from a standard normal distribution, which can reduce execution times for a complex problem. It may be desirable to use this alternative approach in a potential future project under which the capabilities of UNC and CPIEM are combined into a single software program.

**Table 4-1**
**Example of Summary Statistics Output Table from UNC.**

**Summary Statistics for: Formaldehyde 6.1.2**

| Statistic | Mean | Std Dev | Median | 2.5% | 97.5% | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| Arith Mean | 560.93 | 152.24 | 588.28 | 295.62 | 768.61 | 295.62 | 768.61 |
| Arith Std Dev | 372.18 | 101.01 | 390.33 | 196.14 | 509.97 | 196.14 | 509.97 |
| Geo Mean | 412.85 | 112.05 | 432.98 | 217.58 | 565.70 | 217.58 | 565.70 |
| Geo Std Dev | 5.67 | 1.54 | 5.95 | 2.99 | 7.77 | 2.99 | 7.77 |
| Minimum | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Maximum | 2984.94 | 810.09 | 3130.46 | 1573.09 | 4090.04 | 1573.09 | 4090.04 |
| 5% | 186.00 | 50.48 | 195.07 | 98.02 | 254.86 | 98.02 | 254.86 |
| 10% | 219.91 | 59.68 | 230.63 | 115.89 | 301.32 | 115.89 | 301.32 |
| 15% | 267.67 | 72.64 | 280.71 | 141.06 | 366.76 | 141.06 | 366.76 |
| 20% | 310.90 | 84.38 | 326.06 | 163.85 | 426.01 | 163.85 | 426.01 |
| 25% | 340.36 | 92.37 | 356.96 | 179.37 | 466.37 | 179.37 | 466.37 |
| 30% | 368.85 | 100.10 | 386.84 | 194.39 | 505.41 | 194.39 | 505.41 |
| 35% | 408.59 | 110.89 | 428.52 | 215.33 | 559.87 | 215.33 | 559.87 |
| 40% | 426.52 | 115.75 | 447.31 | 224.78 | 584.42 | 224.78 | 584.42 |
| 45% | 472.06 | 128.11 | 495.07 | 248.78 | 646.82 | 248.78 | 646.82 |
| 50% | 492.58 | 133.68 | 516.59 | 259.59 | 674.94 | 259.59 | 674.94 |
| 55% | 520.32 | 141.21 | 545.68 | 274.21 | 712.95 | 274.21 | 712.95 |
| 60% | 539.80 | 146.50 | 566.11 | 284.48 | 739.64 | 284.48 | 739.64 |
| 65% | 574.86 | 156.01 | 602.89 | 302.96 | 787.69 | 302.96 | 787.69 |
| 70% | 616.82 | 167.40 | 646.89 | 325.07 | 845.18 | 325.07 | 845.18 |
| 75% | 684.72 | 185.83 | 718.10 | 360.85 | 938.22 | 360.85 | 938.22 |
| 80% | 764.64 | 207.52 | 801.92 | 402.97 | 1047.73 | 402.97 | 1047.73 |
| 85% | 853.04 | 231.51 | 894.63 | 449.56 | 1168.86 | 449.56 | 1168.86 |
| 90% | 933.38 | 253.31 | 978.88 | 491.90 | 1278.94 | 491.90 | 1278.94 |
| 95% | 1057.90 | 287.11 | 1109.47 | 557.52 | 1449.56 | 557.52 | 1449.56 |
| 100% | 2984.94 | 810.09 | 3130.46 | 1573.09 | 4090.04 | 1573.09 | 4090.04 |

For each model input using the default uncertainty option, UNC generates U sets of parameter values. In the above example, if U = 6, then the output of UNC might be the list

(29.9,5.9);(29.4,5.8);(30.9,6.7);(31.0,4.9);(29.5,5.5);(30.0,6.0);

The U sets of parameter values are then entered into CPIEM. For the example, the first CPIEM run sets the respective model input distribution as normal with a mean of 29.9 and a standard deviation of 5.9. The second CPIEM run sets the model input distribution as normal with a mean of 29.4 and a standard deviation of 5.8.

To compute the default uncertainty distributions, the user needs to supply the default parameter values together with an additional piece of information: the number, n, of observations that the data input distribution is based on. If the distribution is stored in CPIEM, the number of observations (sample size) is included in the database description. The user may instead specify their own selected distribution (normal, lognormal, uniform, triangular, or exponential),

parameter values, and sample size (n) for the model input so that UNC can generate the default uncertainty distributions for that specified distribution.

*Continuous Uncertainty*

If the user chooses the "user-specified continuous uncertainty" option, then he or she will supply UNC with the associated continuous uncertainty distribution for each parameter. For simplicity and consistency with the treatment of input data distributions, each user-specified continuous uncertainty distribution is independently generated by UNC, so that all the uncertain parameter values will be assumed to be independent (and, hence, uncorrelated). The permitted continuous uncertainty distributions in UNC are: normal, uniform, and lognormal. The user will enter into UNC the selected continuous distribution together with the parameters for that selected distribution. The uncertainty distributions for the different parameters of a model input distribution may or may not have the same form.

The output of UNC includes a list of U sets of parameter values, as for the default uncertainty case. Each parameter value is independently randomly generated from the associated uncertainty distribution.

*Discrete Uncertainty*

If the user chooses the "user-specified discrete uncertainty" option, then he or she will supply UNC with two or more sets of parameter values for the selected continuous distribution. For example, suitable sets of parameter values for the normal distribution are (2,1), (3,2), and (4,1). The UNC program will randomly select U of these sets in the same manner as the case name uncertainty option described earlier: For k sets of parameters, if U is divisible by k, the k parameter vectors are randomly re-ordered U/k times. If U is not divisible by k, then the k parameter vectors are randomly reordered [U/k] times, where [U/k] is the next highest integer to U/k, but only the first U vectors are used.

Similarly to the case of mixture weights, the discrete uncertainty option allows the user to specify that each set of parameter values is equally plausible. This situation arises in cases where multiple studies fitted the same distribution but obtained different values for the fitted parameters. If the user gives each study equal "weight" then this uncertainty specification is implemented using the discrete uncertainty option for the given list of parameter sets. As for the case name uncertainty option, some sets of parameter values can be given more "weight" (a higher uncertainty probability) by entering those sets into UNC more than once.

### 4.1.5. Characterizing Uncertainty

The UNC software provides numerous ways for the user to specify uncertainties for the CPIEM model input distributions. In some cases, the user could specify arbitrary uncertainty distributions and obtain strange results. In this section we briefly discuss the best ways to avoid unrealistic specifications of uncertainty.

The case name and discrete uncertainty options are best used to represent cases where several different studies give different sets of distributions for the same model input. The case name option can be used to evaluate the uncertainty about the distribution by giving each study distribution a probability weight. For example, if there are two studies with outdoor concentration distributions and the user is uncertain and ambivalent about which study to use, the user could assign each distribution a 50 % uncertainty. If the type of distribution is known, but the parameters are uncertain, then the case name or discrete options can be used to evaluate the uncertainty about the parameter values, by giving each studies' set of parameter values a

probability weight. For example, one study might assume a normal distribution with mean 4 and standard deviation 1 and the other might assume a normal distribution with mean 5 and standard deviation 1, so the ambivalent user could give the parameter sets (4,1) and (5,1) equal weight.

Using the default uncertainty option also avoids arbitrary uncertainty assumptions. The default uncertainty option accounts for uncertainty attributable to sampling variability only. There may be other sources of uncertainty not represented by the sampling variability of the parameter estimates.

The continuous uncertainty option allows very general uncertainty specifications and is therefore most prone to give unrealistic results due to unreasonable specifications. For example, in most situations the uncertainty about a mean parameter is much less than the uncertainty about a variance parameter, so that if the user specifies a wide uncertainty distribution for the mean and a narrow uncertainty distribution for the variance, the uncertainty of the output distribution is likely to be underestimated because the uncertainty of the variance parameter was underestimated. As another example, if the mean parameter for a lognormal concentration distribution is a small positive value, such as 0.1, and if the uncertainty distribution is normal with a relatively large variance, such as 1, then the mean of the uncertainty distribution will effectively be much greater than 0,1 because the negative part of the uncertainty distribution is truncated. In this example, the user might be surprised to find that the values of the CPIEM output concentration, dose or exposure distributions will tend to be higher than the case where no uncertainty is assumed. Users of the continuous uncertainty option need to be careful to ensure that the uncertainty distributions truly or approximately represent their assessment about the joint uncertainty of the distribution's parameters. Such assessments are usually subjective, representing one or more expert's judgment regarding the uncertainty. The elicitation of expert judgment regarding the uncertainty distributions of parameter values is difficult and it will be hard to verify that the selected uncertainty distributions are consistent with the expert's assessment. The proposed future model enhancement described below for displaying the users' selected uncertainty distributions should help in this regard.

Another important consideration for the continuous uncertainty option is that the UNC model assumes that the uncertainty distributions for each parameter are statistically independent; more realistically, the uncertainty distribution of one parameter will depend upon the values of the other parameters. In principle, an enhanced version of UNC could allow for dependencies between the parameter values. Using the current version, if the parameters are highly dependent, then the user can approximate the uncertainty distribution to any degree of accuracy and precision by using the discrete uncertainty option and a sufficiently large number of parameter sets (drawn from the joint distribution). The same approach of using a discrete approximation can be used if the users' uncertainty distribution is not one of standard distributions provided in the software.

### 4.1.6. Constraints on Parameters of Model Input and Uncertainty Distributions

Note that there are some constraints on the parameters of model input distributions and of the parametric uncertainty distributions. When the user supplies UNC with uncertainty distributions for the model input distribution parameters, one of the constraints could be violated by a simulated parameter value. Also, the parameters of the uncertainty distribution should not violate the constraints. Appendix B presents the treatment for each uncertainty option and parametric distribution.

### 4.1.7. Default Uncertainty Estimates for Parametric Distributions

See Appendix B for details on how the default uncertainty distributions are generated in UNC.

### 4.1.8. Latin HyperSquare Sampling

For the continuous uncertainty option, the user specifies the uncertainty distribution for each parameter of the model input distribution. UNC draws a sample of U values from the user-specified uncertainty distribution, one value per CPIEM run. The UNC software uses the Latin HyperSquare (LHS) sampling method instead of simple random sampling to select the sample of U values from the user-specified uncertainty distribution. This section describes the LHS method and its advantages over simple random sampling.

The simplest way to generate a random sample of U values from a given distribution, F, is to repeatedly select a value at random from the distribution. The U values will be statistically independent. The disadvantage of this method is that it will not cover each part of the distribution equally. For example, although half of the distribution F lies below the median, the fraction of the U simulated values that are below the median could be anywhere from 0 to 1; although the expected fraction, averaged over the infinity of possible sets of U values, is 0.5. If U is large, then the each part of the distribution will be approximately equally represented, so that, for example, there will be about half of the U values below the median, and about one tenth of the U values in each decile of F. If U is small, then the U values will probably not be “evenly spread over the distribution” (this rather vague terminology can be made rigorous). The LHS method ensures an “even spread” because the distribution is divided into U equal slices, and one value is selected from each slice.

To carry out the LHS method, the distribution F is divided into U equal slices. The first slice, covering 1/U of the distribution, ranges from the lower bound to FINV(1/U), where FINV is the inverse distribution to F. The probability that a randomly selected value would be in this slice is 1/U. The second slice, also covering 1/U of the distribution, ranges from FINV(1/U) to FINV(2/U). The probability that a randomly selected value would be in this slice is 1/U. The third slice, also covering 1/U of the distribution, ranges from FINV(2/U) to FINV(3/U). The probability that a randomly selected value would be in this slice is 1/U. This continues to the last slice, ranging from FINV((U-1)/U) to the upper bound, which also has a probability of 1/U. A random arrangement (permutation) of the U slices is selected. This arrangement determines which slice is used for which CPIEM run. Thus the first run might be assigned the first, second, third, … U’th slice. The second run is assigned one of the remaining U-1 slices. And so on. This method defines the selected slice for each of the U simulated values.

The second step of the LHS method selects the value within each slice. If the user selects the option of choosing the midpoint of each slice, then the selected value for slice j is FINV((j-0.5)/U), which is the midpoint (i.e., median) of the distribution within that slice. If the user does not select this option, then a random value within the slice is selected. This value is calculated as FINV((j-1+R)/U), where R is a uniformly distributed value between 0 and 1.

The U values generated using the LHS method will in general tend to be more representative of the user-selected continuous uncertainty distribution than a simple random sample of U values from that distribution. In particular, the U values will include values from the upper and lower tails of the distribution (a random sample could result in no such extreme values), and will not over-represent those tails (a random sample could disproportionately sample from the lower and upper tails).

### 4.1.9. Model Outputs

In general, the CPIEM model will be run U times, once for each uncertainty vector. Each of the U runs will be based on V trials and therefore V or 24V replications. For each run, the CPIEM output provides a histogram, a cumulative distribution plot, various summary statistics and percentiles, and also provides separate files with the raw exposure, dose, and concentration values for each replication. These individual run outputs are described elsewhere in this report. In this section we will describe the model outputs from UNC based on the compiled summary statistics files from the U CPIEM runs. These UNC model outputs summarize the uncertainty and variability of the exposure, dose, or concentration. To make the presentation more concrete, we shall use results from an example of U = 8 sets of V = 200 trials for the time-weighted daily average exposure. The numbers in this example were made up and do not represent a real world scenario.

*1. Uncertainty distribution for the $p^{th}$ percentile of variability and other summary statistics of variability (table)*

Each of the U [= 8] uncertainty simulations gives a different estimate of the variability distribution for the concentrations or exposure/doses. These 8 variability distributions are summarized in the CPIEM output .STE, .STD, or .STC files, for exposure, dose, or concentration, respectively. Each CPIEM output file contains the following percentiles and summary statistics: arithmetic mean, arithmetic standard deviation, geometric mean, geometric standard deviation, minimum, maximum, and percentiles 5%, 10%, 15%, … 100%. These statistics summarize the output variability distribution for the given CPIEM model inputs corresponding to a single uncertainty vector. To use UNC to compute the summary statistics across all U uncertainty simulations, the U output files need to be combined into a single output file with extension .STA. As explained in the Users' Guide, the easiest way to do this is to move all the output files into a new folder and then use the DOS copy command:

copy *.ste filename.sta,

where filename.sta is the desired name of the combined output file and the outputs for each run are given the default .ste extension for exposure distributions. Change the .ste to .std or .stc for dose or concentration outputs.

For each summary statistic and percentile, UNC computes its distribution across the U runs. An example UNC output is presented in Table 4-1, below. For each summary statistic or percentile (column 1), UNC computes the mean, standard deviation, median, $2.5^{th}$ percentile, $97.5^{th}$ percentile, minimum, and maximum across the U runs. For example, the first row of Table 4-1 shows that the arithmetic mean varied from 295.62 (minimum) to 768.61 (maximum) across the 8 uncertainty simulations. The arithmetic mean had a mean of 560.93 (averaged over the 8 runs), a standard deviation of 152.24, a median of 588.28, and the $2.5^{th}$ and $97.5^{th}$ percentiles were 295.62 and 768.61. Similarly, the 40% percentile of variability had a median of 447.31 and the $2.5^{th}$ and $97.5^{th}$ percentiles were 224.78 and 584.42.

The summary statistic values in the median column summarize the variability distribution at the median level of uncertainty. This is an "average" estimate of the variability distribution, taking into account the uncertainties in the model inputs. The summary statistic values in the $2.5^{th}$ and $97.5^{th}$ percentile columns summarize the variability distribution at the 2.5% and 97.5% level of uncertainty, giving "lower" and "upper" bounds for the uncertainty. The interval from the $2.5^{th}$ percentile to the $97.5^{th}$ percentile gives a 95 % confidence interval for each summary statistic.

For example, in Table 4-1, the 40% percentile of variability has an "average" value (median) of 447.31, and a 95 % confidence interval from 224.78 to 584.42.

*2. Variability distribution for various levels of uncertainty (graph)*

The columns of the table described for "Uncertainty distribution for the $p^{th}$ percentile and other summary statistics of variability (table)" give estimates of the variability distribution at various levels of uncertainty. For example, the $2.5^{th}$ percentile column gives a "lower" uncertainty bound for the cumulative distribution of variability, the median column gives an "average" estimate for the distribution, and the $97.5^{th}$ percentile column gives an "upper" uncertainty bound. If there were V = 200 percentile rows, instead of 21, giving all possible percentiles for each level of uncertainty, then the variability distribution for that level of uncertainty could be computed and displayed in detail. However, such detailed outputs would have required significant storage and would probably not be useful for most users. Instead, UNC generates the envelope of the $2.5^{th}$ and $97.5^{th}$ percentile curves plus the median curve.

Figure 4-3 shows the graph created by UNC from the example data.

The envelope of curves for the three uncertainty percentiles provides a valuable summary of the uncertainty and variability distributions. Taken as a whole, the true variability cumulative distribution is estimated to lie somewhere inside the envelope between the $2.5^{th}$ and $97.5^{th}$ percentile curves. The region between the $2.5^{th}$ and $97.5^{th}$ uncertainty percentile curves can be thought of as a 95 percent confidence region for the variability of concentrations or exposure/dose. The median ($50^{th}$ percentile) curve can be thought of as the best point estimate of the variability distribution.

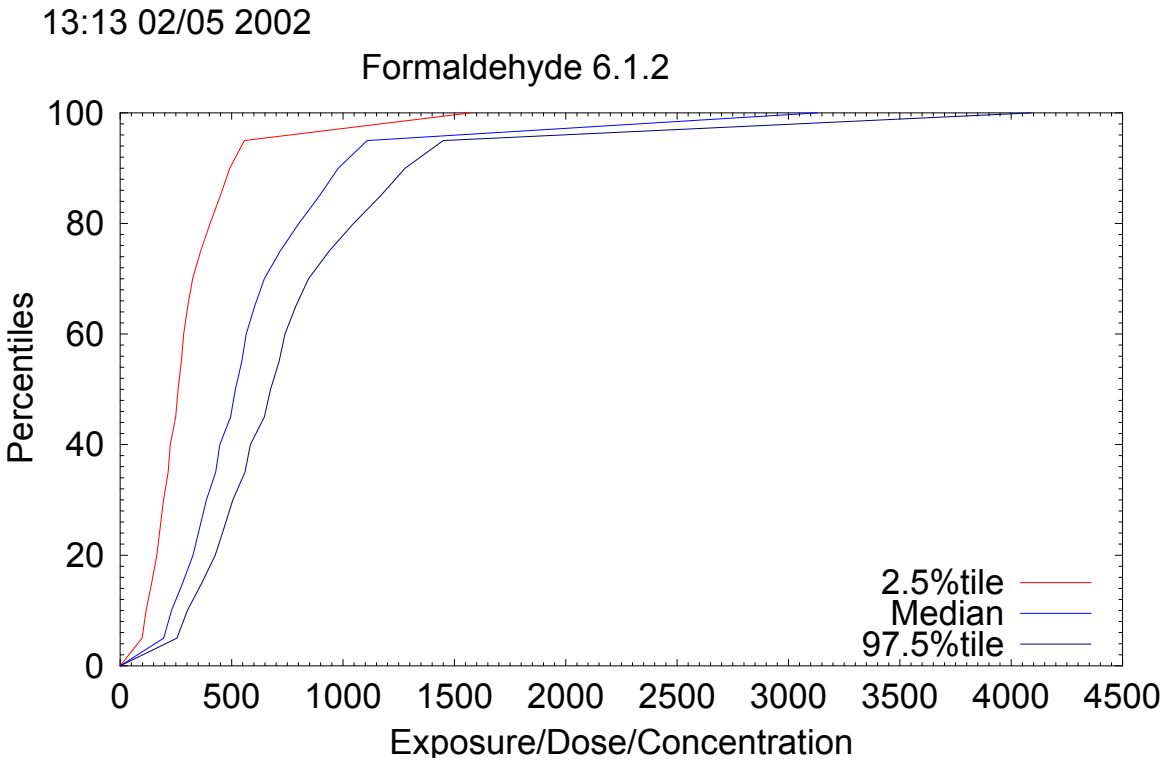
Looking horizontally, for a fixed value of p, the uncertainty of the $p^{th}$ percentile of variability is shown. The interval between the concentrations where the $2.5^{th}$ and $97.5^{th}$ percentile curves meet the horizontal line is a 95 % confidence interval for the $p^{th}$ percentile of variability. In the example graph, at variability percentile 50, the curves meet the line at 260 and 675, which is the 95 % confidence interval for the median, as is also shown in Table 4-1.

Looking vertically, for a fixed level of exposure, dose or concentration x, one finds the estimated probability distribution for the percentage of exposure, dose or concentration less than x. For example, the interval between the percentages p where the $2.5^{th}$ and $97.5^{th}$ percentile curves meet the vertical line at x is a 95 % confidence interval for the percentage of exposure, dose or concentration less than x. For example, Figure 4-3 shows that the 95 % confidence interval for the percentage of exposure below 500 ranges from about 30 % to 95 %.

### 4.1.10. Summary

- The CPIEM model has been enhanced by adding a new module UNC for uncertainty analyses.
- For each CPIEM model input, the UNC program uses the user-selected or default uncertainty distributions to generate the set of CPIEM inputs for each uncertainty simulation.
- The UNC program also uses the compiled CPIEM outputs from the uncertainty simulations to produce summary tables and graphs displaying the uncertainty of the variability distributions.

**Figure 4-3**
**Example of uncertainty output graph created by UNC.**




## 4.2. Adjustments to Activity Pattern Weights

In this section we describe the enhancements made to the set of activity pattern weights.

### 4.2.1. Alternative Weights

The current and previous CPIEM model uses data from two California activity pattern surveys. The time diary survey by Jenkins and others (1992) was a survey of 1579 adults aged 18 or older and 183 adolescents aged 12 to 17; this gave a total of 1,762 responses. A companion survey of 1200 California children aged up to 11 was conducted by Phillips and others (1991). The original survey databases included the two weighting variables TIMEWT and SAMPWT. The variable SAMPWT adjusts the set of activity patterns for the deliberate oversampling in the San Francisco Bay Area and the Rest of State compared to the Southern Coast, and for differences in the numbers of telephones and household members per household. The variable TIMEWT makes the same adjustments as SAMPWT and also adjusts for the oversampling of a weekend day compared to a weekday and of some calendar quarters. The previous version of CPIEM applied the TIMEWT adjustments for all statistical calculations (i.e., mean, standard deviation, percentiles, cumulative distribution function). For the current version, CPIEM was enhanced by allowing the user to select weighting by SAMPWT as an alternative to TIMEWT, since the weekday/weekend and seasonal adjustments applied to compute TIMEWT from SAMPWT were derived using the entire database and may not be applicable for selected activity pattern subsets of interest. For example, the percentages of weekend days sampled may differ for different age groups.

As an additional enhancement, the user is now allowed to supply either an alternative set of activity pattern weights. To accomplish this, the user would make a copy of the original activity pattern file and replace the original values of TIMEWT and SAMPWT by alternative values. To utilize the modified file, the user would then simply name it with the original activity pattern file name, POP.mdb.

An important example where the user may wish to revise the default set of activity weights is to make post-stratification adjustments to match changes in the demographics of the California population since the two activity pattern surveys were carried out. In the following subsection we describe post-stratification adjustments that were made to the TIMEWT and SAMPWT weights so that they reflect the year 2000 California age and gender distributions. The user may wish to further revise the weights to reflect other demographic factors. The user may also want to consider trimming the more extreme weights. The default TIMEWT and SAMPWT weights range up to 12, although only 1 % of the weights are above 6 and most of the weights are below 2. For example, to trim the weights down to the level 2, replace any weights greater than 2 by the value 2. Trimming the weights will tend to increase the precision of the exposure/dose analyses at the cost of a slight bias.

In addition, the enhanced CPIEM allows the user to specify that no weights be used at all, which is equivalent to weighting each activity pattern equally. This option is particularly useful if the user wants to fit statistical models to the CPIEM results, so that the statistical model specifications can do the job of the weights. For example, if the user wants to fit statistical regression models to the exposure distribution outputs as a function of the population characteristics (e.g. age, gender, region), then the no weight option is appropriate.

### 4.2.2. Post-stratification Weights

The default set of TIMEWT and SAMPWT activity pattern weights do not represent the current California census population. For the two California surveys, the weights use the surveyed age and gender distributions in 1989 to 1992 and thus do not represent the current California population fractions in each age and gender group. To combine the children survey (ages 0-11) with the adult and adolescent survey (ages 12 or older), the total survey weight for the children (ages 0 to 11) was made to be 14.0 %, which under represents the fraction of children in the current California population (19.8 % according to year 2000 projections downloaded from the California Department of Finance website). Thus the default CPIEM weights under represent children and disproportionately represent the various age and gender subgroups.

To enhance the CPEIM, an alternative set of activity pattern data files was developed so that the previous (default) set of SAMPWT and TIMEWT weights were adjusted to reflect the current California population by age group and gender. For the year 2000, projected census population data for California were obtained from the California Department of Finance. For each subgroup defined by age group and gender we computed the current population percentage and replaced the CPIEM weights (SAMPWT and TIMEWT) by adjusted weights using the equation:

$$Adjusted\ weight(age, gender) = Original\ weight(age, gender) \times \frac{\frac{2962}{100} \times Current\ population\ percentage(age, gender)}{\sum_{age, gender\ subgroup} Original\ weight(age, gender)} \quad (4\text{-}8)$$

The sum in the denominator is over all persons in the given age and gender subgroup. Age groups were chosen to be 0-4, 5-11, 12-17, 18-29, 30-39, 40-49, 50-65, and 66 or greater.

The 2962 term appears because there are a total of 2,962 set of activity patterns in the CPIEM database. The overall total of the adjusted weights will also equal 2,962.

These post-stratification weights have been provided in revised data files, formatted exactly like the activity current files. To utilize the post-stratification weights the user will re-name the revised file so that they will be selected by the model. The revised User's Guide explains the procedure.

An important caveat must be noted concerning these post-stratification adjustments. These adjustments will appropriately account for the fact that the percentages of the California population within each age and gender subgroup have changed since the surveys were conducted. However, these adjustments cannot account for changes in the activity patterns themselves for the various subgroups. For example, if the California population within a given age and gender subgroup now tends to spend much more time at work compared to the same subgroup in 1990, then the activity patterns for that subgroup are no longer representative. Post-stratification weights cannot adjust for such changes in activity patterns within a demographic subgroup. The best way to correct for changes in activity patterns by demographic subgroup is to carry out an updated activity pattern survey.

### 4.2.3. Summary

- The model has been enhanced so that the user may select a set of activity pattern weights from the following choices:

  1. No weights, i.e. all weights are equal.
  2. TIMEWT, the original model default.
  3. SAMPWT.

- A second file of activity patterns has been provided with post-stratification values for TIMEWT and SAMPWT. The age and gender post-stratification adjustments were computed using California Department of Finance year 2000 projected population counts. Age groups were chosen as 0-4, 5-11, 12-17, 18-29, 30-39, 40-49, 50-65, and 66 or greater.

- This is the default file used by CPIEM 2.0. However, the user may substitute the original file for the default file by re-naming it, according to instructions in the revised User's Guide.

- The user is now able to supply his or her own set of weights or his or her own weighted activity pattern data base by providing a properly formatted file and re-naming it, according to instructions in the revised User's Guide.

## 4.3. Disaggregation of Pollutant Removal Rates (Factor K)

In addition to ventilation and filtration, pollutant mass may be removed from the air by (1) reactive decay with other compounds in the air or with surfaces (irreversible); (2) adsorption by surfaces; or (3) deposition to surfaces. It also may be added to the air by (4) deadsorption from surfaces or (5) resuspension from surfaces. The current version of CPIEM specifies the removal rates for the first three processes with a single variable in the mass balance algorithm, "factor k", as follows.

$$V\frac{d}{dt}C_{in} = pQC_{out} + S - kVC_{in} - QC_{in} \qquad (4\text{-}9)$$

where:

| | | |
|---|---|---|
| $C_{in}$ | = | indoor concentration (mass/volume) |
| $p$ | = | penetration factor (dimensionless fraction) |
| $Q$ | = | air flow rate (mass/time) |
| $k$ | = | pollutant removal rate (1/time) |
| $C_{out}$ | = | outdoor concentration (mass/volume) |
| $S$ | = | indoor generation rate (mass/time) |

Although the factor *k*'s for these removal processes function the same way in the mass balance algorithm, they are quite different and the associated rates correspond to different characteristics of the microenvironment. Common pathways for atmospheric decay indoors are reactions with hydroxyl radicals, ozone, or nitrate. The rates vary widely among pollutants. For example, in relatively unpolluted air (i.e., OH radical concentration of 0.05 ppt, ozone concentration of 0.06 ppm), 1,3-butadiene and cresol have expected atmospheric lifetimes of less than 6 hours, while benzene and carbon tetrachloride have expected atmospheric lifetimes of more than 60 days.

Most VOCs have negligible deposition to surfaces, but highly acidic, semivolatile, and polar compounds can have large deposition rates. Evidence of surface deposition may reflect either a chemical reaction with a surface, resulting in the transformation of the pollutant (an irreversible sink), or "deep" adsorption by surface material. "Deep" adsorption refers to a process by which a VOC is reversibly adsorbed by a surface material, but the rate of desorption is significantly longer than the time constant of the diurnal cycle of air concentrations. A study of adsorption of selected VOCs by common indoor surface materials (Colombo et al. 1993) found evidence such "deep" adsorption in addition to more rapid adsorption-desorption processes, for tetrachloroethylene, $\alpha$-pinene, 1,2,4-trimethylbenzene, 2-butoxyethanol, n-decane, 1,4-dichlorobenzene, 2-ethylhexanol and n-dodecane. The authors concluded that adsorption seems to occur to at least two different sinks with two different rate constants in the same material. For evaluating long-term average concentrations, the rapidly reversible adsorption is probably not relevant, but the impact of the "deep" adsorption may approximate an irreversible sink, unless emissions are intermittent. The Colombo study authors concluded that adsorption generally increases with the boiling point of the compounds, but also depends on other physiochemical properties, such as chemical functionality, as well as on the sorbent material.

Deposition of pollutants onto surfaces reduces average ambient concentrations. This effect may be significant for particles. Dry deposition for most gaseous pollutants is slow, however, with the exception of highly acidic species such as hydrochloric acid (HCl) and polar compounds such as formaldehyde and cresol, as discussed above. Dry deposition rates for particles are primarily a function of particle size, and are much larger for coarse particles (those with diameters between 2.5 and 10 $\mu$m) than for fine particles (those smaller than 2.5 $\mu$m). Deposition of coarse particles is primarily a result of gravitational settling. However, for fine particles, deposition is a more complex phenomenon that depends upon the amount of turbulence near the surface, and the nature of the surface. Available information for deposition rates is often in the form of deposition velocities, i.e. m/sec. In order to estimate a removal rate (1/sec), the deposition velocity must be combined with information on the dimensions of the room. This calculation is currently done outside of the CPIEM model.

Note that the same pollutant may be subject to more than one of these processes. For example, formaldehyde is known to decay rapidly from photolysis (4 to 10 hour atmospheric lifetime), as

well as from reaction with hydroxyl radicals (30 to 36 hour atmospheric lifetime), and possibly from reaction with nitrate radicals. It is also a polar compound subject to dry deposition.

The impacts on concentrations of de-adsorption and resuspension are similar to that of an indoor emission source, and thus they may be represented by similar terms in the mass balance algorithm. However a simpler approach, requiring less input data is to combine de-adsorption with adsorption, and to combine resuspension with deposition, so that the factor k terms represent net removal rates. This is typically how data are presented in the literature. We have implemented this latter approach.

In order to assure that the characterizations of these three removal processes are clear to the user, and to allow for the specification of more than one removal process for a single pollutant, separate terms to represent each process have been added to the mass balance algorithm. Thus, the user now has the option of specifying 3 separate rate constants for pollutant removal in units of 1/time as follows:

$k_1$ = reactive decay rate

$k_2$ = net surface adsorption rate (i.e., adsorption – de-adsorption)

$k_3$ = net deposition rate (i.e., deposition – resuspension)

Thus the governing relationship presented in equation 4-9 is modified as follows:

$$V\frac{d}{dt}C_{in} = pQC_{out} + S - (k_1 + k_2 + k_3)VC_{in} - QC_{in} \qquad \text{(4-9a)}$$

In cases where there is more than one removal process, the value for each *k* is selected independently from the specified distributions, which may be of different types. Thus, this formulation provides increased flexibility for characterizing the effects of removal processes on indoor concentrations.

As noted above, reactive decay and adsorption processes are most relevant to gases, and vary widely among compounds. Adsorption also varies according to the nature of the surface with which the compound makes contact. Deposition is most rapid for particles with the rate dependent on both the particle size and its degree of polarity. Highly polar compounds in the gaseous phase may also be subject to deposition.

The CPIEM database contains data on reactive decay for 3 pollutants, and on adsorption for one pollutant combined with 4 different surfaces. Although no particle deposition rate data that met the inclusion criteria were identified in time to incorporate into CPIEM2.0, such data can easily be added to the database by the user when identified.

.

## 4.4. Exposure Metric

As described in Section 2, one of the output metrics for the Level 1-2 module of CPEIM is the time-integrated exposure concentration encountered by an individual while in an indoor environment, as follows.

$$C_T = \int_0^T C(t)dt \qquad \text{(2-1)}$$

where *C(t)* is the concentration in the environment at time *t*, *T* is the amount of time spent in the environment, and $C_T$ is the time-integrated concentration, measured in units of μg-h/m$^3$. An alternative exposure metric, the time-weighted average concentration, may now be selected by the user. It is given by:

$$C_{TWA} = \frac{C_T}{IT} \qquad (4\text{-}10)$$

where, $C_{TWA}$ is the time-weighted average concentration (μg/m$^3$), and *IT* is the overall integration period. This integration period is the total duration of the activity pattern used for the simulation, including time spent outside the microenvironment(s) analyzed. Depending on the pollutant this may be 1 hour, 8 hours, 12 hours or 24 hours.

# 5. Improvement of Efficiency and Ease of Use

In order to take advantage of newer technology to improve efficiency and ease of use the CPIEM computer code from a QuickBasic platform to a Visual Basic\Windows platform. The Windows platform of this new version of CPIEM greatly improves the software's efficiency and ease of use with standard, easily understood drop-down menus and dialogue boxes. The graphic outputs are presentation quality. Scenarios are easily saved and edited to facilitate sensitivity analysis.

As a practical matter, the literal translation of QuickBasic code to Visual Basic code is much simpler than dissecting the existing code. Several characteristics of the QuickBasic version of CPIEM presented problems requiring substantial additional effort than was anticipated at the start of this project.

- Most of the important data in the program (over 300 variables and arrays) were stored in common, globally-accessible data structures. This meant that data could be accessed and changed by any part of the model code, precluding the possibility of addressing small portions of code independently. Instead, the entire body of code generally needed to be checked in order to modify a variable.

- A related issue was the fact that the original code was not constructed in a strictly modular fashion with discrete parts. Instead, most of the model code was intimately linked together, with control often passing from one large section of code to another via "goto" and "gosub" statements rather than through the use of actual subroutines. This meant that information flowing into and out of different sections of the program could not be readily traced.

- Variable types were generally not explicitly specified and were not named according to a consistent convention. There were also multiple similarly-named variables – sometimes with the same names for locally-defined variables in one program segment as for globally-defined variables in another. Thus a full scan of the model code was required before modifying a single variable, lest a variable that appears to be a local variable for a single procedure turn out to be a global variable used in other parts of the program as well.

- Apparently in response to size constraints, the model was actually constructed as a series of independent programs. CPIEM.EXE included most of the interface. LEVEL12.EXE and LEVEL3.EXE contained most of the model code and output routines. Additional utility programs like POPMATCH.EXE are also included. These programs often include large blocks of code that are nearly, but not quite duplicates of one another, and often "include" code from a range of other files. These programs transfer hundreds of global variables back and forth by storing them on disk and reading them back again, whether they are needed or not. They also had very limited error checking.

- Documentation in the original code was inadequate and at times misleading. In many cases it appeared to have been added after the fact. Extensive use of undescriptive variable names like "xx" and "kk" and "yy1" exacerbated this problem.

To address these issues our fundamental approach in re-writing the CPIEM software has been the separation of the 3 application tiers, i.e. User Interface, Calculation Engine, and Database. This approach inherently minimizes the number of global variables used and further exposes where global constructs are used (i.e., global variables are grouped in a logical Structure constructs and explicitly passed through procedures).

To achieve this objective the software code was systematically disassembled and re-constructed in a modular fashion, retaining calculation and other core algorithms in the model as necessary.

As a result we were able to perform better and tighter testing, both at the module level and at the program level. The calculation engine was initially tested successfully in a stand-alone mode with no interface and no database. The additional tiers where added on in the later phases of testing.

We maintained the same data structure for the existing dBase input files to minimize introduction of new changes and additional testing required to verify the new modifications. To this end we created a set of replacement routines that mimic the third-party dbLib DOS library used by the existing code to read and write dBase files, so that we would not need to modify the existing code. We did however consolidate the bulk of input files into Access database(s) where necessary and efficient. We also created additional new tables to link and store scenario runs and maintain application level settings.

We did not attempt to systematically identify and repair all of the potential problems in the original QuickBasic product, although we have repaired a number of problems discovered in the course of translating and documenting the original code. We tested the Visual Basic product to demonstrate that the key elements of the original model have been ported effectively by duplicating results obtained by the original code for the same inputs. After we added the enhancements described in other sections of this report, we conducted additional testing to assure that the extensions were implemented correctly.

The new Windows-based user interface closely replicates the original one, but provides improved reporting capabilities. Specific improvements include:

- More legible model reports, suitable for reproduction in color or in black and white. This feature will allow the user to prepare graphs and tables directly from the model for use in reports and presentations.

- Elimination of the check-mark model for “visiting” data screens. This feature greatly reduces the time required both for creating scenarios and for modifying them. Thus, systematic changes to input data for senstivity testing is greatly enhanced.

- Implementation of field-level data validation checks to block users from entering values in the wrong formats. This feature improves efficiency of model implementation by minimizing runtime errors due to faulty data inputs.

- Creation of an installation program to install the complete model on a target PC running Windows. This feature provides quick and easy installation for all users, including those with a minimum of computer skills.

# 6. Verification and Validation

For purposes of this report we refer to process by which we confirm that the software code makes specified calculations correctly as verification. The process by which we confirm that the software algorithms correctly represent the real world indoor air exposures is referred to as validation.

In verifying and validating the new CPIEM software we focused on:

- Verifying that the current application code has been translated to Visual Basic without introducing new errors, by duplicating the results obtained by the original version of the model for the same inputs;
- Verifying that new functions created for the revised model work as designed (i.e., the time-weighted average exposure metric in the level 1-2 module, and the disaggregated pollutant removal rate in the level 3 module);
- Verifying that the revised interface supports the input and output requirements of the revised model; and
- Verifying the usability of the revised interface.

## 6.1. Verification of CPIEM2.0

In testing the new application against DOS version, we performed two verification tests for the exposure/dose model (Level12) and two verification tests for the indoor concentration model (level3). Specifically, we used the following sample scenarios from the original User's Guide, which were used to evaluate the original application:

- Example application 6.1.1 for the Exposure/Dose (level12) model - One Environment.
- Example application 6.1.2 for the Exposure/Dose (level12) model - Two Environments.
- Example application 6.2.1 for the Indoor Concentration (level3) model – Long-term source.
- Example application 6.2.2 for the Indoor Concentration (level3) model – Episodic source.
- Diesel exposure analysis run provided by ARB.

These example applications / verification test are described in detail in the CPIEM User's Guide. However, a summary is presented here in Tables 6-1 and 6-2 for a quick comparison. All scenario runs produced exact matching results as compared to the CPIEM 1.4F DOS application.

**Table 6-1**
**Level 1-2 Verification Test Results***

| Scenario | Statistics | CPIEM 1.4 F | CPIEM 2.0 |
|---|---|---|---|
| 6.1.1 | Arithmetic Mean ± SD | 707.61 ± 416.92 | 707.61 ± 416.92 |
| | Geometric Mean ± SD | 469.08 ± 9.97 | 469.08 ± 9.97 |
| | 25%, 50%, 75%, 95% | 437.9, 629.7,864.8,1438.0 | 437.9, 629.7,864.8,1438.0 |
| | Minimum - Maximum | 0.00 – 3646.0 | 0.00 – 3646.0 |
| 6.1.2 | Arithmetic Mean ± SD | 591.24 ± 392.29 | 591.24 ± 392.29 |
| | Geometric Mean ± SD | 435.16 ± 5.98 | 435.16 ± 5.98 |
| | 25%, 50%, 75%, 95% | 358.7,519.2,721.7,1115.0 | 358.7,519.2,721.7,1115.0 |
| | Minimum - Maximum | 0.00 – 3146.2 | 0.00 – 3146.2 |

* *These runs were performed using the original activity pattern weights from CPIEM 1.4F. Use of the new post-stratification weights will alter the final results*

**Table 6-2**
**Level 3 Verification Test Results***

| Scenario | Statistics | CPIEM 1.4 F | CPIEM 2.0 |
|---|---|---|---|
| 6.2.1 | Arithmetic Mean ± SD | 0.70 ± 1.86 | 0.70 ± 1.86 |
| | Geometric Mean ± SD | 0.19 ± 3.97 | 0.19 ± 3.97 |
| | 25%, 50%, 75%, 95% | 0.07, 0.16, 0.35, 2.93 | 0.07, 0.16, 0.35, 2.93 |
| | Minimum - Maximum | 0.03 – 13.83 | 0.03 – 13.83 |
| 6.2.2 | Arithmetic Mean ± SD | 2.07E+03 ± 1.82E+04 | 2.07E+03 ± 1.82E+04 |
| | Geometric Mean ± SD | 1.69E-06 ± 1.75E+05 | 1.69E-06 ± 1.75E+05 |
| | 25%, 50%, 75%, 95% | 0.00, 0.00, 0.00, 3.42E-02 | 0.00, 0.00, 0.00, 3.42E-02 |
| | Minimum - Maximum | 0.00 – 1.81E+05 | 0.00 – 1.81E+05 |
| Diesel | Arithmetic Mean ± SD | 1.88± 0.87 | 1.88± 0.87 |
| (Residential) | Geometric Mean ± SD | 1.68 ± 1.66 | 1.68 ± 1.66 |
| | 25%, 50%, 75%, 90% | 1.24, 1.75, 2.39, 3.03 | 1.24, 1.75, 2.39, 3.03 |
| | Minimum - Maximum | 0.22 – 5.04 | 0.22 – 5.04 |
| Diesel | Arithmetic Mean ± SD | 1.60± 0.73 | 1.60± 0.73 |
| (Office) | Geometric Mean ± SD | 1.44 ± 1.60 | 1.44 ± 1.60 |
| | 25%, 50%, 75%, 90% | 1.06, 1.49, 2.01, 2.57 | 1.06, 1.49, 2.01, 2.57 |
| | Minimum - Maximum | 0.22 – 4.75 | 0.22 – 4.75 |

* *Note that the Level3 validation runs were performed prior to the addition of the new removal rate (factor K) parameters. In the final version of CPIEM2.0, the associated algorithm for these new parameters will be present causing additional calls to the Random Number generator (as compared to CPIEM1.4F); this will alter the outcome of sampling for the subsequent parameters and ultimately the final results.*

## 6.2. Verification of UNC

The UNC module is used to analyze the sensitivity of the CPIEM output variability distributions to various parameters. UNC has two functions: (1) it provides an array of values with a certain distribution for a given parameter to be used as input to CPIEM; and (2) it summarizes the output from successive calculations (runs) of CPIEM into concise tables and charts. This greatly simplifies the assessment of the sensitivity. A detailed discussion of the verification of these two functions is presented in the UNC User's Guide; a summary is provided here.

### 6.2.1. Input Parameter Distributions

A calculation using CPIEM requires certain inputs. UNC generates distributions of input values that are suitable for analyzing, for example, the sensitivity of CPIEM to the precise value of these parameters. Suppose an input is the mean of an exponential distribution and we are interested in the effect of varying it according to different uncertainty distributions. UNC provides the choice of four different continuous distribution types for a CPIEM input parameter: (1) normal; (2) lognormal specifying arithmetic mean and standard deviation; (3) lognormal specifying geometric mean and standard deviation; and (4) uniform between two limits. The distributions for the possibilities (2) and (3) are equivalent but the different specifications are provided for convenience. Several test were performed to verify the generation of data from these distributions, as follows.

For each CPIEM input value UNC generates a user-specified number, n, of these input values. The specified distribution of the parameter (in this example case the mean of exponential distribution) is divided into n slices of equal probability. A value for each slice is either taken as the probability midpoint, or as a random point in that slice interval. For the verification tests we made eight calculations with UNC to provide eight sets of 1000 potential values for the mean of the exponential distribution. Each of the distribution types was requested twice, once with the slice midpoint and one with a random point in the slice. In all cases the distribution of UNC-generated values was nearly indistinguishable from the theoretical distribution.

### 6.2.2. Combined UNC and CPIEM Operations

A second group of verification tests was performed to check the combined operations of UNC and CPIEM. In each case, UNC was used to generate up to 20 sets of CPIEM input parameters, resulting in 20 CPIEM-generated output variability distributions. UNC was then used to generate summary statistics and graphs of the uncertainty and variability. For this group of verification tests, the midpoint slice option for the LHS was used for each case.

For each of these calculations we considered hypothetical residential indoor concentrations and examined the variability distribution for the time-weighted daily average total exposure from the residential microenvironment only. For identification purposes only, the indoor pollutant is called formaldehyde. The inputs, however, do not necessarily correspond to formaldehyde data. For these tests a special version of CPIEM was applied that implemented only activity patterns for which all time was spent in the residential microenvironment. In each CPIEM run, 5,000 trials were used, so that 5,000 activity patterns were selected at random with replacement. However, because only a small subset of the selected activity patterns was used (i.e., those with all time spent in residence) the effective number of trials was approximately 200. It follows that for this special version of CPIEM, the output variability distribution for each CPIEM run should approximately match the selected input distribution if CPIEM is working correctly, with an exact match precluded by the less than infinite sample size.

The calculations were divided into two sets: (1) the residential indoor concentration distribution is a lognormal distribution (CPIEM requires the arithmetic mean and standard deviation as parameters); and (2) the residential indoor concentration distribution is a mixture of normal distributions.

For each case in Set 1, the mean values of the mean and standard deviation parameters were 2 and 1, respectively. Thus for each case in Set 1 the output variability distribution at the median level of uncertainty should be approximately equal to a lognormal (2,1) distribution. Uncertainty was introduced in several ways for the various cases: (a) varying the random number seed only, i.e., no uncertainty in the mean and standard deviation parameters (b) default uncertainty of

parameters, assuming a sample size of 1000 (c) default uncertainty of the parameters, assuming a sample size of 10 (d) distribution of the mean as uniform from 0 to 4, distribution of standard deviation normal with mean 1 and standard deviation 0.01 (e) distribution of the mean as uniform from -3 to 4, distribution of standard deviation normal with mean 1 and standard deviation 0.01, and (f) distribution of the mean as normal with mean 2 and standard deviation 1; and distribution of the standard deviation normal with mean 1 and standard deviation 0.5. The results showed that the relative sizes of the uncertainty bounds among the 6 cases was as expected and comparison of results with SAS-generated distributions for 1000 trials showed a close match for all cases (See Table 6-3).

The second set of cases used a mixture of normal distributions for the residential indoor concentration. Again, uncertainty was introduced in several ways for the various cases: (a) normal distributions with mean values of 4,6,8, and 10, each with a standard deviation of 1; one run each (b) normal distributions with mean values of 4,8,8, and 8, each with a standard deviation of 1; one run each (c) 17 CPIEM runs with normal distributions with means having values 3, 3.5, 4, … ,11, each with standard deviation equal to 1. Again the results matched closely with SAS-generated distributions for 1000 trials in each case (see Table 6-3).

### 6.2.3. Model Validation for Benzo(a)Pyrene

A validation exercise was performed for the original version of CPIEM, using both the Level 1-2 and Level 3 modules, to simulate for daily average benzo(a)pyrene in Riverside residences and compared to values measured as part of the 1990 PTEAM study (Koontz et al. 1998). This validation exercise was repeated using the revised version of CPIEM and the UNC module to account for uncertainties in the model inputs. The findings showed that the PTEAM measurements were all within the CPIEM/UNC predicted uncertainty intervals except at the $90^{th}$ percentile level.

Calculations were made of the distribution of indoor 24-hour concentrations for benzo(a)pyrene. A total of 20 different calculations were made involving a multi-dimensional matrix where 12 different parameters were varied. The development and rationale of the model was discussed in the Koontz report (Section 7.2.2). The starting values for the model parameters were given in Table 7-11 of the report and are reproduced here in Table 6-4.

UNC used the default uncertainty option to calculate the matrix of input values. Thus from each input lognormal distribution, a sample was selected at random and the arithmetic mean and variance parameters were computed from the sample. In order for UNC to calculate default uncertainty distributions, the user must provide the size of the sample along with the parameters of the distribution for the input variables. For this analysis the assumed sample sizes were 100 for the indoor source emission rate, 36 for the ambient concentration, and 10 each for the air-exchange rate and each of the contributing indoor volume distributions (3 of these). Apart from the ambient concentration distribution, these sample sizes are estimates, because the actual sample size values were not reported in Koontz et al. and the original PTEAM data were not readily available. No uncertainty was assumed for the number of indoor sources or for the penetration factor. There are 6 different lognormal distributions and UNC provided 20 sets of arithmetic means and standard deviations for each of these.

## Table 6-3
## Uncertainty Distributions of Selected Summary Statistics

| Set | Summary Statistic | SAS simulations | | | CPIEM simulations | | |
|---|---|---|---|---|---|---|---|
| | | 2.5th | 50th | 97.5th | 2.5th | 50th | 97.5th |
| 1a | Mean | 1.94 | 2.00 | 2.07 | 1.84 | 2.00 | 2.16 |
| 1a | 25% | 1.25 | 1.30 | 1.35 | 1.21 | 1.30 | 1.42 |
| 1a | 50% | 1.73 | 1.79 | 1.85 | 1.58 | 1.77 | 1.97 |
| 1a | 75% | 2.37 | 2.46 | 2.56 | 2.20 | 2.49 | 2.84 |
| 1b | Mean | 1.91 | 2.00 | 2.09 | 1.86 | 1.97 | 2.21 |
| 1b | 25% | 1.23 | 1.30 | 1.38 | 1.16 | 1.29 | 1.45 |
| 1b | 50% | 1.70 | 1.79 | 1.88 | 1.66 | 1.77 | 1.98 |
| 1b | 75% | 2.34 | 2.46 | 2.59 | 2.29 | 2.43 | 2.89 |
| 1c | Mean | 1.47 | 1.96 | 2.68 | 1.74 | 2.07 | 2.57 |
| 1c | 25% | 0.95 | 1.33 | 1.85 | 1.08 | 1.41 | 1.66 |
| 1c | 50% | 1.33 | 1.78 | 2.38 | 1.61 | 1.89 | 2.32 |
| 1c | 75% | 1.76 | 2.38 | 3.33 | 2.15 | 2.59 | 3.28 |
| 1d | Mean | 0.11 | 2.04 | 3.90 | 0.10 | 2.08 | 4.01 |
| 1d | 25% | 0.00 | 1.34 | 3.18 | 0.00 | 1.35 | 3.26 |
| 1d | 50% | 0.02 | 1.84 | 3.77 | 0.01 | 1.79 | 3.80 |
| 1d | 75% | 0.07 | 2.51 | 4.47 | 0.04 | 2.55 | 4.76 |
| 1e | Mean | 0.11 | 2.04 | 3.90 | 0.16 | 2.11 | 3.88 |
| 1e | 25% | 0.00 | 1.34 | 3.18 | 0.00 | 1.41 | 3.13 |
| 1e | 50% | 0.02 | 1.84 | 3.77 | 0.01 | 1.93 | 3.73 |
| 1e | 75% | 0.07 | 2.51 | 4.47 | 0.06 | 2.63 | 4.59 |
| 1f | Mean | 0.35 | 2.02 | 4.00 | 0.31 | 2.01 | 4.10 |
| 1f | 25% | 0.05 | 1.31 | 3.46 | 0.04 | 1.22 | 3.23 |
| 1f | 50% | 0.13 | 1.77 | 3.91 | 0.10 | 1.87 | 3.84 |
| 1f | 75% | 0.36 | 2.43 | 4.60 | 0.35 | 2.66 | 4.95 |
| 2a | Mean | 4.01 | 7.07 | 10.01 | 3.91 | 6.91 | 9.91 |
| 2a | 25% | 3.32 | 6.38 | 9.35 | 3.28 | 6.28 | 9.28 |
| 2a | 50% | 4.02 | 7.08 | 10.01 | 3.89 | 6.89 | 9.89 |
| 2a | 75% | 4.69 | 7.76 | 10.64 | 4.59 | 7.59 | 10.59 |
| 2b | Mean | 4.00 | 7.98 | 8.04 | 3.91 | 7.90 | 7.97 |
| 2b | 25% | 3.38 | 7.30 | 7.31 | 3.28 | 7.23 | 7.32 |
| 2b | 50% | 4.00 | 7.93 | 8.04 | 3.89 | 7.91 | 8.05 |
| 2b | 75% | 4.65 | 8.66 | 8.73 | 4.59 | 8.57 | 8.59 |
| 2c | Mean | 3.02 | 6.98 | 11.02 | 2.91 | 6.96 | 11.02 |
| 2c | 25% | 2.39 | 6.34 | 10.36 | 2.28 | 6.23 | 10.28 |
| 2c | 50% | 3.04 | 6.95 | 11.02 | 2.89 | 6.96 | 11.10 |
| 2c | 75% | 3.69 | 7.69 | 11.73 | 3.59 | 7.56 | 11.66 |

**Table 6-4**
**Summary of Model Inputs for Benzo[a]pyrene (after Koontz et al. 1998, Table 7-11)**

| Input Parameter | Distribution/Value |
|---|---|
| Percent of Residences with Indoor Sources | 28 |
| Number of Indoor Sources | Normal (1,0)[a] |
| Emission Rate, ng/h | Lognormal (390, 1285) |
| Outdoor Concentration, ng/m$^3$ | Lognormal (0.30, 0.36) |
| Penetration Factor | Normal (0.6, 0) |
| Indoor Sink, 1/h | Normal (0,0) |
| Indoor Volume, m$^3$ | |
| • TEAM Study[b] | Lognormal (274.9, 110.6) |
| • SoCal Study | Lognormal (309.5, 159.8) |
| • ADM Study | Lognormal (354, 101) |
| Air Exchange Rate, 1/h | Lognormal (1.25, 1.02) |

[a] Values in parentheses are the mean and standard deviation of the distribution

[b] Inputs for volume from three different studies in Southern California were equally weighted

The mean of the summary statistics for the resulting distributions are similar to those found for the original CPIEM exercise (see Table 6-5). The majority of the calculations gave very small concentrations (less than 1.0 ng/m$^3$), but there are a few that have a relatively large magnitude concentration. For the 1000 trials each, the calculations gave maximum values that extended from 15 to 66 ng/m$^3$. Figure 6-1, a plot obtained using UNC showing the uncertainty distribution for the daily concentration distribution, is dominated by this maximum value.

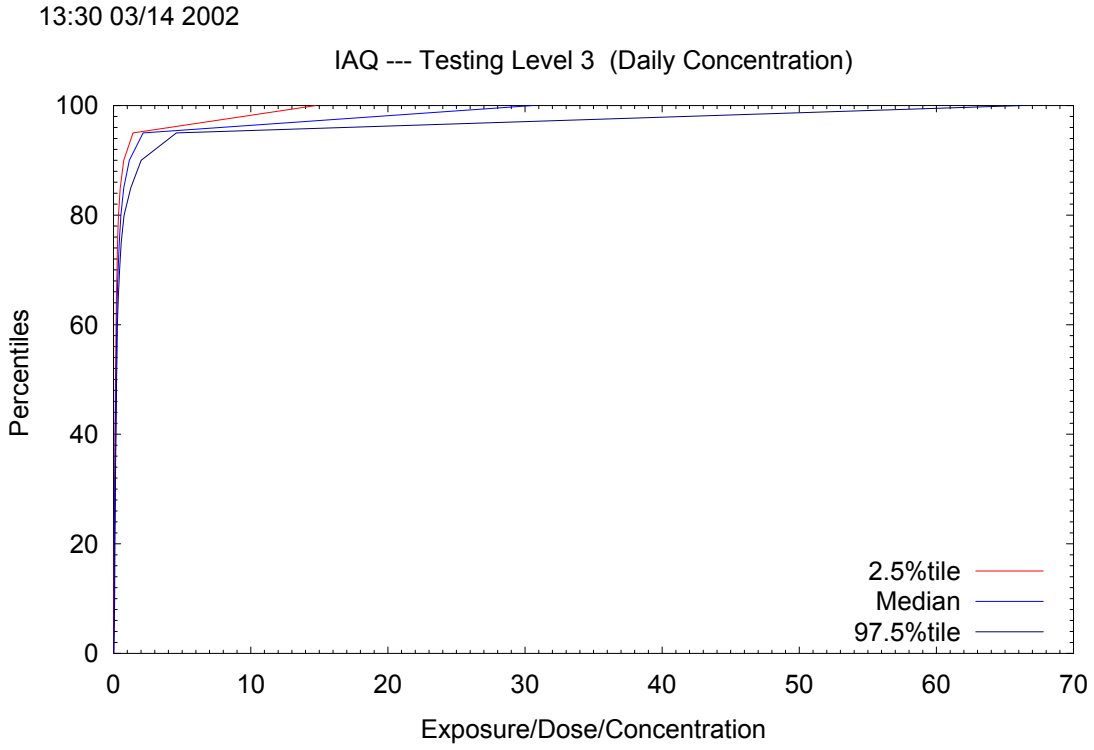
A numerical comparison of the summary statistics from the PTEAM study with the uncertainty intervals is also presented in Table 6-5. Except for the 90$^{th}$ percentile and the standard deviation, all of the PTEAM summary statistics lie within the 95 % uncertainty interval from the 2.5$^{th}$ to the 97.5$^{th}$ uncertainty percentile.

The discrepancy at the upper end of the distribution may be the result of overestimating the unknown sample sizes for the emission rate, the air exchange rate, and/or the building volume inputs to UNC. If the sample sizes were assumed to be smaller, the uncertainty ranges would have been larger, and might have encompassed the measured value at the upper end of the distribution and the observed standard deviation. Another possible cause of the discrepancy is the assumption of no uncertainty for the penetration factor or the number of indoor sources. Accounting for uncertainty in either of these variables would also increase the uncertainty range, and possibly result in a wide enough range to encompass the measured data.

A third possibility is that the default uncertainty calculation underestimates the total uncertainty. As noted in section 4, the default uncertainty option accounts for uncertainty attributable to sampling variability only, i.e., uncertainty of the parameters for the distribution due to the fact that the specified distribution is based on observations of only a subset of the entire population of interest. Other types of uncertainty that would not be reflected in the default uncertainty

distributions include uncertainty about the correct distributional form, uncertainty about the representativeness of the population from which the samples were taken (i.e., proper sampling frame), and uncertainty about the randomness of the sampling procedure. As is the case for any continuous distribution derived from relatively sparse data, the correct distributional forms of all the variables input to UNC are somewhat uncertain. In particular, Koontz et al. assumed that the emission source strength was lognormally distributed because comparison of the arithmetic mean and the range suggested a very skewed distribution. However, they did not have direct information about the standard deviation of the data set. Uncertainty about the distributional form of a highly skewed distribution is likely to be most influential at the upper end of the distribution, which is where our analysis showed a discrepancy. In addition for this analysis there is uncertainty about the sampling frame for the building volumes, since they were derived from studies other than the PTEAM study, from which the observed concentrations were taken.

**Figure 6-1**
**Plot Generated by UNC that Shows the Variation in the Distributions Calculated by CPIEM.**



## Table 6-5
## Model Validation for Benzo(a)pyrene with CPIEM and UNC.

| Statistic | PTEAM Study | Original CPIEM[a] | Uncertainty Distributions from CPIEM and UNC | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | Std Dev | Median | 2.50% | 97.50% | Minimum | Maximum |
| Arith Mean | 0.70 | 0.68 | 0.67 | 0.21 | 0.63 | 0.43 | 1.05 | 0.43 | 1.05 |
| Arith Std Dev | 4.00 | 2.17 | 2.26 | 0.97 | 1.97 | 1.13 | 3.76 | 1.13 | 3.76 |
| Geo Mean | | | 0.20 | 0.03 | 0.2 | 0.16 | 0.27 | 0.16 | 0.27 |
| Geo Std Dev | | | 3.82 | 0.68 | 3.64 | 3.05 | 5.27 | 3.05 | 5.27 |
| Minimum | | 0.02 | 0.01 | 0.01 | 0.01 | 0.00 | 0.02 | 0.00 | 0.02 |
| Maximum | | 17.6 | 34.1 | 15.3 | 30.5 | 14.7 | 66.2 | 14.7 | 66.2 |
| 5% | | | 0.03 | 0.01 | 0.03 | 0.01 | 0.06 | 0.01 | 0.06 |
| 10% | NQ[b] | 0.04 | 0.04 | 0.01 | 0.04 | 0.02 | 0.08 | 0.02 | 0.08 |
| 15% | | | 0.06 | 0.02 | 0.05 | 0.03 | 0.09 | 0.03 | 0.09 |
| 20% | | | 0.07 | 0.02 | 0.06 | 0.04 | 0.10 | 0.04 | 0.10 |
| 25% | 0.08 | 0.08 | 0.08 | 0.02 | 0.07 | 0.05 | 0.12 | 0.05 | 0.12 |
| 30% | | | 0.09 | 0.02 | 0.09 | 0.07 | 0.13 | 0.07 | 0.13 |
| 35% | | | 0.11 | 0.02 | 0.11 | 0.09 | 0.15 | 0.09 | 0.15 |
| 40% | | | 0.13 | 0.02 | 0.13 | 0.10 | 0.18 | 0.10 | 0.18 |
| 45% | | | 0.15 | 0.02 | 0.15 | 0.12 | 0.20 | 0.12 | 0.20 |
| 50% | 0.19 | 0.15 | 0.17 | 0.02 | 0.17 | 0.13 | 0.22 | 0.13 | 0.22 |
| 55% | | | 0.20 | 0.03 | 0.20 | 0.16 | 0.25 | 0.16 | 0.25 |
| 60% | | | 0.23 | 0.03 | 0.23 | 0.18 | 0.28 | 0.18 | 0.28 |
| 65% | | | 0.27 | 0.04 | 0.27 | 0.21 | 0.34 | 0.21 | 0.34 |
| 70% | | | 0.33 | 0.06 | 0.33 | 0.25 | 0.44 | 0.25 | 0.44 |
| 75% | 0.36 | 0.36 | 0.42 | 0.08 | 0.41 | 0.29 | 0.55 | 0.29 | 0.55 |
| 80% | | | 0.54 | 0.13 | 0.53 | 0.35 | 0.76 | 0.35 | 0.76 |
| 85% | | | 0.78 | 0.22 | 0.74 | 0.49 | 1.25 | 0.49 | 1.25 |
| 90% | 0.65 | 1.15 | 1.25 | 0.4 | 1.15 | 0.74 | 2.02 | 0.74 | 2.02 |
| 95% | | | 2.51 | 0.95 | 2.16 | 1.42 | 4.58 | 1.42 | 4.58 |
| 100% | | | 34.14 | 15.25 | 30.53 | 14.74 | 66.21 | 14.74 | 66.21 |

a Averages across 10 model runs with different random number seeds

b Not quantifiable

*This page intentionally left blank.*

# 7. Recommendations for Future CPIEM Enhancements

Not all the potential improvements identified for CPIEM during this project were implemented. In this section, we discuss enhancements that we recommend for future projects.

## 7.1. Breathing Rates

The breathing rates in the current CPIEM are assigned by three population groups: Adult (12 or older) male, adult female, or child. They are also assigned by activity level: heavy, moderate, light, or resting. The current software only allows the user to specify a single breathing rate value for each population group and activity level combination; default values are also provided. A more realistic model would take into account the variation of the breathing rate, which varies a) by person, b) between the specific activities of an activity level, c) during the time period of the activity. Based on a review of the OEHHA's *Technical Support Document for Exposure Assessment and Stochastic Analysis* (OEHHA, December 1996), and of other literature, the variation by person is primarily explained by variation in age, gender, and body mass or body surface area. We recommend that the model be enhanced so that the user is permitted to supply data input distributions for the breathing rate by population group and activity level. The set of options for defining these breathing rate distributions could be expanded to the same as for the current model inputs for concentrations, and the user could also be allowed to define uncertainty distributions for these breathing rate data input distributions (as described in our recommendations for adding uncertainty modeling).

We recommend that default variability distributions be developed and used in CPIEM in the form of data sets of simulated breathing rates, based on simulating age, gender, body mass (as a random function of age and gender), and, finally, breathing rate per kg body mass. These data sets of simulated values would be developed external to the model, and supplied as default distributions. Our recommended approach is given in detail in Appendix A.

## 7.2. Population Activity Data

The US EPA has compiled the Consolidated Human Activity Database (CHAD) from several survey studies conducted in the US, including those sponsored by ARB. The database also contains other studies that include subjects residing in California. Data on these additional California residents could be extracted from CHAD for incorporation into the CPIEM activity pattern database. However, this would require a substantial amount of data processing both to cross-match the activity coding of the two databases and to format the CHAD data for CPIEM, and thus could not included in the current project.

Options for Characterizing Data Input Distributions

In the current version of the model, the variability of a model input can be defined by the user by several different methods. The user can:

1. Select from a list of default distributions;
2. Specify the parameters for one of a set of several parametric distributions (e.g., normal, lognormal, uniform);
3. Supply a data set of possible values, all assumed to have equal probabilities, i.e., an empirical distribution;

4. Supply a set of percentiles (including the 100$^{th}$ percentiles = maximum), i.e., a cumulative frequency distribution giving the (100p)$^{th}$ percentile values for selected values of p; or

5. Specify a set of fixed probability weights (adding up to 100 %) for two or more variability distributions (default and/or user-specified), defining a distribution mixture.

Table 7-1 lists the options currently available to the user for specifying the distribution of each model input. Table 7-1 shows that the options for the exponential distribution, user-supplied data set, and user-defined weights are not available for every model input. In principle, it would be desirable to enhance the model so that all these options are available for all of the model inputs, except for breathing rate (discussed elsewhere in this report) and start time. We recommend implementing this modification in the future.

In principle, options for additional statistical distributions such as a gamma, chi-square, or Weibull distribution could be added to the CPIEM. These additional distributions are often used by researchers to characterize distributional shapes. We recommend implementing this modification in the future. In the meantime, we note that any required distribution can be simulated to an arbitrary accuracy by inputting a suitably large data set of values simulated from that distribution. These distributions can also be approximated using the available percentile representation option.

Another useful improvement to the model would be allowing the user the convenient alternative of specifying the parameters for the lognormal distribution as the geometric mean and standard deviation in addition to the existing option of specifying the parameters as the arithmetic mean and standard deviation.

## 7.3. Uncertainty Analysis Functions

In this subsection we will briefly summarize recommendations for further enhancing the capabilities of CPIEM for uncertainty assessment. The various recommendations are listed below:

1. Incorporate all the features of UNC (uncertainty distribution solicitation, simulations, and statistical analyses) into CPIEM.

2. Display histograms of default and user-specified continuous uncertainty distributions on input. Display uncertainty/variability graphs.

3. Provide an option for displaying additional uncertainty percentiles in the tables and graphs.

4. Provide an option for a table displaying the percentages of exposure, concentrations, or dose less than a threshold.

5. Provide an option for uncertainty analysis of data sets using a nonparametric bootstrap.

*1. Incorporate all the features of UNC (uncertainty distribution solicitation, simulations, and statistical analyses) into CPIEM.*

This recommendation would create a seamless piece of software that would eliminate the requirement for the user to separately enter uncertainty distributions into UNC, hand enter the selected parameter values into CPIEM, and then compile the statistics summary files for entry into UNC. In particular, the uncertainty and variability distributions would be solicited at the same time.

We recommend a screen design for soliciting variability and uncertainty distribution information as follows. Each screen that solicits distribution type and parameters for the variability analysis would include a checkable box for performing an uncertainty analysis. If (and only if) that box is checked, appropriate secondary fields would appear on the same screen to specify the distribution type and parameters for the uncertainty distribution. When those fields appear, they would be populated with the system-generated default uncertainty distributions, and users can modify them, subject to some basic feasibility checking. The user would be able to specify discrete or continuous uncertainty distributions or to use the case name uncertainty option for mixtures. This design is uncluttered, keeps logically related information close together, and allows the users to see the default values.

*2. Display histograms of default and user-specified uncertainty distributions on input. Display uncertainty/variability graphs.*

For each model input distribution, the default uncertainty option produces continuous uncertainty distributions for the parameters of the input distribution based on the sampling variability. A user may instead select his or her own uncertainty distributions for the input distribution parameters. This recommendation would allow for the automatic display of a histogram of these default and/or selected distributions so that the user can examine whether those distributions match the user's uncertainty beliefs.

These histograms would be very useful because the user can immediately see if the distributions are highly inconsistent with the selected parameter values (such as a distribution with a large percentage of negative values being used for the uncertainty of a parameter constrained to be positive), or with the user's uncertainty beliefs.

A further, related enhancement is for the program to generate a graph of the variability distribution for various levels of uncertainty for each model input based on the user's selected or default uncertainty distribution. This graph could be computed and displayed immediately after the selections are made for that model input. The graph would be similar to the model output graph described above, showing the variability distribution at the median, $2.5^{th}$ and $97.5^{th}$ percentiles of uncertainty. Such a graph would give the user an indication of the relative roles of variability and uncertainty for the given CPIEM input. If the range of uncertainty is absurdly wide or narrow because of the uncertainty assumptions made, the user will see this using the graph, and may wish to revisit those assumptions.

*3. Provide an option for displaying additional uncertainty percentiles in the tables and graphs.*

The model output tables include the $50^{th}$, $2.5^{th}$, $97.5^{th}$, $0^{th}$ and $100^{th}$ percentiles of uncertainty. For some users, a more detailed analysis would be desirable. The user could be given the option to select additional percentiles to be tabulated and/or graphed.

*4. Provide an option for a table displaying the percentages of exposure, concentrations, or dose less than a threshold.*

This proposed output focuses on the uncertainty of the percentages of concentrations or exposure/doses below given thresholds, rather than on the uncertainty of the percentiles. First, the set of thresholds (concentrations or exposure/doses) of interest will need to be selected. The most convenient option would be for the user to select the thresholds.

Second, as before, the CPIEM model runs will provide V (or 24V) concentrations or exposure/doses for each of U randomly selected uncertainty vectors. For each threshold x, and

each of the U uncertainty vectors, the percentage of simulated values less than x is computed from the V (or 24V) replications. The arithmetic mean and standard deviation of these U percentages is tabulated against x.

*5. Provide an option for uncertainty analysis of data sets using a nonparametric bootstrap.*

The default uncertainty distributions provided in the current model are the sampling distributions of the parameter estimates for each continuous model input distribution using a parametric bootstrap. These uncertainties represent the uncertainty due to sampling error of the selected continuous distributions. In a similar manner, the uncertainty due to sampling error for a discrete distribution can be accounted for using the nonparametric bootstrap: For each data set sample of n values, U new data sets of n values each are created by sampling at random with replacement from the original data set. These U data sets are the model input data sets for the U uncertainty simulations. The same approach could be extended for distributions defined by percentiles.

**Table 7-1**
**Options for model inputs (X means that this option is currently available)**

| Input | Normal | Lognormal | Triangular | Uniform | Exponential | Percentile | Data set | Weights |
|---|---|---|---|---|---|---|---|---|
| Concentration | X | X | X | X | | X | X | X |
| Breathing Rate | | | | | | | Enter values or use defaults | |
| Load Factor | X | X | X | X | | X | | |
| When Installed | X | X | X | X | | X | | |
| Initial Emission Rate | X | X | X | X | | X | | |
| Decline In Rate | X | X | X | X | | X | | |
| Quantity Present | X | X | X | X | | X | | |
| Time Since Use | X | X | X | X | X | X | | |
| Duration of Use | X | X | X | X | | X | | |
| Episodes Per Day | X | X | X | X | | X | Enter frequency distribution | |
| Start Time | | | | | | | Enter percent share by hour | |
| Penetration Factor | X | X | X | X | | X | X | X |
| Indoor Sinks Decay Rate | X | X | X | X | | X | X | X |
| Building Volume | X | X | X | X | | X | X | X |
| Air Exchange Rate | X | X | X | X | | X | X | X |

# 8. References and Attachments


Avol, Edward L., Navidi, William C., and Colome, Steven D. Modeling Ozone Levels in and around Southern California Homes. *Environmental Science & Technology*, FEB 15 1998, v 32 n 4, p 463+.

Colombo, A., De Bortoli, M., Knoppel, H., Pecchio, E., and Vissers, H. 1993. Adsorption of selected volatile organic compounds on a carpet, a wall coating, and a gypsum board in a test chamber. *Indoor Air* 3:276-282

Colome, S.D., Fung, K., Behrens, D.W., Billick I.H., Tian, Y., and Wilson A.L. 1994. *Benzene and toluene concentrations inside and outside of homes in California.* Presented at the Air and Waste Management Association 87th Annual Meeting and Exhibition. Cincinnati, OH. 94-WP90.03.

Daisey, J.M.; Hodgson, A.T.; Fisk, W.J.; Mendell, M.J.; Brinke, J.T. Volatile Organic Compounds in Twelve California Office Buildings: Classes, Concentrations and Sources, Report Number: LBL-33686 1993. *Atmospheric Environment*, Vol. 28, No. 22, pp. 3557-3562, 1994

Koontz, M.D., W.C. Evans, and C.R. Wilkes. 1998. *Development of a Model for Assessing Indoor Exposure to Air Pollutants.* California Air Resources Board Contract No. A933-157

Lewis, C.W. 1991. Sources of air pollutants indoor: VOC and fine particulate species. *Journal of Exposure Analysis and Environmental Epidemiology* 1(1) 31-44.

Ligocki, M.P., L.G. Salmon, T. Fall, M.C. Jones, W.W. Nazaroff and G.R. Cass. 1993. Characteristics of airborne particles inside Southern California museums. *Atmos. Environ*. 27A(5):697-711.

Offermann, F. J.; S.A. Loiselle; J.M. Daisey; L.A. Gundel; and A.T. Hodgson. 1990. "A Pilot Study to Measure Indoor Concentrations of Polycyclic Aromatic Compounds." In: *Indoor Air '90, Proceedings of the International Conference in Indoor Air Quality and Climate, Ottawa, Canada,* 2: 379-384.

Peters, J.M. 1997, Epidemiological Investigation to Identify Chronic Health Effects of Ambient Air Pollutants on Southern California. USC-LA, Final Report to CARB, NTIS No. PB 98-140833/XAB

Rodes, C., L.Sheldon, D.Whitaker, A.Clayton, K, Fitzgerald, J. Flanagan, F.DiGenova, S. Hering, and C. Frazier. 1998. *Measuring Concentrations of Selected Air Pollutants Inside California Vehicles*. Prepared for the California Air Resources Board, Contract No, 95-339.

Spicer, C.W., Kenny, D.V., Ward, G.F., and Billick, I.H., 1993, *Journal of the Air & Waste Management Association*, v 43 n 11 , p1479+.

Underwood, M.C., 1996, Assessing the indoor air impact from a hazardous waste site: A case study. *Toxicology and Industrial Health*, 12: 179-188.

Wilson, A.L., S.D.Colome, Y.Tian, E.W.Becker, P.E.Baker, D.W.Behrens, I.H. Billick, C.A. Garrison. California residential air exchange rates and residence volumes. *J Expo Anal Environ Epidemiol*; Vol 6(3):311-326.

Womble, S.E., Girman, J.R., Ronca, E.L., Brightman, H.S., and McCarthy, J.F. (1995). "Developing Baseline Information on Buildings and Indoor Air Quality (BASE '94): Part I-Study

Design, Building Selection, and Building Descriptions." Proceedings of Healthy Buildings '95 Milan, Italy, Vol. 3, 1995, pp. 1305-1310.

## Attachment A

Benzaldehyde
Benzene
Butyl acetate
2-Butoxyethanol
n-Decane
Dicholoromethane
n-Dodecane
Ethanol
Ethyl acetate
Ethylbenzene
2-Ethyltoluene
3- & 4-Ethyltoluene
n-Heptane
Hexanal
n-Hexane
Limonene
Methylcyclopentane
Methylcyclohexane
3-Methylhexane
n-Nonane
n-Octane
n-Pentanal
n-Pentane
1-Phenylethanone
2-Propanol
2-Propanone
Styrene
Tetracholoethane
Toluene
Trichlorofluoromethane
Trichloroethene
1,1,1-Trichloroethane
1,2,3-Trimethylbenzene
1,2,4-Trimethylbenzene
1,3,5-Trimethylbenzene
2,2,5-Trimethylhexane
n-Undecane
m,p-xylene
o-xylene
TVOC

## Attachment B

Naphthalene
2- Methylnaphthalene
Methylnaphthalene
Biphenyl
Acenaphthalene
Flourene
Phenanthrene
Anthracene
2-Methylanthracene
9-Methylanthracene
Fluoranthene
Pyrene
Chrysene

## Attachment C

| VOCs | Elements | Other |
|---|---|---|
| Isobutylene | Lead | PM10 |
| 1,3-Butadiene | Cadmium | PM2.5 |
| Acetonitrile | Chromium | CO |
| DCM | Manganese | Black Carbon |
| MTBE | Nickel | |
| Benzene | Sulfur | |
| Toluene | | |
| Ethylbenzene | | |
| m,p-xylene | | |
| o-xylene | | |
| TCFM | | |
| Formaldehyde | | |

## Attachment D

| VOCs | Elements |
|---|---|
| Ethylene | Silicon |
| Acetylene | Sulfur |
| 2-Methypentane | Calcium |
| Methylcyclopentane | Iron |
| Benzene | Lead |
| Cyclohexane | Zinc |
| 3-Methylhexane | |
| 2,2,4-Trimethylpentane | |
| Toluene | |
| Ethylbenzene | |
| m,p –xylene | |
| o-xylene | |
| paraffins | |
| olefins | |
| aromatics | |
| TNMHC | |

# Appendix A:
# Development of Default Breathing Rate Distributions

OEHHA (1996) describes various approaches to estimating breathing rates and summarizes the results of various studies. Information from that report has been summarized in the following discussion. This technical support document (OEHHA, 1996) was downloaded from the OEHHA website. Although this "draft" document is marked "Do not cite or quote," the fact that the comment period ended several years ago and yet the document is still on the public website suggests that the information in that document can now be used freely. Furthermore, the citation and location for this document was provided to us by the ARB contract manager.

A commonly used approach begins with a compilation of various laboratories and field studies on measured breathing rates for different activities. Typically, the various activities are grouped into activity levels, such as resting, light activity, moderate activity, heavy activity and the compiled studies are averaged together to provide a single breathing rate for each activity level (e.g., Snyder et al, 1975; USEPA, 1985; USEPA, 1989; AIHC, 1994). If sufficient data are available, then a representative distribution of breathing rates by activity level can also be developed (e.g., AIHC, 1994). To estimate daily breathing rates, estimates of the time spent at the various activity levels are used to weight the breathing rates by activity level; these activity pattern distributions can be obtained either by making simple assumptions or by direct surveys. For enhancing the CPIEM, it is inappropriate to apply daily breathing rates averaged over all activities (or activity levels) since the model itself time-weights the breathing rate by activity level.

An important consideration is the variation of breathing rate between different people doing the same activity, or at the same activity level. Some researchers have addressed this issue by expressing the breathing rate as $m^3$/hour-kg of body mass, or, similarly, normalize by body surface area rather than body mass. These normalizations adjust for the observed increases of the breathing rate with body size.

Layton (1993) developed three alternative approaches for estimating breathing rates based on energy expenditure, EE (e.g., kilocalories per hour). The ventilation rate equals EE multiplied by the volume of oxygen consumed per unit energy (H) and the ventilatory equivalent (VQ), which is the ratio of the ventilation rate of air to the oxygen uptake rate:

Ventilation Rate ($m^3$/hr)

= Energy Expenditure (kCal/hr) × Energy Conversion Factor ($m^3$/kCal) × VQ.

First, he developed a model for EE based on food caloric intake. Second, he developed a model where EE is estimated by multiplying the basal metabolic rate (BMR) by a constant energy expenditure factor (EEF) that represents the relative energy expenditure for normal activity compared to the basal metabolic rate:

EE = BMR × EEF.

A third approach is similar to the second approach but applies different EEFs for different activity levels.

More recently, McCurdy (2000) has recommended a similar energy expenditure rate model that is used in the pNEM/CO model currently being developed by ICF Consulting, TRJ Environmental, and James Capel for US EPA. For pNEM/CO, the ventilation rate of interest is

the alveolar ventilation rate, representing the portion of breathed-in air that enters the alveoli and exchanges with the blood air. The alveolar ventilation rate equals EE multiplied by H and AVQ, the alveolar ventilatory equivalent, which is the ratio of the ventilation rate of alveolar air to the oxygen uptake rate. The energy expenditure for a specific activity is estimated by multiplying the resting metabolic rate (RMR) by an activity-specific ratio (MET, the "metabolic equivalent of work") that represents the relative energy expenditure for the activity compared to the resting metabolic rate. The distribution of RMR is estimated from the body mass, which, in turn, is estimated from distributions of body mass by age and gender found in the literature. More specifically, the model is built up from the following 5 steps:

1. Body Mass (kg) is lognormally distributed with a mean and variance that depend on age and gender (Brainard and Burmaster, 1992).

2. Resting Metabolic Rate (kCal/hr) = RMR

   RMR = a + b × Body Mass + error,

   where the errors are normally distributed with mean zero. The parameters a, b and the error variance depend on the age and gender, as given in Scholfield, (1985).

3. Energy Expenditure (kCal/hr) = EE = RMR × MET

   MET, the "metabolic equivalent of work," is a dimensionless, activity-specific ratio. The MET distributions compiled by McCurdy (1998) for each activity, age group, and gender are also specified in EPA's Consolidated Human Activity Database (CHAD).

4. Oxygen Uptake Rate ($m^3$/hr) = EE × H,

   where

   H = Energy Conversion Factor ($m^3$/kCal), assumed to be approximately uniformly distributed between 0.20 and 0.21 (based on Esmail et al., 1995).

5. Alveolar Ventilation Rate ($m^3$/hr) = Oxygen Uptake Rate × AVQ.

   AVQ, the alveolar ventilatory equivalent, is approximately 19.63 (based on Journard et al., 1991).

OEHHA (1996) also summarize results from an ARB sponsored laboratory and field study of breathing rates by activity (Adams, 1993). They developed breathing rate per kg body mass distributions using the data from that study combined with the California activity pattern data for children, adolescents, and adults that is included in the current CPIEM (Phillips et al, 1991; Wiley et al, 1991a, Wiley et al, 1991b). These normalized distributions are given in the form of summary statistics for: sample size, mean, standard deviation, skewness, kurtosis, and selected percentiles. Separate distributions for liters per minute per kg body mass are provided to represent: adult men resting, adult women resting, combined men and women resting, combined men and women in light activity, combined men and women in moderate activity, and combined men and women in heavy activity.

This review leads to a couple of options for obtaining default breathing rate distributions for use in CPIEM instead of the current single values. The simplest, and recommended, option is based on the distributions developed by OEHHA for breathing rate by kg body mass by activity level. For each population group (adult males, adult females, and children), the Californian distribution of age and gender can be obtained from available California census data. The age and gender

of that person is selected at random from that distribution. Based on the age and gender, the body mass is sampled from the lognormal distributions available in Brainard and Burmaster (1992). Finally, the breathing rate per kg for the specific activity level is sampled from the distributions given in the OEHHA report (assuming the distribution is suitably smoothed out between the available percentiles). The gender-specific distributions for resting will be used for male and female adults resting. The distributions combined for men and women would be used for the other combinations of population group and activity level (including children). The simulated breathing rate in $m^3$ per hour is the product of the sampled normalized breathing rate and the sampled body mass.

Breathing Rate ($m^3$/hr) = Normalized Breathing Rate ($m^3$/hr-kg) × Body Mass (kg).

Repeating this set of simulations sufficiently many times gives a dataset of breathing rates for each combination of activity level and population group that can be used in CPIEM.

A more accurate, but more resource-intensive, approach is based on the pNEM/CO simulation model. Due to the availability of resources, this approach is not recommended for this research project. The first step would be to the use the set of CHAD distributions for the MET ratio (McCurdy, 1998) to assign MET distributions for each specified activity, age, and gender in the California activity pattern data. For each person in the activity pattern database, a MET value is sampled for every activity and a time-weighted average MET is computed for the resting, light, moderate, and heavy activity levels. Using the same distributions used in pNEM/CO, summarized above, for each activity level and person, values are sampled for body mass, hence resting metabolic rate (RMR), hence the energy expenditure (EE), using the average MET, and finally the oxygen uptake rate. This is the same as the pNEM/CO approach, steps 1 to 4 above, except that the individual activities are aggregated to the four activity levels.

In CPIEM, the breathing rate of interest is overall ventilation not alveolar ventilation. To simulate breathing rate from oxygen uptake rate, a reasonable approach would be to use the results of ICF's analysis of the Adams (1993) data that was developed for an interim version of the pNEM/CO model. The ICF analysis showed that the logarithm of the breathing rate per kg could be treated as an intercept plus a multiple of the logarithm of the oxygen uptake rate per kg plus a normally distributed error term.

Log (Breathing Rate / Body Mass)

= a + b × Log (Oxygen Uptake Rate / Body Mass) + Error.

The intercept, slope, and error variance parameters are age group and gender-specific. Using these equations, the breathing rate can be simulated from the oxygen uptake rate. This gives a simulated breathing rate for each person in the activity pattern database and each activity level. The datasets of simulated breathing rates by activity level, age group, and gender can be obtained by simply combining the set of simulated breathing rates by activity level for all persons in each population group (adult males, adult females, children), based on one or more simulated breathing rates per person. A more representative analysis would be based on selecting survey respondents at random using the age and gender distributions in the current California population.

## References


AIHC (1994). Exposure Factors Sourcebook. American Industrial Health Council, pp. 6.39 -6.43.

Adams (1993). Measurement of Breathing Rate and Volume in Routinely Performed Daily Activities. Final Report. Human Performance Laboratory, Physical Education Department, University of California, Davis. Prepared for the California Air Resources Board, Contract No. A033-205, April 1993.

Brainard, J, and D. Burmaster. 1992. A Bivariate Distributions for Height and Weight of Men and Women in the United States. Risk Analysis, Vol. 12, No. 2, pp. 267 - 275.

Esmail, Bhambhani, and Brintnell. 1995. A Gender Differences in Work Performance on the Baltimore Therapeutic Equipment Work Simulator. Amer. J. Occup. Therapy. Volume 49, pp. 405 - 411.

Layton D. (1993) Metabolically consistent breathing rates for use in dose assessments. Health Phys 64: 23-36.

McCurdy, T. 1998. Personal Communication. Information on distributions available upon request to author. U.S. Environmental Protection Agency, National Exposure Research Laboratory, Research Triangle Park, North Carolina 27711.

McCurdy, T., 2000. A Conceptual Basis for Multi-Route Intake Dose Modeling Using an Energy Expenditure Approach. Journal of Exposure Analysis and Environmental Epidemiology, Volume 10, pp. 1 - 12.

OEHHA (1996). Technical Support Document for Exposure Assessment and Stochastic Analysis. Draft for Public Review (December, 1996).

Phillips TJ, PL Jenkins, EJ Mulberg (1991). Children in California: Activity patterns and presence of pollutant sources. Proceedings of the 84th Annual Meeting and Exhibition of the Air and Waste Management Association, Vancouver, British Columbia. Vol. 17.

Schofield, W. N. 1985. A Predicting Basal Metabolic Rate, New Standards, and Review of Previous Work. Hum. Nutr. Clin. Nutr. Volume 39C, Supplement 1, pp. 5 - 41.

Snyder WS, MJ Cook, ES Nasset, LR Karhausen, GP Howells, IH Tipton (1975). Report of the Task Group on Reference Man, International Commission on Radiological Protection No. 23, Pergamon Press: Oxnard, 1975, pp. 338-347.

U.S. EPA (1985). Development of Statistical Distributions or Ranges of Standard Factors Used in Exposure Assessments. U.S. Environmental Protection Agency, Office of Health and Environmental Assessment, Washington, D.C. EPA/600/8-85/010

U.S. EPA (1989). Exposure Factors Handbook, U.S. Environmental Protection Agency, Office of Health and Environmental Assessment, Washington D.C., PB90-106774

Wiley JA, JP Robinson, T Piazza, K Garrett, K Cirkensa, YT Cheng, G Martin (1991a). Activity Patterns of California Residents. Final Report. Survey Research Center, University of California, Berkeley. Prepared for California Air Resources Board, Contract No. A6-177-33, May 1991.

Wiley JA, JP Robinson, YT Cheng, T Piazza, L Stork, K Pladsen (1991b). Study of Children's Activity Patterns, Final Report. Survey Research Center, University of California, Berkeley. Prepared for California Air Resources Board, Contract No. A733-149, September 1991.

*This page intentionally left blank.*

# Appendix B: Uncertainty Topics

## Number of Simulations

The optimum number of simulations (U and V) that the user should select cannot be determined in advance, since it depends on several considerations.

The aspect of this problem that is easiest to deal with in general terms is the variation of the "Monte Carlo precision" with U and V. The "Monte Carlo precision" compares the model results for a finite U and V with the limiting case where U and V are both infinite; in the infinite case, the only possible errors are in the specification of the model inputs and in the equations used to represent the physics of indoor air exposure. Except in the extreme case where all the model inputs have only one possible value, rather than a distribution of two or more values (infinitely many for continuous distributions), increasing U or V will always improve the precision of the Monte Carlo simulation. This might suggest that the user should make U and V as large as possible, subject to limitations in execution time (due to the possible cost of computing time and the urgency of getting final results). However, since the standard deviation of an arithmetic mean of a simple random sample of n values is inversely proportional to the square root of n, the precision of the arithmetic mean, median, percentiles, and other summary statistics will be inversely proportional to the square root of n. Thus, for example, doubling V will only improve the precision by a factor of $\sqrt{2}$, i.e., about a 40 percent increase, while quadrupling V would double the precision. For U, this calculation does not apply, because UNC uses the more efficient (i.e., more precise) method of Latin HyperSquare sampling to select the uncertainty vectors instead of simple random sampling, but one can expect the precision to be even less sensitive to changes in U. Once U and V are suitably large, then further increases in U and V will make negligible improvements (the law of diminishing returns).

The sizes of U and V needed for sufficient Monte Carlo precision will depend upon the pollutant of interest and the variances in the associated model inputs. If all the model data inputs have small variances, and if the input uncertainty distributions are also narrow, then small U and V values (say, 50) will be sufficient. If the data inputs have large variances, then V should be relatively large to capture the full ranges of those distributions. If the uncertainty distributions have large variances, then U should be relatively large to capture the full range of input uncertainty.

Note that increasing U or V will improve the numerical estimates of the overall mean variability distribution (averaged over all levels of uncertainty) and of the amount of uncertainty in the model outputs, but it will not reduce the variability or uncertainty of the results.

For selecting U, a critical issue is the effort required for data entry. Resource constraints prohibited making it possible for UNC to produce an electronic file of uncertainty vectors that could automatically be entered into CPIEM. Therefore, for each of the U uncertainty vectors, the user needs to manually re-enter into CPIEM a new set of parameter values for each uncertain CPIEM model input.

A practical approach is to first run the model with U and V set to be a reasonably small value such as 20. Now double V, keeping U at 20, and examine whether the summary statistics of exposure or dose changed significantly (based upon the expected health impact). The incremental data entry effort required to double V should be small since the U sets of model inputs can be saved. Relevant summary statistics of interest might be the mean, median, $5^{th}$ and $95^{th}$ percentile of variability, all evaluated at the median, $2.5^{th}$ and $97.5^{th}$ percentiles of

uncertainty. Keep on doubling V until those summary statistics become reasonably stable. Now keep V fixed and increase U by a small fixed amount, such as 10. Because of the Latin HyperSquare sampling used in UNC, it will be necessary to regenerate and reenter a new set of U + 10 uncertainty vectors into CPIEM rather than simply combining the results from the original U uncertainty vectors with results from 10 additional uncertainty vectors. Keep on increasing U by the same amount until the summary statistics become reasonably stable.

# Mixtures

If the input is a mixture of two or more distributions, then the user decides whether to: specify uncertainty about which distribution to use; specify uncertainty about the weights; or specify no uncertainty about the mixture. The user may also wish to specify uncertainty about the distributions themselves.

In the first case, the user is uncertain about which distribution to use. For example, this situation would arise if different studies reported different distributions for that input (e.g., the outdoor concentration distribution) and the user was uncertain about which studies to believe. In this case, the mixture weights represent uncertainty rather than variability, so they should NOT be entered as user-supplied weights in CPIEM. Instead, assign a case name label to each distribution and use the case name uncertainty option in UNC to determine which distribution is selected for each CPIEM run of V trials.

For example, suppose there are three distributions labeled as NORM1, LOGNORM, NORM2, and these distributions are equally likely, so that the weights are all equal. Then the case names would be entered into UNC as:

NORM1;LOGNORM;NORM2;

UNC will generate a list of U case names by randomly reordering the list of three case names U/3 times. (If U is not divisible by 3, then the number of re-orderings is the next highest integer to U/3, but only the first U case names are used). If U = 8, UNC might generate the list

NORM1;NORM2;LOGNORM;LOGNORM;NORM2;NORM1;NORM1;LOGNORM;

so that the distribution labeled as NORM1 is used for the first CPIEM run, NORM2 for the second run, LOGNORM for the third run, and so on. If the weights are not all equal, then some of the case names would be entered more than once into UNC. For example, entering

NORM1;LOGNORM;NORM2;LOGNORM;

into UNC assigns uncertainty probability weights of 50 % to LOGNORM, 25 % to NORM1, and 25 % to NORM2, since LOGNORM appears twice. For U = 8, we might get

NORM1;NORM2;LOGNORM;LOGNORM;NORM2;LOGNORM;LOGNORM;NORM1;

In the second case, the user is uncertain about the weights. For example, this situation would arise if the distributions were indoor concentration distributions for kitchens with gas, electric, or wood stoves and different studies reported different fractions of residences with gas, electric, or wood stoves. In this case, the mixture weights represent variability across California residences, so they should be entered as user-supplied weights in CPIEM. Since the weights themselves are uncertain, assign a case name label to each set of weights and use the case name uncertainty option in UNC to determine which set of weights is selected for each CPIEM run of V trials. For example, suppose there are three sets of percentage weights labeled by the user in

an obvious manner as 33-33-33, 60-20-20, and 40-40-20, and these sets of weights are equally likely. Then the case names would be entered into UNC as:

33-33-33;60-20-20;40-40-20;

For U = 4, UNC might generate the list

60-20-20;40-40-20;33-33-33;60-20-20;

showing that the 60%, 20% and 20% weights should be used for the first CPIEM run.

In the third case, the possible distributions and user-supplied weights are treated as having no uncertainty and so this model input mixture would not be entered into UNC.

In the fourth case, one or more of the distributions within the mixture might be uncertain. For example, the model input might be a normal and a uniform distribution each with 50 % variability weights, but the parameters of the normal and uniform distributions might be uncertain. The mean of the normal distribution might be close to 2, but expert judgment assesses that the true mean is uniformly distributed from 1.9 to 2.1. To deal with the fourth case, treat each distribution within the mixture as a separate CPIEM model input. For the Figure 4-2 flowchart, the mixture of two distributions is treated as three separate CPIEM model inputs: the mixture, the first distribution, the second distribution. For the mixture, the user decides whether to specify case name uncertainty about which distribution, case name uncertainty about the weights, or no uncertainty. For the individual distributions, the user may specify no, default, continuous, or discrete uncertainty, as described below.

# Constraints on Parameters of Model Input and Uncertainty Distributions

There are some constraints on the parameters of model input distributions and of the parametric uncertainty distributions. When the user supplies UNC with uncertainty distributions for the model input distribution parameters, one of the constraints could be violated by a simulated parameter value. Also, the parameters of the uncertainty distribution should not violate the constraints. This section presents the treatment for each uncertainty option and parametric distribution.

### Case Name Uncertainty

If the case name uncertainty option is selected, then the distributions are labeled by the case name and there are no parameter constraints that could be violated.

### Default Uncertainty

If the default uncertainty option is selected, then the user enters into UNC the name (e.g., normal, uniform) and parameters of the model input distribution together with the sample size. If the selected parameters violate the parameter constraints for that distribution, an error message is immediately given after the data entry and the user will need to reenter a consistent set of parameter values. The sets of parameter values generated by UNC by resampling from the model input distribution cannot violate the parameter constraints.

### Discrete Uncertainty

If the discrete uncertainty option is selected, then the user enters into UNC the name of the distribution and the possible sets of parameter values. If the selected parameters in any of the

sets violate the parameter constraints for that distribution, an error message describing the violation is immediately given after the data entry. The user will need to reenter consistent sets of parameter values. UNC then randomly selects sets of parameter values from these consistent sets.

**Continuous Uncertainty**

If the continuous uncertainty option is selected, then the user enters into UNC the name of the distribution and specifies uncertainty distributions for each parameter. There are two ways in which the parameter constraints could be violated:

First, the parameters of a specified uncertainty distribution could violate the constraints for that distribution. For example, if the model input distribution is normal, then specifying the uncertainty distribution for the mean parameter as being a uniform distribution with a minimum of 3 and a maximum of 2 violates the constraints for the uncertainty distribution (since a uniform distribution has minimum <= maximum). If the selected parameters for an uncertainty distribution violate the parameter constraints for that distribution, an error message is immediately given after the data entry and the user will need to reenter a consistent set of parameter values.

Second, the simulated parameter values from the uncertainty distribution could violate the constraints for the model input distribution. For example, if the model input distribution is normal, then specifying the uncertainty distribution for the standard deviation parameter as being a uniform distribution with a minimum of –3 and a maximum of 2 could result in impossible negative values for the standard deviation parameter. UNC treats the situation in two ways depending on the type of model input distribution:

If the model input distribution is normal, lognormal, or exponential, then UNC automatically truncates the uncertainty distribution so that the truncated range of values does not violate the constraints. For the example, the uniform uncertainty distribution from –3 to 2 is replaced by a uniform distribution from 0 to 2, so that negative standard deviations are not generated. The UNC output includes a message that warns the user that the distribution of a specified model input distribution parameter was truncated and also reports the probability of getting an impossible value (60 % for the example).

If the model input distribution is uniform or triangular, then UNC generates the parameter values from the selected uncertainty distributions, but replaces impossible sets of parameter values. For the uniform distribution, if the simulated value for the minimum parameter exceeds the simulated value for the maximum parameter, then both parameters are replaced by the lower of the two simulated values. For the triangular distribution, if the constraint minimum <= mode <= maximum is violated by the simulated values, then all three parameters are replaced by the lowest of the three simulated values. The UNC output includes a message that warns the user that the parameters of a specified model input distribution were replaced because the constraints were violated and also reports the number of violations.

The warning messages should alert the user that they have poorly specified uncertainty distributions. For the normal, lognormal, or exponential model inputs, a large truncation probability (20 % or more) means that the truncated uncertainty distribution used by UNC is very different to what the user specified and so the user should carefully check their selections to make sure they reflect their uncertainty about the input distributions. For the uniform or triangular model inputs, the user should carefully review their selections if there are too many cases where constraint violations were reported.

### Normal Distribution

The mean of a normal distribution is unconstrained. The standard deviation of a normal distribution cannot be negative or zero. Note that the UNC program does not allow zero standard deviations although CPIEM does allow zero standard deviations. Instead of using a normal distribution with mean m and standard deviation 0, the user can specify the mathematically equivalent uniform distribution with minimum = m and maximum = m.

### Lognormal Distribution

Similarly, the lognormal distribution can be specified by its arithmetic mean, $\alpha$, and arithmetic standard deviation, $\beta$, neither of which can be negative or zero. UNC refers to this specification as the arithmetic lognormal distribution.

In UNC, the lognormal distribution may instead be specified by its geometric mean and geometric standard deviation. UNC refers to this specification as the geometric lognormal distribution. The geometric mean cannot be negative or zero. The geometric standard deviation cannot be one or less than one. The current version of CPIEM does not allow the lognormal distribution to be specified by the geometric mean and geometric standard deviation.

### Triangular Distribution

The triangular distribution is specified by three parameters: minimum, mode, maximum. These parameters cannot be independently defined since minimum $\leq$ mode $\leq$ maximum.

### Uniform Distribution

The uniform distribution is specified by two parameters: the minimum, *a*, and the maximum, *b*. The maximum has to be greater than or equal to the minimum.

### Exponential Distribution

The exponential distribution is specified by its mean, $\mu$, which cannot be negative or zero.

# Default Uncertainty Estimates for Parametric Distributions in the UNC Software

### Normal Distribution

Let *X* have a normal distribution with mean $\mu$ and standard deviation $\sigma$. Assume that the parameters $\mu$ and $\sigma$ are estimated from a random sample of size *n* from *X*, as follows.

$$\mu \approx \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad \text{(B-1)}$$

$$\sigma \approx S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}. \qquad \text{(B-2)}$$

The sampling distributions of $\overline{X}$ and *S* for independent samples of size *n* can be thought of as defining the uncertainty of the parameter values. This is a frequentist, objective, non-Bayesian

approach based on confidence intervals, instead of a subjective, Bayesian approach based on an expert's assessment of the probabilities for alternative model input distributions. (The Bayesian approach can be applied using the options for user-specified uncertainty distributions.) In UNC, the default uncertainty distributions are obtained directly by generating samples of size *n* from the given normal distribution, and computing the sample mean and standard deviation for each sample. The resulting distributions of sample means and sample standard deviations are used as estimates of the uncertainty of the true mean and true standard deviation of the model input distribution. This is the parametric bootstrap approach.

If the features of UNC and CPIEM are combined into a single piece of software in a future project, then, to avoid additional simulations that might substantially lengthen the run time, we recommend using an approximation to the theoretical sampling distribution. The details are given in C.

Appendix C shows that the approximate default uncertainty distributions for a normal distribution with mean m and standard deviation s based on n values are:

$\mu$ is normal, with mean = m, standard deviation = s/$\sqrt{n}$

$\sigma$ is s|W|$^{1.5}$,

where W is normal, with mean = 1 – 2/{9(n-1)}, standard deviation = $\sqrt{\{2/\{9(n-1)\}}$

As an example, consider the case of a normal distribution with mean 30, standard deviation 6, fitted to 31 values. The $\mu$ default uncertainty distribution is normal, with mean 30, standard deviation 6/$\sqrt{31}$ = 1.08. The W distribution is normal, with mean 1 – 2/270 = 0.993, standard deviation $\sqrt{\{2/\{270\}}$ = 0.086. The $\sigma$ default uncertainty distribution is 6|W|$^{1.5}$.

**Lognormal Distribution**

In UNC, the default uncertainty distributions for the arithmetic mean and arithmetic standard deviation are obtained by generating samples of size *n* from the given lognormal distribution, and computing the sample mean and standard deviation for each sample. The default uncertainty distributions for the geometric mean and geometric standard deviation are obtained by transforming the arithmetic parameters to the geometric parameters:

$$GM = \frac{\mu}{\sqrt{1+\frac{\sigma^2}{\mu^2}}} \qquad \text{(B-3)}$$

$$GSD = \exp\left(\sqrt{\ln\left(1+\frac{\sigma^2}{\mu^2}\right)}\right) \qquad \text{(B-4)}$$

(ln denotes natural logarithms.) Substitute the sample mean for $\mu$ and the sample standard deviation for $\sigma$ to obtain the geometric mean and geometric standard deviation parameters for that sample.

To obtain approximate default uncertainty distributions, a transformation can be applied to the default uncertainty distributions for the normal case. The details are given in Appendix C.

**Uniform Distribution**

The uniform distribution is specified in CPIEM by two parameters: the minimum, *a*, and the maximum, *b*. If the default uncertainty option is selected, then the uncertainty distribution for *a* and *b* will be defined by the sampling distribution of the usual (maximum likelihood) estimates of *a* and *b*:

$$\hat{a} = \min(X_1, X_2, \ldots . X_n) \qquad \text{(B-5)}$$

$$\hat{b} = \max(X_1, X_2, \ldots . X_n) \qquad \text{(B-6)}$$

For given values of *a*, *b*, and *n*, the sampling distributions of the statistics $\hat{a}$ and $\hat{b}$ can be derived by direct simulation of samples $(X_1, X_2, \mathrm{K}\ X_n)$ as in UNC. To reduce run times in a future project, the theoretical calculations in Appendix C can be used.

**Exponential Distribution**

To obtain the default uncertainty distribution, UNC simulates samples of size n and uses the sampling distribution of the sample mean, $\overline{X}$. To reduce run times in a future project, the theoretical calculations in Appendix C can be used.

**Triangular Distribution**

To obtain the default uncertainty distribution, UNC simulates samples of size n. For each sample, the values of the minimum and maximum parameters are the sample minimum and maximum. To estimate the mode parameter, the theoretical relationship between the mode and mean of a triangular distribution is used:

$$\text{mode} = 3 \times \text{mean} - \text{minimum} - \text{maximum} \qquad \text{(B-7)}$$

Thus the mode parameter is found by substituting the sample mean, minimum, and maximum into this equation. (The sample mode is a poor estimate of the true mode since with probability one all the sample values will be different and the sample mode is undefined).

*This page intentionally left blank*

# Appendix C:
# Default Uncertainty Distributions

## Normal Distribution

Let *X* have a normal distribution with mean $\mu$ and standard deviation $\sigma$. Assume that the parameters $\mu$ and $\sigma$ are estimated from a random sample of size *n* from *X*, as follows.

$$\mu \approx \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad \text{(C-1)}$$

$$\sigma \approx S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}. \qquad \text{(C-2)}$$

For the default uncertainty distributions of $\mu$ and $\sigma$ we recommend using the sampling distributions of $\overline{X}$ and *S*, which can be computed as follows. The exact sampling distribution of $\overline{X}$ is well known to be a normal distribution with mean $\mu$ and standard deviation $\sigma / \sqrt{n}$. The exact sampling distribution of *S* is also well known:

$S = \sigma \sqrt{\frac{C}{n-1}}$, where *C* has a chi-square distribution with $n - 1$ degrees of freedom. (C-3)

Furthermore, $\overline{X}$ and *S* are statistically independent. To avoid the need to generate data from a chi-square distribution, which requires numerical approximation, we recommend using the reasonably accurate Wilson-Hilferty approximation to the chi-square distribution:

$Z = \dfrac{\left(\dfrac{C}{n-1}\right)^{1/3} - 1 + \dfrac{2}{9(n-1)}}{\sqrt{\dfrac{2}{9(n-1)}}}$ approximately has a standard normal distribution. (C-4)

This gives

$\overline{X}$ is normal, mean = $\mu$, standard deviation = $\sigma / \sqrt{n}$, and

S is distributed approximately as the function $\sigma |W|^{3/2}$,

where $W = \left( \frac{C}{n-1} \right)^{1/3}$ is normal, with mean = $1 - \frac{2}{9(n-1)}$, and standard deviation = $\sqrt{\frac{2}{9(n-1)}}$.

To generate values from the default uncertainty distributions, the values of $\mu$ and $\sigma$ are set at the values specified for the variability distributions, and the given normal distributions for $\overline{X}$ and *W* are generated. Finally, *S* is calculated from *W* and $\sigma$ using the 3/2 power function.

## Lognormal Distribution

If *Y* is lognormal, and has mean $\alpha$ and standard deviation $\beta$, then it follows (after some algebra) that *X* = log (*Y*) is normally distributed with a mean $\mu$ and standard deviation $\sigma$, where

$$\begin{cases} \mu = \log \alpha - \frac{1}{2} \log \left( 1 + \left( \frac{\beta}{\alpha} \right)^2 \right), \\ \sigma = \sqrt{\log \left( 1 + \left( \frac{\beta}{\alpha} \right)^2 \right)} \end{cases} \quad \text{(C-5)}$$

All logarithms are assumed to be natural logarithms (base *e*). The uncertainty distributions for $\mu$ and $\sigma$ are generated using the equations for the normal distribution:

$\overline{X}$ is normal, mean = $\mu$, standard deviation = $\sigma / \sqrt{n}$. (C-6)

S is distributed approximately as the function $\sigma |W|^{3/2}$, (C-7)

where W is normal, with mean = $1-\frac{2}{9(n-1)}$, and standard deviation = $\sqrt{\frac{2}{9(n-1)}}$.

The inverse equations are:

$$\begin{cases} \alpha = e^{\mu + 0.5\sigma^2}, \\ \beta = e^{\mu}\sqrt{e^{\sigma^2}(e^{\sigma^2}-1)} \end{cases} \qquad \text{(C-8)}$$

Thus the generated values of ***a*** and ***b*** are given by

$$\begin{cases} a = e^{\overline{X} + 0.5S^2}, \\ b = e^{\overline{X}}\sqrt{e^{S^2}(e^{S^2}-1)} \end{cases} \qquad \text{(C-9)}$$

To generate values from the default uncertainty distributions: First, the values of $\mu$ and $\sigma$ are computed from equations (C-5), based on the specified $\alpha$ and $\beta$. Then equations (C-6) and (C-7) are used to generate the pair $\overline{X}$ and *S*. Finally, the simulated values of ***a*** and ***b*** are calculated from equations (C-9).

## Uniform Distribution

Let $(X_1, X_2, \mathrm{K}\ X_n)$ be a sample of *n* values from a uniform distribution on the interval (*a*, *b*). The maximum likelihood estimates of *a* and *b* are given by:

$$\begin{aligned} \hat{a} &= \min(X_1, X_2, \mathrm{K}\ X_n), \\ \hat{b} &= \max(X_1, X_2, \mathrm{K}\ X_n). \end{aligned} \qquad \text{(C-10)}$$

The cumulative distribution function for $\hat{a}$ is given by the equation:

$$F(x) = P(\hat{a} \le x) = 1 - P(X_1 > x, X_2 > x, \text{K}\ , X_n > x)$$
$$= 1 - P(X_1 > x)^n = 1 - \left(\frac{b-x}{b-a}\right)^n, \quad a \le x \le b. \tag{C-11}$$

Given the value of $\hat{a}$, the cumulative distribution function for $\hat{b}$ is:

$$G(x \mid \hat{a} = y) = P(\hat{b} \le x \mid \hat{a} = y) = nP(\hat{a} = X_1 \mid \hat{a} = y)P(\hat{b} \le x \mid \hat{a} = y \text{ and } \hat{a} = X_1)$$
$$= n\left(\frac{1}{n}\right)P(X_1 \le x, X_2 \le x, X_3 \le x, \Lambda\ X_n \le x \mid X_1 = y, X_2 > y, X_3 > y, \Lambda\ , X_n > y)$$
$$= \left\{P(y < X_2 \le x) / P(y < X_2)\right\}^{n-1} = \left(\frac{x-y}{b-y}\right)^{n-1}, \quad y \le x \le b.$$

Thus, the pair $\hat{a}$ and $\hat{b}$ has the same distribution as the pair *asim* and *bsim*, defined as follows:

$$asim = F^{-1}(1-U) = b-(b-a)U^{1/n},$$
$$bsim = G^{-1}(V \mid asim) = asim + (b-asim)V^{1/(n-1)}. \tag{C-12}$$

where *U* and *V* are independent samples from a standard uniform distribution on the interval (0, 1).

To generate values from the default uncertainty distributions, two independent samples, *U* and *V*, are drawn from a standard uniform distribution on the interval (0, 1), and *asim* and *bsim* are calculated from *U*, *V*, *a*, and *b*.

## Exponential Distribution

For the exponential distribution with population mean $\mu$, the standard (maximum likelihood) estimate of $\mu$ is the sample mean, $\overline{X}$, for a sample of size *n*. The exact sampling distribution of $\overline{X}$ is a multiple of a chi-square distribution:

$C = 2n\overline{X} / \mu$ has a chi-square distribution with 2*n* degrees of freedom. (C-13)

As for the normal case, we recommend employing the Wilson-Hilferty approximation to the chi-square distribution. This gives:

$\overline{X}$ is distributed approximately as the function $\mu|W|^3$,

where W = $\left(\frac{C}{2n}\right)^{1/3}$ is normal, with mean = $1-\frac{2}{9(2n)}$, and standard deviation = $\sqrt{\frac{2}{9(2n)}}$.

To generate values from the default uncertainty distribution, a sample is drawn from the given normal distribution *W*, and $\overline{X}$ is calculated from the samples *W* and $\mu$, which is set at the value specified for the variability distribution.

*This page intentionally left blank*