Appendix 3A. Expert Elicitation Data Development Procedures

Sources and Structure of Raw Data

Two data sources were used as the basis for generating sample ecological data to be used in the expert elicitation process. Both datasets were generated by the North Carolina Department of Environmental Quality (DEQ). The first source is a data set that includes a wide range of ecological measurements across 68 creeks and rivers within the Haw River and Upper Neuse River watersheds, and was used as the source of measurements for five biophysical metric variables: fecal coliform, specific conductance, total nitrogen, total phosphorus, and turbidity.

The second data source is a set of macroinvertibrate data collected from creeks and rivers in the Upper Neuse River watershed and the Cape Fear River watershed, of which the Haw River is a major tributary. This data set served as source of biotic index measurement data.

Data Cleaning and Formatting

The first water quality dataset was processed to specify the name of the creek or river and the watershed in which it runs. Measurement type was filtered to include only six types: fecal coliform, inorganic nitrogen, Kjeldahl (organic) nitrogen, specific conductance, total phosphorus, and turbidity, and unit conversions were done to ensure all measurements in each category had the same unit. Measurements noted as being below the analytical threshold for that technique were included as zero values. The data were also filtered by date so that all measurements included were taken during the growing season, which in North Carolina occurs between March and November.

The macroinvertibrate data were filtered to include only the measurements of biotic index along with geographic and temporal identification data. These data were then formatted to match the other ecological measurement data and the two datasets were merged. A small portion of this formatted data can be seen below.

In order to determine the relationships between each of the water quality variables, the data were organized and formatted by determining each instance where all six measurements were taken in the same location on the same day. These events are identified by creek name, date, and coordinates of the testing site. Once these events are identified, the water quality dataset is filtered to include only data collected from them. These data are organized so that each row represents a measurement event with six columns, one for each nutrient measurement.

Analysis of Ecological Data

The nitrogen measurements in the dataset skew much higher than expected for typical waterways in the included watersheds. This is likely due to an overrepresentation of testing sites downstream of wastewater treatment plants. A K-means cluster analysis was performed to identify testing events that are likely affected by wastewater treatment plant discharge. The water quality dataset was filtered to exclude these suspected outliers.

The fecal coliform data featured several outliers with extremely high values (Z-value > 45) which have an outsize influence on summary statistics and on the sample data generation process. A similar k-means cluster analysis was performed, and suspected outliers were removed from the dataset.

Biotic index (BI) data are added to the dataset; however BI measurements were not made during the same measurement events and at the same sites as the nutrient values were collected. BI measurements are much more consistent for a particular stream than nutrient levels. Average BI measurements were calculated for each stream, and are joined to each row of nutrient data by stream.

Sample Data Generation

Data are prepared for the generation algorithm by eliminating zero values so that the logarithm of the dataset can be taken, since the distribution is assumed to be log-normal. The logarithm of the dataset is taken to normalize the distribution, a sample dataset with 1000 rows is generated, and the data is converted back to the original log-normal distribution.

Selection of Sample Data Rows

Z-Values are calculated for each ecological data point in the generated sample data. For each variable, the 50 data rows with greatest magnitude Z-value are selected. These rows are narrowed by excluding any row which features in the top 50 Z-values for more than one variable, and then 12 rows are selected at random for each variable, making up the first 72 data rows.

The remaining rows are selected by filtering the data to include only the lowest 80 percent of measurements for each water quality variable, and randomly selecting 40 observations from this data set. The 72 data rows representing the extremes of each variable and the 40 data rows that represent the more normal range of each variable are combined to form the final set of 112 generated data rows.