SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 21, No. 8, pp. 1570–1577, 2002

Printed in the USA
0730-7268/02 $9.00 + .00

USING MONTE CARLO TECHNIQUES TO JUDGE MODEL PREDICTION ACCURACY: VALIDATION OF THE PESTICIDE ROOT ZONE MODEL 3.12

WILLIAM WARREN-HICKS,*† JOHN P. CARBONE,‡ and PATRICK L. HAVENS§
†The Cadmus Group, 1920 Highway 54, Suite 100, Durham, North Carolina 27713, USA
‡Rohm and Haas, Toxicology Department, 727 Norristown Road, Spring House, Pennsylvania 19477, USA
§Dow AgroSciences, Global Environmental Chemistry Laboratory, 9330 Zionsville Road, Indianapolis, Indiana 46268, USA

(*Received* 19 *March* 2001; *Accepted* 21 *January* 2002)

Abstract—Individuals from the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Environmental Model Validation Task Force (FEMVTF) Statistics Committee periodically met to discuss the mechanism for conducting an uncertainty analysis of Version 3.12 of the pesticide root zone model (PRZM 3.12) and to identify those model input parameters that most contribute to model prediction error. This activity was part of a larger project evaluating PRZM 3.12. The goal of the uncertainty analysis was to compare site-specific model predictions and field measurements using the variability in each as a basis of comparison. Monte Carlo analysis was used as an integral tool for judging the model's ability to predict accurately. The model was judged on how well it predicts measured values, taking into account the uncertainty in the model predictions. Monte Carlo analysis provides the tool for inferring model prediction uncertainty. We argue that this is a fairer test of the model than a simple one-to-one comparison between predictions and measurements. Because models are known to be imperfect predictors prior to running the model, the inaccuracy in model predictions should be considered when models are judged for their predictive ability. Otherwise, complex models can easily fail a validation test. Few complex models, such as PRZM 3.12, would pass a typical model validation exercise. This paper describes the approaches to the validation of PRZM 3.12 used by the committee and discusses issues in sampling distribution selection and appropriate statistics for interpreting the model validation results.

Keywords—Federal Insecticide, Fungicide, and Rodenticide Act Uncertainty analysis Model validation Monte Carlo Pesticide root zone model

INTRODUCTION

Ecological risk assessments are uncertain because of the complexity of ecological systems and the costs of collecting the data required to predict the behavior of such systems. This is true for both the exposure and the effects components of ecological risk assessment as outlined by the U.S. Environmental Protection Agency [1]. Yet the vast majority of ecological risk assessments conducted to date have been based on conservative and deterministic quotients that have not been supported by a quantitative uncertainty analysis. An uncertainty analysis, if performed, is typically restricted to a list of sources of uncertainty and perhaps qualitative statements of believability or confidence in the estimated quotients. As a result, risk managers and interested parties are not aware of the extent of uncertainty in the risk assessment and its consequences to the decision-making process.

A properly constructed uncertainty analysis can be used directly in the risk calculations, but it can also be used to judge the validity of both the exposure and the effects estimates independently. In particular, Monte Carlo analysis of complex models can be an integral tool for judging a model's ability to predict accurately. Monte Carlo analysis can be used for model validation. The model is judged on how well it predicts measured values, taking into account the uncertainty in the model predictions. Monte Carlo analysis provides the tool for inferring model prediction uncertainty. We argue that this is a fairer test of the model than a simple one-to-one comparison between predictions and measurements. Because models are known to be imperfect predictors prior to running the model, the inaccuracy in model predictions should be considered when models are judged for their predictive ability. Otherwise, complex models can easily fail a validation test. Few complex models, such as the pesticide root zone model, would pass a typical model validation exercise.

In most model validation frameworks, the model is asked to accurately predict a measured value, and the validity of the model is judged on the basis of a statistical estimate of the difference between the model prediction and the measured value. Typical estimators of model accuracy are the mean squared error, paired *t* statistic, correlation statistics, and others. While these statistics may or may not be valid indicators of statistical accuracy, a larger issue arises in that these statistics do not reflect the uncertainty in model use, such as the decisions made in model calibration, model structure, or choice of time step. In some sense, the statistics are bottom-line integrators of the results of the many decisions made prior to running the model. But, we argue, a simple comparison of observations and predictions is a naive approximation of the usefulness of the model or the expected inferences that can be drawn from the model output.

Do standard validation statistics reflect a degree of belief in the model output? They do, but not a comprehensive one. For example, a classical paired *t* statistic comparing model predictions and field measurements uses the variance in the paired differences as a basis for the test statistic. The variance reflects the range of paired differences within the data set. Is this an appropriate estimator for model validation? One perspective is that the estimator integrates all sources of uncertainty into the paired difference and is directly related to the

* To whom correspondence should be addressed (billwh@mindspring.com).

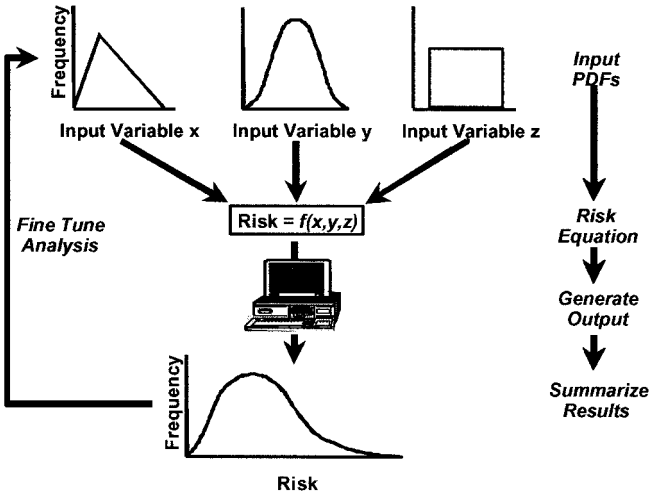

Fig. 1. The Monte Carlo process.

hypotheses under evaluation (H_0: $\mu = 0$, where μ is the true average difference between the paired model predictions and measurements). However, this estimator cannot directly incorporate sources of uncertainty, such as sampling error, temporal and spatial components, operator error, and other issues that are not a direct statement about the model itself, but reflect the available information and operator proficiency at the time the model was run. Monte Carlo analysis, on the other hand, provides a tool for incorporating all types of uncertainty. We argue that a true test of the model should reflect the total uncertainty in the modeling procedure and the model inputs at the time the model was run. The Monte Carlo outputs can be directly compared to field measurements as a method of model validation, thus incorporating the many sources of uncertainty into a decision about model validity.

Monte Carlo analysis is a conceptually simple tool that requires a great deal of thought to implement properly. The method is a process by which a degree of belief is inferred about the uncertainty in model predictions. Monte Carlo analysis is a series of cascading choices resulting in an estimate of model prediction error. In most real-world problems, a large degree of uncertainty is inherent in the choice of data set, treatment of outlying data points, choice of model, choice of spatial and temporal scales, choice of sampling distribution and associated parameters, and so on. The analyst is faced with many decisions before implementing the Monte Carlo analysis and is subsequently faced with the challenge of interpreting the final output. Each choice the investigator makes plays a role in the interpretation of the Monte Carlo predictive distribution and in the expectation that decisions made based on the analysis are indeed correct.

MONTE CARLO ANALYSIS

The underlying theory of Monte Carlo analysis is grounded in the long-run frequency interpretation of statistics and, in this sense, is an inherently frequentist (classical statistics) concept. In Monte Carlo analysis, samples are drawn from a distribution. As more and more samples are drawn, the mean of the samples is assumed to converge to the most likely value of the parameter (expected value). This convergence assumption is the basis for Monte Carlo theory and, in practice, is implemented by the repeated drawing of samples from the parameter sampling distribution (see Fig. 1).

Monte Carlo sampling is discussed extensively in Hammersley and Handscomb [2], Kloek and Van Dijk [3], Hammersley and Morton [4], and Wilson [5]. For Monte Carlo results to be believable, the convergence properties of the Monte Carlo estimators must be met. Several statistical and practical limitations exist in this regard. The most important practical limitations of Monte Carlo are misspecification of the sampling distribution; use of Monte Carlo sampling with a large number of parameters, particularly when the parameters are represented by different classes of distributions; and implementation with a relatively small number of iterations. For example, the distribution from which the samples are drawn is assumed to be the true distribution of the parameter of interest. To the degree that the sample distribution differs from the actual distribution, the confidence in the Monte Carlo results is decreased. Just how close these distributions must be is a complicated statistical issue that is frequently unclear. In a practical sense, if misspecification of a sampling distribution occurs for a very sensitive parameter in a multiparameter model, then the confidence in the Monte Carlo results would be greatly diminished because the model prediction would be greatly influenced by that parameter.

What is clear, however, is that the ''garbage in, garbage out'' adage applies. For example, many risk assessment studies use complicated exposure and population models with little or no field measurements available for parameterizing the model. The investigators make up distributions for some or all of the model inputs as part of a conceived Monte Carlo analysis. The investigators then initiate the Monte Carlo run, often with a small number of iterations, and examine the resultant distribution of the model predictions. Frequently, the scientists find the upper 95th percentile of the model predictions and use this value in a decision-making context. While the investigator is willing to expend many hours performing the Monte Carlo analysis, little time is given to activities that would increase the confidence in the Monte Carlo results. Warren-Hicks and Butcher [6] showed that small changes in the distributional assumptions of a Monte Carlo analysis performed using a typical population model (~12 input parameters) can drastically change the shape of the resultant Monte Carlo distribution. In particular, sampling from a series of independent normal distributions results in a very different Monte Carlo result than if a multivariate normal distribution (with an appropriate covariance matrix) is used. The underlying statistical theory behind Monte Carlo assumes that enough iterations are implemented for the convergence properties of the Monte Carlo estimators to hold. Again, the number of iterations required is not clear, particularly with disparate distributional assumptions among a large number of parameters. In hindsight, the investigator may actually have greater confidence in a small number of data samples on the parameter of interest in lieu of performing Monte Carlo analysis on a model for which basic parameterization and verification studies have not been implemented.

Burmaster and Anderson [7] have proposed 14 ''principles of good practice'' for using Monte Carlo techniques. They suggest that ''before an analyst undertakes an MC [Monte Carlo] risk assessment . . . she or he should read widely in the growing literature on probabilistic risk assessment.'' Principles for a properly conducted Monte Carlo analysis have also been proposed by the U.S. Environmental Protection Agency [8].

Conventional Monte Carlo methods are used only to un-

derstand how a model's parameter uncertainty may affect the model's prediction. This approach accounts for only a part of the total uncertainty. Uncertainty due to the data used for model calibration is not considered. In addition, almost all applications of Monte Carlo methods in model uncertainty analysis assume that the parameter distributions are given. New data are usually not used for updating information on parameter uncertainty. This practice is inefficient and sometimes may be misleading.

Many analysts do not understand the mathematics underlying the Monte Carlo method. While simple in concept, the underlying theory is very complex. An understanding of the theory is important from the following perspectives: The analyst is better able to judge the effect of decisions made during the course of the analysis, the analyst is better able to explain and communicate the results of the Monte Carlo analysis and the statistical endpoints, and the analyst is better equipped to combine the Monte Carlo results with other analyses in a complex risk framework (e.g., combining exposure and effects distributions into a risk distribution).

The Monte Carlo method provides approximate solutions to a variety of mathematical problems by performing statistical sampling experiments on a computer. The modern Monte Carlo method originated during the development of atomic energy in the post–World War II era, when it was used to provide solutions to the integral-differential equations. Later, the concept of using sampling experiments on a computer came to prevail in many scientific disciplines. Compared to other numerical methods, the Monte Carlo method is efficient with regard to computing time and easy to implement and understand. Using Monte Carlo methods for simulating the propagation of input errors through model predictions was initiated by O'Neill [9] and McGrath and Irving [10].

The most common applications of the Monte Carlo method in numerical computation are for evaluating integrals. Monte Carlo methods can also be used in solving systems of equations. All instances of Monte Carlo simulation can be reduced to the evaluation of a definite integral like the following:

$$\mu = \int_a^b f(x)\, dx \tag{1}$$

Formally, suppose we have a random variable, X, which has measurements over the range a to b. Also, assume that the probability density function of X can be written as $p(x)$. In addition, assume a second function, g, such that $g(x)p(x) = f(x)$. For example, $g(x)$ could represent a dose–response function on concentration and $p(x)$ is the density on concentration. The expected value (which is the most likely value or the mean value) of $g(x)$ is μ:

$$E(g(X)) = \int_a^b g(x)p(x)\, dx = \int_a^b f(x)\, dx = \mu \tag{2}$$

Notice that Equation 2 can be reduced to the same form as Equation 1. Estimating the expected value of $g(x)$ is a familiar statistical problem. A natural way of doing this is to take a random sample from x_i with distribution $p(x)$ and use the sample mean of $g(x_i)$ as an estimate of μ, that is,

Step 1. Draw random samples from $p(x)$:

$$x_i \sim p(x), \quad \text{for } I = 1, \ldots, n$$

Step 2. Calculate the sample mean:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} g(x_i) \tag{3}$$

This estimate has a variance of

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\int_a^b \left(g(x) - \int_a^b g(t)\, dt\right)^2 dx \tag{4}$$

As a simple example, suppose that X is a random variable with a uniform density over the interval $[a, b]$ with $p(x) = 1/(b - a)$. As a result, $g(x) = (b - a)f(x)$. The integral is estimated by

$$\mu = (b - a)\, E(f(X)) \tag{5}$$

The sample mean is calculated as

$$\mu = \frac{(b - a)}{n}\sum f(x_i) \tag{6}$$

where x_i are values of a random sample of size n from a uniform distribution over (a, b). The estimate is unbiased [11], and the variance of the estimate is

$$\mathrm{Var}(\hat{\mu}) = \frac{b - a}{n}\int_a^b \left(f(x) - \int_a^b f(t)\, dt\right)^2 dx \tag{7}$$

The estimated μ is based on a sample of simulated data; as a result, sampling error is always associated with the estimate. The law of large numbers states that the sample mean converges to the true mean in probability as the sample size increases:

$$\lim_{n\to\infty} \Pr(|\hat{\mu} - \mu| < \varepsilon) = 1 \tag{8}$$

In other words, a large sample size is necessary to reduce this sampling error.

In addition to increasing the sample size, reducing the sampling error can be done through efficient sampling. The Latin hypercube sampling is the most frequently used sampling technique for reducing Monte Carlo sampling error [12–14].

This method is designed to reduce sampling variance when sampling from several covariates. The technique uses a balanced or partially balanced fractional factorial design to sample, such that the sampling variance would be small at a given number of sample size. The Latin hypercube method was developed by McKay et al. [15] for providing input to a computer experiment. If the k covariates are uniform $U(0, 1)$, the ith sample of the jth variate is sampled by

$$v_j^i = \frac{p_j(i) - 1 + u_j}{n} \tag{9}$$

where $p_j(\cdot), j = 1, \ldots, k$ are permutations of the integers $1, \ldots, n$, sampled randomly, independently, and with replacement from the set of $n!$ possible permutations, $p_j(I)$ is the ith element of the jth permutation, n is the sample size, and u_j is an independent sample from $U(0, 1)$. Many researchers show that using Latin hypercube sampling can reduce the variance of the Monte Carlo estimator [12–14].

Using sampling experiment terminology, uncertain model parameters can be regarded as factors that contribute to the variation of the response. A Monte Carlo simulation is to sample all possible outcomes of the factors and study the impact they have on the response. The result from a Monte Carlo simulation is a collection of response data. These numbers can be summarized to produce statistics of interest, such as the

mean and variance. Let $g(x)$ be the model with uncertain variable x with probability distribution $p(x)$. The model output $y = g(x)$ is a random variable. A typical Monte Carlo simulation procedure starts with a sample of uncertain model parameters q randomly generated from the density $p(x)$. The sample is then used as realizations of true model parameters, and the corresponding model responses are evaluated. The mean and variance of the resulting model responses are often examined and used as estimates of the true mean and variance of y. The arithmetic mean of the Monte Carlo samples of model response is equivalent to the integral

$$\bar{y} = \int_{\Theta} g(x)\pi(x)\, dx \tag{10}$$

and the variance of the model output is equivalent to the integral

$$\mathrm{Var}(y) = \int (g(x) - \bar{y})^2 \pi(x)\, dx \tag{11}$$

The first integral (Eqn. 10) can be evaluated by first generating a sample of x and then calculating the sample mean of $g(x)$. The second integral (Eqn. 11) can be evaluated by first generating a sample of x and then calculating the sample variance of $g(x)$, which is the mean of $(g(x) - \bar{y})^2$. In other words, the mean of the model prediction is the integral of $f(x) = g(x)p(x)$. The importance of selecting the correct parameter distribution functions is self-evident. If an incorrect probability distribution for the uncertain parameter is used, the resulting estimates of the mean and variance are wrong.

The FEMVTF Statistics Committee process

As indicated in the previous discussion, Monte Carlo analysis is a multifaceted process and requires a great deal of consideration prior to implementing the model. The process by which Monte Carlo was implemented by the FEMVTF Statistics Committee provides an excellent case study on the steps and dynamics of implementing a Monte Carlo analysis with limited data.

Individuals from the FEMVTF Statistics Committee met to discuss the mechanism for conducting an uncertainty analysis of the PRZM 3.12 model and to identify those model input parameters that most contribute to model prediction error. This activity is part of a larger project evaluating the PRZM 3.12 model. The goal of the uncertainty analysis is to compare site-specific model predictions and field measurements using the variability in each as a basis of comparison.

To identify the most sensitive model input parameters, the committee relied on the results of a FEMVTF Plackett–Burhman analysis [16]. The Plackett–Burhman technique identifies those model inputs that cause the greatest change in model predictions as the values of the input parameters are varied. The committee discussed the Plackett–Burhman results and developed a final list of model inputs for evaluation using Monte Carlo techniques. For each of the final parameters, the committee attempted to define the nature of the sampling distribution for use in the Monte Carlo uncertainty analysis. Several of the committee members agreed to supply data or analyses that provide insight into the sampling distributions of the PRZM 3.12 input parameters selected for evaluation.

The team developed criteria for establishing sampling distributions of the PRZM 3.12 inputs. These criteria were used to ensure consistency in the procedures for evaluating model prediction error across sites. The criteria also ensure that the sampling distributions represent, to the degree possible, the actual site-specific uncertainty and variation in the parameters. Therefore, the criteria effectively increase the confidence that the Monte Carlo uncertainty analysis results reflect the true model error associated with a specific site and parameter set. In addition, the criteria provide a record against which the sampling distributions can be judged. Criteria for input parameter sampling distributions follow.

First, the sampling distributions must explicitly reflect within-site variation of the input parameters. This criterion ensures that intra- and intersite variation are explicitly identified and that any confounding of these types of variation is avoided (unless explicitly stated). For example, it would be inappropriate to have one input parameter distribution reflect within-site variation and the distribution for a second parameter to reflect between-site variation. The interpretation of the Monte Carlo output is difficult with such a parameterization.

Ideally, the input distribution should represent the range of possible values of the parameter for the explicit application of the model at a specific site. Preferably, actual field measurements of the parameter should be used to establish the distribution. Contributions to the prediction variance of intersite and interchemical components of uncertainty should not be used explicitly to judge model prediction accuracy. However, model runs that incorporate such variance components can be used to test the sensitivity of the model to the largest possible input parameter variance. In fact, incorporating the intersite and interchemical components of variation can be used to evaluate the expected model prediction error with small or nonexistent site-specific data sets.

Second, the form of the sampling distribution should be consistent between sites for a specific parameter. However, the sufficient statistics of the distribution may change. For example, if a normal distribution is chosen for a parameter at one site, then a normal distribution should be used at all other sites. However, the mean and variance of the normal distribution can be site specific.

This criterion ensures consistency in the interpretation of the Monte Carlo outputs between sites. It also provides a foundation for dealing with sparse data sets for specific parameters at some sites. In many cases, as few as two or three observations of the parameter are available at one site, with more data available at other sites. Therefore, we can use the site with the most data to determine the form of the distribution, with the sufficient statistics calculated on a site-specific basis. In addition, a consistent interpretation of the shape and spread of the Monte Carlo outputs between sites requires a consistent use of parameter-specific sampling distributions. The shape of the Monte Carlo prediction distribution is generally a function of the input distributions. The use of consistent input distribution forms allows the shape of the Monte Carlo output distributions between sites to be compared. For example, the output distribution may be fatter at one site than another. And finally, no scientific or modeling reason exists to believe that the form of the distribution for a specific input parameter should change between sites.

Third, the form of the distribution should reflect the magnitude, range, and interpretation of the parameter. Many of the input parameters have restricted ranges. For example, application rate cannot be negative. The sampling distribution should reflect the restricted range, with no chance of randomly drawing a negative value. The effect of this criterion is to

restrict the use of a normal distribution and increase the use of uniform, lognormal, beta, and custom distributions (random draws of actual measurements substitute for a formal distribution).

In addition, this criterion ensures that the expected site-specific range of a parameter is covered by the selected distribution. It also ensures that values outside the expected range are not overemphasized. For example, use of uniform distributions over a narrow range may be appropriate when the probability of occurrence of any parameter value is equal over the range.

Finally, expert judgment in establishing site-specific distributions is appropriate when few data are available, but a sensitivity test of the choice of distribution is required. For most input parameters, expert judgment is involved in the selection and calibration of the sampling distributions. We will perform sensitivity tests to evaluate changes in the Monte Carlo outputs with choice of distribution.

The FEMVTF Statistics Committee paid close attention to the procedural and statistical pitfalls of Monte Carlo analysis. The committee implemented the following activities as an effort to ensure the correct implementation of the Monte Carlo analysis: Strict guidelines were developed for the selection of sampling distributions for the input parameters (see the previous discussion); numerous information sources, databases, and experts were identified and consulted in the course of selecting the input parameter sampling distributions; a rigorous evaluation of statistical correlation among the input parameters was undertaken (the committee concluded that no statistical correlation exists between the parameters selected for evaluation); and a comprehensive sensitivity testing of the Monte Carlo outputs is planned in an effort to ensure results that are not overly dependent on the committee's assumptions and interpretations.

Site-specific sampling distributions for 11 parameters were generated for model applications at four case study sites. Groundwater measurements were available from U.S. sites in Georgia (GA1L) and North Carolina (NC4L), while runoff measurements were available from sites in Iowa (IA2R) and Georgia (GA1R). Appendix 1 presents the final sampling distributions and sources of information. A complete discussion of the experimental protocol is found in Carbone et al. [17]. A discussion of the methods for generating Monte Carlo predictions from PRZM 3.12 is found in Havens [18].

Of particular interest is the committee's use of the uniform distribution to represent many of the random parameters in the model. Information on the chemical decay rate was available for three of the four case study sites. The data sets at each site typically had 30 to 40 observations. Standard distribution fitting techniques on these relatively large data sets consistently yielded a beta distribution as the best fit. In addition, the committee agreed that the resulting shape of the beta distribution at the case study sites consistently matched the expert opinion as to the shape and scale of the distributions. However, for all other parameters, the amount of site-specific data was much less, with data sets containing from two to eight values. The committee decided that such small data sets were inadequate for establishing the shape of the sampling distributions but were adequate for establishing the parameter range for a specific location. Several important reasons exist for the use of the uniform distribution. First, the distribution reflects the degree of belief of the committee members with regard to the shape of the sampling distribution. For most case study locations, the frequency of occurrence for specific values of the parameter is unknown, and the existing data sets were not large enough for estimation. Second, from a model validation perspective, the major role of the sampling distribution is to bound the model predictions. The probability (or frequency) of occurrence within the bounds is of little interest. The objective is to evaluate the probability that field-specific measured values fall within the model predictions. Finally, as discussed earlier in this paper, the degree of belief in the Monte Carlo output is directly related to the understanding and belief in the parameter sampling distribution. The shape of the distribution is unknown and not easily estimable at most sites for most parameters, but the scale is reasonably defined. Therefore, the uniform distribution reflects the amount of knowledge and degree of belief that the committee had in the parameter sampling distributions, thus providing a basis for understanding and interpreting the final Monte Carlo predictive distributions. The width of the uniform distributions did not mask the ability to interpret site-to-site differences in model predictions. The authors feel that uniform distributions are more reflective of the true amount of knowledge in many model uncertainty analyses with limited data. The uniform distribution provides a reasonable alternative to guessing at the shapes of sampling distributions for limited data or limited knowledge situations.

Statistical comparisons of model predictions and field measurements

The choice of test statistic will directly influence the interpretation of the model validation results. A variety of statistical estimators and approaches can be used to compare model predictions with field observations, each test statistic possibly resulting in a different interpretation of model performance. Several test statistics and approaches were considered, as described in the following.

First, a one-to-one comparison of a model prediction to a field measurement was considered. A number of statistics are available, including mean squared error, absolute value of the differences, and others. In this approach, the model will be found to perform well only if the model provides a close estimate of individual measured values for a specific time step, depth (for leaching runs), or integrated runoff estimate (runoff comparisons). For complicated models like PRZM 3.12, asking the model to be accurate for small increments of space and time may not be a reasonable test of the model. In addition, the test statistics are highly influenced by individual data points where the modeled and measured values are far apart. Also, this approach does not consider model prediction error or measurement error in the interpretation of model performance. Finally, the PRZM 3.12 model predictions are generally not used on a point-by-point basis. Typically, the predictions over space or time are compiled, and sufficient statistics are used for interpretation and evaluation. Therefore, a one-to-one model validation approach does not reflect the practical use of the model outputs.

A probabilistic comparison of model predictions and field measurements was also considered. In this approach, model error is incorporated into the interpretation of model performance. Monte Carlo methods are used to provide uncertainty estimates for the model predictions at each time step and depth (leaching only). A comparison of the model predictive distribution at each time step and depth to a single measurement directly incorporates parameter uncertainty into the interpretation of model performance. We note that this approach does

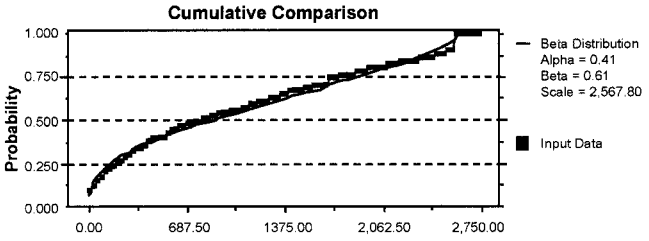

Fig. 2. Monte Carlo output: Runoff volume (m^3): 1992, day 195.

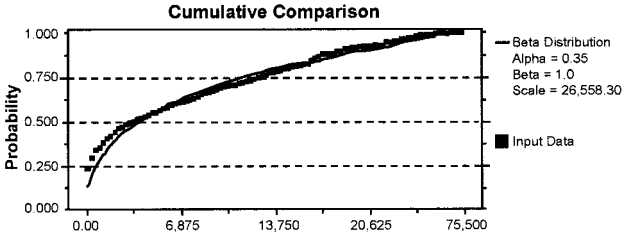

Fig. 3. Sediment yield (kg): 1992, day 195.

not incorporate other important variance components, such as model structure error, between-user error, and measurement error. However, if estimates of these variance components are available, they can easily be incorporated into the uncertainty analysis framework [19]. In this study, estimates of these variance components were not available and are not included in the analytical approach. We note, however, that the addition of such components will serve to increase the model prediction variance, resulting in the interpretation of increased model performance.

The objective of the validation exercise is to ascertain model performance under conditions of typical use. Therefore, the committee decided to use the second approach described previously for model validation. The first approach was judged to be an overly stringent analysis, serving to increase the chance that the model would not pass the validation exercise.

The test statistic for the second approach can be written as

$$PE = \frac{\sum_{i=1}^{n} X_i\{X_i = 1 \text{ if } P_i \geq M_j, \text{ else } X_i = 0\}}{n} \cdot 100 \quad (12)$$

where PE is percentage exceedence, n = number of Monte Carlo iterations, M_j $\{j = 1$ to the number of field observations$\}$ is an individual field measurement, P_i $\{i = 1$ to the number of Monte Carlo iterations$\}$ is an individual model prediction, and X_i is an indicator variable.

The expected value of PE is 50%, indicating that half the model predictions are above the measurement and half are below the measurement. Model accuracy is evaluated by examining the percentage of model predictions below and above the measured value. When the measured field value is shown to be in the general center of the prediction distribution, the model can be considered to be reasonably predictive. When the measured value occurs in the lower or upper portions of the prediction distribution, the model can be considered less accurate (within the bounds of uncertainty) but acceptable given the variability in the model parameters. If the entire prediction distribution is above or below the measured value, then the model may be considered to be inaccurate for those given circumstances. In some circumstances, however, this latter interpretation does not hold. In particular, for very small measured values (near the level of quantification), the model is frequently shown to predict into the range below the detection level or only slightly above the detection value. Carbone et al. [17] provide a discussion of this issue.

VALIDATION RESULTS

A complete discussion of the model validation results is found in Carbone et al. [17]. Examples of the Monte Carlo predictions are shown in Figures 2 and 3 of this paper. As shown in the figures, the beta distribution for the chemical decay rate is controlling the shape of the predictive distributions, while the scale of the resulting distributions is heavily influenced by the uniform distributions assigned to the remaining random parameters. In general, the model is shown to be a reasonable predictor of groundwater and runoff pesticide concentrations.

CONCLUSIONS

This paper describes an approach for model validation using Monte Carlo methods. This project was unique because of the interaction of committee members representing both the government and industry and the process by which a consensus was built as to the philosophical and statistical approach to model validation. Monte Carlo methods are shown to be a useful tool for model validation, incorporating important sources of uncertainty into the interpretation of model performance. A distinct advantage exists to the use of Monte Carlo analysis for model validation over typical model validation approaches. The uncertainty in the model input parameters is directly incorporated into the interpretation of model performance. The choice of the uniform distribution to reflect the committee's degree of belief is an extremely important point of this exercise and may be at odds with other, similar publications on model uncertainty analysis. In addition, the use of the percentage exceedence statistic as an appropriate statistic for model validation is an important contribution of the paper. Finally, establishing criteria for model validation, and reflecting those criteria in the choice of Monte Carlo sampling distributions and validation statistics, is shown to be a viable group-oriented process for model validation.

Acknowledgement—We are especially grateful to Blanche Dean for graphics assistance.

REFERENCES

1. U.S. Environmental Protection Agency. 1998. Guidelines for ecological risk assessment: Risk assessment forum. EPA/630/R-95/002F. Washington, DC.
2. Hammersley JM, Handscomb DC. 1964. *Monte Carlo Methods.* Chapman & Hall, New York, NY, USA.
3. Kloek T, Van Dijk HK. 1978. Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrics* 46:1–20.
4. Hammersley JM, Morton KW. 1956. A new Monte Carlo technique: Antithetic variates. *Proceedings of the Cambridge Philosophical Society* 52:449–475.
5. Wilson JR. 1984. Variance reduction techniques for digital simulation. *Am J Math Manag Sci* 4:277–312.
6. Warren-Hicks WJ, Butcher JB. 1996. Monte Carlo analysis: Classical and Bayesian applications. *Human Ecolog Risk Assess* 2: 643–649.
7. Burmaster DE, Anderson PD. 1994. Principles of good practice for the use of Monte Carlo techniques in human health and ecological risk assessments. *Risk Anal* 14:447–481.
8. U.S. Environmental Protection Agency. 1997. Guiding principles for Monte Carlo analysis. EPA/630/R-97-001. Office of Research and Development, Washington, DC.

9. O'Neill RV. 1973. Error analysis of ecological models. In Nelson DJ, ed, *Radionuclides in Ecosystems,* Conference 710501. National Technical Information Service, Springfield, VA, USA.
10. McGrath EJ, Irving DC. 1973. Techniques for efficient Monte Carlo simulation. AD 762 721–723. National Technical Information Service, Springfield, VA, USA.
11. Gentile JE. 1998. *Random Number Generation and Monte Carlo Methods.* Springer-Verlag, New York, NY, USA.
12. Stein M. 1987. Large sample properties of simulations using Latin hypercube sampling. *Technometrics* 29:143–151.
13. Beckman RJ, McKay MD. 1987. Monte Carlo estimation under different distributions using the same simulation. *Technometrics* 29:153–160.
14. Tang B. 1993. Orthogonal array-based Latin hypercubes. *J Am Stat Assoc* 88:1392–1397.
15. McKay MD, Conover WJ, Beckman RJ. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21:239–245.
16. Wolt J, Singh P, Cryer S, Lin J. 2002. Sensitivity analysis for validating expert opinion as to ideal data set criteria for transport modeling. *Environ Toxicol Chem* 21:1558–1565.
17. Carbone JD, Havens P, Warren-Hicks WJ. 2002. Validation of pesticide root zone model 3.12: Employing uncertainty analysis. *Environ Toxicol Chem* 21:1578–1590.
18. Havens PL. 2002. Development of a Monte Carlo sampling shell for the pesticide root zone model and its application by the Federal Insecticide, Fungicide, and Rodenticide Act Environmental Modeling Validation Task Force. *Environ Toxicol Chem* 21:1566–1569.
19. Warren-Hicks WJ, Moore DRJ, eds. 1998. *Uncertainty Analysis in Ecological Risk Assessment.* SETAC, Pensacola, FL, USA.
20. U.S. Department of Agriculture Research Service and Soil Conservation Service. 1998. Crop Parameter Intelligent Database. National Soil Erosion Research Laboratory, West Lafayette, IN.
21. Robertson WK, Hamond LC, Johnson JT, Boote KJ. 1980. Effects of plant water stress on root distribution of corn, soybeans and peanuts in sandy soil. *Agron J* 72:548–551.
22. Jung YS, Taylor HM. 1984. Differences in water uptake rates of soybean roots associated with time and depth. *Soil Sci* 137:341–350.
23. Borst HL, Thatcher LE. 1931. Life history and composition of the soybean plant. Research Bulletin 494. Ohio Agricultural Research and Development Center, Ohio State University, Columbus, OH, USA.
24. Mayaki WC, Teare ID, Stone LR. 1976. Top and root growth of irrigated and nonirrigated soybeans. *Crop Sci* 16:92–94.
25. Wischmeier WH, Smith DD. 1978. *Predicting Rainfall Erosion Losses—Guide to Conservation Planning.* Agricultural Handbook 537. U.S. Department of Agriculture, Washington, DC.

APPENDIX

Sampling distributions for selected uncertain parameters in Version 3.12 of the pesticide root zone model

1. Chemical Decay Rate (d^{-1})

Distribution: Beta

Because these are log10 transforms of the dissipation rate constant data, the following procedure was employed where the rate constant was set using the following formula: $k = -\ln(0.5)/(10^b)$ (where the superscript b is the sampled value from the beta distribution) (this assumes first-order degradation kinetics):

GA1L: $\alpha = 4.00$ $\beta = 9.92$ scale = 3.37(log10data)

NC4L: $\alpha = 4.58$ $\beta = 0.68$ scale = 2.75(log10data)

IA2R: $\alpha = 2.33$ $\beta = 0.46$ scale = 1.53(log10data)

GA1R: No data available.

Data source: Registrant chemical specific data package.

2. Rooting Depth (cm)

The pesticide root zone model (PRZM) has a limitation that the maximum rooting depth cannot be deeper than the total depth of the soil profile. For example, the total soil depth in the IA2R scenario is 91 cm, even though maximum rooting depth can range up to 122 cm. The Monte Carlo application does an error check, and if the rooting depth sampled is greater than the soil profile depth, the rooting depth is set to the depth of the soil profile (e.g., to 91 cm in IA2R).

Distribution: Uniform

Corn, Midwest: 0.457–1.219 m
Corn, Southeast: 0.32–0.9 m
Soybeans: 0.65–0.90 m

Data sources: Crop Parameter Intelligent Database [20]; Robertson et al. [21]; Jung and Taylor [22]; Borst and Thatcher [23]; Mayaki et al. [24].

3. Curve Numbers

Distribution: Uniform

GA1R:

Fallow: 82–88
Cropping: 73–91
Residue: 75–81

IA2R:

Fallow: 82–88
Cropping: 45–100
Residue: 75–81

Data source: Site specific based on measured rainfall and runoff data; PRZM 3.12 user manual.

4. K_d (cm^3/g)

Distribution: Uniform

GA1L: 0.25–0.36

Data source: Registrant chemical specific data package. Measured K_{oc} was used to generate a chemical specific regression equation relating K_{oc} and organic carbon (OC) to K_d. The regression equation was then used in a Monte Carlo analysis in conjunction with measured soil organic carbon to generate a distribution of potential K_d values across the site.

NC4L: 0.02–0.19

Data source: Registrant chemical specific data package. Measured K_{oc} was used in a Monte Carlo analysis in conjunction with measured soil OC [$K_d = K_{oc} \cdot OC/100$] to generate a distribution of potential K_d values across the site.

IA2R: 18.7–208

To set K_d for the lower soil horizons, the following procedure was used: K_{oc} was calculated from the sample K_d value ($K_{oc}(1) = K_d(1)/0.0183$; horizon 1 has 1.83% OC):

$$K_d(2) = K_{oc}(1) \cdot 0.0135 \text{(horizon 2 has 1.35\% OC)}$$

$$K_d(3) = K_{oc}(1) \cdot 0.0093 \text{(horizon 3 has 0.93\% OC)}$$

$$K_d(4) = K_{oc}(1) \cdot 0.0057 \text{(horizon 4 has 0.57\% OC)}$$

Data source: Registrant chemical specific data package. Measured K_{oc} was used in a Monte Carlo analysis in conjunction with measured soil OC ($K_d = K_{oc} \cdot OC/100$) to generate a distribution of potential K_d values across the site.

GA1R: No site-specific information is available.

5. Bulk Density (g/cm^3)

Distribution: Uniform

Note: The bulk density distributions are depth specific. When models are run at depths that do not match the measured field data, field data associated with the nearest reasonable depth are used to parameterize the model.

GA1R: No data available

Site	Depth (cm)	Range of bulk density
IA2R	10	1.10–1.19
	30	1.07–1.27
	60	1.02–1.36
	90	1.09–1.28

Data source: Registrant chemical specific data package.

Site	Depth (cm)	Range of bulk density
GA1L	15	1.49–1.56
	30	1.49–1.60
	45	1.49–1.57
	60	1.49–1.59
	75	1.54–1.59
	90	1.54–1.57
	105	1.54–1.57
	120	1.54–1.57
	135	1.54–1.62
	150	1.54–1.59

Data source: Registrant chemical specific data package.

Site	Depth (cm)	Range of bulk density
NC4L	0	1.45–1.54
	15	1.38–1.52
	30	1.42–1.53
	45	1.40–1.54
	60	1.39–1.50
	75	1.40–1.43
	90	1.37–1.43
	105	1.39–1.45
	120	1.41–1.49
	135	1.40–1.49
	150	1.48–1.49
	165	1.45–1.48
	180	1.42–1.43
	195	1.43–1.48
	210	1.46–1.54
	225	1.46–1.56
	240	1.43–1.53
	255	1.43–1.47
	270	1.44–1.45
	285	1.44–1.47
	300	1.45–1.46
	315	1.43–1.45

Data source: Registrant chemical specific data package.

6. Bulk Density

Global variability: 13–16.2% RUSTIC user manual.

7. Pan Factor (%)

Distribution: Uniform

GA1L and GA1R: 75–77
IA2R: 71–73
NC4L: 75–77

8. Application Rate (kg/ha)

Distribution: Uniform

NC4L: No site-specific data.
GA1L: 0.15–0.32
GA1R: 0.13–0.22
IA2R: 0.94–2.12

Data source: Registrant chemical specific data package.

9. Management Factor (%)

Management factors are taken from predicting rainfall erosion losses [24]. Each matrix below is crop and crop practice specific. Crop-specific USLEC value ranges were selected from those presented by Wischmeier and Smith [25] to most closely approximate plant growth stages as constrained by PRZM input requirements reflecting fallow, cropping, and residue conditions.

Distribution: Uniform

GA1R: Annual cotton, conventional moldboard plow and disk

Fallow period:	36–42
Seedbed period:	59–68
Crop stage 1 (establishment):	59–63
Crop stage 2 (development):	43–49
Crop stage 3 (maturing crop):	22–44

IA2R: Corn after corn in meadowless systems, spring moldboard plow, crop residues left on field

Fallow period:	31–51
Seedbed period:	55–68
Crop stage 1 (establishment)	48–60
Crop stage 2 (development):	38–45
Crop stage 3 (maturing crop):	20–33
4L (residue):	23–47

GA1L: Corn after corn
as for GA1R

NC4L: Soybeans after corn, spring moldboard plow, crop residues left on field, plow disk and harrow for seedbed

Fallow period:	33–45
Seedbed period:	60–68
Crop stage 1 (establishment):	52–60
Crop stage 2 (development):	38–43
Crop stage 3 (maturing crop):	17–29