SAP Report No. 2000-01
May 25, 2000

PARTIAL REPORT [*]

FIFRA Scientific Advisory Panel Meeting, February 29-March 3, 2000, held at the Sheraton Crystal City Hotel, Arlington, Virginia

Sets of Scientific Issues Being Considered by the Environmental Protection Agency Regarding:

Session II - ***Dietary Exposure Evaluation Model (DEEM) and MaxLIP (Maximum Likelihood Imputation Procedure) Pesticide Residue Decompositing Procedures and Software***

Session III - ***Dietary Exposure Evaluation Model (DEEM)***

Session IV - ***Consultation on Development and Use of Distributions of Pesticide Concentrations in Drinking Water for FQPA Assessments***

[*] **Session I - Food Allergenicity of Cry9C Endotoxin and Other Non-digestible Proteins -** ***Report to be provided at a later date***

NOTICE

This report has been written as part of the activities of the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), Scientific Advisory Panel (SAP). This report has not been reviewed for approval by the United States Environmental Protection Agency (Agency) and, hence, the contents of this report do not necessarily represent the views and policies of the Agency, nor of other agencies in the Executive Branch of the Federal government, nor does mention of trade names or commercial products constitute a recommendation for use.

The FIFRA SAP was established under the provisions of FIFRA, as amended by the Food Quality Protection Act (FQPA) of 1996, to provide advice, information, and recommendations to the EPA Administrator on pesticides and pesticide-related issues regarding the impact of regulatory actions on health and the environment. The Panel serves as the primary scientific peer review mechanism of the EPA, Office of Pesticide Programs (OPP) and is structured to provide balanced expert assessment of pesticide and pesticide-related matters facing the Agency. Food Quality Protection Act Science Review Board members serve the FIFRA SAP on an ad-hoc basis to assist in reviews conducted by the FIFRA SAP. Further information about FIFRA SAP reports and activities can be obtained from its website at http://www.epa.gov/scipoly/sap/ or the OPP Docket at (703) 305-5805. Interested persons are invited to contact Larry Dorsey, SAP Executive Secretary, via e-mail at dorsey.larry@.epa.gov.

TABLE OF CONTENTS

Session II: Dietary Exposure Evaluation Model (DEEM) and MaxLIP (Maximum Likelihood Imputation Procedure)

Session III: Dietary Exposure Evaluation Model (DEEM)

Session IV: Consultation on Development and Use of Distributions of Pesticide Concentrations in Drinking Water for FQPA Assessments

SAP Report No. 2000-01B, May 25, 2000

REPORT:

FIFRA Scientific Advisory Panel Meeting,
March 1, 2000, held at the Sheraton Crystal City Hotel,
Arlington, Virginia

Session II - A Set of Scientific Issues Being Considered by the Environmental Protection Agency Regarding:

Dietary Exposure Evaluation Model (DEEM) and MaxLIP (Maximum Likelihood Imputation Procedure) Pesticide Residue Decompositing Procedures and Software

Ms. Laura Morris	Christopher Portier, Ph.D.,
Designated Federal Official	Session Chair
FIFRA/Scientific Advisory Panel	FIFRA/Scientific Advisory Panel
Date:____________________	Date:____________________

Federal Insecticide, Fungicide, and Rodenticide Act
Scientific Advisory Panel Meeting
March 1, 2000

SESSION II - Dietary Exposure Evaluation Model (DEEM) and MaxLIP (Maximum Likelihood Imputation Procedure) Pesticide Residue Decompositing Procedures and Softwares

PARTICIPANTS

FIFRA Scientific Advisory Panel Session Chair

Christopher Portier, Ph.D., National Institute of Environmental Health Sciences,
Research Triangle Park, NC

FIFRA Scientific Advisory Panel Members

Fumio Matsumura, Ph.D., Professor, Institute of Toxicology and Environmental Health, University of California at Davis, Davis, CA

Herbert Needleman, M.D., Professor of Psychiatry and Pediatrics, University of Pittsburgh, School of Medicine, Pittsburgh, PA

Mary Anna Thrall, D.V.M., Professor, Department of Pathology, College of Veterinary Medicine & Biomedical Sciences, Colorado State University, Fort Collins, CO

Food Quality Protection Act Science Review Board Members

Christopher Frey, Ph.D., Associate Professor, Civil Engineering, North Carolina State University, Raleigh, NC

David Gaylor, Ph.D., Associate Director for Risk Assessment Policy and Research, U.S. Department of Health and Human Services/FDA, National Center for Toxicological Research Jefferson, AR

Steve Heeringa, Ph.D., Statistician, Institute for Social Research, Ann Arbor, Michigan

Peter D. M. MacDonald, D. Phil., Professor of Mathematics and Statistics, Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada

Nu-may Ruby Reed, Ph.D., Staff Toxicologist, California Environmental Protection Agency, Department of Pesticide Regulation, Sacramento, CA

John Wargo, Ph.D., Associate Professor of Environmental Policy and Risk Analysis, Yale University, New Haven, CT

Designated Federal Official

Ms. Laura Morris, FIFRA Scientific Advisory Panel, Office of Science Coordination and Policy, Environmental Protection Agency, Washington, DC

PUBLIC COMMENTERS

Oral statements were made by:
Christine F. Chaisson, Ph.D., on behalf of Science, Strategies and Analysis Systems
Kim Travis, Ph.D., on behalf of Zeneca Agrochemicals
Leila Barraj, D.Sc., on behalf of Novigen Sciences, Inc.

Written statements were received from:
Christine F. Chaisson, Ph.D., on behalf of Science, Strategies and Analysis Systems

INTRODUCTION

The Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), Scientific Advisory Panel (SAP) has completed its review of the set of scientific issues being considered by the Agency regarding issues pertaining to the assessment of the Dietary Exposure Evaluation Model (DEEM) and the Maximum Likelihood Imputation Procedure (MaxLIP) pesticide residue decompositing procedures and software. Advance notice of the meeting was published in the *Federal Register* on February 4, 2000. The review was conducted in an open Panel meeting held in Arlington, Virginia, on March 1, 2000. The session was chaired by Christopher Portier, Ph.D. Ms. Laura Morris served as the Designated Federal Official.

In estimating dietary exposure to pesticides, the Agency uses several sources for monitoring data of pesticide residues in foods. These monitoring data, however, are in the form of pesticide residues on composited samples and do not directly represent concentrations of pesticide residues in single food items. For acute dietary exposure estimation, it is the residues in single items of produce that are of interest rather than “average” residues measured in composited samples. The decomposition module in the DEEM 7.0 software uses a statistical procedure and the MaxLIP software uses a maximum likelihood estimation procedure in order to “decomposite” composited monitoring data to estimate residues in single items. This presentation described the decomposition module in the DEEM software and the MaxLIP software.

CHARGE

The specific issues to be addressed by the Panel are keyed to the background documents, *"Overview and Statistical Basis for the Maximum Likelihood Imputation Procedure (MaxLIP) for Imputing Single-Item Residue Concentration Distributions from Composite Samples,"* and *“Maximum Likelihood Imputation Procedure for Imputing Single-Serving Residue Concentration Distributions from Composite Samples”* memorandum dated February 17, 2000, and are presented as follows. In addition, supplemental background documents from the May 1999 SAP meeting were provided for consideration by the Panel.

1. The current Pesticide Data Program (PDP) collects residue data on approximately 5 lb. composite samples, whereas the residue values of interest in acute risk assessment are associated with residue concentrations in single items of produce. In order to make better and fuller use of the current PDP data, OPP is currently using its own decomposition method in an effort to convert residues from a “composite” basis to a “single-item” basis which was presented to the SAP in May, 1999. Two additional methods for decompositing pesticide residue data have been presented to the SAP (RDFgen and MaxLIP).

What are the overall strengths and weaknesses of each method with respect to their ability to adequately represent pesticide residues in single unit items ?

2. The OPP comparison attempted to gauge each decomposition method's performance against several standard sets of data which reflected differences in number of samples, degree of skewness, amount of censoring, and number of distributions. Each method may be sensitive to various "imperfections" , limitations, or characteristics of real-world data. For example, often data from many fewer than 30 composite samples are available for decomposition. Frequently, the data are censored and/or are heavily left-skewed. Many times, the composite samples may have been collected from a multitude of separate and distinct pesticide residue distributions.

How sensitive are the two methods being presented to the SAP for consideration to these different factors? Does each method being presented to the SAP have an adequately robust statistical underpinning?

3. Despite an adequate statistical underpinning and overall robustness, there may be specific situations in which characteristics of available data may make it unreasonable to expect a method to adequately deconvolute a data set comprised of composite samples and decomposition should be avoided as it may produce invalid or questionable output data.

What limitations does the Panel see in the decomposition methodologies being presented to the SAP (e.g., minimum number of samples, degree of censoring, etc.)? In what specific kinds of situations might each presented methodologies fail or be likely to fail?

4. In contrast to OPP's original decomposition method which was presented to the SAP in May 1999, the MaxLIP and RDFgen methods being presented to the current Panel do not assume that PDP residue measurements are derived from one overall lognormal distribution of residues. MaxLIP permits up to five distinct residue distributions, while RDFgen permits any number of residue distributions and assumes that each composite measurement is derived from its own distribution. The MaxLIP method is able to account for only up to five separate distributions of residues and the user must use the Likelihood Ratio Test to determine if an adequate number of distributions is modeled.

Does the Panel have any comments on this aspect of the program and how might this affect the adequacy of the decompositions which are performed? In contrast, RDFgen assumes that each composite is derived from a separate and distinct distribution and decompositing is performed by using the standard deviation of composite value measurements and assuming (once adjusted) that this applies to each composite. Does the Panel have any comments on these differences in approach and assumptions?

5. Although limited in scope, OPP's comparison of each method's ability to accurately predict individual item residue levels based only on information in residue levels in composite samples did not appear to provide any clear evidence of systematic over- or under-estimation of residues in decomposited samples. All three methods did not necessarily perform equally well (particularly at the upper and lower tails of the distribution) under all circumstances in predicting single-item *residue levels*, but differences in predicted *exposure levels* (and therefore risk levels) appeared to differ to a much lesser extent. This situation is not unexpected: it is often not the

extreme upper tail of a *residue* distribution which is responsible for driving the 99.9th or 99th percentile *exposure* levels, but rather a combination of reasonable (but high end) consumption and reasonable (but high end) residue levels of one or two frequently consumed agricultural commodities. That is, it is not necessarily true that significant differences in predicted residue levels in the upper tail (e.g., >95th percentile) of the residue distribution will as a matter of course result in significant differences in predicted exposure levels at the upper tails of the exposure distribution, since it is a combination of both consumption and residue levels over a wide variety of commodities which determine high-end exposure levels.

Does the Panel have any thoughts, insights, or concerns about the potential for underestimation or overestimation (or other biases) of residue levels by each of the two decomposition procedures being presented for consideration? Does any concern regarding over/under estimation extend to concern about over/under estimation of exposures (and therefore risks)? Can any characteristic statements be made about over/underestimation at various percentile levels (e.g., median, 75th, 90th, 99th 99.9th percentiles)?

PANEL RECOMMENDATIONS

- The Agency is commended for the effort to evaluate decompositing tools. Overall, the MaxLIP procedure is the preferred method at this time. Although it has limitations associated with making inferences at the upper tail of a distribution of single-item variability in residue concentrations, it is the only method that has a capability for simulating intra-class correlation among single units that form a composite sample. The MaxLIP method also may be slightly less dependent upon parametric assumptions than the RDFgen method. The MaxLIP method deals with censoring in a more rigorous manner than does the RDFgen method.

- The Panel recommends that the Agency consider additional simulation studies to investigate the behavior of the MaxLIP and RDFgen when the numbers of composite samples are small.

- There are no bright lines that determine the minimum number of samples, degree of censoring or size of intra-cluster correlation that will distinguish success or failure of a simulation run. MaxLIP's method seems like the more satisfactory method in this regard, but additional numerical simulations studies and validation using actual samples of observed single-unit residue concentrations are encouraged to develop a more complete understanding of the performance of both algorithms under real world conditions and restrictions for sample sizes, censoring, distributional assumptions and intraclass correlation among the single units that form PDP composite measures.

- The Panel encourages more test examples to capture a wide range of statistical conditions portraying the factors that shape a residue profile. In addition, model design can also benefit from more understanding of the residue database to which they are applied. A logical next step would be to characterize the general pesticide residue profile with respect to the overriding controlling factors (e.g., spatial, temporal, agricultural practices, chemical properties, specific characteristics of a residue monitoring program).

- It is important to keep in mind that concern over MaxLIP's and RDFgen's ability to accurately simulate extreme percentiles must be interpreted in the context of how these data will used to estimate chronic and acute exposures. Questions to research or study in actual applications are: How often do extreme values contribute to extreme values of simulated acute exposures? Do extreme values or outliers have a significant impact on estimated distributions of chronic exposures? The Panel also recommends to the Agency that it study not only the accuracy (unbiasedness) of the imputed distribution of single unit residues, but also the variability in these distributions from one simulation run to the next.

- The Panel encourages the developers of MaxLIP and the DEEM RDFgen module to open their code to enable review of testing by scientists and the user community.

DETAILED RESPONSE TO THE CHARGE

1. The current Pesticide Data Program(PDP) collects residue data on approximately 5 lb. composite samples, whereas the residue values of interest in acute risk assessment are associated with residue concentrations in single items of produce. In order to make better and fuller use of the current PDP data, OPP is currently using its own decomposition method in an effort to convert residues from a "composite" basis to a "single-item" basis which was presented to the SAP in May, 1999. Two additional methods for decompositing pesticide residue data have been presented to the SAP (RDFgen and MaxLIP).

What are the overall strengths and weaknesses of each method with respect to their ability to adequately represent pesticide residues in single unit items ?

The MaxLIP and RDFgen algorithms simulate pesticide residue concentrations for single unit food items. Each algorithm generates a simulated distribution of single unit residues by using 1) actual data on residue concentrations measured for composite samples and 2) a model of the relationship of means and variances for composite samples to that for the imputed single unit values. Since there is limited single-unit pesticide residue data (three studies are presented) to empirically validate the performance of the two algorithms, the Panel's comparison of strengths and weaknesses of the two algorithms focuses heavily on: 1) the theoretical basis of the approach to the problem, 2) properties of the algorithms that are needed to address known features of real world single unit pesticide residue distributions, and 3) simulation results for known distributions.

The "simulation" of residue concentration values for single-unit servings based on observed concentrations for pooled composite samples is a statistical problem of imputation based on grouped or "coarsened" data. The observed data are the PDP composite measures of concentration, the Limit of Detection (LOD) and Limit of Quantitation (LOQ) values of measurement censoring, and the proportion of the food item units that are untreated. The imputed values are the single-unit concentration values that are generated by the algorithms and combined with survey data on consumption for use in large scale simulations of consumer exposures. Theoretically, MaxLIP incorporates many elements of a maximum likelihood approach to the simulation problem including the estimation of model parameters (mean and variance) under the left censoring model for lognormal data. The exception to a strict maximum likelihood approach is the rejection sampling step in which samples of generated single unit values are composited and compared to actual observed composite measures. If the simulated composite of generated single unit values does not fall within +/- 5% of any observed composite value, the set of single unit samples is rejected. This is certainly a restriction on the algorithm's search for the maximum likelihood solution. The Panel recommends that MaxLIP's +/- 5% restriction for simulated composites of single-unit samples be reevaluated. The Panel cannot judge it's practical impact on the speed of convergence of the maximum likelihood (ML) algorithm. The operating characteristic of this restriction to the maximum likelihood (ML) algorithm is that it will introduce "coarseness" into the distribution and restrict imputation of extreme values when the number of composite samples is small.

In contrast to MaxLIP, RDFgen's simulation method is more ad hoc. RDFgen assumes the underlying model for single-unit residues is a single, truncated lognormal distribution with a common mean and variance that are computed based on the observed composite values (the "Central Limit Theorem" method) or modeled via a user-specified relationship between the value of the single-unit value and its variance (the Coefficient of Variation method). The simulation of single-unit samples is closely governed by the means and variance properties of the individual observed composites. RDFgen also uses a rejection sampling approach that forces the composite means of simulated single-unit values to lie within +/- 5% of an observed composite value. The current version of RDFgen assumes independence of single-unit values within composites. The general consensus of the Panel is that the statistical theory underlying the RDFgen method needs to be formalized. A suggested possibility would be to consider a Bayesian model relating the distribution of composite means (with an appropriate prior on the distribution of their values) to the likelihood for the distribution of single-unit values.

Since Question 1 is a complex question with many dimensions, the Panel's response to this question will be organized as a series of comparisons:

***Comparison 1**: Do the methods reflect the possibility that composites are derived from single units from the same field or region of similar pesticide application practice? Can user-supplied empirical data on intraclass correlation of single-unit residue concentrations within the composite samples be incorporated in the simulation?*

Both methods reflect to some extent the possibility that single items within a composite have a distribution of intra-composite single-item variability that reflects a common origin of all of the items in the composite. MaxLIP uses a maximum likelihood type approach to estimate the means, variances, and mixture parameters for the underlying lognormal distributions of single unit concentration values. The MaxLIP procedure takes into account the possibility of subpopulations on a national scale and, through the screening process, takes into account, at least indirectly, the possibility that single items within a composite may have a relatively narrow range of values compared to the inferred population distribution. MaxLIP allows the user to supply a measure of the assumed intraclass correlation among the single-unit concentrations that form the composite samples. The MaxLIP method also has a procedure for sampling rank correlated random values in simulating the single items that comprise a composite. Therefore, "intra-class" (intra-composite) correlation can be accounted for with this method. MaxLIP assumes that this correlation is constant for all composites. This probably is not the case in the real world, but the tests results using empirical data on residues suggest that MaxLIP results are not highly sensitive to varying the values of the correlation parameter about the population value. The Agency should consider constraining the components of the lognormal mixture in MaxLIP to have a common standard deviation or a common coefficient of variation.

The RDFgen method is based upon an a priori assumption that variability in residue concentrations among single items is lognormally distributed, with the mean value of the residues among the single items in a composite equal to the observed composite residue concentration. Intra-composite variability can be specified on either an absolute or relative basis. Novigen, the developers of RDFgen, presented information during the SAP meeting indicating that the use of relative standard deviations produces more accurate results. Thus, operationally, it appears that the RDFgen method is based upon a unique mean value for each distribution of intra-composite single-item variability, with the same relative standard deviation (coefficient of variation) used for each composite. Therefore, the method used by RDFgen appears to simulate lognormally distributed clusters of residue concentrations for single items, with each lognormal distribution centered upon the observed composite concentration. It is not clear that this method would ever lead to an extreme scenario in which all but one of the single items has essentially no residue, with only one single-item containing all of the residue in the composite. As noted above, RDFgen currently assumes independence of all single-unit residues within a composite sample. There is no capability at this time to introduce intra-class correlation in the process of generating single-unit residue concentrations.

It is doubtful that the coefficient of variation is the same for all composites, in the tests conducted by EPA, varying the coefficient of variation did not appear to have much effect on the estimates of the percentiles of the distribution of pesticide residues. This is likely due to a larger variance between composites relative to the variance within composites. In the absence of an estimate of the coefficient of variation among units within composites, RDFgen estimates the variance among units within composites, V(u), to be “n” times the variance among composites, V(c), where “n” is the number of single units in a composite. This is likely to be a gross overestimate of V(u). This can be illustrated through standard ANOVA relationships. If V(a) represents the component of variance between composite means due to differences in agricultural practices, weather, processing, etc., then $V(c) = V(a) + V(u)/n$. The variance for a single unit, i.e., $n = 1$, is $V(s) = V(a) + V(u)$. If V(a) is much larger than V(u), then V(s) and V(c) are approximately equal. In this case V(u) could be ignored and decompositing may not be necessary. At the other extreme, discussions from the previous FIFRA/SAP meetings indicate that V(u) is not likely to be greater than V(a). Even in the unlikely case that V(u) is as large as V(a),i.e., $V(u) = V(a)$, $V(c) = V(u) [1 + 1/n]$ giving $V(u) = V(c)/[1+1/n]$. For $n = 1$, $V(u) = V(c)/ 2$. If n is large, V(u) is approximately equal to V(c) and V(s) is approximately equal to 2V(c). In no case does $V(u) = n \bullet V(c)$. It appears that in the absence of information about the coefficient of variation of units within composites, the maximum value of V(u) is not likely to exceed V(c). It is recommended that using this maximum value should be evaluated for RDFgen when no information is available for the coefficient of variation of units within composites. It appears that the estimated variance of a single serving will almost always lie between one and two times the variance measured among composites. These two extremes can be used to evaluate the potential range of the distribution of residues without using decompositing techniques.

Overall, the MaxLIP procedure is a more theoretically rigorous approach for simulating single-item values that may arise from subpopulations, be constrained by observed composite values, and be correlated with other single-item residue concentrations within a composite.

Comparison 2*: Can the algorithm simulate single-unit concentrations at upper percentiles, perhaps beyond the range that can be directly imputed from observed composites?*

Both methods produce an imputed distribution of single-item residue concentrations. The RDFgen method is based on assuming individual lognormal distributions for single-item variability for each composite, whereas the MaxLIP procedure involves a screening process based upon an inferred parametric population distribution comprised either of a single lognormal distribution or a mixture of up to five lognormal distributions.

Consider a hypothetical experiment involving five composites each of which has five single items. The available data are observations regarding the residue concentrations among the five composites. Using either RDFgen or MaxLIP, one can make inferences regarding the variability in residue concentrations among the 25 single items represented by the five composites. The maximum percentile represented by the empirical distribution in this case would be calculated, using a common plotting position formula, as:

$$F(x) = \Pr(X \le x_i) = (i-0.5)/n,$$

Where i = rank of each data point, n = number of data points, and Pr(x<=X) is the probability that the random variable X has values less than or equal to some specified value x. The rank of a data point is estimated by ordering the data in ascending order as follows:

$$x_1 \le x_2 \le x_3 \le \dots \le x_n$$

to obtain the rank i of x_i.

In the case of our hypothetical experiment, the maximum percentile of the inferred empirical distribution that we can estimate directly would be for i=25. Since the data set has n=25 data points, we obtain:

$$Pr(X \leq x_{25}) = (25-0.5)/25 = 0.98$$

Thus, in this case, we can make direct inferences regarding only the 98th percentile and lower.

In order to make a direct inference regarding the 99.9th percentile, we would require a sample size of:

$$n = 1/(1-F(x)) = 1/(1-.999) = 1/0.001 = 1,000$$

This could correspond, for example, to 100 composites with 10 single items per composite. In many cases, we do not have this many composites or this many single items represented by the available composite data.

The MaxLIP procedures involve over-sampling or repeated sampling of the same composites in order to obtain an empirical distribution with 10,000 single-item values. For example, if there were 100 composites with 10 single items each, then each composite would be randomly simulated an average of 10 times in order to yield a total of 10,000 single items. However, the method can make inferences regarding no more than 1,000 single items based upon the observed 100 composite values. It would be more correct to interpret the 10,000 samples as approximately 10 different possible realizations of the estimated empirical population distribution of inter-item residue concentration variability.

From the background report it appears that the RDFgen method generates an empirical distribution based upon random simulation of the number of single items contained in the observed composites, which is often much less than 1,000.

Based upon the information described here, it appears that neither the MaxLIP or the RDFgen method includes a procedure for estimating the upper tail (e.g., 99.9th percentile) of the inferred population distribution of inter-single-item variability unless the number of composites and number of single items per composite is such that well over 1,000 single items are represented by the available composite data.

One comment from this analysis is that both methods should avoid presenting predictions of percentiles of the distribution of variability in single unit concentrations that cannot reasonably be inferred from the available data. However, it would be possible to extrapolate beyond the range of direct inference either by fitting a parametric distribution (or mixture of parametric distributions) to the data generated for a given realization of the empirical single-item distribution or using a "mixed empirical-parametric" distribution approach to make inferences regarding the upper tail.

A key shortcoming of both methods is that they will never generate single-item residue estimates larger than the largest possible single-item concentration represented by the available composite data. For example, if a composite has a residue concentration of 1 ppm and contains

10 single items of equal size, the largest possible concentration among the single items would be if one item had a residue of 10 ppm and the other nine items had residues of 0 ppm. However, since the observed composites are only a sample from a population, it is possible that there may be an unobserved composite with a residue larger than 1 ppm. Neither method extrapolates beyond the range of maximum single-item values based upon observed composites. Thus, the results of either of the methods at the upper tail restricts the imputation of extreme values.

A comment made during the discussion was that the upper tail does not necessarily govern predictions of high end exposures or individual risk. The highest exposures or risks may be associated with residue concentrations that might be in the mid or upper portion of the residue concentration distribution but not necessarily in the upper tail. Therefore, the degree of precision needed regarding predictions of the upper tail of the population distribution of variability among single items may not be high. The Panel encourages the Agency and the software developers to further explore the sensitivity of the ultimate exposure estimates to the extreme values and outliers in the generated sample distributions.

Comparison 3*: Does each method deal with non-detects (values below the Limit of Detection - LOD)?*

Both methods appear to deal with non-detects in making inferences regarding the distribution of residues among single items. The MaxLIP procedure uses a more rigorous method than does RDFgen. In the MaxLIP procedure, maximum likelihood estimation (MLE) is used to make inferences regarding the distribution of inter-item variability in the censored portion of the distribution (below LOD). The MLE formulation is a general one that can accommodate multiple LODs, even for a single composite. In contrast, the RDFgen approach uses one-half the LOD to represent each value below LOD. As an approximation, this latter approach may be adequate. The MLE method is likely to be less biased and is more appealing from a theoretical perspective. The practical difference in final results using these two approaches is less clear. Clearly, values below the LOD will have little or no influence on the upper tail of the distribution of exposure.

Comparison 4*: Does each method appropriately deal with data below the Limit of Quantitation (LOQ)?*

As a practical matter, it appears that the distinction between data below LOD versus data below LOQ has not been a significant issue for actual PDP data sets; specifically, data below LOQ have typically also been below LOD.

The MaxLIP method is able to deal with situations in which a data point may be above the LOD but below the LOQ, using a generalized formulation of the likelihood function. The RDFgen method does not have this capability.

Therefore, the MaxLIP procedure is preferred over the RDFgen procedure as a method for dealing with this type of situation. However, as noted, this type of situation appears to arise rarely, if at all, in actual data sets.

***Comparison 5**: Does the method make statistical inferences that are driven more by the observed data than by additional assumptions embedded in the method (e.g., parametric assumptions)?*

Both the RDFgen and MaxLIP methods employ the assumption of parametric distributions. The RDFgen method assumes that the variability among single items within a composite is distributed as a lognormal distribution, whereas the MaxLIP procedure begins with an assumption that the population of single-item residue concentrations is distributed either as a lognormal distribution or as a mixture of lognormal distributions. The MaxLIP procedure employs a screening (or windowing) procedure to generate what appears to be an empirical distribution of variability among single items. However, as reported by Dr. Sielken during the SAP meeting, the final result for the inferred population distribution of single-item residues is sensitive to the initial parametric assumption regarding the population distribution. Therefore, the MaxLIP procedure is not truly non-parametric, although it may be less sensitive to parametric assumptions than the RDFgen procedure.

The lognormal distribution is, to an approximation, a reasonable distribution for modeling concentrations. Concentrations must be non-negative, and the distribution of concentrations is typically positively skewed. The lognormal distribution captures these characteristics. However, the lognormal distribution can tend to be a "tail-heavy" distribution compared to, say, a gamma or Weibull distribution. (Tail-heavy refers to having more probability mass in the upper tail compared to other distributions.) Since the tail of the distribution is typically a portion of the distribution for which there is little data and considerable uncertainty, it is often mostly a matter of judgment as to which parametric distribution to use to represent a given dataset. Either the RDFgen or MaxLIP procedures could be reprogrammed to allow the user to select a particular type of parametric distribution for use in the analysis, such as Lognormal, Gamma, or Weibull. The sensitivity of the predictions of these methods to alternative parametric assumptions should be explored.

The RDFgen method uses Latin Hypercube Sampling (LHS) to generate "random" samples from the assumed distributions for intra-composite single-item variability. LHS is a stratified sampling technique. Therefore, by definition, it does not and cannot generate *simple random* samples. The stratified sampling from the modeled distribution of single-unit values will in many cases actually reduce the variability of summary statistics based on the generated data set. Random Monte Carlo simulation should be used in RDFgen, not LHS.

***Comparison 6**: Can each method deal with data of various number of composites?*

This question is aimed at the robustness of the methods. For example, in their presentation, the Agency indicated that a possible concern was whether each method would be able to perform calculations in situations with a small number of composites. It appears that both methods are able to produce results. However, there are some issues to consider in interpreting the results, especially if the number of composites simulated is small. The limit of the highest percentile for which a method may directly infer variability among single units will depend upon the number of composites and number of units within the composites, as described in **Comparison 1**. If the number of parameters (e.g., mean, standard deviation) used to "fit" a distribution or assign distributions regarding inter-unit variability is larger than the number of observed composites, then there are essentially no degrees of freedom in the estimation of parameters and the models are "over-fit" to the available data. It is possible that some numerical methods, such as the solution for the maximum likelihood estimators used in MaxLIP, may not be robust if the sample sizes are too small. However, the maximum likelihood parameters estimates for the lognormal distribution are typically relatively robust. Since the RDFgen method is based upon matching mean values of intra-composite variability among single units to the composite residue concentration and assigning a standard deviation based upon a simple rule (e.g., constant coefficient of variation among all composites), this method should be capable of calculating results even for very small numbers of composites.

***Comparison 7**: Can each method deal with various numbers of single units within composites?*

There are two possible interpretations to this question. One is whether the methods are robust if the number of units within each composite is small for all composites. Another is whether it is possible to simulate a different number of units from one composite to another.

Both methods should be able to simulate either small or large numbers of single items from every composite. While the level of uncertainty in the predicted population distributions of inter-unit variability would be a function of the total number of single items simulated, from a computational/numerical perspective both methods should be able to reliably simulate an arbitrary number of single units.

For the RDFgen procedure, it should be the case that the method is capable of simulating essentially any arbitrary number of single units for any composite, although it is not set up this way at this time. The MaxLIP procedure already appears to be capable of simulating a different number of single units from each composite.

***Comparison 8**: Are the methods robust to various levels of censoring or to other challenges to the methods?*

As noted previously, the MaxLIP procedure has a stronger theoretical basis for dealing with censoring. MLE in general is often capable of making statistical inferences even for highly censored data sets (e.g., where perhaps well over half of the data are below LOD). While there are possibly situations in which a particular statistical estimator will fail because of anomalies in a particular data set, the MLE approach is a reasonably robust method.

The RDFgen method does not account for censoring as rigorously as does the MaxLIP method.

Overall Comparison

Overall, the MaxLIP procedure is the preferred method at this time. Although it has limitations associated with making inferences at the upper tail of a distribution of single-item variability in residue concentrations, it is the only method that has a capability for simulating intra-class correlation. The MaxLIP method also may be slightly less dependent upon parametric assumptions than the RDFgen method. The MaxLIP method deals with censoring in a more rigorous manner than does the RDFgen method.

More real-life data on single-unit residues and their relationship to properties of composited samples are needed to clearly justify the underlying assumptions in these models.

Nevertheless, decompositing tools are necessary for constructing a range of residues in commodities that are monitored in a group (e.g., 10-15 apples) but typically consumed as a single unit (e.g., 1 apple) within a eating occasion and a unit time frame (e.g., a day) . These models are designed for use in assessing the acute exposures. The Agency is commended for its effort to evaluate decompositing tools. It appears that the estimated variance of a single serving will almost always lie between one and two times the variance measured among composites. These two extremes can be used to evaluate the potential range of the distribution of residues without using decompositing techniques.

2. The OPP comparison attempted to gauge each decomposition method's performance against several standard sets of data which reflected differences in number of samples, degree of skewness, amount of censoring, and number of distributions. Each method may be sensitive to various "imperfections" , limitations, or characteristics of real-world data. For example, often data from many fewer than 30 composite samples are available for decomposition. Frequently, the data are censored and/or are heavily left-skewed. Many times, the composite samples may have been collected from a multitude of separate and distinct pesticide residue distributions.

How sensitive are the two methods being presented to the SAP for consideration to these different factors? Does each method being presented to the SAP have an adequately robust statistical underpinning?

Many of these items were addressed in the answer to Question 1. A few are elaborated here. There is no bright line regarding a required number of samples. It is possible to impute serviceable single-unit values from fewer than 30 composite samples. However, as the sample size decreases, the variance of distributional properties for simulated single-unit samples (due to imputation) will increase. Since the simulated single-unit samples produced by RDFgen (and to a lesser extent MaxLIP) are linked to the observed composite values, smaller sample sizes will limit the amount of distributional smoothing and simulation of extreme values that can occur in the model-based simulations.

The MaxLIP method has a stronger statistical underpinning for dealing with censored datasets. Maximum likelihood methods work best when the samples used for model fitting are large. If sample sizes are small and the level of censoring of single unit residue concentration is high, the maximum likelihood estimates of the parameters of the mixture of lognormal distributions will tend toward instability. The Panel recommends that the Agency consider additional simulation studies to investigate the behavior of the MaxLIP and RDFgen when the numbers of composite samples is small.

It would be useful to conduct numerical experiments in which the test cases are based upon different distributions than the lognormal mixtures assumed by MaxLIP and RDFgen. For example, single-item residue concentrations could be simulated from assumed population distributions that are normal, gamma, Weibull, or other parametric distributions or from mixtures of such distributions, with and without censoring and with or without correlations. The ability of the methods to infer, on average, the "true" population distribution of single-unit residues would be a measure of how robust the methods are under these different situations. Simulation studies should apply each method to multiple samples generated under the trial assumptions to establish the average or typical performance of the method.

The Panel also noted that, while the test examples are very helpful for studying the performance of these methods under different distributional assumptions, the impact on final DEEM exposure estimates was not large. For a single-commodity consumption, such as the case used in the test examples, the lack of apparent impact on exposures could partly be due to expressing the percentile of exposures on a “per-capita” basis (i.e., include eaters and non-eaters of the commodity in the exposure population). For a single-commodity exposure analysis, the inclusion of non-eaters would likely drive the high-end exposure points further up the higher percentile. Thus, the impact of decompositing for a single commodity could be better illustrated on a “per-eaters” basis.

3. Despite an adequate statistical underpinning and overall robustness, there may be specific situations in which characteristics of available data may make it unreasonable to expect a method to adequately deconvolute a data set comprised of composite samples and decomposition should be avoided as it may produce invalid or questionable output data.

What limitations does the Panel see in the decomposition methodologies being presented to the SAP (e.g., minimum number of samples, degree of censoring, etc.)? In what specific kinds of situations might each presented methodologies fail or be likely to fail?

Many of the points addressed in the responses to Questions 1 and 2 also apply in the response to this question. Small sample sizes, a high degree of censoring, composite sample concentrations below the limit of quantitation (LOQ) or limit of detection (LOD), and intra-composite correlation among residues on single items in the composite samples will all contribute to greater variability and simulation bias in both MaxLIP and RDFgen. There are no bright lines that determine the minimum number of samples, degree of censoring or size of intra-cluster correlation that will distinguish success or failure of a simulation run. MaxLIP’s method seems to be the more satisfactory method in this regard, but additional numerical simulations studies and validation using actual samples of observed single unit residue concentrations are encouraged to develop a more complete understanding of the performance of both algorithms under real world conditions and restrictions for sample sizes, censoring, distributional assumptions, and intra-class correlation among the single units that form PDP composite measures.

4. In contrast to OPP's original decomposition method which was presented to the SAP in May 1999, the MaxLIP and RDFgen methods being presented to the current Panel do not assume that PDP residue measurements are derived from one overall lognormal distribution of residues. MaxLIP permits up to five distinct residue distributions, while RDFgen permits any number of residue distributions and assumes that each composite measurement is derived from its own distribution. The MaxLIP method is able to account for only up to five separate distributions of residues and the user must use the Likelihood Ratio Test to determine if an adequate number of distributions is modeled.

Does the Panel have any comments on this aspect of the program and how might this affect the adequacy of the decompositions which are performed? In contrast, RDFgen assumes that each composite is derived from a separate and distinct distribution and decompositing is performed by using the standard deviation of composite value measurements and assuming (once adjusted) that this applies to each composite. Does the Panel have any comments on these differences in approach and assumptions?

Theoretically, the capability in the MaxLIP or RDFgen algorithm to fit mixtures of distributions is desirable. Practically, there are limits to fitting these mixture models and limits to the utility of the simulated data that are generated from highly or over-parameterized models. Practically, the greatest value of introducing the possibility of true mixtures into the MaxLIP algorithm is the ability to fit multimodal distributions of single unit pesticide concentrations. An empirical observation from the work of one panel member and from Dr. Sielken's comments, is that a mixture of five lognormals is likely to be rarely needed. One is likely to obtain a good fit with as few as one lognormal or perhaps a mixture of two or three lognormals. With three lognormals, there are two weights and six parameters to be estimated, for a total of eight parameters. One concern is that this could be an "over-fitting" if the number of observed composites is less than forty (i.e., at least five observations per parameter). If the number of composites is "m", the RDFgen method involves essentially m+1 parameters, since the mean is estimated separately for the single-unit distribution associated with each composite and the standard deviation is assigned based upon a simple rule (e.g., use of the same relative standard deviation for all intra-composite distributions). Thus, one concern is that both methods may have a tendency to overfit distributions. In the case of RDFgen, this is always the case. For MaxLIP, this happens if the number of parameters approaches the number of observed composites.

It should be noted that while a mixture of two-parameter lognormals is intended to model a mixture of food sources and treatment histories, this does not correspond to what you usually get with an unconstrained maximum likelihood fit. Typically, one component will fit the overall distribution and others will be used to describe specific details such as single spikes or long tails. This is overfitting. Constraining the standard deviations or the coefficients of variation of the components to be equal will help ensure that the fit is more likely to correspond to the conceptual model.

The likelihood ratio test to determine the number components of the lognormal mixture in MaxLIP is not valid because the regularity conditions for the test statistic to follow a chi-squared distribution do not hold (Titterington et al., 1985, pp. 154-156)[1] .

The comparison scenarios showed that the last test set using real-life data (i.e., PDP data) resulted in the greatest degree of disagreement between the model output and the original data, indicating that factors other than what have been statistically addressed by the models are also at play. However, without a good general understanding of the pesticide residue database and specific patterns of the relationship between the composites and their single units, it is difficult to pinpoint and address all major contributing factors in designing and effectively applying a decompositing model.

Here again, the Panel encourages more test examples to capture a wide range of statistical conditions portraying the factors that shape a residue profile. In addition, model design can also benefit from more understanding of the residue database to which they are applied. A logical next step would be to characterize the general pesticide residue profile with respect to the overriding controlling factors (e.g., spatial, temporal, agricultural practices, chemical properties, specific characteristics of a residue monitoring program). This would then provide the context for specifying the roles of decompositing models for improving the accuracy of dietary exposure assessment. Eventually, these steps could lead to a development of guiding criteria for the model use and enable a better grasp of the basis for determining the model limitations and assumptions. Experiences from scenarios in which decompositing fails to satisfactorily match the parent distribution would also be useful for defining conditions under which these models would not be applicable.

5. Although limited in scope, OPP's comparison of each method's ability to accurately predict individual item residue levels based only on information in residue levels in composite samples did not appear to provide any clear evidence of systematic over- or under-estimation of residues in decomposited samples. All three methods did not necessarily perform equally well (particularly at the upper and lower tails of the distribution) under all circumstances in predicting single-item *residue levels*, but differences in predicted *exposure levels* (and therefore risk levels) appeared to differ to a much lesser extent. This situation is not unexpected: it is often not the extreme upper tail of a *residue* distribution which is responsible for driving the 99.9th or 99th percentile *exposure* levels, but rather a combination of reasonable (but high end) consumption and reasonable (but high end) residue levels of one or two frequently consumed agricultural commodities. That is, it is not necessarily true that significant differences in predicted residue levels in the upper tail (e.g., >95th percentile) of the residue distribution will as a matter of course result in significant differences in predicted exposure levels at the upper tails of the exposure

[1] Titterington, D.M., A.F.M. Smith and U.E. Makov (1985). Statistical Analysis of Finite Mixture Distributions, Wiley, New York.

distribution, since it is a combination of both consumption and residue levels over a wide variety of commodities which determine high-end exposure levels.

Does the Panel have any thoughts, insights, or concerns about the potential for underestimation or overestimation (or other biases) of residue levels by each of the two decomposition procedures being presented for consideration? Does any concern regarding over/under estimation extend to concern about over/under estimation of exposures (and therefore risks)? Can any characteristic statements be made about over/underestimation at various percentile levels (e.g., median, 75th, 90th, 99th 99.9th percentiles)?

Both RDFgen and MaxLIP should provide reasonably good descriptions of the distribution of residues for units within the range of the observed composite means. When information is available from only a few composites, a good fit may be obtained with MaxLIP, but extrapolation into the highest quantiles (98%, 99%) of the residue distribution when the numbers of observed composites is fewer than 50-100 is tenuous. With information available from a limited number of composites, the extreme tails of the distribution of residues are not likely to be included in the measurements at hand. Since the variation among composites with RDFgen is limited to the observed range, the extreme tails of the distribution may be underestimated. It is important to keep in mind that concern over MaxLIP's and RDFgen's ability to accurately simulate extreme percentiles must be interpreted in the context of how these data will used to estimate chronic and acute exposures. Questions to research or study in actual applications are: How often do extreme values contribute to extreme values of simulated acute exposures? Do extreme values or outliers have a significant impact on estimated distributions of chronic exposures? If the answer to these questions is no, then secondary concern over fine-grained accuracy in extreme upper tail simulation of single-item residues should not be the primary concern. If correct simulation of extreme values for the population of single-unit residues is essential, focused studies of concentrations for known high-residue samples may be an option to truly understand the properties of the extremes.

The Panel also recommends that the Agency study not only the accuracy (unbiasedness) of the imputed distribution of single unit residues but also the variability in these distributions from one simulation run to the next. It is particularly important to incorporate the variability due to imputing single-unit values when these simulated observations are combined with sample data on consumption to estimate overall exposures (see Panel's responses to questions for Session 3 pertaining to DEEM exposure estimation). To estimate the variability associated with the imputing the distribution of single-item unit residues from composite sample values, multiple imputation methods could be used (Rubin, 1987).[2]

A member of the Panel also noted that bootstrapping methods could be used for this same purpose. The bootstrapping approach alluded to at the meeting could be conceptualized as

[2] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons. New York.

follows:

Let N = number of composites

Let *i* be a composite index from 1 to N

Let n_i be the number single units per composite identified by index *i*.

We have observations regarding the residue concentration in composite, *i*, which we denote as C_i.

Let $c_{i,j}$ be the residue concentration for the j^{th} unit in composite i.

Thus, the total number of single units, n_t, for which we may make an inference from the N observed composite residue values is the summation:

$$n_t = \sum n_i \quad \text{from } i = 1 \text{ to N}$$

Both of the decomposition methods proposed feature generation of synthetic values of individual-item residue values based upon some type of numerical sampling procedure. The way that both methods work is to make some assumptions about either a population distribution of inter-unit variability or a composite-specific population distribution of inter-unit variability for that composite only. Therefore, synthetic values are generated. The mean of the synthetic values for the n_i units associated with composite *i* is compared to the observed residue value of composite *i*. If the mean is within plus or minus 5 percent, that synthetic sample is retained and stored for use in characterizing an estimated empirical population distribution of inter-unit variability.

The Panel wishes to infer the residue concentration in the n_t number of individual units that comprise the composite. One cannot make an inference regarding a sample from the population of single-unit residues any larger than n_t. This is because the data from which statistical estimation was performed is based upon this sample size.

A bootstrapping procedure would be based upon repeatedly generating synthetic samples of n_t single-item residue concentrations to characterize an empirical distribution for each replication. If an index b = 1 to B is assigned for each replication, one would perform B bootstrap replications to obtain B alternative possible realizations of the estimated empirical distribution of inter-item variability. From these replications, one can infer distributions for any statistic of the distribution (e.g., mean, standard deviation, percentiles, etc.). A probability distribution for a statistic is a *sampling distribution*. These sampling distributions reflect the lack of knowledge or uncertainty associated with any particular statistic as a function of random sampling error. While this is not the only source of uncertainty in the simulation, it can be a dominant source of uncertainty in some cases.

The outcome of dietary exposure is dependent not only on the residue profiles but also on the consumption patterns. The impact of a decomposited residue profile may be masked when more commodities are added to the analysis and the population basis widens (i.e., increasing the percentage of “users”). Multiple sets of exposure analysis may be necessary to sort out the dynamic interaction of factors that could shift the overall exposure distribution one way or the other, resulting in a corresponding shift in the exposure level associated with a specific percentile (e.g., 95th, 99th, 99.9th) that the Agency has deemed critical for the evaluation of food safety.

SAP Report No. 2000-01C, May 25, 2000

REPORT:

FIFRA Scientific Advisory Panel Meeting,
March 2, 2000, held at the Sheraton Crystal City Hotel,
Arlington, Virginia

Session III - A Set of Scientific Issues Being Considered by the Environmental Protection Agency Regarding:

Dietary Exposure Evaluation Model (DEEM)

Ms. Laura Morris	Christopher Portier, Ph.D.,
Designated Federal Official	Session Chair
FIFRA Scientific Advisory Panel	FIFRA Scientific Advisory Panel
Date:______________________	Date:________________________

Federal Insecticide, Fungicide, and Rodenticide Act
Scientific Advisory Panel Meeting
March 2, 2000

SESSION III - Dietary Exposure Evaluation Model (DEEM)

PARTICIPANTS

FIFRA Scientific Advisory Panel Session Chair
Christopher Portier, Ph.D., National Institute of Environmental Health Sciences, Research Triangle Park, NC

FIFRA Scientific Advisory Panel Members
Fumio Matsumura, Ph.D., Professor, Institute of Toxicology and Environmental Health, University of California at Davis, Davis, CA
Herbert Needleman, M.D., Professor of Psychiatry and Pediatrics, University of Pittsburgh, School of Medicine, Pittsburgh, PA
Mary Anna Thrall, D.V.M., Professor, Department of Pathology, College of Veterinary Medicine & Biomedical Sciences, Colorado State University, Fort Collins, CO

Food Quality Protection Act Science Review Board Members
Christopher Frey, Ph.D., Associate Professor, Civil Engineering, North Carolina State University, Raleigh, NC
David Gaylor, Ph.D., Associate Director for Risk Assessment Policy and Research, U.S. Department of Health and Human Services/FDA, National Center for Toxicological Research, Jefferson, AR
Steve Heeringa, Ph.D., Director, Statistical Design and Analysis, Institute for Social Research, University of Michigan, Ann Arbor, MI
Peter D. M. Macdonald, D.Phil., Professor of Mathematics and Statistics, Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
Nu-may Ruby Reed, Ph.D., Staff Toxicologist, California Environmental Protection Agency, Department of Pesticide Regulation, Sacramento, CA
John Wargo, Ph.D., Associate Professor of Environmental Policy and Risk Analysis, Yale University, New Haven, CT

Designated Federal Official
Ms. Laura Morris, FIFRA Scientific Advisory Panel, Office of Science Coordination and Policy, Environmental Protection Agency, Washington, DC

PUBLIC COMMENTERS

Oral statements were made by:
Christine F. Chaisson, Ph.D., on behalf of Science, Strategies and Analysis Systems

Written statements were received from:
None

INTRODUCTION

The Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), Scientific Advisory Panel (SAP) has completed its review of the set of scientific issues being considered by the Agency regarding issues pertaining to the components and methodologies used by the Dietary Exposure Evaluation Model (DEEM). Advance notice of the meeting was published in the *Federal Register* on February 4, 2000. The review was conducted in an open Panel meeting held in Arlington, Virginia, on March 2, 2000. The session was chaired by Christopher Portier, Ph.D. Ms. Laura Morris served as the Designated Federal Official.

A major component of assessing risks of pesticide substances is the estimation of dietary exposure to pesticide residues in foods. The Agency currently uses the DEEM exposure assessment software in conducting its dietary exposure and risk assessment. The purpose of this session was to describe the components and methodologies used by the DEEM software.

CHARGE

The specific issues to be addressed by the Panel are keyed to the background document, "Background Document for the Sessions: Dietary Exposure Evaluation Model (DEEM) and DEEM Decompositing Procedure and Software," memorandum dated January 19, 2000, and are presented as follows:

1. Quality audits and validation of the algorithms used by the DEEM software have been conducted via comparisons with published consumption estimates and duplication of the DEEM outputs using other software (e.g., Microsoft Excel and Crystal Ball).

What additional types of validation or steps need to be conducted, or are the audits that have been conducted sufficient?

2. The (pseudo) random number generator used by the Novigen DEEM program is an integral part of the Monte-Carlo procedure used by the software and is critical to its proper functioning.

Does the Panel have any thoughts or comments on the randomization procedure used by the DEEM software?

3. DEEM offers two options for estimating acute daily exposures. The first option ("Daily Total") combines the distribution of total daily consumption levels with the distribution of residue values. The second option ("Eating Occasion"), combines the consumption levels corresponding to each eating occasion with the distribution of residues and sums the resulting estimated exposures to produce an estimate of the daily exposures. For example, if an individual reported consuming a given food twice during the day (say 100 gm and 120 gm), the "Daily Total" option would

combine the total daily consumption of that food (220 gm) with a randomly selected residue value. On the other hand, the "Eating Occasion" option would combine the first amount consumed (100 gm) with a randomly selected residue value and the second amount consumed (120 gm) with another (possibly different) randomly selected residue value and compute a total daily exposure estimate. EPA currently uses the "Daily Total" approach in its exposure assessments.

Under what circumstances should the EPA consider using the "Eating Occasion" approach?

DETAILED RESPONSE TO THE CHARGE

1. Quality audits and validation of the algorithms used by the DEEM software have been conducted via comparisons with published consumption estimates and duplication of the DEEM outputs using other software (e.g., Microsoft Excel and Crystal Ball).

What additional types of validation or steps need to be conducted, or are the audits that have been conducted sufficient?

It is difficult to determine when adequate quality audits and validation have been conducted but the Panel identified several areas where it felt that more could be done. The Panel would also like to have a better sense of the limitations of the model. Could the output from DEEM indicate when predictions are interpolations, well within the domain of the available data, and when they are extrapolations, outside the experience of the database and/or the assumptions of the model?

The Panel addressed three different aspects of validation:

Verification: the process of making sure that the computer code and its algorithms are doing what they are supposed to be doing.

Validation: the process of comparing predictions of the model to the real world.

Usability: the process of ensuring that the software will be used correctly when it is taken from the developers and given to the users.

The Panel was pleased that Novigen had released much of the source code. The evaluation of scientific software is facilitated when the software is public domain, or at least open code, and subject to scientific peer review. This keeps the assumptions "up front" and allows a greater range of expertise to evaluate the validity of the model and its implementation.

The Panel was very concerned that DEEM calculations and all the numerical validations that have been done depend on system software by Microsoft that is not publicly documented. This is discussed in more detail below, under Issue #2.

Verification

Verification may include procedures such as:

Dimensional analysis. Make sure that the units (e.g., grams) associated with numbers are correct, and that all conversions are performed correctly (e.g., mass amounts are calculated correctly from concentration data). Dimensional analysis was not mentioned in the DEEM documentation provided to the Panel. Therefore, it is recommended that dimensional analysis be performed. Furthermore, reporting of units associated with all numbers, in a clear manner, is encouraged. While the DEEM user interface appears to provide information regarding units in most cases, DEEM outputs should be reviewed to make sure that units are properly and consistently reported.

Comparison to alternative calculation schemes. The calculations of specific algorithms or code segments could be compared with similar calculations made with independently developed software. A number of activities along these lines are reported by Novigen, and further opportunities for such comparisons should be identified by EPA and pursued in future work. The Panel did not consider Microsoft Excel and Microsoft Crystal Ball to be sufficiently independent from Microsoft Visual Basic and recommended that software other than those be used for verification.

Comparison to hand-calculated values. For some simple situations, it may be possible to compare model predictions with values calculated by hand. This type of verification, while potentially time consuming, is often the most revealing. Errors in dimensional analysis, the sign of numbers (+ or -), coding mistakes, mistakes in the formulation of equations or algorithms, etc., can be identified using this approach, because it involves intensive scrutiny of the code and computational procedures. Note that it would be relatively simple to check the Monte Carlo algorithm for combining residues and consumption to obtain exposures. If lognormal distributions are specified for residue and consumption, the distribution of the log of exposure is a normal distribution with a variance that is the sum of the variance of log residues and variation of log consumption.

Sensitivity Analysis. Sensitivity analysis can reveal the operational characteristics of the model; that is, how does the model perform when inputs are varied over reasonable ranges? Is the response of the model to changes in its inputs reasonable? Does it behave in some counter-intuitive way that is revealing and insightful? Does it behave in some counter-intuitive way that reveals a problem with the model itself? Opportunities for performing sensitivity analyses should be identified and pursued. Sensitivity analyses could indicate what aspects of DEEM are in most need of validation. Sensitivity analysis may also indicate where the predictions DEEM provides are safe and where they are extrapolations from the model and the available data.

Random Numbers. Verification of the random number generator is important. It is discussed in more detail below, under Issue #2.

Validation

Validation involves comparing of predictions of the model or components of the model (e.g., subroutines or specific algorithms) to an independent data set for the quantities that the model is intended to predict. Some examples of validation activities include:

Do the model predictions of total pesticide residue exposures compare reasonably with external measures of pesticide use? Taking into account reasonable adjustment factors?

Are there external data sets regarding actual pesticide exposures to which model predictions might be compared (such as CDC data or others)?

Are there other reality checks that might be employed for specific components of the model; e.g., does the predicted consumption of a particular type of food agree with actual food production/consumption data? How does the estimate of total exposure of the population to a pesticide compare to total production of the pesticide?

Would it be possible to undertake even a limited study to measure directly what DEEM attempts to impute? For example, select several individual person-days of consumption from the food consumption survey, purchase the foods from normal retail sources, prepare as specified in the consumption survey, then measure total pesticide consumption for the day. This sample should be replicated at the purchase and preparation stages to determine the variance. The mean and variance over these replications can then be compared to model predictions.

Validation primarily has been limited to children 1-6 years of age. This certainly is a group of high interest but perhaps another age group also should be considered, e.g., the elderly.

*Peer-Review Publication***.** While it may not be possible to validate all possible model predictions, a combination of partial validation and exploration of the response of the model to changes in its inputs is important to pursue. Wherever possible, it is desirable to compare model predictions with actual real world data to assess model performance and to identify needs for changes to the model or for new input data. The process of scientific peer review is one method for obtaining feedback regarding verification and validation of the model.

*Openness of the Code***.** Related to the issues of verification and validation are the accessibility of the code for peer review and to the public. The Panel strongly encourages EPA and Novigen to submit for peer-reviewed journal publication the algorithms used in DEEM and some case studies that illustrate the key features of the model for realistic applications.

Models that are used in the public policy process should be open to review by the general public and should be made readily available. For example, air quality models endorsed by EPA for rule-making and permitting purposes are available on an EPA website. The information available from the web site includes an executable file, the source code, example input and output files, and documentation of the model and how to use it. This level of information and openness is important from a scientific perspective. The availability of open code will enhance both verification and validation activities and is strongly encouraged.

Estimating uncertainty in model predictions. EPA and the software developers are strongly encouraged to include prediction of uncertainty regarding statistics predicted by the model, including quantiles of distributions (e.g., the 90th percentile value of exposure, the 95th percentile of exposure, etc.). Information regarding uncertainty in predictions of exposure is important to take into consideration when comparing model predictions with real world data. For example, if the model predictions have an uncertainty of plus or minus 50 percent for a particular value, and if a validation procedure indicates that the model prediction is within 30 percent of the "true" value, then there can be a level of confidence associated with the validity of the model. If, instead, the model predictions are more than 50 percent different than the real world value in this example, then there would be evidence that the model was inadequate at making this particular prediction and either the input data and/or the structure of the model should be carefully evaluated and updated or modified as appropriate.

Usability

There has been no formal study of usability; however, Novigen has kept in close touch with DEEM users and has implemented many of their suggestions. One panelist has had extensive experience with students running simulations and reports no problems.

2. The (pseudo) random number generator used by the Novigen DEEM program is an integral part of the Monte-Carlo procedure used by the software and is critical to its proper functioning.

Does the Panel have any thoughts or comments on the randomization procedure used by the DEEM software?

The Panel was concerned that DEEM uses the Microsoft Visual Basic system random number generator to generate pseudorandom U(0, 1) values. The algorithm used is not publicly documented, but there is reason to suspect that at some point it uses a short integer (2 bytes) and hence may not have an adequately long period. This random number generator is possibly related to the one used by Microsoft Excel which is also not publicly documented, but has been deemed inadequate by McCullough & Wilson (1999). In some versions of Excel, the U(0, 1) generator returns an exact 0 or 1 about once in every 32,000 instances and this has serious implications for the periodicity of cycling. These exact 0s and 1s will also make it unsuitable for generating random numbers from other distributions by the "inverse probability integral transformation" method, as used by DEEM (e.g. Code Segments #20, #21, #22).

There are a number of pseudo-random number generator algorithms available in the literature. It should be possible to select one with known performance that meets or exceeds specifications regarding uniformity of the sequence of random numbers, independence of the sequence of random numbers (lack of auto-correlation), and periodicity of the cycling of the random numbers (i.e., how many random numbers are generated before the same sequence of random numbers is repeated). Barry (1996) evaluates several algorithms and provides information on the tests that can be done to evaluate the three characteristics of uniformity, independence, and periodicity. One algorithm that appears to be good is the "combined multiple recursive random number generator."

The algorithms used to generate random numbers for specific distributions should be evaluated. For example, the procedure for generating samples from a normal distribution (an inverse probability integral transformation method using a simple approximation to the inverse probability integral and a user-supplied maximum absolute value for the result) is not the best available procedure. Standard normal random variates can be generated more accurately and more efficiently using the Box-Muller polar method (Law & Kelton, 1991).

Because DEEM uses so many random numbers in each simulation, serial independence and the length of the cycle period are critical. The use of an arbitrary user-specified upper limit (e.g., DEEM Code Segments #20, #21, #22) should be avoided.

It appears that the user can specify either a user-defined seed or a random seed based on clock time. The implications of this choice should be explained in the documentation: a user-specified seed may be a poor starting value, but if the seed is based on clock time then it will not generally be possible to replicate the simulation with the same random numbers.

The Panel recommends that the random number generators be re-written to use publicly available state-of-the-art algorithms. The system random number generator in Microsoft Visual Basic or Microsoft Excel should not be used.

If the Agency has existing guidelines for evaluating a random number generator for use in Monte Carlo analysis, any generator used in DEEM should be tested accordingly.

3. DEEM offers two options for estimating acute daily exposures. The first option ("Daily Total") combines the distribution of total daily consumption levels with the distribution of residue values. The second option ("Eating Occasion"), combines the consumption levels corresponding to each eating occasion with the distribution of residues and sums the resulting estimated exposures to produce an estimate of the daily exposures. For example, if an individual reported consuming a given food twice during the day (say 100 gm and 120 gm), the "Daily Total" option would combine the total daily consumption of that food (220 gm) with a randomly selected residue value. On the other hand, the "Eating Occasion" option would combine the first amount consumed (100 gm) with a randomly selected residue value, and the second amount consumed (120 gm) with another (possibly different) randomly selected residue value, and compute a total daily exposure estimate. EPA currently uses the "Daily Total" approach in its exposure assessments.

Under what circumstances should the EPA consider using the "Eating Occasion" approach?

In terms of the actual exposures, the "Daily Total" approach seems most appropriate in situations where an individual would have multiple servings from a single unit of food (e.g., several slices of a single watermelon) over the course of day.

If an individual had multiple units of a particular food item (e.g., apple) over the course of a day, then it is possible that the residue concentration may differ among the two or more apples consumed by that individual over the course of the day. In this latter case, the "Eating Occasion" approach would be more reflective of reality.

Since there may be high correlation in the residue concentrations of the apples that a particular individual possesses, for example, because they were obtained simultaneously from the same supermarket, then an "Eating Occasion" approach may be appropriate only if the intra-class correlation is properly accounted for unless, of course, the level of analysis is the exposure for single eating occasions.

If one is looking for acute toxicity in a fast-clearing pesticide, then only the "Eating Occasion" approach is appropriate. However, DEEM does not consider "binge" and other special eating habits, and data on rarely-eaten foods will come from relatively few individuals, and these factors may limit the validity of "Eating Occasion" estimates.

For the analysis of total daily exposure, the expected value of simulated total exposures for the eating occasion and total daily consumption approaches should be the same but the variance properties of the two methods will differ. Given the general observation in the OPP residue test sampling that intra-class correlations of residues are positive and high, the Panel recommends conducting simulations using draws from the distribution of total daily residues. To make independent draws of residue concentrations for separate eating occasions when positive intra-class correlation is present would result in underestimation of the total variability of simulated total daily exposures. Thus, the "Daily Total" approach may be a more appropriate default procedure to use until such time as intra-class correlations are accounted for in DEEM.

Since the Agency staff also confirmed that the differences in outcome from the two approaches is indeed small, for simplicity sake, the "Daily Total" approach should suffice as a default. On the other hand, the choice of the two options would likely be more critical when there is a need (e.g., toxicity driven) to assess the risk based on eating occasions within a period of less than 24 hours. When treating each eating occasion as a separate exposure scenario, the second option of "Eating Occasion" would understandably be the choice, where the exposure from eating occasions would be summed within a specified range of time.

ADDITIONAL COMMENTS PROVIDED BY THE PANEL MEMBERS

Dietary exposure analysis is an extremely complex process. It utilizes many pieces of data from different sources, each carrying its own limitations and deficiencies for the purpose. Therefore, a careful documentation of the database limitations and the uncertainties associated with the estimated exposure is essential for a proper interpretation of the exposure estimates.

Since the session is mainly intended for the review of DEEM, an attempt is made to separate the comments pertaining to DEEM from comments that pertain more generally to dietary exposure assessment using an analysis logistics such as in DEEM. It is understood that a clear distinction of the two categories is not always possible.

Comments on DEEM

Sufficient documentation of the characteristics and limitations of the "hard data" should accompany the software. The Agency is commended for generating a list of food commodities that are included in the 1989-91 and 1994-96 CSFII. This list provides a valuable overview of the consumption database for dietary exposure estimations. It shows what commodities are included or unreported in each database. The unreported commodities are not included in the dietary exposure analysis. Further documentation on the limitations of the database is needed. One critical area that could have significant impact on exposure estimates involves the commodities that have low frequency of reported eating occasions during the days of the consumption survey. Unlike those commodities with no consumption (i.e., not included in the dietary exposure analysis), these low frequency commodities often reflect the variations of eating habits, and they may be well-liked and consumed in substantial quantities by some individuals but are not favorite foods to many others in a population. In the acute analysis, the contribution of these commodities to the overall exposure would tend to be masked when many commodities are included in one analysis. However, the impact is especially serious in a chronic exposure scenario when the program assumes that the average food consumption of all surveyed individuals on the day of survey equals the long term consumption level for all individuals. The analysis from this assumption would likely lead to a significant underestimation of exposure for those who favor the consumption of these commodities. A clear documentation would allow a user to apply needed precaution in using these data and focus the attention on any associated uncertainties. Documentation is also needed for other areas that were brought up during this session's presentation and discussions, such as weighting factors and the exclusion of any survey data points.

Dietary Exposure Analysis

An uncertainty analysis should accompany a dietary exposure analysis. The complexity of dietary exposure estimates underscores the importance of presenting the commodity contribution and uncertainties associated with an analysis. In light of the lack of a built-in uncertainty analysis tool in DEEM, it is recommended that multiple sets of dietary exposure analyses be routinely conducted to capture the impact of the critical factors that are identified in the steps leading up to the dietary analysis (e.g., the choice of residue data, whether to combine residue data from regions, seasons, or years, differences in eating habits and preferences).

A simple hand-calculation test is recommended for testing the reality of the exposure and risk estimates from a dietary exposure software program. Select a high-consumption commodity that has a high detected residue level in a composite sample or a commodity identified as a high contributor to the overall exposure. Calculate the exposure from this single commodity by multiplying the residue level with a "reasonable high consumption" (based on the consumption data or a commonsense estimate). Compare this exposure level both to the toxicity threshold (e.g., acute RfD, NOEL, or acute "risk cup") and the exposure estimates from DEEM that account for the exposures from all commodities. If the single commodity exposure comes close to the toxicity threshold, further safety analysis would be needed to estimate a reasonable background level of exposure from the rest of the commodities. In a similar way, this simple procedure can provide a reference to the DEEM output, especially when the single commodity exposure comes close to the DEEM output at the critical percentile determined by the Agency as appropriate for safety evaluation (e.g., 95th, 99.9th percentile).

This simple hand-calculation exercise can also provide a quick estimate of the potential risk for commodities that are commonly eaten as a single unit (e.g., a single apple) but the available residue data are from composite samples (e.g., 10-15 apples). According to the Agency's current tier approach for dietary risk assessment, the initial screening tiers of analysis assume that residues are at the tolerance or the field trial levels. The subsequent tiers of analysis, including using monitoring data (e.g., USDA Pesticide Data Program) in a probabilistic analysis, are performed when the screening tiers show a potential risk of concern. In the probabilistic analysis, monitoring data of composite samples are decomposited into a distribution of residue for single units. The tier approach is based on the assumption that the initial screening tiers are "conservative," over-estimating the exposure. However, this does not take into consideration the composite nature of these residue levels. According to the previous day's SAP discussions regarding decompositing procedures (FIFRA SAP meeting, March 1, 2000; Session II), the residue in a single unit stone fruit could reasonably be 7-fold higher than the composite sample to which it belongs. The substantially higher residue in a single unit may prompt further discussions regarding the adequacy of the tier approach and the tolerance evaluation. For example, when the residue level for a composite sample is found to be near the tolerance (e.g., within 30% of the tolerance), the residue in a single unit within this sample could be substantially above the tolerance. Given that the residue level at the tolerance represents a criterion of safety, it is important to evaluate the risk above the tolerance. Before launching into a complex probabilistic analysis that includes the full set of commodities, a simple hand-calculation (multiplying the residue level with the consumption rate) can provide a quick preliminary assessment for a particular commodity or food form of concern.

REFERENCES

Barry, T.M. (1996) "Recommendations on the testing and use of pseudo-random number generators used in Monte Carlo analysis for risk assessment," Risk Analysis, 16(1):93-105.

Law, A.M., and W.D. Kelton (1991) *Simulation Modeling and Analysis 2nd ed.*, McGraw-Hill: New York

McCullough, B.E. and B. Wilson (1999) On the accuracy of statistical procedures in Microsoft Excel 97, *Computational Statistics & Data Analysis,* 31:27-37.

SAP Report No. 00-01D, May 25, 2000

REPORT:
FIFRA Scientific Advisory Panel Meeting, March 3, 2000, held at the Sheraton Crystal City Hotel, Arlington, Virginia

Session IV - A Set of Scientific Issues Being Considered by the Environmental Protection Agency Regarding:

Consultation on Development and Use of Distributions of Pesticide Concentrations in Drinking Water for FQPA Assessments

Mr. Paul Lewis
Designated Federal Official
FIFRA Scientific Advisory Panel
Date:____________________

Christopher Portier, Ph.D.,
Session Chair
FIFRA Scientific Advisory Panel
Date:______________________

Federal Insecticide, Fungicide, and Rodenticide Act
Scientific Advisory Panel Meeting
March 3, 2000

SESSION IV - Consultation on Development and Use of Distributions of Pesticide Concentrations in Drinking Water for FQPA Assessments

PARTICIPANTS

Session Chair
Christopher Portier, Ph.D., National Institute of Environmental Health Sciences, Research Triangle Park, NC

FIFRA Scientific Advisory Panel
Charles C. Capen, DVM, Department of Veterinary Biosciences, The Ohio State University, Columbus, OH
Herbert Needleman, M.D., University of Pittsburgh, School of Medicine, Pittsburgh, PA

FQPA Science Review Board Members
Jeffrey G. Arnold, Ph.D., USDA-ARS, Temple, TX
Mr. Michael Battaglia, Abt Associates Inc., Cambridge, MA
Bernard Engel, Ph.D., Purdue University, West Lafayette, IN
Steve Heeringa, Ph.D.,University of Michigan, Institute of Social Research, Ann Arbor, MI
Mark Nearing, Ph.D., USDA/ARS, National Soil Erosion Research Laboratory West Lafayette, IN
R. Peter Richards, Ph.D., Water Quality Laboratory, Heidelberg College, Tiffin, OH
Ali Sadeghi, Ph.D., USDA ARS, Beltsville, MD
Mark J. Schervish, Ph.D., Department of Statistics, Carnegie Mellon University Pittsburgh, PA
Harold Van Es, Ph.D., Cornell University, Department of Crop and Soil Sciences, Ithaca, NY
John Wargo, Ph.D., Yale University, New Haven, CT

Designated Federal Official
Mr. Paul Lewis, FIFRA Scientific Advisory Panel, Office of Prevention, Pesticides and Toxic Substances, Environmental Protection Agency, Washington, DC

PUBLIC COMMENTERS

Oral statements were made by:
Peter Hertl, Ph.D. on behalf of the American Crop Protection Association
Warner Phelps, Ph.D. on behalf of Novartis Crop Protection
David Gustafson, Ph.D. on behalf of Monsanto
Mr. Ed Gray, on behalf of the FQPA Implementation Working Group
Mr. David Esterly, on behalf of the Spray Drift Task Force

Written statements were received from:
Mr. Ed Gray, on behalf of the FQPA Implementation Working Group
David Gustafson, Ph.D., on behalf of Monsanto

INTRODUCTION

The Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), Scientific Advisory Panel (SAP) has completed its review of the set of scientific issues being considered by the Agency regarding issues for the development and use of distributions of pesticide concentrations in drinking water for FQPA assessments. Advance notice of the meeting was published in the *Federal Register* on February 4, 2000. The review was conducted in an open Panel meeting held in Arlington, Virginia, on March 3, 2000. The meeting was chaired by Christopher Portier, Ph.D. Mr. Paul Lewis served as the Designated Federal Official.

The Panel was provided with a progress report on the Agency's efforts to implement the drinking water component of the Food Quality Protection Act (FQPA) aggregate exposure assessment. Aggregate exposure is defined to encompass multiple potential sources of exposure to pesticides and includes exposure from pesticide residues in food, in drinking water, and in the home. In order to combine the drinking water component with the population based distribution of pesticide residues on food items in a statistically rigorous manner, the data should be developed with the same general structure. In this way, the Monte Carlo procedure used for the risk assessment for food stuffs can be extended to the drinking water component. Ronald Parker, Ph.D. (EPA, Office of Pesticide Programs) opened the session providing an overview of the Agency's efforts. Mr. Robert Gilliom (USGS) and Gregory Schwarz (USGS) presented the SPARROW model and regression approaches to distributions for drinking water concentrations. Elizabeth Doyle, Ph.D. (EPA, Office of Pesticide Programs) reviewed the Agency's approach for incorporating drinking water concentration estimates in aggregate and cumulative risk assessment. Ronald Parker, Ph.D. (EPA, Office of Pesticide Programs) concluded the Agency's presentation by reviewing questions to be posed to the Panel.

CHARGE

1. The FQPA requires the Agency to consider “all anticipated dietary exposures and all other exposures for which there is reliable information” in development of an aggregate exposure assessment. For FQPA exposure assessment in food, the Agency now develops population-based regional distributions of pesticide residues in the diet. Are we correct in our assumption that population-based regional (or national) distributions are also the most appropriate representation of pesticide residues (concentrations) in drinking water for incorporation into aggregate and cumulative risk assessment as defined by FQPA?

2. We are exploring the use of regression models and the SPARROW model developed in large part by the USGS for development of distributions of pesticide concentrations at drinking water facility intake locations. This is likely to include a nested set of water drawn for drinking from large rivers, from smaller streams, from shallow ground water, and from reservoirs. Does the SAP believe that these approaches are sufficiently rigorous and promising at this time to warrant further developmental efforts?

3. FQPA requires the Agency to address “major identifiable subgroups of consumers” who might be impacted more than the population at large. How could the EPA use this method of developing regional and national distributions of concentrations across hundreds or thousands of sites to identify such subgroups and estimate their exposures?

4. Are the regression approaches for shorter term (95th percentile) maximum annual concentration values likely to produce values that are useful for acute risk assessment? If not, are there other ways of estimating the upper level percentiles. For example, would it be appropriate to adjust predicted values to higher percentiles based on ratios taken from other extensive pesticide monitoring data sets for which upper percentiles could be more accurately calculated? Are there additional approaches which the panel could recommend?

5. The ability to estimate distributions of atrazine using the regression approaches including SPARROW is due partially to the availability of extensive monitoring data for the chemical in both surface and groundwater. Although good monitoring data sets are available for some older pesticides, OPP will also be required to conduct drinking water exposure assessment for pesticides for which there are little or no field data. A method will be needed to estimate distributions of concentration values for those chemicals for which there are sufficient monitoring data to use a normal regression approach but which have failed the screening level exposure assessment using a single, conservative, high exposure site.

a. Would the panel support an effort to build a level of predictive capability into the regression approaches presented, based upon adding pesticide use and important environmental fate properties as additional regression variables?

b. Would it be appropriate to use national or regional distributions of atrazine concentration data adjusted for use area and environmental fate properties as a conservative benchmark for evaluating other compounds in a regulatory setting?

c. • Does the Panel have other suggestions for developing distributions of pesticide concentrations in drinking water?

6. One of the results that is most difficult to address in performing statistical manipulations of pesticide concentration values in water is that of handling concentration data below the detection limit, often called non-detects. Are the methodologies presented for working with the non-detects in these regression approaches sufficiently rigorous to develop accurate and useful concentration distributions?

PANEL RECOMMENDATION

- The use of population-based regional (or national) distributions to represent pesticide residues (concentrations) in drinking water is very appropriate.

- The SPARROW model and the regression approach being explored by the USGS are promising, and further development is warranted. Each approach has some limitations, at least in its current state of development, but until better approaches are identified, these approaches are of value because they allow development of exposure distributions. Before pursuing this approach further though, the limitations and implications of this approach need to be fully understood.

- The idea of "building a level of predictive capability" into the regression approaches has merit and should be explored further. Included should be chemical properties and management factors.

- The limitations of extrapolation should be recognized, especially for geographically-targeted, minor-use chemicals. One problem is that detection limits are calculated on the basis of concentrations in the extract analyzed. If the volume extracted varies, the detection limit in the sample varies from sample to sample. This raises the problem of multiple censoring levels in its worst form.

DETAILED RESPONSE TO THE CHARGE

The specific issues addressed by the Panel are keyed to the Agency's background document, "Development and Use of Distributions of Pesticide Concentrations in Drinking Water for FQPA Exposure Assessments: A Consultation", dated February 4, 2000, and are presented as follows:

The Panel highly commends the Agency for the significant progress that has been made toward estimating pesticide exposure to the U.S. populace and was impressed with the increased sophistication of the assessment of pesticide exposure, especially its move from deterministic to stochastic methods. The approaches involving the use of population-based distributions and the use of regression-type models based on real-world monitoring data are, in the Panel's view, a significant step forward and address some of the concerns raised during earlier SAP meetings. The Agency was commended for its willingness to incorporate biophysical nuances into the estimation process and for its focus on the use of real monitoring information when available. The Panel also applauded the Agency for involving the USGS in this effort, as their experience in regional and national water quality monitoring and modeling is proving to be very helpful to the implementation of the FQPA.

1. The FQPA requires the Agency to consider "all anticipated dietary exposures and all other exposures for which there is reliable information" in development of an aggregate exposure assessment. For FQPA exposure assessment in food, the Agency now develops population-based regional distributions of pesticide residues in the diet. Are we correct in our assumption that population-based regional (or national) distributions are also the most appropriate representation of pesticide residues (concentrations) in drinking water for incorporation into aggregate and cumulative risk assessment as defined by FQPA?

The use of population-based regional (or national) distributions to represent pesticide residues (concentrations) in drinking water is very appropriate and the Agency is commended for trying to find ways to move beyond point estimates of drinking water exposure to pesticides. We support a realistic approach that incorporates the diversity of biophysical conditions in the U.S. and addresses the variable exposure and risk associated with various subpopulations in various regions and at various time intervals. This is perhaps a complex approach, but given the diversity of conditions and populations, it is necessary to evaluate true exposure risks. It reduces the level of conservatism which is necessary to deal with uncertainty when point estimates are used. An important additional benefit is the better understanding that is gained of the complexity of exposure risk by incorporating the diversity issues.

A question that needs to be addressed as this work progresses is the level of quality and representation of the monitoring data that are available for this effort, especially as it relates to water quality monitoring data for limited-use compounds. The examples presented to the FIFRA SAP involved mostly high-use pesticides that are widely applied and for which extensive monitoring data are available. The Agency needs to move forward with this approach but, along the way, needs to remain aware and critical of the appropriateness of the procedures, and the level of confidence associated with estimates of pesticide exposure. This especially relates to the higher-percentile estimates of pesticide distributions for short-term exposure assessment of minor-use or new pesticides and is especially important when using logarithmic values in regression models. Ultimately, it is important that estimates are bound by real-world exposure levels. It is apparent that the Agency is aware of this issue.

One unusual aspect of drinking water distributions exists – how do we deal with the population which uses bottled water? It clearly has a geographic distribution which cuts across regions which might otherwise be established. Perhaps the DEEM model can help deal with this problem (FIFRA SAP Reports 2000-1B and 2000-1C).

One Panel member commented that the Agency’s overall goals and objectives should be used to determine the most appropriate methods to use. If the Agency seeks to have a national estimate of the levels of pesticide concentration in drinking water, one could draw a national probability sample of community water systems (CWSs) and do direct observations of that national sample. This would yield a direct national estimate. If, on the other hand, the goal is to have a screening tool that can be applied to each and every (or most) CWS in the U.S. to identify those systems that have higher than desired levels of pesticide concentrations, some type of model-based approach is probably better to use because it would be costly and time consuming to do direct observations at all CWSs in the U.S. In the development of model-based (i.e., indirect) estimates of pesticide concentrations, one can develop the regression model using a national monitoring sample of CWSs. On the other hand, if it is thought that regional variation exists which cannot be explained by the available predictor variables in the model, then the development of separate regional models could lead to more accurate predictive tools. The development of a model for each region would of course require a larger monitoring sample of CWSs. It seems appropriate to use the model-based approach in order to develop predictions for all CWSs in the U.S., and to use this to determine which may have high pesticide concentrations.

2. We are exploring the use of regression models and the SPARROW model developed in large part by the USGS for development of distributions of pesticide concentrations at drinking water facility intake locations. This is likely to include a nested set of water drawn for drinking from large rivers, from smaller streams, from shallow ground water and from reservoirs. Does the SAP believe that these approaches are sufficiently rigorous and promising at this time to warrant further developmental efforts?

The SPARROW model and the regression approach being explored by the USGS are promising and further development is warranted. Each approach has some limitations, at least in its current state of development, but until better approaches are identified, these approaches are of value because they allow for the development of exposure distributions. Before pursuing this approach further though, the limitations and implications of the approach need to be fully understood. Explicitly documenting the strengths, assumptions, and limitations of this approach would be very helpful to those using this approach and its output. Several of the background documents discuss some of these factors; however, documenting these factors in a single location would be useful. For instance, Larson and Gilliom (2000), in the Agency's background documents, discuss some of the limitations. One of the most limiting sources of information is the need for adequate observed water quality data for each pesticide.

The SPARROW model, if correctly understood, would be able to produce national or regional distributions of estimated annual mean concentrations at drinking water intakes. The regional mean exposure could be estimated with some confidence, because many values contribute to its calculation. According to some of the background literature, considerably less confidence would be associated with the estimate for any particular location. If this is true, how much confidence can be placed in the accuracy of the *distribution* of exposure concentrations? If the distribution is not trustworthy, how much trust can we place in its use in overall exposure.

Hierarchical models can help address extrapolation to new chemicals. These models give a framework in which you can build in uncertainty about the features that might make the new chemical different from the ones you have already studied. Of course predictions will be less certain than they would for a better understood chemical, but the models explicitly quantify that uncertainty.

There are some technical considerations. At the calibration stage, when a basin contains monitoring stations upstream, the load at each of those points is lumped together as a single input at that point, and the details are lost. Does this mean that SPARROW cannot make predictions of concentrations upstream from those points? If so, how serious a problem is this for its use in estimating spatial distributions of exposures through drinking water? Also, converting load estimates to concentration estimates by dividing by reach discharge creates a flow-weighted mean estimate. Time-weighted concentrations are more appropriate for drinking water assessment. How different are these two estimates likely to be? Which is likely to be higher?

Model structure concerns were also addressed. Land management impacts on in-stream concentrations are not considered in the regression model. Including land management effects in the model is important because the resultant model can then be used to encourage positive land management strategies. The lack of inclusion of management factors in the model appears to be due to the fact that data necessary are not available at the necessary required scale. There is considerable management data available. However, it is difficult to include in a regression model and is more easily simulated in a deterministic model. As a potential solution, the Agency might consider using other, more specific, models, including process-based ones, to develop relationships which can be appended or/ incorporated into the principal regression model. A similar task is taken in other regulatory applications. For example, in regulating soil conservation practices in the United States, statistically-based USLE/RUSLE technology is used, but often modifications to the input parameters or model relationships are applied using more process-based models such as CREAMS or WEPP. These models have been developed and tested on more specific experimental data sets for specific effects.

The above suggested approach can also be used for other model modifications and extensions. For example, Monte-Carlo simulations of process-based simulation models can be used to evaluate and test confidence limits on the frequency of occurrence curves derived from the regression model.

The Panel also identified specific data and data analysis concerns with the Agency's approach as presented below:

Data concerns:

The data used to develop the regression model are relatively short term. One would expect that data of more than a decade would be necessary to develop reasonable concentration distributions, particularly for the upper end of the data which is of major significance here. Also, there are concerns about the frequency of data collection and the associated possibility of missing peak and infrequent concentrations. One suggestion, of course, is to collect more data for a longer period and more frequently. Another conjunctive possibility would be to use a deterministic model calibrated to the measured data distributions and use the deterministic model to extend and fill in the missing data.

Care must be taken in terms of high-end data, and the use of log values. Also, it is necessary to recognize the importance of temporal variability, especially as it relates to poorly-timed extreme events (e.g. a 10-year storm occurring within several days after application). The inclusion (or lack thereof) of such events may significantly affect the resulting distributions. This is a drawback of the regression approach over the deterministic modeling approach which allows for better temporal upscaling based on climate data. Perhaps more effort should be put into regression approaches that focus on such unusual events for estimation of high-end concentrations, especially for vulnerable watersheds.

Specifically, the atrazine in runoff sampling data that were used to develop the regression approaches and SPARROW may have some properties that limit the usefulness of the resulting models. For example, the "temporal" resolution of the atrazine detection data is such that significant numbers of peak concentrations were missed. Although the atrazine data is a "long-term" data set compared to other pesticides, it is certainly not long-term from a hydrologic standpoint. What are the implications of these issues with respect to the ability of the resulting SPARROW and regression models to provide useful information for FQPA analyses?

The Agency already identified other factors that it wants to consider as dependent variables, such as tile drainage and wetland areas. These have a significant influence on pesticide transport and should be investigated. Also, pesticide-specific physical and chemical data should be included as regression parameters to allow for estimation of new compounds. Another issue that affects the potential for pesticide loss is the specific timing and application method of a pesticide, and care needs to taken with extrapolation of results from one chemical to another chemical which is managed very differently, i.e. pesticides with similar inherent loss potentials may actually pose different risks due to management factors.

The currently available data for the SPARROW approach limits the size of areas that can be analyzed with confidence to relatively large watershed areas. Many drinking water utilities obtain their water from smaller watersheds than can currently be analyzed with SPARROW. To extend the use of the SPARROW approach to smaller areas, significant additional data will be needed. The Agency's background document "Development and Use of Distributions of Pesticide Concentrations in Drinking Water for FQPA Exposure Assessments: A Consultation" indicates that the current SPARROW approach is only valid for source areas larger than about 1,000 km^2 (the background document indicates this is 10,000 but discussion with the model developers indicates this is really 1,000). To extend the approach to smaller areas, better digital stream networks are needed. The other spatial data used in SPARROW may also need to be improved to extend the model to smaller areas. Obtaining these data should be a high priority, if economically feasible.

Currently limited observed pesticide concentration in water data are available for applying the SPARROW approach. Significant additional data collection may be needed to extend this approach to other pesticides unless techniques are used that apply the approach to other pesticides such as attempting to account for the properties of pesticides that do not have significant amounts of observed water quality data. Pesticides that would seem to have enough data to apply SPARROW are atrazine, metolachlor, cyanazine, alachlor, and trifluralin. Additional observed water quality data may be needed for these pesticides and other pesticides. Approaches that attempt to extend the regression and SPARROW approaches to other pesticides by considering pesticide properties will likely result in confidence limits that are potentially quite large. This may limit the usefulness of this approach.

The Agency's background document indicates that the current SPARROW approach can estimate only average long-term pollutant concentrations. Are values for acute situations needed? If so, it will be difficult to extend this framework.

Data analysis concerns:

For the case where the concern is the upper tail of the distribution curves (high concentrations), the use of the log transform may not be appropriate. The Agency should find a transformation, if one is necessary, which better focuses on the data range of interest. Perhaps a different analysis is in order depending on whether the interest is long-term exposure (where averages and medians may be of more significance) or short term exposure (where emphasis definitely is on the high concentrations).

Because the SPARROW model is not capable of estimating exposures at a specified point very accurately, it would appear not to be of use in identifying water supplies that are “at risk”, something which is a logical extension of the current issues. Is it an accurate characterization to say that the SPARROW model could produce useful national and regional exposure distributions, and identify regions where concentrations are likely to be higher than elsewhere, but not identify individual sites with high concentrations?

The regression approach outlined by Larson and Gilliom is considerably more simple than the SPARROW model, but nonetheless it has merit because it leads to estimated distributions of specified percentiles of exposure concentration across multiple locations. The issue here is the reliability of the concentration estimates. Atrazine, metolachlor, and trifluralin are said to be estimated within a factor of 10, and alachlor and cyanazine within a factor of 30. This applies when the regression models are applied to the same sites used to develop them. Their reliability is likely to be less when they are applied to sites external to the “training” set. As an example, alachlor was, in general, underpredicted in the training data set. Hopefully the uncertainties associated with predictions from these regression approaches can be substantially reduced.

The Agency is cautioned to avoid methods that are extremely novel in approach. First, the current approach is to regress several percentiles from the same distributions on covariates to develop the model. Doing each percentile separately ignores the multivariate nature of these data and it is suggested that a simple additional complexity would be to utilize a multivariate approach. Secondly, it seems obvious that some of the covariates may interact to produce a particular response and the Agency is encouraged to use some type of stepwise algorithm to consider the inclusion of interaction and possibly nonlinear terms in the regression equation. Inclusion of covariates used in the SPARROW model into the simple regression or direct inclusion of regression parameters into the SPARROW model will bring the two closer together possibly leading to a path with a single analytical tool. Finally, the longitudinal nature of the data is lost when only percentiles are evaluated and the Agency should consider formal time-series analyses with covariates as another approach, possibly using quasi-likelihood methods as currently developed by researchers at Harvard University and The Johns Hopkins University.

3. FQPA requires the Agency to address "major identifiable subgroups of consumers" who might be impacted more than the population at large. How could the EPA use this method of developing regional and national distributions of concentrations across hundreds or thousands of sites to identify such subgroups and estimate their exposures?

The idea of identifying subgroups of consumers who might be impacted more adversely than the population at large means that we want to identify subpopulations residing in CWSs that have high pesticide concentration in their drinking water. Subpopulations take many forms: geographic, socio-economic, temporal. Pesticides in water will be more closely tied to consumption location than residues on foods. The current effort benefits heavily from the GIS capability, hydrological inventories and measurements. Sample-based searching for hot spots is inefficient and unlikely to meet the objective of identifying subpopulations at high risk. Models may need to be used that focus the search and understanding of the mechanisms. Assessments that connect individuals to their water supply and predicted pesticide intake will benefit from the model, but many subpopulations (shallow wells on farms and nearby areas) are at higher risk. Model predictions need to be linked with a program of on-site sampling and measurement. Composite estimation that combines model results with stratified sampling to establish true risk for subpopulations assumed to be at high risk can then be applied.

Demographic subgroups may be added to this list. This may be a complex task to undertake. If we assume that we have predicted pesticide concentrations for all CWSs in the U.S., we can partition them, for example, into CWSs with high concentrations versus all remaining CWSs. From the EPA SDWIS file we know the population of the CWS, but that is all that is available in that file. To determine the socioeconomic, demographic, and occupational characteristics of the population in the two groupings of CWSs, we need local level information on population characteristics. For example, the 1990 Census publishes a rich set of characteristics of the population at the ZIP Code level. If one could overlay the ZIP Codes with the CWSs, then one can establish the characteristics of the population living in CWSs with high pesticide concentrations. Ultimately, what one would like to be able to state is that X% of people who belong to subgroup A reside in CWSs with a high concentration, compared to Y% of people who belong to subgroup B.

4. Are the regression approaches for shorter term (95th percentile) maximum annual concentration values likely to produce values that are useful for acute risk assessment? If not, are there other ways of estimating the upper level percentiles. For example, would it be appropriate to adjust predicted values to higher percentiles based on ratios taken from other extensive pesticide monitoring data sets for which upper percentiles could be more accurately calculated? Are there additional approaches which the panel could recommend?

The need to be able to estimate peak exposure to chemicals is obvious, but one Panel member wondered about the fixation on particular quantiles. It would be good to know something about toxic levels or accumulations. For example, if the 60th percentile of the distribution provides a toxic dose in one day, then what is the value of the 95th and 99.7th percentiles. On the other hand, if toxic levels are never observed over years of data, then no reliable estimated quantiles will be useful. It is presumed that neither of these extremes represents reality, but it would be good to know what interesting levels are before worrying too much about which quantile to estimate. Indeed, once key levels or ranges are identified, it is then useful to determine which quantiles correspond to these levels.

The distribution characteristics of model outputs are often known *a priori*, and may not need to be checked for each run. This is definitely not the case with monitored data. As the Agency moves toward the incorporation of distributions rather than point estimates, and toward incorporating monitoring data in various ways, it is critical that any exercise that uses monitoring data to estimate upper percentiles of exposure include an analysis and correction for biases built into the sampling strategy, and a check to see that the log-probit model (if used) is appropriate for that data.

It may be of interest to consider how the 99.7th percentile concentration compares to the 95th percentile. If the standard log-probit regression and inverse prediction approach is used, the

ratio of the 99.7th to the 95th percentile is a function only of the slope of the regression line, and is given by

$$\log(\text{ratio}) = k/m,$$

where m is the slope and k is a constant equal to the difference between the probit values for the two percentiles. FIFRA SAP member Peter Richards calculated regressions of this sort for more than 400 raw water data sets for atrazine. The sites are concentrated in the Midwest, but span the country and include a few Canadian sites. A quick preliminary analysis of these results shows that the ratios range from 1.1 to 34 with a median of 4 and a mean of 5. The highest ratios tend to be associated with low absolute concentrations. Lakes and reservoirs appear to typically have slightly lower ratios than rivers and streams, although this was not formally analyzed.

The ratios of these percentiles for a major subset of this data were also calculated, but in this case deriving the percentiles non-parametrically (Excel PERCENTILE function), by interpolation between the values whose percentiles bracket the desired percentile. NAWQA and NASQAN data were not included, because it was organized in a way that prevented this being done quickly. In this case the ratios *of the same percentiles* ranged from 1 to 7 with a mean of 1.3 and a median of 1.6 (75th percentile 1.7). Because many of the data sets are not normally distributed as assumed in the approach in the preceding paragraph, the rations just listed are correct. If so, the 99.7th percentile is likely to be less than twice the 95th percentile in most cases.

The Panel provided specific comments on calculating percentiles. The log-probit approach is generally preferred because it is parametric, and therefore includes the possibility of confidence intervals around estimated values. It can also be used (with caution) to extrapolate to percentiles beyond that represented by the highest value in a small data set; how to extrapolate with the non-parametric approach is unclear (some programs (e.g. Excel) do it, others (e.g. Data Desk) refuse to). However, the log-probit approach is very sensitive to the distribution of the data, and, if the data are not log-normally distributed as assumed, the estimates of upper percentiles can easily be grossly in error. This problem of model mis-specification probably represents a larger issue than the issue of spreading confidence intervals about the regression line raised in this question.

Even if a given process is log-normally distributed, the process is generally sampled non-randomly, and the sampling program often places greater emphasis on the periods of time when high concentrations are expected. The data from such a sampling program will *not* be log-normally distributed even if the parent process is. Estimates derived by blindly submitting such data to log-probit regression analysis will be very untrustworthy, and will probably be severely biased high. Even estimates based on interpolation will tend to be high, because high values contribute disproportionately to the data and the ranks fail to take that into account.

Several examples illustrate this issue as presented by FIFRA SAP member Peter Richards. The Ohio rivers and streams listed in Table 1 below are monitored by Heidelberg College, Tiffin, Ohio, with samples three times per day during storm runoff in the pesticide runoff season (May through August), two samples per week during low flow in the pesticide runoff season, and a sample every two weeks or every month in the fall and winter months – a highly biased sampling program. The data sets span 13 years (1983 through 1995) except for Lost Creek, which spans 11 years. Daily means were calculated on a time-weighted basis when more than one sample per day was available, estimated by interpolation for days without samples, and in some cases approximated using a fixed value (0.1 µg/L) in fall and winter months when the gap between samples was longer than a week. The estimates are of the 90th percentile for atrazine for the entire period of record.

Results based on the raw data are higher than results based on evenly-spaced data by factors ranging from 2.3 times to almost 12 times. Biases for the regression approach are systematically higher than those for the interpolation approach.

A different situation is illustrated in Figure 1. The sampling program was intended to characterize concentrations during the post-application period, and most of the samples define a pretty good straight line. Four samples fall far from the trend; these are the earliest four samples, and belong to the time before the pesticide runoff began. In a sense, they belong to a different year, and to a part of the annual distribution not intended to be sampled. The presence of these *low* outliers rotates the regression line, causing it to substantially *over*estimate the upper percentiles of interest.

Table 1. Monitoring of Rivers in Ohio

	Regression		Interpolation	
	Raw Data	Daily Means	Raw Data	Daily Means
Lost Creek near Definance, OH	15.830	1.351	14.179	1.638
Rock Creek at Tiffin, OH	16.307	1.816	13.110	2.714
Honey Creek at Melmore, OH	26.194	3.539	16.582	4.489
Sandusky River near Fremont, OH	17.038	2.783	11.426	3.400
Maumee River at Bowling Green, OH	11.838	2.827	8.258	3.544

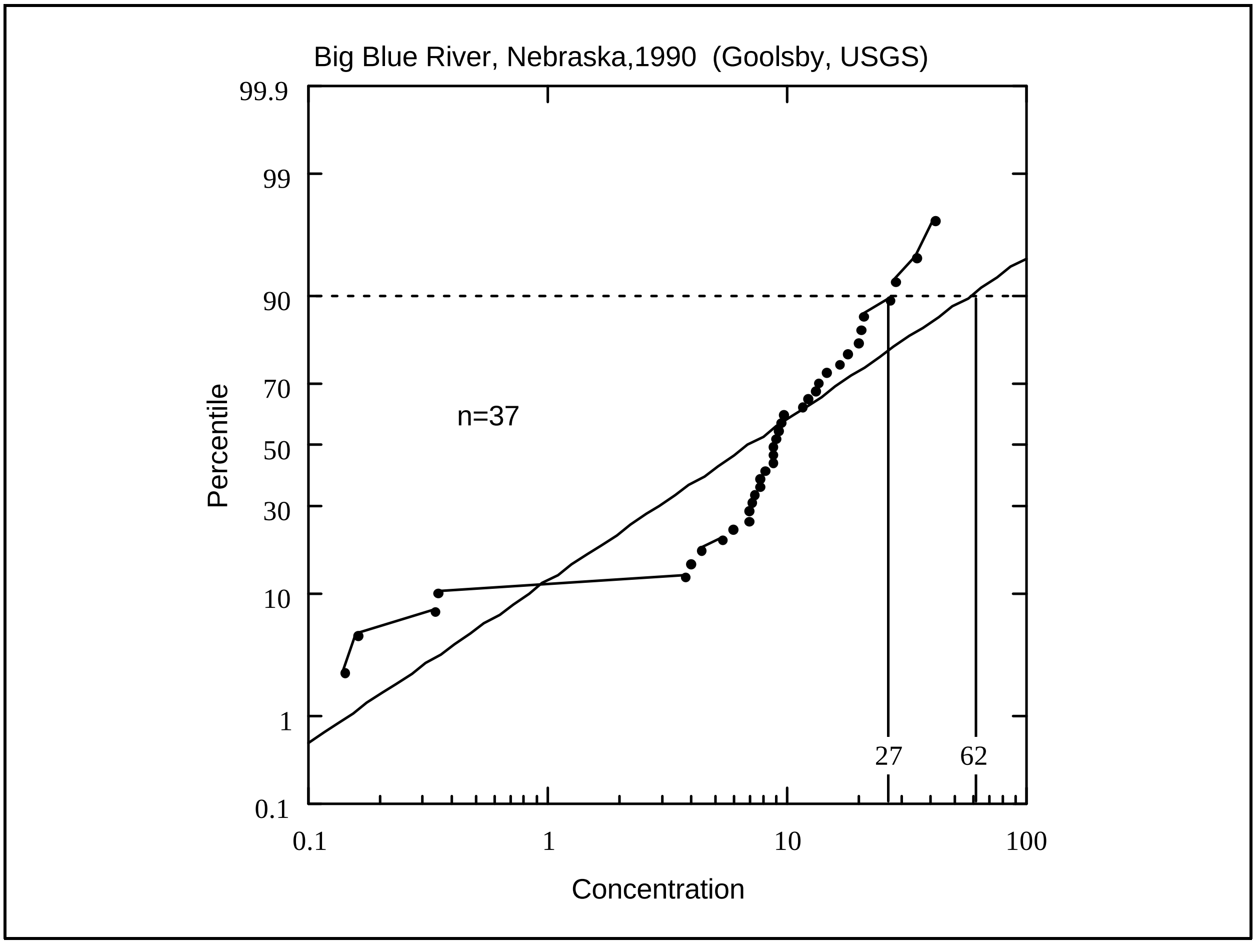

Figure 1. Corruption of log-probit regression fit by four data points from samples taken before the pesticide runoff season began. Note that inappropriate low values have led to overestimation of upper percentiles.

Other Panel members commented that the sample sizes seem too small for reliable estimation of extreme quantiles. Indeed some have apparently been too small for maximum likelihood estimation altogether. Hierarchical modeling can help to overcome some of the small sample size problems. Such models have been used successfully in many spatial estimation problems. These models are not alternatives to the regression models presented by the Agency, but rather they are augmentations that allow regions with low data density to borrow strength from regions with more data. The definition of region does not have to be geographical. The regions can be defined by the same covariates that are used for predicting in the current models. It seems that the first stage of SPARROW, in which daily values are estimated before computing an annual average load, can be used to predict annual peaks as well.

There was not much discussion about the impact of measurement error and sampling variability. These correspond to repeated measurements either on the same samples or on samples from the same site at the same time. If such variability is significant, it can make the precision of estimates and predictions much lower. Studies could be designed to get estimates of these levels of variation in order to facilitate more precise prediction based on the main body of data.

Finally, methods like maximum likelihood and significance testing on coefficients in regression models are notorious for understating the uncertainty about what model should be used. Bayesian techniques and model averaging often restore more reasonable uncertainty levels to our inferences. For example, if a model is rejected because some coefficient is not statistically significant, or because some other model predicts slightly better, it might be the case that the rejected model predicts almost as good and that its predictions, although different from those of the chosen model, should also be considered plausible when stating what is a reasonable level of uncertainty. Model averaging helps to incorporate additional models that are almost as good as the chosen model into our inference.

Determining where extrapolation begins and interpolation ends is not an easy exercise. One simple tool found useful is Box plots of the ratio of the upper bound of a prediction to its point estimate when evaluating hundreds of endpoints. Looking at these Box plots as you increase from the 50-th percentile to the 90-th to the 95-th, etc. will illustrate when the bounds get extremely large relative to the point estimate. This approach can be in error if the bounds are chosen to be constant relative changes (bounds on a curve can be developed in a number of ways; the classic approach results in bounds that are tight near the center of the data and expand toward the tails).

5. The ability to estimate distributions of atrazine using the regression approaches including SPARROW is due partially to the availability of extensive monitoring data for the chemical in both surface and groundwater. Although good monitoring data sets are available for some older pesticides, OPP will also be required to conduct drinking water exposure assessment for pesticides for which there are little or no field data. A method will be needed to estimate distributions of concentration values for those chemicals for which

there are sufficient monitoring data to use a normal regression approach but which have failed the screening level exposure assessment using a single, conservative, high exposure site.

a. Would the Panel support an effort to build a level of predictive capability into the regression approaches presented, based upon adding pesticide use and important environmental fate properties as additional regression variables?

The idea of "building a level of predictive capability" into the regression approaches has merit and should be explored further. Included should be chemical properties and management factors. It is recommended to recognize the limitations of extrapolation, especially for geographically-targeted minor-use chemicals. The process should allow for recognition of the fact that the exposure cannot be reasonably estimated and the emphasis may need to be on targeted intensive monitoring of the chemical.

With regards to the usefulness of adding additional regression predictor variables, it will be important to look for predictor variables that would be available for all CWSs in the U.S. In other words, a strong predictor of pesticide concentrations might exist, but if you can only obtain it for the monitoring sample, and not for all units (i.e., CWSs or water basins) in the population, then it cannot be used to come up with predicted values for the units that are not in the monitoring sample.

An alternative is to develop versions of physically-based models such as SWAT or AGNPS for this purpose. In either case, it is critical that enough monitoring data be gathered that the representativeness of the model predictions can be verified.

One Panel member had several specific concerns with regard to this question:

1) As we know, one model cannot fit all pesticides and all conditions since pesticides generally fall into three categories in terms of their toxicity (acute, chronic, cancerous). Thus, would it be more appropriate, if you generate three separate regression models for each category of pesticides, based on their toxicities?

2) In a non normal distribution, which is the case most of the time for pesticide concentrations in soil and water; when you take geometric means or log values of the observations to calculate the means, you minimize the weight (influence) of the less frequent, but have high value observations, for the calculation of the means. For pesticides, if our goal is to be conservative in our estimation, then the Agency should take general means of the original measured observations and then start to build up different percentile distributions.

3) With regard to the question of whether the Agency should use national or regional distributions, the Panel member suggested using regional distributions and that this selection should not be random, but based on factors such as geographic differences, differences in water

intake systems, sensitivity of various populations with regard to age and the state of health, etc.

b. Would it be appropriate to use national or regional distributions of atrazine concentration data adjusted for use area and environmental fate properties as a conservative benchmark for evaluating other compounds in a regulatory setting?

The use of a “transfer function” to estimate distributions of other compounds from those of atrazine is an interesting one. It is unclear whether it will work, particularly for compounds with very different half-lives, partition coefficients, etc. Two important questions are: What are the critical environmental fate properties and do they differ from compound to compound? Do we have adequate knowledge of the use areas and amounts for both atrazine and the new compound to benefit from this variable?

The approach will be especially difficult if the compound being evaluated is one for which acute exposures are of concern, rather than chronic exposures. In this case, one would apparently be attempting to estimate the distribution of annual *maximum* (or 99.7th percentile) concentrations of the new compound from the distribution of annual *average* atrazine concentrations. It is not obvious that there would be a predictable relationship between these two distributions. On the other hand, one could develop the distribution of annual maximum concentrations for atrazine and use it as the basis for estimating the distribution of the other compound. This would probably produce better results.

It would be helpful to evaluate this approach first by attempting to estimate the percentiles of interest for other relatively well studied compounds not too different from atrazine in their properties, such as simazine, alachlor, or acetochlor. If the approach is not successful with these, it is unlikely to work with very different compounds such as OPs. If such an approach is attempted, what are the criteria for success? How can success be evaluated if there are no data for the new compound? How can we be sure that the results would be “conservative”?

c. Does the Panel have other suggestions for developing distributions of pesticide concentrations in drinking water?

There is some discomfort about any approach which does not meet the test of empirical validation. The approaches being proposed seem to assert that we may not be able to predict the value of interest for any one station very accurately, but we can establish the average and indeed the distribution of values in spite of this – the errors basically cancel out in the aggregate. This may very well be true, but it should perhaps not be *presumed* to be true. Therefore, data from a well-designed monitoring network is and must remain an important component of the development of these distributions.

6. One of the results that is most difficult to address in performing statistical manipulations of pesticide concentration values in water is that of handling concentration data below the detection limit, often called non-detects. Are the methodologies presented

for working with the non-detects in these regression approaches sufficiently rigorous to develop accurate and useful concentration distributions?

One problem is that detection limits are calculated on the basis of concentrations in the extract analyzed. If the volume extracted varies, the detection limit in the sample varies from sample to sample. This raises the problem of multiple censoring levels in its worst form.

The Tobit model is appropriate for maximum likelihood estimation with normal data and a known truncation/censoring point--generally for left censoring. The efficiency of the method will depend on the quantile of the censoring threshold. Fitting a tobit regression to normal data that are censored at the 30th percentile will be more efficient than if the same data were censored at the 70th percentile. At some level of censoring it may be better to look at treating the data as a mixture and use maximum likelihood to fit the regression to a distributional form for the detected observations. Uncertainty over the censoring limit is a second issue. Under the Tobit model, incorrect assumptions concerning the censoring threshold will bias estimation. Nelson (1977) discusses Tobit models with unobserved censoring limits. Bayes methods may be applied.

When there is uncertainty about what levels of censoring were used, one can use information about what censoring levels are available as prior information in performing a Bayesian analysis, treating the unknown censoring level as an unknown parameter.

Literature Cited

Nelson, F.D. (1977). "Censored regression models with unobserved, stochastic censoring thresholds". *Journal of Econometrics*. Vol. 6, pp. 309-327.