5.0 Verification and Validation of 3MRA Site-Based Data Collection and Processing

The 3MRA modeling system allows users to model specific facilities using site-based data. In the representative national data set, site-based data include spatial data defining the site layout (watersheds, waterbodies, and receptor types and locations), WMU data, soil data for the watershed surface soils and the vadose zone, and a few other characteristics based on slope and land use patterns. Figure 5-1 provides an overview of the steps taken to ensure the quality of the site-based data. This section summarizes the data collection methods for the representative national data set (Section 5.1), verification of the data (Section 5.2), and validation of the data (Section 5.3).

5.1 Site-Based Data Collection Methods

Site-based data are data that are assigned to sites based on characteristics specific to the site's location. A significant component of the site-based data is the site layout, i.e., the physical arrangement features at the site. Other components of the site-based data include

- WMU characteristics (e.g., WMU area, capacity, and waste volumes), which are based on site-specific Industrial D survey data,
- Soil texture, which is assigned based on the predominant soil texture in a watershed from the State Soil Geographic (STATSGO) database, and
- Regional assignments to a meteorological station, hydrogeologic environment, or U.S. Geological Survey (USGS) hydrologic region, which are made based on site location or site-specific data. For example, hydrogeologic environment assignments are assigned based on the subsurface conditions at the site.

For the representative national data set, EPA collected and processed site-based data for 201 Industrial Subtitle D[1] facilities using a combination of geographic information system (GIS) and database programs. In general, GIS Arc Macro Language (AML) programs (for ArcInfo and ArcView) were used to establish the spatial frame of reference at each site, collect spatial data within this frame, and interrelate different spatial data coverages using overlays and spatial relationships to create the data necessary to populate the site layout variables required by the

[1] An Industrial Subtitle D facility is an industrial facility that generates solid wastes that are not defined as hazardous wastes under Subtitle C of the RCRA. The focus of the 3MRA modeling system representative national data set is on the management of wastes in WMUs designed to manage nonhazardous Industrial Subtitle D wastes.

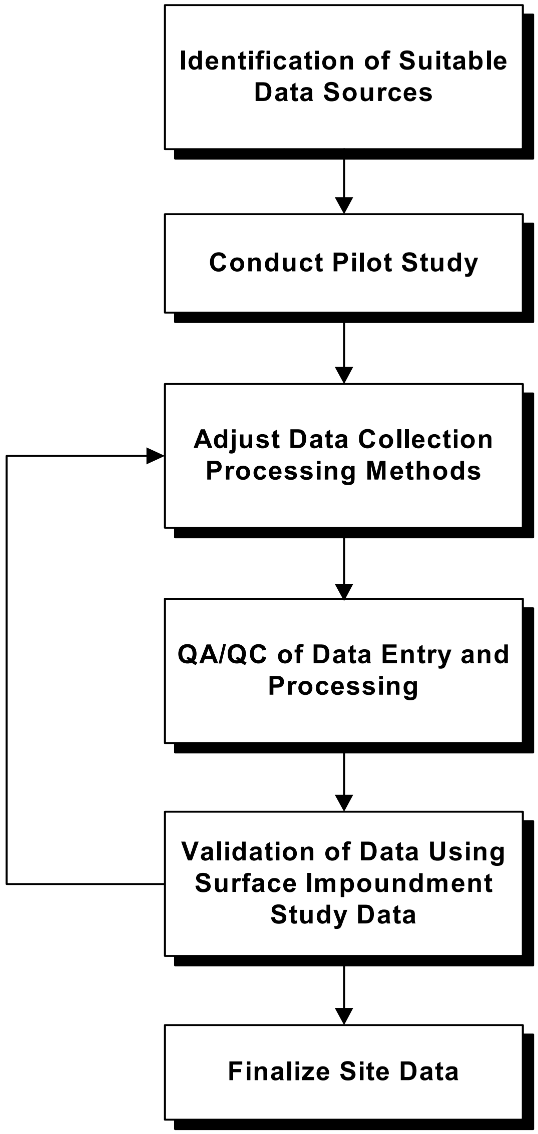

Figure 5-1. Overall approach to ensure the quality of the 3MRA site-based data.

3MRA modeling system. EPA used GIS AML programs to create the following spatial data coverages:

- Waterbody and watershed layout,
- Human receptor locations, including farms and residences, and
- Ecological habitats and receptor home ranges.

Figure 5-2 shows these coverages. GIS AML programs also assign soil map units to each watershed; assign meteorological stations and hydrologic regions to each site; populate the human receptor points using U.S. Census data; establish *x,y* (grid) locations for each spatial feature; overlay and relate spatial features; and export these site-based spatial data as a series of Access database tables.

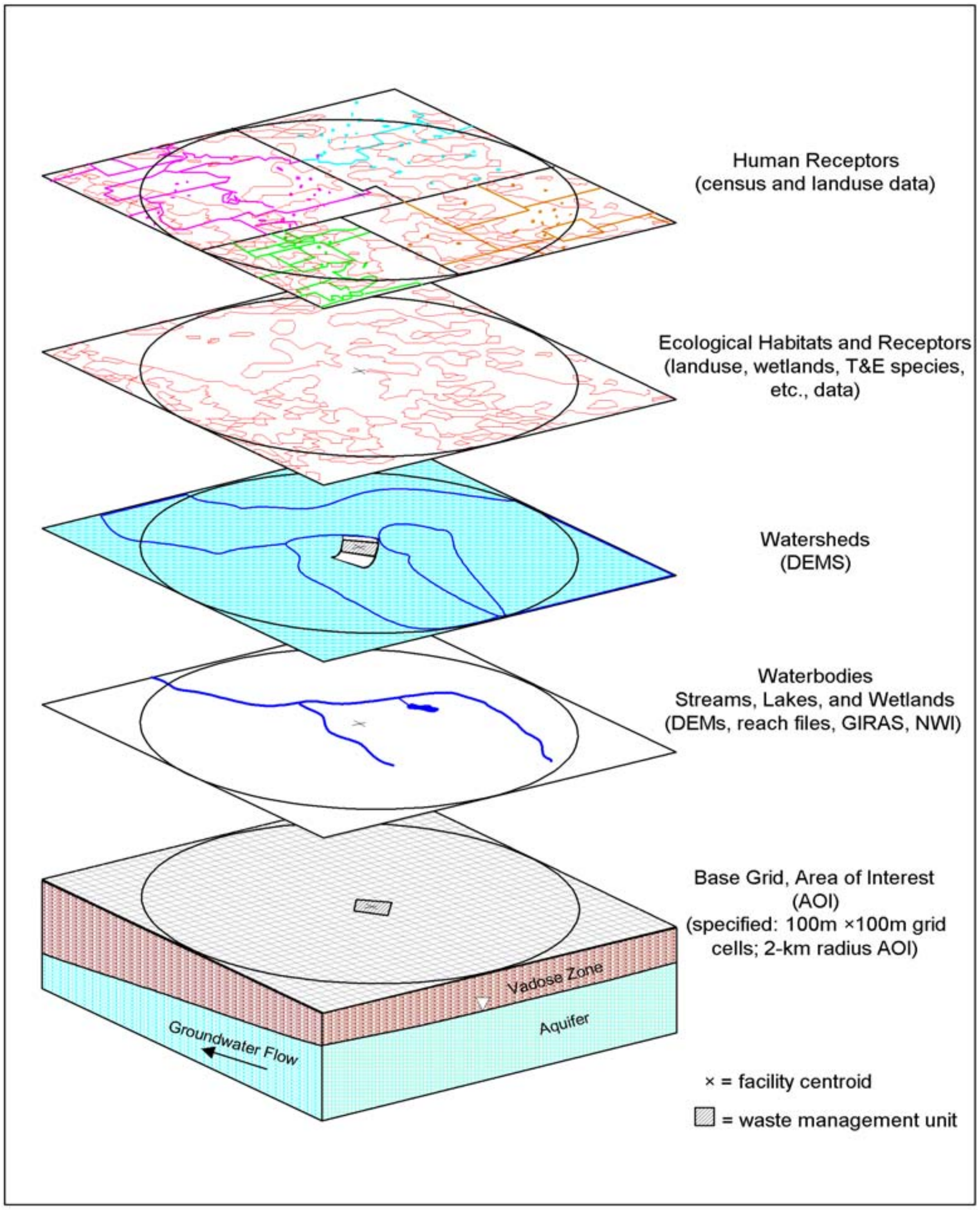

Figure 5-2. Site-based spatial overlays.

Following GIS processing, EPA used database programs (in structured query language [SQL] and Visual Basic) to perform quality control (QC) checks and process site-based data from the GIS to meet 3MRA modeling system variable and database specifications and requirements. These database processing programs produce two primary files:

- The 3MRA modeling system input database, including site-based, regional, and national input data tables, along with data tables for references, variable correlations, user-defined empirical distributions, and general facility information.

- The grid database,[2] including six data tables containing *x,y* coordinates for watersheds, waterbodies, farms, human receptor points, drinking water wells, and ecological habitats.

Table 5-1 summarizes the site-based data collection process for the 3MRA representative national data set, including methodology and data sources. The following sections provide additional detail on the major steps involved in creating the sample site-based data set.

5.1.1 Conduct Pilot Study

EPA conducted a pilot data collection study for five Industrial Subtitle D sites to provide information needed to plan the data collection effort. The pilot study had three components:

1. Desktop methods, such as manual delineation of watersheds using USGS topographic quadrangle maps,

2. Site visits, using global positioning system (GPS) technology, to obtain accurate facility locations and verify features (e.g., receptor locations) and conditions of interest around the site through windshield surveys, and

3. State office visits, to collect available permit information on facility location; WMU location, design, and operation; and hydrogeology (where available).

Table 5-2 summarizes, by data type, the information that was collected during the pilot study, along with the methods developed and used to collect the data for the 201-site representative national data set. In many cases, the pilot study enabled EPA to identify, develop, and use efficient and effective methodologies during the data collection effort.

5.1.2 Establish Spatial Framework/Initial Setup

Given the location (latitude and longitude) of the WMU to be modeled, the AOI for an assessment is determined in the GIS by drawing a radius (2 km for the representative national data set) extending from the corners of the WMU (all WMUs are assumed to be square for the sample data set). This determines the spatial frame of reference for the analysis by defining the area to be characterized by the site-based data collection effort and modeled by the 3MRA modeling system. The AOI may be further subdivided into distance rings. For the representative national data set, these rings were set at 500 m, 1,000 m, and 2,000 m from the edge of the WMU for human risk and at 1,000 m and 2,000 m for ecological risk.

[2] The grid database is used by the 3MRA modeling system Site Layout Processor (SLP) to create the air modeling points for the Air Module and add them to the 3MRA modeling system input database.

Table 5-1. Site-Based Data Collection for Representative National Data Set: Summary by Major Activity

Activity	Methodology Summary	Data Sources
Conduct pilot study	Desktop methods of data collection Site visits State office visits	See Table 5-2
Establish spatial framework/ initial setup	Locate site and place WMU Delineate area of interest and distance rings from WMU edge Establish base grid Assign meteorological station, hydrologic region, hydrogeologic environment Collect GIS source data	Location: Industrial Subtitle D database, EnviroFacts Base grid coverage Meteorological station coverage Hydrologic region coverage Surficial geology and aquifer coverages Source data: see below
Delineate waterbodies	Programs use Reach File 3 (RF3) stream networks to adjust digital elevation models (DEMs) to be consistent with actual stream networks (automated delineations only) Use National Wetlands Inventory (NWI), RF3, and Geographic Information Retrieval and Analysis System (GIRAS) for lakes and wetlands Attribute with type, connectivity, length, area (lakes and wetlands)	Stream network: RF3 Lakes: RF3, NWI Wetlands: NWI, land-use coverage (GIRAS)
Delineate watersheds	Delineate using programs or manually Digitize manual delineations Attribute with area, slope, soil map units, land-use data, waterbody connectivity	Waterbody delineation coverage Topography: DEMs Soils: STATSGO Land use: GIRAS
Place human receptors	Place receptors at centroid of Census block/ring polygons Place county median farm in every block group with beef or dairy farmers and cropland/pasture land use Place wells at receptor points in block groups with wells Attribute with Census population data by receptor type, age cohort	U.S. Census (1990) U.S. Agricultural Census (average of 1987 and 1992) Land use: GIRAS (1975–1985)
Place ecological receptors	Manually delineate habitats using GIS ArcView tool Place home range bins in habitats Assign receptors to home range bins by habitat Attribute with type, area	Waterbody delineation coverage Topography: DEMs Land use: GIRAS Habitat/receptor data: literature
Overlay GIS coverages	Overlay and relate GIS coverages Create data tables on connectivity, area of overlap Create data tables with *x,y* coordinates of waterbodies, watersheds, human receptors, farms, wells, and receptor home ranges	Distance ring coverage Waterbody delineation coverage Watershed delineation coverage Human receptor and farm coverages Habitat and home-range coverages Base grid coverage
Process data for 3MRA modeling system	Import GIS data tables and run QC programs and protocols Create derived values/distributions for soil and land-use variables Populate home ranges and attribute ecological receptors Convert connectivity to 3MRA site layout format, including correct and complete indices Convert Universal Transverse Mercator (UTM) coordinates to site coordinates Run QC programs on completed data set	GIS data tables and grid (coordinate) files STATSGO soil database Land use and soil property look-up tables (from literature) 3MRA data specifications

Citations and detailed descriptions of the data sources and methodologies shown in this table can be found in Volume II, Data Collection.

Table 5-2. Comparison of 3MRA Modeling System Representative National Data Set Data Collection Activities with Pilot Study Approach

Data Type	Approach for 201 Sample Data Set Sites	Approach for Pilot Study
Facility location	Industrial Subtitle D facility match to Locational Reference Tables (LRT) (facility centroid, front gate, zip code centroid locations) Address matching software for unmatched sites	GPS front gate, facility boundaries, WMUs State office visits (facility maps, locations)
Waste management unit	Industrial Subtitle D Screening Survey data (WMU area, capacity, waste loading) National estimates based on model unit approach for other inputs	State office visits (facility maps; WMU locations, dimensions, operation) Site visits to confirm characteristics where WMUs are visible
Land use	GIS analysis (GIRAS coverage)	Confirm land use, details within GIRAS area coverages during site visits
Topographic	Watershed delineation (area, flow length, slope, streams) using GIS DEMs or manual methods (flat sites)	Desktop delineation of watersheds (area, flow length, slope, streams) on USGS topographic maps Confirm general watershed characteristics, topographic features, man-made drainage during site visits
Waterbody	Waterbody network delineation using RF3, NWI, GIRAS and DEM coverages Assume lakes, order 4 streams are fishable	Desktop location on topographical map During site visits, GPS waterbody locations (streams, lakes, ponds), confirm general characteristics and add details on fishable waterbodies
Aquifer	Hydrogeologic Database (regional analysis), national distributions for aquifer properties Assume water flows downhill, towards surface water	Desktop data review and compilation from state office visits and hydrogeologic setting analysis During site visits, ground-truth setting and detail on residential well use
Human receptor information	U.S. Census data (block centroids, areal averages within radius of interest) and other data sources (county agricultural census, national home gardener percentage)	Ground-truth locations (GPS), areal averages, farms, exposure pathways Provide detail on home gardeners, farmers, subsistence activities, maximum exposed individual
Ecological receptor information	Habitat within AOI delineated using GIRAS land-used data, topographic, other data	Ground-truth habitats, exposure pathways

Once the AOI and rings were established, the site-based source data were collected for each of the spatial data layers to be created. (The GIS data sources for these data are summarized in Table 5-1.) Regional assignments, including meteorological station and USGS hydrologic region, were assigned using GIS overlays during this initial step of the analysis.

Site-based WMU data include the WMU dimensions, capacities, and 1985 waste volumes obtained from EPA's *Screening Survey of Industrial Subtitle D Establishments* (Schroeder et al., 1987) and the size-related WMU variables imputed from the Industrial Subtitle D survey data. The imputed WMU inputs were derived from the site-specific Industrial Subtitle D data using national relationships developed from literature, personal communications, and best engineering judgments.

5.1.3 Delineate Waterbodies and Watersheds

Creating the waterbody and watershed layout requires GIS processing to delineate the waterbodies and watersheds and to obtain spatially related parameters (Figure 5-3, Steps 1 through 8) and then conducting database processing of the GIS data to provide the exact data for the 3MRA modeling system (Figure 5-3, Step 9). For the representative national data set, waterbodies (lakes, streams, and wetlands) within the AOI were delineated using nationally available coverages, including EPA's RF3, land use/land cover data (GIRAS), and, where available, the NWI. Watershed subbasins were delineated around each waterbody either automatically, using DEMs of topography, or manually, using topographic maps created from the DEMs.

After waterbody and watershed delineations were complete, waterbody type was attributed using the original GIS coverages. Watershed and surface water connectivity, as well as stream order for manually delineated sites, were assigned manually. A local watershed[3] was identified for the WMU and divided into two to three subareas, depending on the WMU type and topography. Because the waterbody layout data were highly variable from site to site, the resulting waterbody data were visually QC-checked for every site before final processing.

5.1.4 Place Human Receptors

For the representative national data set, EPA used GIS programs to place human receptors within the AOI using U.S. Census block centroids (or block/ring centroids if a block was divided by a distance ring) and to randomly place beef and dairy farms based on U.S. Census block group, agricultural census, and land-use data. Table 5-3 shows the primary data sets used to derive the human receptor data.

The representative national data set includes two basic receptor types: residential receptors (residents and home gardeners) and farmers. Residential receptors may be recreational fishers in addition to being a resident or home gardener. Farmers may be beef or dairy farmers, and either type of farmer may also be a recreational fisher. Within each of the two basic receptor types, the 3MRA modeling system considers five age cohorts: infants (aged 0 to 1 year), children aged 1 to 5 years, children aged 6 to 11 years, children aged 12 to 19 years, and adults (aged 20 years and older).

---

[3] The drainage subbasin that includes the WMU.

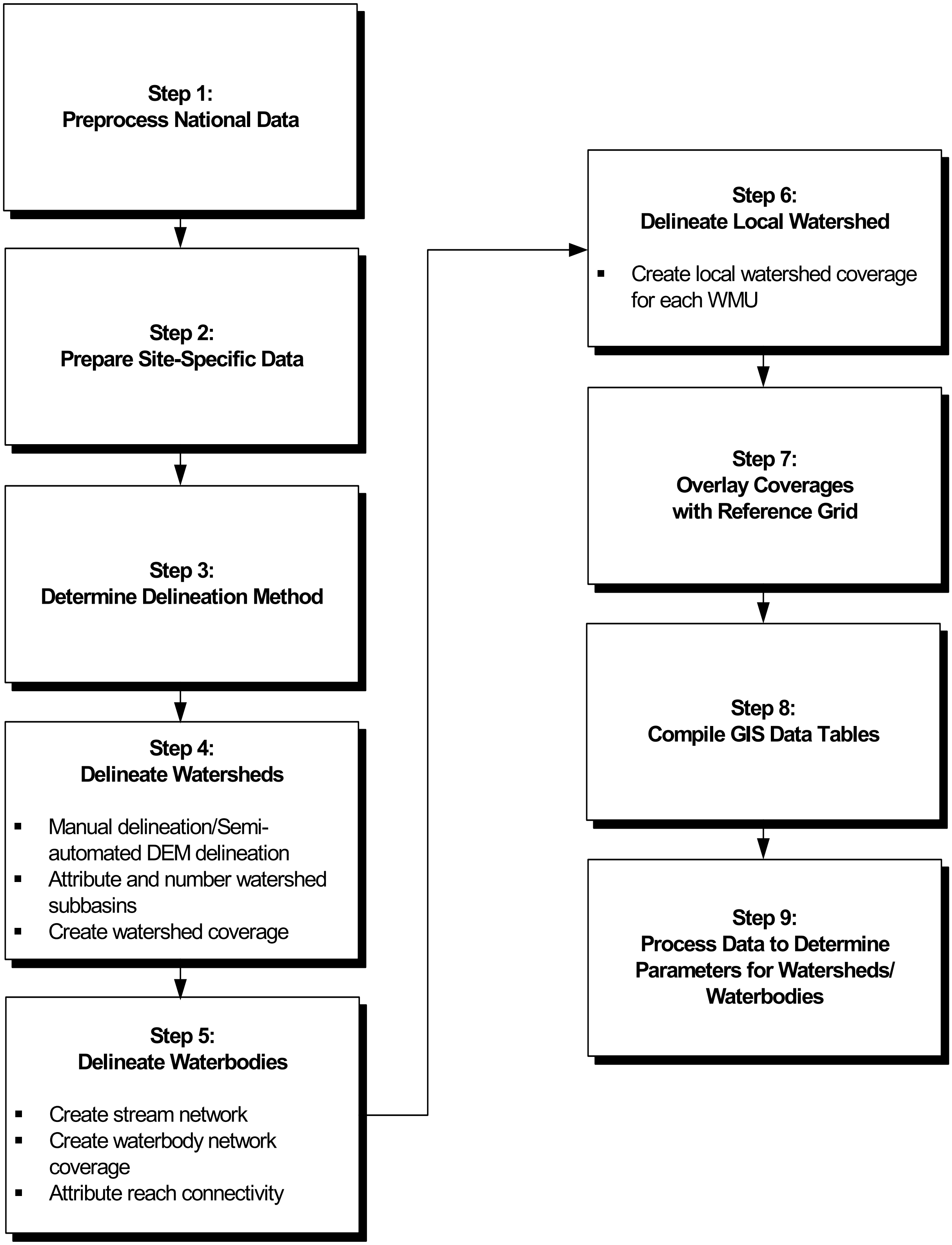

Figure 5-3. Overview of watershed and waterbody layout processing.

Table 5-3. Primary Human Receptor Data Sets, Date, and Scale

Data/ Receptor Type	Data Set	Date	Scale
Residents, home gardeners	U.S. Census block data Summary Tape File (STF) 1B attribute data U.S. Census Topologically Integrated Geographic Encoding and Referencing System (TIGER)/Line block coverages Exposure Factors Handbook (home gardener)	1990	1:100,000 scale mapping
Beef and dairy farmers, farm size	U.S. Census block group data STF 3A attribute data U.S. Census TIGER/Line block group coverages	1990	1:100,000 scale mapping
	GIRAS land-use data	1975–1985	1:250,000 scale mapping
	U.S. Census of Agriculture	1987 and 1992 (avg.)	County level
Recreational fishers	National Survey of Fishing, Hunting, and Wildlife	1992	State level

Citations and detailed descriptions of the data sources shown in this table can be found in *Volume II, Data Collection.*

Following the placement of receptors, each receptor point and farm was populated by receptor type and age cohort using 1990 U.S. Census data (and other data shown in Table 5-3) and assuming a uniform distribution of the population across the entire block.

5.1.5 Place Ecological Receptors

The representative national data set includes 12 terrestrial, wetland, and waterbody margin habitats, each with assigned ecological receptors that were developed to represent ecological receptors nationwide. For each site, EPA delineated ecological habitats by manually applying an ArcView GIS tool that included GIRAS land use data, the representative national data set waterbody and watershed coverages and attributes, and DEM topographic data. Habitats were delineated by selecting grid cells and coding them with one of the 12 habitat codes.

After the manual habitat delineation was completed, EPA generated receptor home ranges using automated GIS AML programs. Because placing a home range for each receptor within every habitat proved inordinately time consuming (even though automated), EPA used four home range bins to reduce the number of home range placement iterations. Within a habitat, each receptor was assigned to one of four bins based on its average home range size. The GIS programs delineated the four home ranges for each habitat by randomly placing the largest bin within the habitat and then randomly placing subsequent, smaller bins within the next largest bin.

Following the placement of home ranges, both the habitats and home ranges were limited to the area within the AOI and overlaid with the coverages of waterbodies, watersheds, local watersheds, and distance rings to determine the spatial relationships between them. Ecological receptors were added to the habitat using two look-up tables: receptors by habitat type and receptors by Bailey's section.

5.1.6 Overlay GIS Coverage

Because the GIS coverages cannot be exported directly to the 3MRA modeling system, spatial layers are defined in terms of their relationships to each other and in terms of a base grid composed of 100 × 100 m cells. This resolution roughly corresponds to the minimum resolution of several of the site-based data sources used to develop the representative national data set. To create the grid files, GIS coverages for watersheds, waterbodies, farms, and ecological habitats were converted to UTM coordinates and overlaid on the master 100 × 100 m grid file to create grid tables containing the *x,y* UTM coordinates of each grid cell occupied by a feature. The UTM coordinates for human receptor points and wells (human receptor points with drinking water wells) were also sent as part of the grid database.

In the 3MRA modeling system, spatial features are related to each other using variables in the site layout data group. These variables include indices that identify the features that connect to or overlap with another type of feature (e.g., which watersheds overlap with a farm, which watersheds contain a human receptor) and the fraction of overlap. EPA used GIS programs to overlay the coverages (such as farms over watersheds, or receptor points over distance rings) to determine the spatial relationships between coverages (e.g., watershed occupied by human receptor or farm, waterbody used by farm, etc.).

5.1.7 Process Data for 3MRA

EPA used an Access database to process GIS and other data to create the 3MRA modeling system representative national data set input database. This processing database includes a series of SQL and Visual Basic programs to automatically QC-check and process the site-based, regional, and national data necessary to run the 3MRA modeling system. For the GIS data, the programs check the incoming data for completeness and consistency. Programs then convert the site-based data to the 3MRA modeling system format. This conversion includes

- Creating and attributing the waterbody network variables
- Connecting watersheds to waterbodies
- Providing data for soil- and land-use-derived variables
- Matching ecological receptors in each habitat to the appropriate home range bin
- Assigning indices and ensuring all indices are correct and complete
- For the grid database, converting real-world UTM coordinates to a set of metric *x,y* points centered about 0,0 (at the facility centroid).

The database processing performs any necessary calculations and/or data manipulations to produce the final variables and format required by the 3MRA modeling system.

5.2 Data Verification

Data verification includes the activities undertaken prior to and during the data collection effort to ensure that data of the correct type, amount, and quality (i.e., data that meet data quality objectives) are provided in the 3MRA modeling system representative national data set. These activities included

- A data collection plan that specified data sources and how data would be collected from those sources;

- Quality assurance (QA)/QC protocols that were specified as part of the data collection plan, including data entry checks, independent calculations to verify that data were processed correctly in all circumstances, and automated checks of critical parameters, formats, and processes; and

- Independent testing of the major site-based data elements.

EPA updated the data collection plan and it is now the documentation for the overall data collection effort (Volume II). Section 5.2.1 describes the QA/QC protocols that EPA planned and implemented during the 3MRA modeling system representative national data set data collection effort. Section 5.2.2 describes independent testing of the data.

5.2.1 Quality Assurance/Quality Control

Prior to data collection, EPA developed a basic QA/QC protocol for each data type and distributed it to all staff working on the data collection. In addition, EPA developed certain QC protocols common to many data types (see below). Any necessary deviations from these protocols during data collection were discussed with and approved by the team leader and the QA officer. The specific QA/QC protocols used during the data collection effort are discussed in detail by data type in Volume II, *Data Collection*. EPA conducted QA to ensure that an adequate QC methodology was in place and was correctly implemented and recorded.

The common QC protocols outlined in the data collection plan were followed during the data collection effort with minor changes. Common QC protocols for site-based data include

- Conducting a senior review and manually checking 100 percent of data entered from hardcopy sources.
- Recording the name of the staff member performing QC checks and the date QC checks were performed as part of the QC record.

- Maintaining files documenting QC activities. These files were used to track information such as data sources, data entry, and changes to data, and included copies of hardcopy data sources.

- Keeping metadata electronically for all electronic data sources

- Validating the data extraction and processing system for automated import of data from electronic sources and automated data processing in the electronic data sources before use through hand checks and calculations. After initial system validation, a sufficient fraction of the data (usually 5 to 10 percent) was manually checked to ensure that the data processing system was functioning properly. When possible, automated checks were also built into the system to detect data inconsistencies

- Validating the 3MRA modeling system input data processor system by manually checking a portion (usually 5 percent or more) of the processed data for each variable to ensure that the system is functioning properly. The system also included automated checks to identify inconsistencies

- Using the 3MRA modeling system Site Definition Processor (SDP) to check each database update for missing data and for consistency with the 3MRA modeling system input data specifications (i.e., the site specification file dictionary files).

For certain types of problems discovered during the QC of automated processing, it was possible to create automated checks that would be performed during the processing to catch any similar errors. For instance, these checks included searching for problems that violated model specifications, as well as performing other logical checks of the data. Table 5-4 contains a list of automated checks performed on the final model input database, all of which resulted from errors found during QC. Checks for certain errors, such as duplicate rows of data, checks of the indices, and other routine checking could uncover errors resulting from a number of different processing problems.

During the data collection process, EPA made minor changes to the protocols set forth in the data collection plan for individual data types. Many of the changes resulted from changes in the data collection methodology or further QC needs based on the initial checks of the data or data processing. Changes from the planned QC protocol to those actually employed are summarized by data type in the following sections.

5.2.1.1 Site/WMU locations. The data collection plan outlined manual and automated QC activities designed to obtain better locations for the Industrial Subtitle D facilities. EPA used automated matching of zip codes and addresses between the Industrial Subtitle D database and EPA's LRT database to identify spurious facility locations and duplicate locations, which EPA then screened manually to eliminate duplicates and mismatched data. EPA conducted manual checks of the matched sites to verify the automated matching process. When zip codes or addresses did not agree, EPA conducted manual verification. Although none of the planned QC protocols were changed, the need for additional review of site locations was identified during the initial watershed delineations of the sites. Some locations put WMUs in rivers or other waterbodies. Other sites ended up in areas of inappropriate land use (e.g., large surface impoundments in residential land use areas). As a result, each site/WMU location underwent a visual review and manual relocation as necessary to ensure reasonable location prior to GIS processing.

Table 5-4. Automated QC Performed on Model Input Database

Data Element	Automated QC
All data	Data type checked against data type in data dictionary (DIC) file
All data	Units checked against units in DIC file
All data	Central tendency value checked against maximum in the DIC file
All data	Central tendency value checked against minimum in the DIC file
All data	Maximum value checked against maximum in DIC file
All data	Minimum value checked against minimum in DIC file
All site-based data	Checked for duplicate rows in the site-based data table
All site-based data	All indices in the site-based data table checked to make sure that the largest index is equal to the central tendency of the Num variable to which the index corresponded based on DIC file
Grid tables and site centroid	Checked that the average *x* and *y* coordinates from the combined WBNRch and WSSub grid tables are less than 100 meters
Local watersheds	Checked the number of local watershed subareas to make sure they are correct based on WMU type
Ecological habitats	Checked to make sure that the number of habitats for different WMUs at the same facility are the same
Ecological habitats	Checked to make sure that the number of habitats in EcoRing 3 (which covers the entire site AOI) is equal to the value of NumHab and that the number of habitats in Rings 1 and 2 does not exceed the number in EcoRing 3

5.2.1.2 Waste Management Unit. The QC plan for site-based WMU data included 10 percent checks of automated data extraction with checks being conducted across all WMU types to ensure that all calculations in the database were checked and correct. Internal reviews consisted of senior engineer review of individual parameter values for realism and review of overall model system designs to ensure that parameter estimates within the model were internally consistent. External reviews of model facility designs and parameter estimates were also conducted to ensure that these were representative of typical industry practices. All of the QC protocols in the data collection plan were implemented during the data collection.

5.2.1.3 Watershed and Waterbody Layout. The data collection plan outlined a largely automated methodology for delineating watershed subbasins and waterbodies. However, as EPA developed programs and reviewed actual coverages, it became apparent that manual interaction would be required to delineate the watersheds and waterbodies accurately. For some of the sites, the 1-degree DEMs were not of sufficient quality to be used in the automated watershed delineation, so EPA developed a semi-manual methodology. Because of these difficulties in the delineation, EPA greatly expanded the focus of the QC protocols for these data. The QC protocols in the data collection plan primarily focused on checking the DEMs for problems, reviewing and verifying programming, and documenting and storing programs and metadata. EPA modified some of these protocols and added many new ones to accommodate the updated methodologies. Because the new method of delineating watersheds involved manual interaction with programs, the QC checks focused more on checking those manual interactions and less on checking the automated programs, since problems with the programs would typically be discovered during the manual interactions. Many new QC protocols were added during the data collection effort because the data were more complex and variable than were expected. QC protocols were added to check the GIS processing and the database processing that processed the raw GIS data to fill the variables required by the 3MRA modeling system. The GIS QC protocols required a GIS staff member, other than the one doing the delineations, to visually examine each site and run an interactive macro to verify that everything was done correctly. The database processing QC protocols required that all waterbody layout variables be checked for each site following processing. The watershed layout variables were more straightforward and consistent, and, therefore, only about 10 percent of the data were checked following processing.

5.2.1.4 Soil Data. In general, EPA used the common QC protocols for the soil data. EPA checked 100 percent of manually entered data following review by a senior staff member, and performed manual checks of automated calculations on processed data. To check data compilation from STATSGO and CONUS[4] tables within the soil database, EPA checked all data processing operations using hand calculations for soil map units randomly selected from the database. When special processing rules applied to certain data categories, at least one of each type was chosen and checked.

5.2.1.5 Hydrogeologic Environment. The site-based hydrogeologic assignment and aquifer temperature were transferred in electronic format from the previous EPA models. EPA spot-checked these data after transfer to ensure that they were processed into the 3MRA modeling system format without error. For the GIS-derived spatial layout variables generated automatically, a percentage of the final values were checked manually to ensure that the GIS and database programs were processing data accurately and consistently. One exception to the spatial layout variables was the aquifer flow direction, assumed to flow downhill toward waterbodies, which was checked manually for 100 percent of the data. The scale of the local watershed at many of the sites made the use of the DEMs somewhat unreliable in determining the downhill flow direction.

5.2.1.6 Human Receptors. The QC plans for human receptor data included developing automated programs to check individual block/group population values in the Census coverages for population outliers or mistakes in the data. The programs compare all Census population-type values against urban/rural status and polygon size to check the reasonableness of the data. EPA also conducted random checks against different data sources (e.g., Census CD, a commercial CD of Census data) and performed manual calculations for a few sites to validate the programs used in the automated processing. EPA followed all of the QC protocols in the data collection plan and developed additional protocols based on problems discovered during processing. Because of the variability of Census data overlaid with land use data, visual inspection of farm placement at each site was the safest way to ensure correct farm placement at every site that had farmland use and farmers. To further check the population numbers in the final representative national data set input database, EPA compared population totals from the database (summed across all receptor points, receptor types, and age cohorts), to population totals (within the 2 km AOI) obtained from GIS Census coverages. Additionally, EPA created an Excel spreadsheet that repeated the calculations of the human farm receptor GIS program.

[4] Data set for the conterminous United States.

This spreadsheet expedited QC by enabling staff to enter raw data into the spreadsheet and calculate expected results for comparison with the output of the GIS programs. Prior to use, the methodology and logic for the formulas in the spreadsheet were reviewed and checked by senior staff.

5.2.1.7 Ecological Receptors. The QC performed for ecological receptor placement focused on the delineation and GIS processing of the habitats and ranges. Because habitat delineation consisted largely of subjective evaluations and hand delineation of habitat boundaries, two designated senior ecologists performed all habitat delineations. Limiting the delineators to two individuals helped limit the degree of variation in interpretation of spatial data. Both delineators adhered strictly to the crosswalk tables developed for delineation. Both delineators kept records of any delineations that involved unusual circumstances or conflicting issues. These notes were reviewed by both delineators to maintain consistency. Following delineation of the habitats at all sites, the delineators performed checks on 75 percent of each other's delineations. Before final data processing, automatic QC programs were used to query the GIS data tables to ensure that each habitat had no more than four home range bins and that there were no grid cells in the home range that were not also in the habitat containing the home range.

5.2.1.8 Grid Database. To ensure the accurate transfer of grid information to the 3MRA modeling system grid database, EPA performed the following QC activities during data collection:

- Automatic regeneration of the grid template for a site prior to GIS postprocessing of spatial data to ensure proper correspondence with the facility centroid;
- Generation and visual review of thumbnail images of all spatial data for every site to ensure accurate registration and collocation of all data layers; and
- Visual checks of a subset of sites in the 3MRA modeling system grid database against the original GIS coverages to ensure accurate data processing and transfer.

5.2.2 Independent Data Testing

EPA conducted independent data testing to assess the accuracy of the site-based data in the 3MRA modeling system representative national data set. In addition to independent testing of the data collection methodologies and the 3MRA modeling system representative national data set, QA/QC procedures performed during data collection were also reviewed. The independent data testing addressed site-based data collection in the following areas:

- Spatial layout
- WMU data
- Watershed and waterbody layout
- Soil data
- Human receptor data
- Ecological habitats and receptors.

Specific items were checked for each data area. Testers carried out the following checks:

- Reviewed and compared data collection and model documentation for consistency.
- Reviewed QA/QC history for completeness. Data preparers were contacted to review information on QA/QC history to identify any potential gaps in the QA/QC of the data.
- Reviewed selected data methodologies, programs, and results. Techniques used depended on data type, but often included comparing the input data to the original source, independently re-creating input data from raw data and comparing them to the model input data, and reviewing and checking QC procedures and records.

Testing confirmed that most of the data were accurate, but some errors were identified during the testing and later corrected in an updated 3MRA modeling system representative national data set (in September 2000). Table 5-5 summarizes the results of independent testing of site-based data. For more detail on the independent data testing, see *Independent Data Testing for the Hazardous Waste Identification Rule—Draft* (RTI, 2000a).

Table 5-5. Summary of Results of Independent Testing of Site-Based Data

Data Element	Data Testing
Spatial framework	Coordinates in the grid database were plotted using ArcView for 10 sites and compared to the jpeg images provided in the data docket. Coordinates from all tables— AquWell, Farm, Habitat, HumRcp, WBNRch, and WSSub— were plotted.
Landfills	Landfill data were extracted from the Industrial Subtitle D database, used to derive the landfill depth, and used in a statistical regression to calculate replacement capacity values.
Waste piles	Waste pile data were extracted from the Industrial Subtitle D database and used to derive replacement values for missing waste quantities. A statistical regression was done using facilities that had waste quantity data and met certain height criteria.
Land application units	LAU data were extracted from the Industrial Subtitle D database and used in a statistical regression of facilities that reported waste quantity data and met the waste application rate constraints.
Surface impoundments	Extracted Industrial Subtitle D surface impoundment data were used to derive the depth of the surface impoundment. To calculate replacement values, two statistical regressions were done: one using facilities that reported waste quantity data and one using facilities that reported capacity data and also met the unit constraints.
Watershed and waterbody layout	Watersheds were delineated manually on USGS 1:24,000 topographic maps using the same guidelines used in the data collection documentation.

(continued)

Table 5-5. (continued)

Data Element	Data Testing
Soil data	A review was conducted of QA/QC history associated with use of soil parameters for various projects.
Human receptor data	Review of receptor placement and attribution was conducted using the Census data and methodologies described in the documentation.
Ecological receptors and habitats	Habitats were manually delineated and, according to the same methodology, described in the data collection documentation and compared to the 3MRA modeling system representative national data set delineations. Receptors were manually assigned for each of these habitats and compared with the list of receptors in the 3MRA modeling system database.

Table 5-6. Comparison of Industrial D and SIS Survey Data for Common Sites

Facility Name,[1] Address, SIC[2]	1985 Industrial Subtitle D Survey Data	1999 SIS Survey Data
Crown Paper Co. (Crown Zellerbach Corp.) St. Francisville, LA 70775 2611 (Pulp mills)	0832903: 3 impoundments, 121 acres (1,355,075 metric tons) total; 1985 waste volume is 40,824 metric tons per year	3062: 2 impoundments. ■ ASB (Aeration Stabilization Basin; aerated biological treatment), 43 acres (1,158,100 metric tons); 42,860,356 metric tons per year ■ WSI (West Sludge Impoundment; sedimentation and anaerobic biological treatment), 32 acres (419,323 metric tons), 172,824 metric tons per year
Cenex Harvest States Cooperatives; Laurel Refinery (Farmers Union Central Exch) Laurel, MT 59044 2911 (Petroleum refining)	1230111: 3 impoundments, 4.13 acres (33,877 metric tons) total; 1985 waste volume is 557,928 metric tons per year	2418: 2 impoundments (aggressive aerated biological treatment) ■ North Aerated BioPond, 1.50 acres (11,089 metric tons); 1,106,074 metric tons per year ■ South Aerated BioPond, 1.49 acres (9,476 metric tons); 1,077,845 metric tons per year

[1] Industrial D Survey name in parentheses
[2] SIC = Standard Industrial Classification Code

5.3 Data Validation

EPA validated the accuracy of the site-based data in the representative national data set by comparing those data with data and model results for two of the sites where more recent data were independently collected during EPA's 1999 Surface Impoundment Study (SIS) Survey (U.S. EPA, 2001). The following section describes this validation activity in detail.

5.3.1 Surface Impoundment Study Data

The 1985 Industrial Subtitle D Survey and the 1999 SIS Survey were statistically designed to characterize the same universe: facilities managing nonhazardous industrial wastes in onsite WMUs. The two surveys overlap in their coverage of surface impoundments, but differ in the level of detail of information collected. The Industrial Subtitle D Survey characterized number of surface impoundments and the total area, capacity, and waste volumes of those surface impoundments for more than 1,800 facilities with surface impoundments across the country. The SIS Survey collected very detailed information about waste characteristics, design and operating conditions, and the surrounding environment (including the location of residences and drinking water wells) for 220 facilities. Because the SIS survey was statistically designed, it should be valuable in validating and corroborating the older Industrial Subtitle D data, as well as the supplementary data collected for the 3MRA modeling system representative national data set.

Although more than 60 of the SIS facilities overlap with the Industrial Subtitle D survey sites, only two of the 201 Industrial D sites randomly selected for the 3MRA modeling system representative national data set are also in the SIS data set. These sites are shown in Table 5-6, which also compares the waste management information supplied with each survey. For each site, three impoundments were reported in the 1985 survey, compared with two in the 1999 SIS. With respect to the impoundment size, area and total capacity compare reasonably well, as do annual waste volumes for the Montana refinery (1230111). However, the Industrial D 1985 annual waste volume for the Louisiana pulp mill (0832903) appears to be very low when compared with the SIS data.

5.3.2 Development of SIS Data Sets

The two overlapping sites provided an opportunity to validate the Industrial Subtitle D survey data used for the representative national data set against independent 3MRA modeling system data sets developed using the more recent and more detailed SIS information available for these sites.

Table 5-7 compares the data sources and collection methodologies used by the 3MRA modeling system sample national data collection effort and SIS. The primary differences are that the SIS contains more detailed information of surface impoundment operation characteristics and waste properties and the SIS survey responses provide more accurate human receptor locations by marking individual residences on topographic maps. The 1999 SIS data are also more recent when compared to the data sources used for the 3MRA modeling system representative national data set, which range from the 1985 Industrial Subtitle D Survey to the 1990 U.S. Census.

Table 5-7. Comparison of Site-based Data Sources: SIS and 3MRA Modeling System Representative National Data Set

Data Category	3MRA Modeling System Data Source / Methodology	SIS Data Source / Methodology
Facility location	Zip code centroids (from Industrial D survey), improved using EPA Envirofacts preferred locations, address matching, and visual placement using GIS coverages	Respondent-marked topographic maps and diagrams showing impoundment locations and areal dimensions
WMU dimensions	1985 Industrial D survey data: number of units, total area, total capacity, total 1985 waste volume (totals of all impoundments)	1999 SIS Survey data (by in-scope[1] impoundment): area, capacity, average flow rates (wastewater and sludge), diagrams (depth)
WMU operating characteristics	Imputed from Industrial D dimensions using published and derived engineering relationships; aeration assumed for all units	1999 SIS Survey detailed data on impoundment type, function, aerators, mixers
Waste properties	Assumed using national distributions	1999 SIS Survey data on contaminant concentrations and other waste properties (pH, temperature, BOD, COD, TSS,[2] etc.)
Human receptor locations	1990 Census block/ring centroids; wells placed based on block group data	Respondent-marked topographic maps showing residences and wells within 2 km
Hydrogeologic environment	Assigned based on zip code and national atlas of conditions	Assigned based on SIS survey subsurface information; national GIS coverages of aquifers, soils, and surficial geology

[1] Impoundments that have contaminants of concern present or a pH below 2 or above 11.
[2] BOD = biological oxygen demand; COD = chemical oxygen demand; TSS = total suspended solids

5.3.3 Preliminary Comparisons

Although the SIS-based 3MRA modeling system data sets are not complete, some comparisons may be made using the underlying data. Table 5-8 compares WMU and waste property data from the 3MRA modeling system representative national data set and SIS sources. Significant differences include the following:

- Impoundment areas and depths are similar, but the Industrial Subtitle D waste flows are significantly lower than the SIS data, especially for site 0832903/3062

- The degree of aeration assumed in the 3MRA modeling system representative national data set is significantly higher than reported in the SIS survey, especially for site 1230111/2418

- The SIS data contain actual site-based waste property data compared to the national distributions or fixed values assumed for the 3MRA modeling system representative national data set.

Table 5-8. Comparison of Waste Management Data—3MRA Modeling System Representative National Data Set versus SIS

	Laurel Refinery (MT)			Crown Paper (LA)	
	3MRA	SIS		3MRA	SIS
Variable	1230111	2418North	2418South	832903	3062ASB
Number of surface impoundments	3	2	2	3	2
Dimensions					
Area (m^2)	5,571	6,056	6,043	163,229	174,015
Depth (m)	2.77	1.52	1.63	2.77	4.55
Operating Volumes					
Q_wmu (m^3/s)	0.0059	0.035	0.034	0.0004	1.36
d_setp[1]	0.41	0.26	0.14	0.57	0.31
Aeration Characteristics					
d_imp (cm)	61	14.2	14.2	61	49.5
w_imp (rad/s)	126	377	377	126	126
O_2eff	distribution	NA[4]	NA[4]	distribution	0.85
n_imp[2]	8	4	2	55	32
Power (hp)	556	40	20	5,000	2,400
F_aer[3]	0.52	0.034	0.017	0.17	0.072
Waste Properties					
pH	distribution	8.09	7.64	distribution	7.6
Temp (C)	8.64	35.3	18.2	19.4	41
BOD (g/cc)	distribution	0.000006	0.00001	distribution	0.000243
Density	0.998	1.001	1.001	0.998	NA
TSS (g/cc)	distribution	0.000084	0.00001	distribution	0.000133
foc	distribution	0.55	0.55	distribution	0.74

[1] fraction of impoundment occupied by sediments
[2] number of impellers
[3] fraction aerated
[4] NA = not available

Measurement of the impacts of these differences on model results will be possible once the SIS data sets are ready for use in the 3MRA modeling system. Depending on the results of these comparative runs, it may be desirable to use the SIS data set to develop more reasonable site-based data from the existing Industrial D data. These improvements could include more appropriate operating characteristics and waste properties for the specific industries represented by the Industrial Subtitle D and SIS surveys.

Figures 5-4 and 5-5 compare the site layouts generated for the 3MRA modeling system with the SIS data. Preliminary observations include the following:

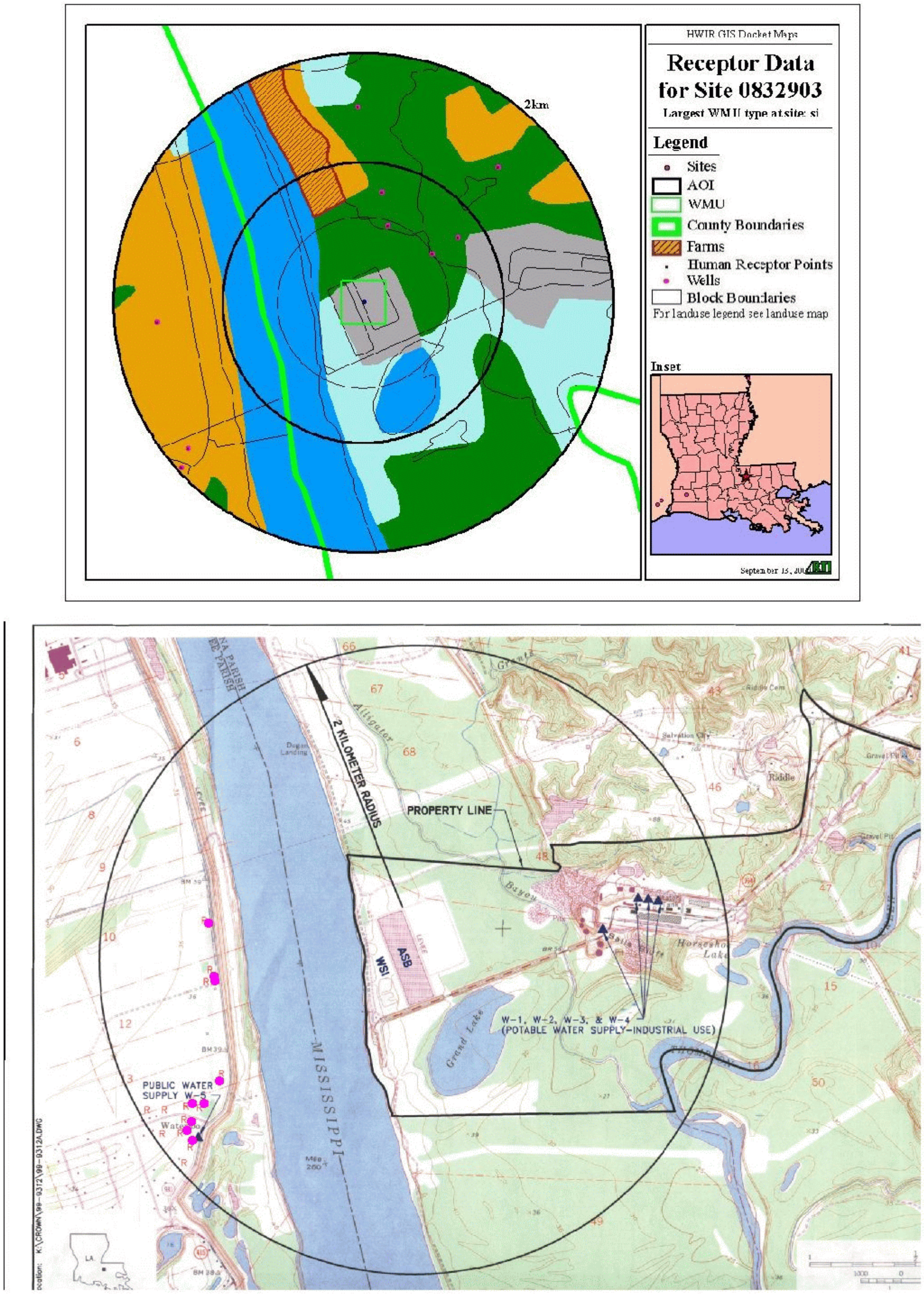

Figure 5-4. Crown Paper, St. Francisville, LA: Comparison of 3MRA modeling system (0832903, upper) and SIS (3062, lower) site layouts. (Red dots indicate receptor locations.)

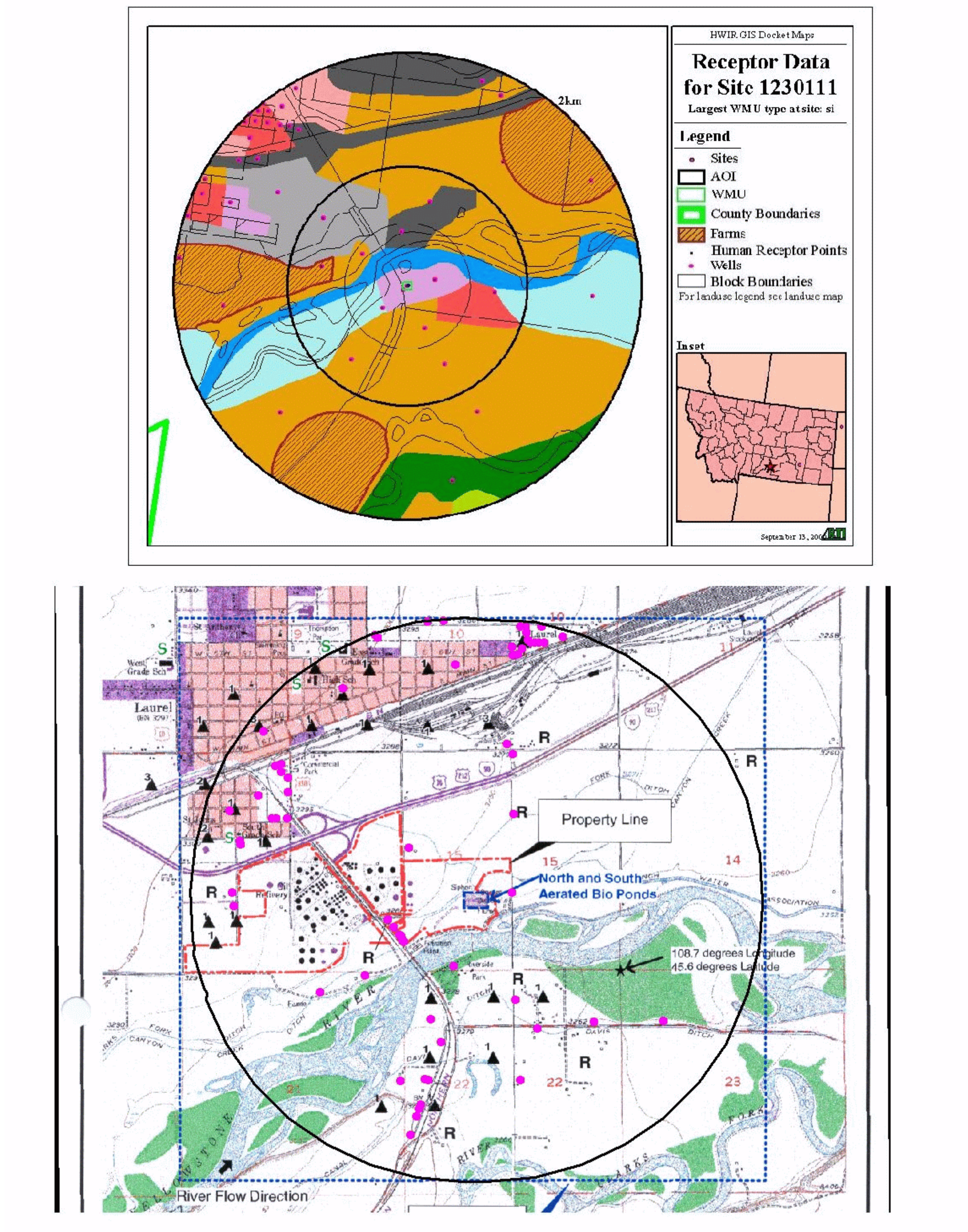

Figure 5-5. Cenex Refinery, Laurel, MT: Comparison of 3MRA modeling system (1230111, upper) and SIS (2418, lower) site layouts. (Red dots indicate receptor locations.)

- The location for the Louisiana site was very accurate, but the 3MRA modeling system location for Montana was across the river from the actual location.

- Receptors were similar for the Montana site, but the SIS data provide more highly resolved placement of residences within Census blocks than the block/ring centroids used in the 3MRA modeling system.

- For the Louisiana site, receptors placed at block/ring centroids on the east side of the Mississippi river were not present in the SIS data (i.e., residents in these blocks are actually outside of the 2 km-radius AOI).

As with WMU characteristics, the SIS data sets will allow for these differences to be quantified with respect to the 3MRA modeling system's risk estimates. To separate impacts of the WMU differences from the impacts of site layout differences, hybrid data sets can be prepared by using the same WMU characteristics for the different site layouts or by using different WMU data at each site layout.

Aquifer assignments also differed between the 3MRA modeling system representative national data set and the SIS data set. Based on SIS survey subsurface data and GIS coverages of surficial geology, soil, and aquifers, both sites were assigned to a river alluvium hydrogeologic environment (GWClass 6), versus the 3MRA modeling system representative national data set assignments of sand and gravel (GWClass 4) for the Louisiana site and alluvial basins (GWClass 5) for the Montana site. This difference can be attributed to the fact that the 3MRA modeling system representative national data set assignments (which date to before 1995) were based on zip code centroid locations and a fairly low-resolution national atlas of principal aquifers. As a result, the narrow alluvial aquifer environments (GWClass 6 and 7), which should be fairly common given that industrial facilities are often located along rivers, are not represented in the 201-site representative national data set.

[This page intentionally left blank]