

(Occurrence Summary Document) (USEPA, 2002f), were used to inform the modeling effort.

EPA Response: EPA's occurrence model development work was significantly revised to reflect peer review comments prior to the March 2002 Occurrence Methodology Document (USEPA, 2002e) and the April 17, 2002, **Federal Register**. The additional work involved the development of a detailed simulation study to evaluate the Bayesian model. EPA evaluated the performance of the Bayesian estimator and an alternative occurrence estimation approach, the Regression on Ordered Statistics (ROS) method, against synthetic data (*i.e.*, data developed with known national contaminant occurrence distributions). This simulation study also enabled an explicit evaluation of the validity of the assumption of a log-normal distribution of the data.

The simulation study was conducted using varying conditions of a correctly and incorrectly specified model, and synthetic data sets developed with high and low amounts of non-detected data. The study findings indicated that the Bayesian estimator performed well at estimating the distributions of contaminant concentration means (especially in the upper tails), performed better than the alternate approach (*i.e.*, the ROS method), and accurately estimated the uncertainty of the distributional estimates. The Agency believes that this analysis supports the validity of EPA's analytical approach. The Bayesian model was tested against the ROS approach because the ROS method is an accepted drinking water contaminant occurrence estimation approach and was used to estimate occurrence for the recent arsenic rule. These findings were all included and described in the Six-Year Review's Occurrence Methodology Document.

EPA has attempted to make its occurrence analysis as clear as possible. In response to the concerns raised by the peer reviewers, a less technical description of the occurrence estimation methodology, aimed at the general reader, was added to the main body of the document. A detailed description of the analysis, intended for readers with technical expertise, including the complete computer code used for model analysis, was incorporated into an appendix of the document. EPA agrees that its estimation methodology is complex, but also believes that it is as transparent as possible while still providing a technically accurate description of the Agency's analysis. The use of simple national occurrence (statistical) assessments is not possible

at this time because there is no national database with a complete collection of regulated contaminant occurrence data. Thus, there is no ideal basis for comparison of national occurrence studies (*i.e.*, the true system contaminant means and national distributions of contaminant occurrence are not, and cannot, be known). The validation approach suggested by the commenter (*i.e.*, basing the model on a portion of the data set and using the remainder to test the model) is intended for a regression-type of model using observed system means to develop a model for system-specific predictions. This approach is not possible for the six-year occurrence assessments, since, to the best of EPA's knowledge, data on the true individual system contaminant mean concentrations and national distributions are not available.

Regarding the other survey studies included in the Occurrence Summary Document, few, if any, provide the quantitative analytical results and national, representative coverage that would enable direct comparison to, or inclusion in, the Six-Year Review estimation analyses conducted with the 16-State cross-section occurrence data.

c. Other Issues Related to the Occurrence Technical Review. One commenter stated that the Agency's current approach to estimate occurrence, employing a conservative methodology and making conservative simplifying assumptions in the absence of definitive data, was appropriate. On the other hand, the commenter argued that it was not appropriate for the Agency to conduct as massive a data collection and analysis project as was undertaken without clear quantitative objectives for the analysis identified a priori. The commenter noted that it was not apparent from either the April 17, 2002, **Federal Register** or the Occurrence Methodology Document (USEPA, 2002e) that the Agency undertook an effort to set performance objectives for the occurrence estimation.

The commenter felt that the Occurrence Methodology Document does not allow the reader to determine if the data are well apportioned among the categories for which results are reported. They also noted that they were unable to find indications in the support document that such an analysis was undertaken in preparation for constructing the Bayesian model. The commenter stated that the support document does not include actual numeric counts or ranges of detected values and suggested that it would be useful to have this information by contaminant, State, system size category, and water type, as well as an

explicit count of non-detects by this same matrix.

EPA Response: There are several general approaches when undertaking and designing studies that require large amounts of data. As the commenter states, a priori data quality objectives are part of one research approach where study objectives (including technical statistical performance measures) are set, determinations are made on how to meet those objectives, and then the study is designed and implemented accordingly. This ideal was not practical for the national occurrence study conducted for the Six-Year Review because EPA did not have the resources to generate original data, and was thus dependent on the data that could be obtained from the States. The approach taken by the Six-Year Review was to gather a large amount of data that, in aggregate, was expected to be indicative of national contaminant occurrence, develop an occurrence estimation model that built upon what has been learned from recent regulatory development work, and then evaluate how good the resulting model estimates are.

As discussed in section IV.A.6.b of today's action, the true national distributions of contaminant occurrence cannot be known. The 16-State national cross-section data set used for the Six-Year Review is the largest compliance monitoring database for drinking water compiled by EPA to date. The database represents approximately 37 percent of the total number of public water systems and 43 percent of the total population served by public water systems in the United States. External peer reviews assessed the approach for developing the national cross-section and its "representativeness" separately under the Chemical Monitoring Reform (CMR) project (in 1998/1999) (USEPA, 1999c) and the Six-Year Review project (USEPA, 2002e), and provided generally favorable comments.

The data management and cross-section development have been described in detail in the support documents for the CMR and the Six-Year Review. Further tabulations of the data have been generated and presented, as the commenter requested, in the final Occurrence Methodology Document (USEPA, 2003d). This information includes the numbers and percentages of analytical detections and non-detections for each contaminant in each of the system size and source water type categories. Generally, because of the large amount of data and the manner in which the Bayesian model handles data, the distribution of observations across the various categories does not significantly affect EPA's estimates. The