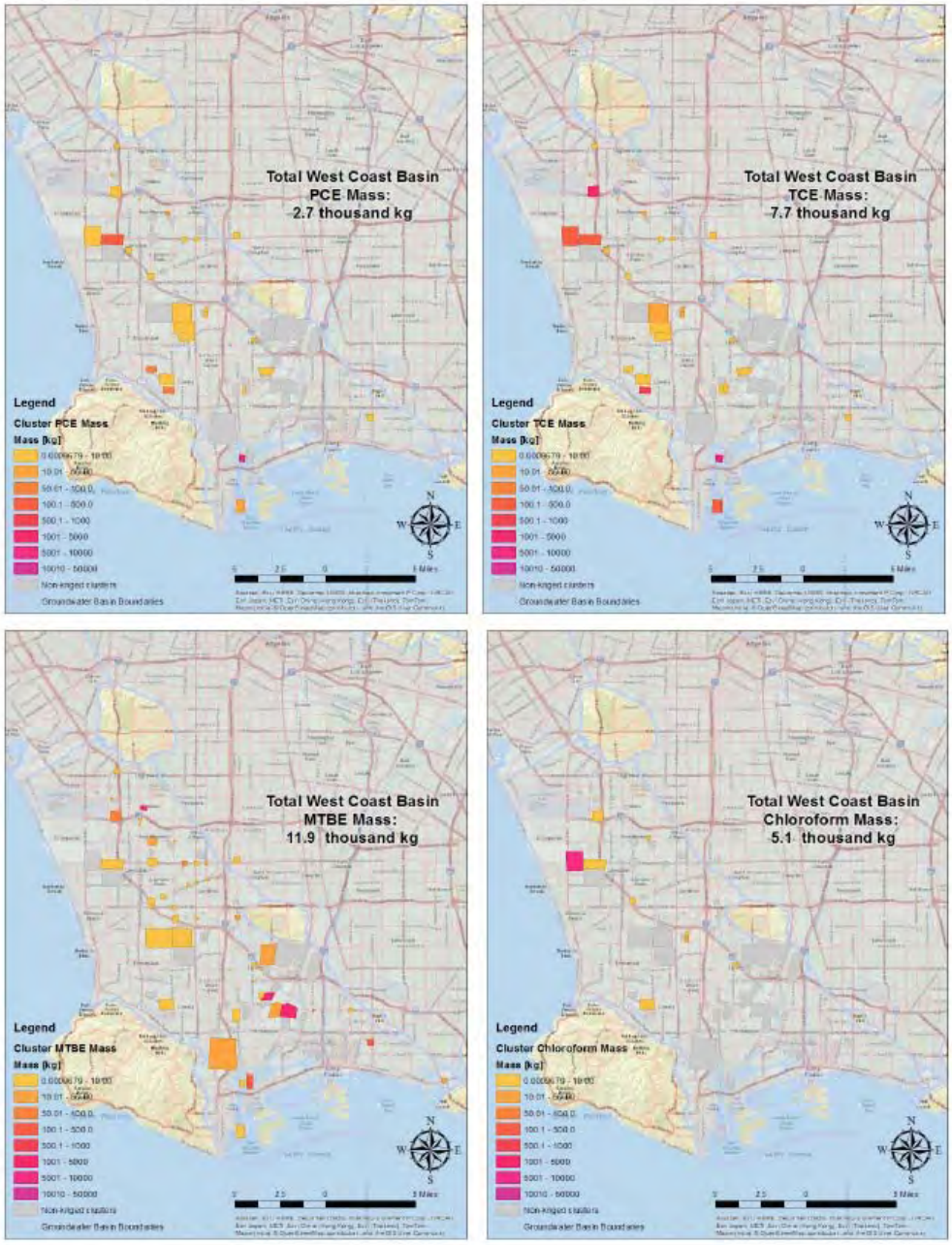
Estimating the Mass of VOC Pollutants in the West Coast Los Angeles Groundwater Basin



Client: Yue Rong, Los Angeles Regional Water Quality Control Board
Advisors: Dr. Travis Longcore and Travis Brooks
Group Members: Soseh Baboumian, Maya Bruguera, Albert Lu, Emilio Ronquillo, Evan Tang, Ziqi Wan and Shihao Zhu

Table of Contents

1. EXECUTIVE SUMMARY

The Los Angeles Regional Water Quality Control Board (LARWQCB) is interested in exploring the possibility of regulating groundwater pollution using estimates of contaminant mass, in addition concentration-based regulations. As a step toward establishing mass-based regulation, the LARWQCB tasked a UCLA senior Environmental Science Practicum team with building a pollutant mass-estimate model for the West Coast section of the Los Angeles Groundwater Basin. The project's goal was to calculate the mass of volatile organic compounds (VOCs), some of which have been associated with adverse health effects ranging from neurological issues to cancer.

The model construction was based on well-sample concentration and geographic coordinate data from GeoTracker GAMA database and the development of a mass estimate equation, known as **Equation 1**: **Mass of VOC = C_{XY} * A * h* α_S* β_Z .** The first two terms, which represent the contaminant concentration (C_{XY}) and predicted area of concentration for a subset of the well-sample points throughout the basin, were obtained through the geostatistical interpolation technique called kriging. The next two terms, aquifer height (h) and aquifer storage capacity (α_S), came from georeferencing the well-sample data onto a grid created by a collaboration of the United States Geological Survey (USGS) and Water Replenishment District (WRD). **Equation 1's** final variable, called the z-coefficient (β_Z), accounts for the change of pollutant concentration as a function of depth.

By calculating VOC mass at numerous clusters of wells throughout the basin and summing them together for each individual pollutant, the masses of four prominent contaminants in the region were estimated: perchloroethylene (PCE), trichloroethylene (TCE), chloroform and methyl tertiary butyl ether (MTBE). The total masses (in kilograms) of PCE, TCE, chloroform and MTBE were, respectively, as follows: 2,699; 7,657; 5,102; and 11,872. The accuracy of the mass estimates require further validation by comparison to other model estimates of plume mass and empirical case study data; but, the heavily concentrated distribution of each VOC mass throughout the basin demonstrated the capability of the model to help regulators identify hot spots of pollution. For instance, three or fewer clusters of well points accounted for at least 85 percent of the predicted contamination for any given pollutant throughout the whole basin.

A key product of the project was a Python code and tool that together automate nearly the entire mass calculation process. These two items combine to make conversion between concentration and mass as straightforward as downloading data in the appropriate format from GeoTracker and running those inputs through the code.

While the model only accounts for basins with soil and aquifer properties that are similar to those of the West Coast Basin, it appears to be generalizable to at least the adjacent Central Coast Basin within Los Angeles. Application to other regions throughout California would require aquifer storage data similar to the geospatial data available from existing USGS and WRD work products. Application to areas across the nation would also necessitate having extensive well-monitoring information similar to the data taken from GeoTracker.

This type of model allows water agencies to take a step forward in achieving mass-based groundwater regulation. With this new approach, stakeholders have additional insight into the condition of the regional groundwater and can use model data to inform the establishment of monitoring and remediation priorities. More informed regulation will lead to more effective and efficient protection of a major resource in a place and time where the availability of local water supplies is of major concern.

2. INTRODUCTION

Groundwater accounts for more than 40 percent of California's drinking water, with dependence rising to as much as 60 percent during drought years (Beckman *et al.*, 2001). Remediation of contaminated groundwater therefore serves as a major issue for water regulatory agencies and policymakers. While pollutant source control programs remain important objectives, effective and efficient remediation is crucial because groundwater has proven difficult to restore once contaminated (Asanoa and Cotruvo, 2004).

The Los Angeles Regional Water Quality Control Board (LARWQCB) seeks to improve groundwater policies by employing mass-based regulation of groundwater pollution, such as to better inform planning for remediation and resource management.

This project assists the LARWQCB in working toward its goals by providing a methodology that uses sampling well and aquifer data to estimate the total mass of any volatile organic compound (VOC) within the groundwaters of Los Angeles' West Coast Basin.

3. BACKGROUND

3.1 RECENT REGULATORY HISTORY OF GROUNDWATER

No comprehensive groundwater monitoring program existed in California as recently as 15 years ago (Beckman *et al.*, 2001). Six state and three federal agencies implemented monitoring programs prior to 1999 (State Water Resources Control Board, 2003), but major inconsistencies and data gaps stemmed from how these entities had different goals and objectives. The National Resources Defense Council assessed the availability of information pertaining to California groundwater quality as frequently unreliable and "extremely limited" (Beckman *et al.*, 2001).

A foundation for more cohesive California groundwater monitoring began taking shape in 1999. The Supplemental Report of the 1999 Budget Act included a section calling for a comprehensive program establishing "baseline ambient surface and groundwater quality monitoring" in the state (Legislative Analyst's Office, 1999).

California's State Water Resources Control Board (SWRCB) responded to the report by establishing the Groundwater Ambient Monitoring and Assessment Program (GAMA) in 2000 (Borkovich, 2012). Assembly Bill 599 — the Groundwater Quality Monitoring Act of 2001 — required expansion of the nascent GAMA program to incorporate and enhance existing efforts to characterize California's surface and groundwater quality (Borkovich, 2012).

The wealth of data gathered by a variety of different agencies was to be shared, integrated and evaluated by an interagency task force (SWRCB, 2003). Special attention was to given to identifying trends in groundwater issues and giving the public a tool to assist in the establishment of regulatory practices and priorities on a basin and sub-basin level (SWRCB, 2003). In addition to compiling existing groundwater monitoring data and enhancing hydrogeologic knowledge, members of GAMA also aimed to contribute additional public supply well sampling for dating groundwater and detecting low-level VOCs (SWRCB, 2003).

Today, 95 percent of California's groundwater use is derived from basins evaluated by the GAMA program. See **Appendix A** for descriptions of the GAMA program's four projects (Borkovich, 2012), one of which is the publicly accessible GeoTracker online database that

serves as the primary input for this project's pollutant mass estimate model.

3.2 Pollutant Mass Estimation: A Future Direction for Groundwater Regulation

Groundwater pollution has historically been regulated on the basis of concentration, as seen in the data provided by GeoTracker GAMA. But the LARWQCB, whose responsibilities include regulating groundwater, seeks a shift to a mass-based estimate for pollution.

Precedent for mass-based pollutant estimation exists, perhaps most notably in the form of the Total Maximum Daily Load (TMDL). A component of the Clean Water Act and a national standard, the TMDL is a measure that describes how much pollutant a surface water body can experience before losing function (Cahn and Hartz, 2014).

Concentration studies carry a considerable deal of uncertainty. Groundwater volume changes less predictably over space and time when compared to surface waters, and such changes are more pronounced for concentration studies (Cahn and Hartz, 2014). Two bodies of water with different volumes may have the same concentration for a pollutant, even though one receives more contaminant inputs than the other. Though more expensive than concentration-based evaluation (Cahn and Hartz, 2014), mass-based estimates could allow for regulatory agencies to better detect the degree of pollutant required for concentrations to surpass environmental health standards (Burke *et al.*, 2013).

3.3 The Environmental and Public Health Threat of VOCs

The LARWQCB hopes to better characterize the amount and distribution of pollutants within Los Angeles' West Coast Groundwater Basin. VOCs pose a particular problem to the state, as they comprise the "vast majority" of organic compounds contaminating California's groundwater reserves (Beckman *et al.* (2001). Gasoline, paints, paint thinners and solvents for cleaning and degreasing provide major inputs to the subsurface environment's total VOC mass (Beckman *et al.*, 2001).-According to Moran *et al.* (2007), most VOCs have short half-lives and degrade rapidly, but others threaten water quality long after release because some may experience little biodegradation and thus persist in the environment for decades.

Health measures for VOCs are similar to those of other pollutants. Carried out by the state and federal Environmental Protection Agency (EPA) under the Safe Water Drinking Act, maximum contaminant levels (MCLs) define the highest concentration permissible for a contaminant within water used by a public system (United States EPA, 2013b). A pollutant's MCL is decided upon after the EPA reviews health studies and sets a maximum contaminant level goal (MCGL), which determines the maximum concentration of a contaminant in drinking water at which no adverse health effects are observed. MCGLs are set conservatively and without consideration for detection limits, allowing for a margin of safety that protects human health. MCGLs therefore may not be realistic for many water systems (United States EPA, 2013b). Health benchmarks are generally conservative in projections, such as to better ensure people's safety.

The United States Geological Survey (USGS) National Water-Quality Assessment (NAWQA) program found that 8 of 55 monitored VOCs had concentrations surpassing human health benchmarks in 1 percent of more than 3000 drinking-water supply well samples taken from throughout the nation between 1985 and 2002 (USGS, 2014). The NAWQA characterized ambient groundwater and groundwater reaching drinking-water supply wells, as opposed to

treated drinking water delivered to consumers.

Sample concentrations, with one assigned to each studied aquifer, were compared to a VOC's MCL when possible, while samples for VOCs lacking MCLs were evaluated against health-based screening levels (HBSLs) derived from peer-reviewed toxicity studies and EPA Office of Water methodologies (Hamilton *et al.*, 2006). Not enough toxicity information was available to assign an HBSL to every VOC lacking an MCL.

With VOC concentrations mostly well beneath health-based limits, the USGS report concluded that even if water with the observed concentrations would be ingested over the course of a lifetime, adverse effects were, in general, unlikely to occur (USGS, 2014).

Though low concentrations found by the NAWQA study suggest that VOC exposure may be rare, some VOCs nonetheless have associations with cancer and adverse health effects on the nervous, circulatory, reproductive, immune, cardiovascular, and respiratory systems (Moran *et al.*, 2007).

3.4 VOCs OF INTEREST IN THE WEST BASIN

GAMA's Priority Basin projects, which also use drinking-water quality benchmarks to give context to groundwater contamination, found three VOC classification groups at moderate to high levels of concentrations within the 860-square mile Coastal Los Angeles Basin, which includes our West Coast Basin study area: chlorinated solvents, trihalomethanes (THMs) and gasoline additives (Belitz and Fram, 2012). Various physical and chemical traits of VOCs account for differing migration levels in the subsurface environment.

Chlorinated solvents tend to be dense, non-aqueous phase liquids (DNAPLs) and so they often extend beyond shallow groundwater and persist for long periods of time underground when they accumulate in plumes above rock and soil layers of low permeability (Interstate Technology and Regulatory Council, 2002). Such solvents are hydrophobic but soluble enough to be effectively transported through water. With high mobility to go with slow biodegradation rates, chlorinated solvents are found at high concentrations in 4 percent of the Coastal Los Angeles Basin's primary aquifer system and at moderate concentrations within 11 percent of the system; perchloroethylene (PCE) and trichloroethylene (TCE) are among the five solvents most frequently detected at high levels (Belitz and Fram, 2012).

Most solvents detected in Los Angeles today are deemed "legacy contaminants" that were pushed by recharge waters over the course of decades to reach the layers of soil from which water is pumped (Belitz and Fram, 2012). That these VOCs can impact water quality long after release emphasizes the importance of persistent monitoring efforts.

THMs rank as the second-most highly detected VOC group in the Coastal Los Angeles Basin, detected at moderate concentrations in about 2 percent of the primary aquifer system (Belitz and Fram, 2012). These compounds migrate long distances into the ground, especially in areas with high dissolved oxygen and low organic carbon content (USGS, 2014).

THMs typically form as byproducts of water disinfection, resulting from reactions between natural compounds within the water and chlorine and chloramine (United States EPA, 2013a). Infiltration of irrigation water and leakage of distribution systems facilitate movement into subsurface waters (Belitz and Fram, 2012).

Gasoline additives come in a variety of forms, such as oxygenates and hydrocarbons, so their mobilities demonstrate marked variability (USGS, 2014). The USGS found gasoline additives, a group including methyl tertiary butyl ether (MTBE) and aromatic compounds like benzene, toluene and xylene, at moderate concentrations in less than 1 percent of the Los

Angeles basin aquifer system (Belitz and Fram, 2012).

Oxygenates like MTBE tend to have higher water solubility and lower natural degradation rates than its hydrocarbon gasoline additive counterparts, and do not tend to get caught in soil, leading to considerable transport in the soil's unsaturated zone (USGS, 2014).

Gasoline oxygenates' presence in the environment stems from their use as additives to allow for cleaner fuel burning, and thus reduced toxic emissions from vehicle tailpipes (Lidderdale, 2000). Meanwhile, hydrocarbons, although the most widely used and produced VOCs, are seldom detected because of high biodegradation and volatilization rates, along with a tendency to have subsurface movement stifled by sorption to organic carbon (USGS, 2014).

PCE and chloroform, two of the three VOCs most frequently detected by the NAWQA program (USGS, 2014), were found by the Agency for Toxic Substances and Disease Registry (ATSDR) to cause liver and kidney tumors in mice, and are "reasonably anticipated" to be human carcinogens (ATSDR, 2012). The ATSDR also views TCE exposure as a potential risk factor for liver, kidney or lung cancer (2012). MTBE,though not considered a human carcinogen, has been shown to damage rodental nervous systems, livers and kidneys (ATSDR, 2012).

3.5 Study Area: West Coast Basin of Los Angeles

The Water Replenishment District (WRD), which manages "two of the most utilized groundwater basins in California" — the Central and West Coast Basins — in serving southern Los Angeles County, identifies urban sources as the dominant contributors to groundwater pollution in our study area (Matsumoto, 2009). The likes of landfills, gas stations, refineries and chemical processing facilities therefore comprise the bulk of contaminant sources in the West Basin (Matsumoto, 2009). VOCs are driven further into the soil from their source through gravity and, in the West Coast Basin, water recharge from Central Basin flow and the injection of imported and reclaimed water into seawater intrusion barriers. (Belitz and Fram, 2012).

The extent to which groundwater can be contaminated depends on numerous characteristics of the soil and rock through which it flows. Implications of some soil factors on the degree of groundwater contamination possible are summarized in **Table 1**.

Table 1. Soil Characteristics' Implications for Groundwater Contamination

Soil Trait	Relationship to Contamination
Texture	Finer soils trap contaminants better than coarser soils.
Porosity	More porous soils allow for further contamination.
Specific Yield	Higher specific yield leads to a higher degree of contamination, as more pollutant-carrying water can pass through.
Filtration	Increased filtration capability reduces contamination.

Within the West Coast Basin, the USGS has classified the aquifers into four layers. Each layer is a system separated from one another by an aquitard, or a low permeability layer. The upper aquifer Lakewood system is largely comprised of unconsolidated sand and gravel deposits, so hydraulic conductivity is generally high (Crawford *et al.*, 2003). The upper aquifer system consequently sees most of its contamination occur within the first two layers, going as deep as

200 feet below the ground surface, to the extent of the Gage Aquifer.

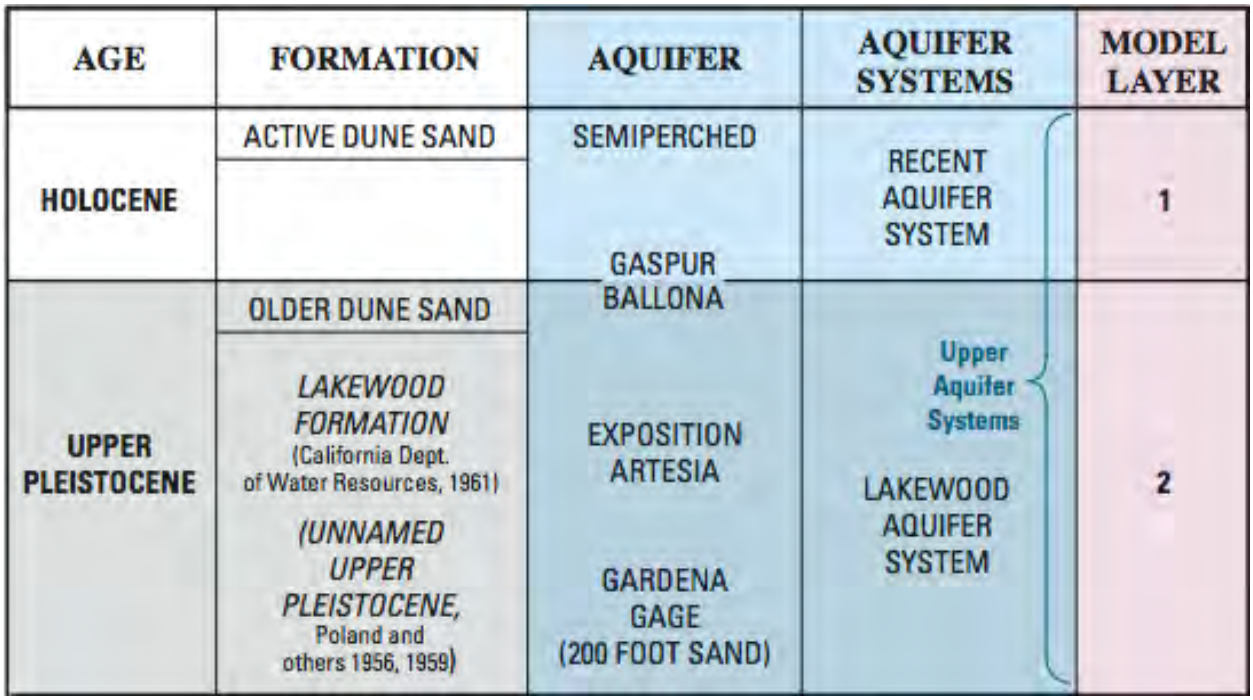


Figure 1. A diagram showing Layer 1 and 2, the West Basin's two upper aquifer systems. Formations of the two systems are listed vertically starting from shallow to deep. Water can move easily in these layers, making groundwater contamination possible as far as 200 feet below the surface.

The lower aquifer systems are the Upper San Pedro and Lower San Pedro, which respectively extend about 400 and 1200 feet below sea level. The Lower San Pedro aquifer system may be a marker for where groundwater contamination is possible. The underlying Pico Formation is non-water bearing and its contact forms the boundary of the groundwater basin (Crawford *et al.*, 2003).

Most aquifers in the West Coast Basin are confined (Crawford *et al.*, 2003). The first confining bed is a layer of low-permeability clays called the Bellflower aquiclude, which may inhibit complete downward vertical migration of contaminated plumes from semi-perched, or unconfined, aquifers. TCE migration, however, is observed where there are windows of permeable deposits in the confining layer (Lyles, 1998). The Bellflower Aquiclude does not restrict groundwater movement between aquifer strata (United States EPA, 2012).

3.6 Foundations from Past Research

Burke *et al.* (2013) used concentration as well as longitudinal and latitudinal coordinate data from the GeoTracker GAMA database to quantify the mass of a single contaminant, PCE, within a test site and the entire West Basin.

A major component of the project was the use of an interpolation technique called kriging to generate a prediction surface of concentration values between and around data points, which in this case are the reported concentration values of a contaminant at different groundwater sampling well sites. The model developed in this paper would make use of the same general methodology to generate concentration prediction surfaces, but apply the concept to multiple

contaminants.

Another important aspect of the Burke *et al.* model incorporated into this model was the creation and summation of a series of smaller prediction surfaces of smaller data subsets, or clusters of well points. Breaking up the study area into smaller sections for kriging software to process yielded for Burke *et al.* a more accurate model than one obtained from generating a single, bigger prediction surface.

Burke *et al.* also acknowledged two key areas to address for improved models. One suggestion was to better characterize the vertical distribution of contaminants into the subsurface environment, and consequently the total volume of a contaminant plume. Burke *et al.* assumed that contaminants traveled no further than a depth of 20 feet underground, the average length of well screens and a measure seen by the LARWCQB as standard within groundwater monitoring.

Such a model could significantly underestimate plume volume. Using screen lengths to represent vertical plume extent would not account for contaminant concentrations existing beyond the well screening interval. The resulting underestimation would likely be considerable, given how groundwater contamination in the West Coast Basin can be found anywhere from the shallow subsurface to depths of up to 1000 feet below the surface in the Lower San Pedro aquifer system (Crawford *et al.*, 2003). We also expected the model to overestimate volume of dense contaminants in areas where the screen lies too deep into the confining unit.

Another suggestion by Burke *et al.* was to develop a more complex way to address contaminant concentrations below a threshold detection level, which we will refer to as "non-detect data" for the rest of this paper.

3.7 THE ISSUE OF NON-DETECT DATA

Non-detect data pose challenges that must be accounted for in constructing a mass estimation model. More than half of the PCE well samples in the GeoTracker GAMA have concentrations values of zero, meaning that many pollutants were either absent from the wells at the time of monitoring, or, more likely, present at a concentration below the detection limit for a given sampling well.

Detection data exists for only 4,347 of 9,435 measurements in the Los Angeles coastal plain for PCE. If these 5,000-plus zeroes remain untreated before input into a mass estimation model, resulting projections would underestimate contaminant concentration.

Substitution provides one approach for dealing with non-detect data. Burke *et al.* employed substitution by downloading minimum detection limit (MDL) information from an Environmental Sensitivity Index (ESI) monitoring report for their single pollutant of interest, and took half of that MDL as a substitution number to assign to all non-detect values in a specific well. The halfway point between zero and an MDL gives a standard-practice estimate of non-detect data (Burke *et al.*, 2013). But using such substitution introduces an external input into the dataset, potential altering the data in a way that Helsel and Lee warn against (2006).

Another downside to the ½-MDL method is that many MDLs exist for any one contaminant because detection limits are set according to instruments, operators and methods specific to certain wells. Finding each MDL for a specific pollutant at a particular well site would be demanding, especially given how the project detailed in this paper was intended to analyze multiple contaminants.

Substitution with a uniform value would also decrease the standard deviation of the database and affect related statistical tests. A high percentage of non-detect data in the GeoTracker database may lead to big changes in standard deviation (Pitt, 2007). An indicator of data variation, standard deviation will be an important parameter in later analysis, particularly for modeling through kriging. Substitution, while widely used, is not an ideal method of data treatment.

Alternative methods for addressing non-detect data include maximum likelihood estimation (MLE), the Kaplan-Meier (K-M) estimator, and regression on order statistics (ROS) offered by the Nondetects and Data Analysis (NADA) package (Helsel and Lee, 2006).

In addition to choosing an improved substitution method, another component of our project would deal with whether ½-MDL substitution values would even improve upon the accuracy established by simply keeping concentrations of zero within our model.

4. GOALS & OBJECTIVES

The LARWQCB seeks to regulate groundwater pollution based on total mass limits, such as to create more flexible and efficient regulation for long-term water resources planning. That agency set a goal for this project to estimate the total mass of VOCs in the West Coast section of the Coastal Los Angeles Basin. To allow for enough time to refine a mass estimation model, the scope of this project was narrowed from all VOCs to just four priority pollutants in groundwater: PCE, TCE, chloroform and MTBE.

The project's objectives were to:

(a) develop a model with reproducible results that estimates groundwater pollutant concentration and uncertainty in the estimates of concentration,

(b) optimize this model by varying data manipulation processes and comparing error outputs,

(c) run our optimized model for the West Coast Basin,

(d) convert concentrations output to mass using groundwater storage capacity data and

(e) program a script and tool that, together, automate the entire process of calculating VOC mass.

5. METHODS

The following formula, **Equation 1**, was developed to form the basis for calculating the total mass of a given VOC within the West Basin:

Equation 1: Mass of VOC = C_{XY} * A * h* α_S* β_Z ,

where

C_{XY} = predicted XY concentration for a single cluster (obtained from kriging)
A = prediction area for a single cluster (obtained from kriging)
h = aquifer height (subtracting the aquifer layer heights from each other)
α_S = aquifer storage coefficient (for confined aquifers) or specific yield (for unconfined aquifers)
β_Z = z-axis variation coefficient (from the concentration curves).

This equation would be applied to a number of smaller areas comprising the West Coast Basin. The masses would then be summed to produce the total amount of contaminant within the study area. **Figure 2**, shown below, summarizes the process of obtaining the value for each of **Equation 1**'s five variables. Following sections will detail each step and identify important assumptions required to make these calculations.

5.1 GENERATING A CONCENTRATION PREDICTION SURFACE

5.1-a Data Source: GeoTracker GAMA

The input data for this model consisted of coordinate and concentration information for VOCs of interest within the West Coast Basin. To create a snapshot of the current groundwater contamination situation and keep the data processing manageable within kriging, only data from the past year was utilized. Locations not represented in the database were assumed to be free from contamination, given that well-monitoring efforts are probably committed to urban areas best believed to actually have VOC problems. The model was thus only representative of the wells put into it, and therefore did not generate any predicted well points.

Concentrations for wells with multiple readings were averaged out to produce a single concentration at each well. The locations of wells sampled for PCE are shown in **Figure 3**.

Examination of whether or not there were any well points sampled in previous years that the current data set did not account for showed that almost all well points were present in multiple datasets. Points not present in the current dataset were non-detects in prior years.

The option to download data from the past three years was not utilized because in the process of generating a prediction surface, kriging does not take time into account. Samples taken at the same well points at different times would both be represented in the prediction surface, and such a scenario would not accurately represent reality.

Among the project's initial goals was to develop an efficient and reliable technique for breaking up the downloaded GeoTracker GAMA dataset into smaller groups. The creation and implementation of a clustering method would make for more accurate models: Running the entire data set demonstrated error magnitudes greater than that of a selected test-site model.

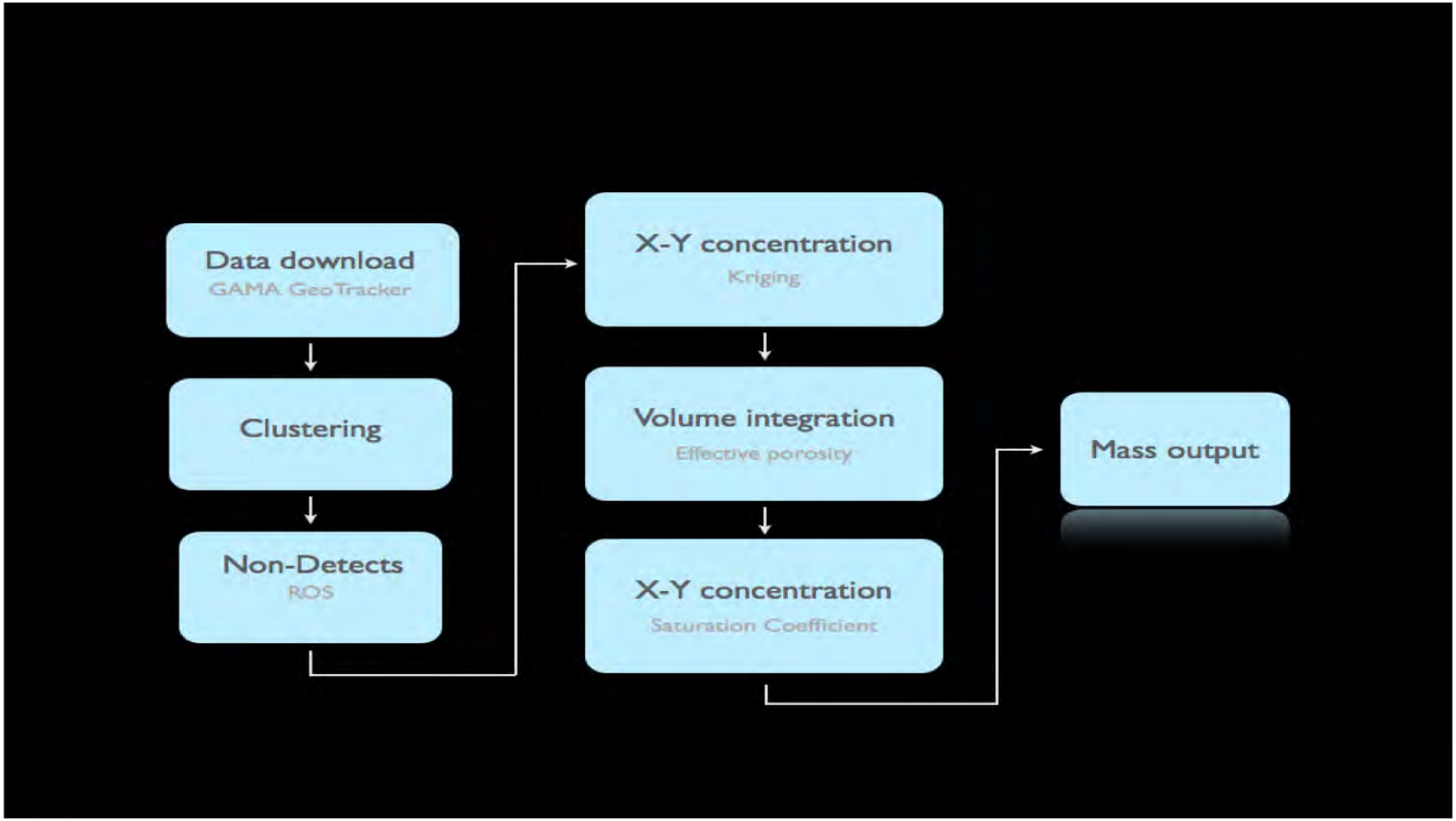


Figure 2. A flowchart summarizing this project's methodology for calculating the mass of a single contaminant within the West Coast basin.

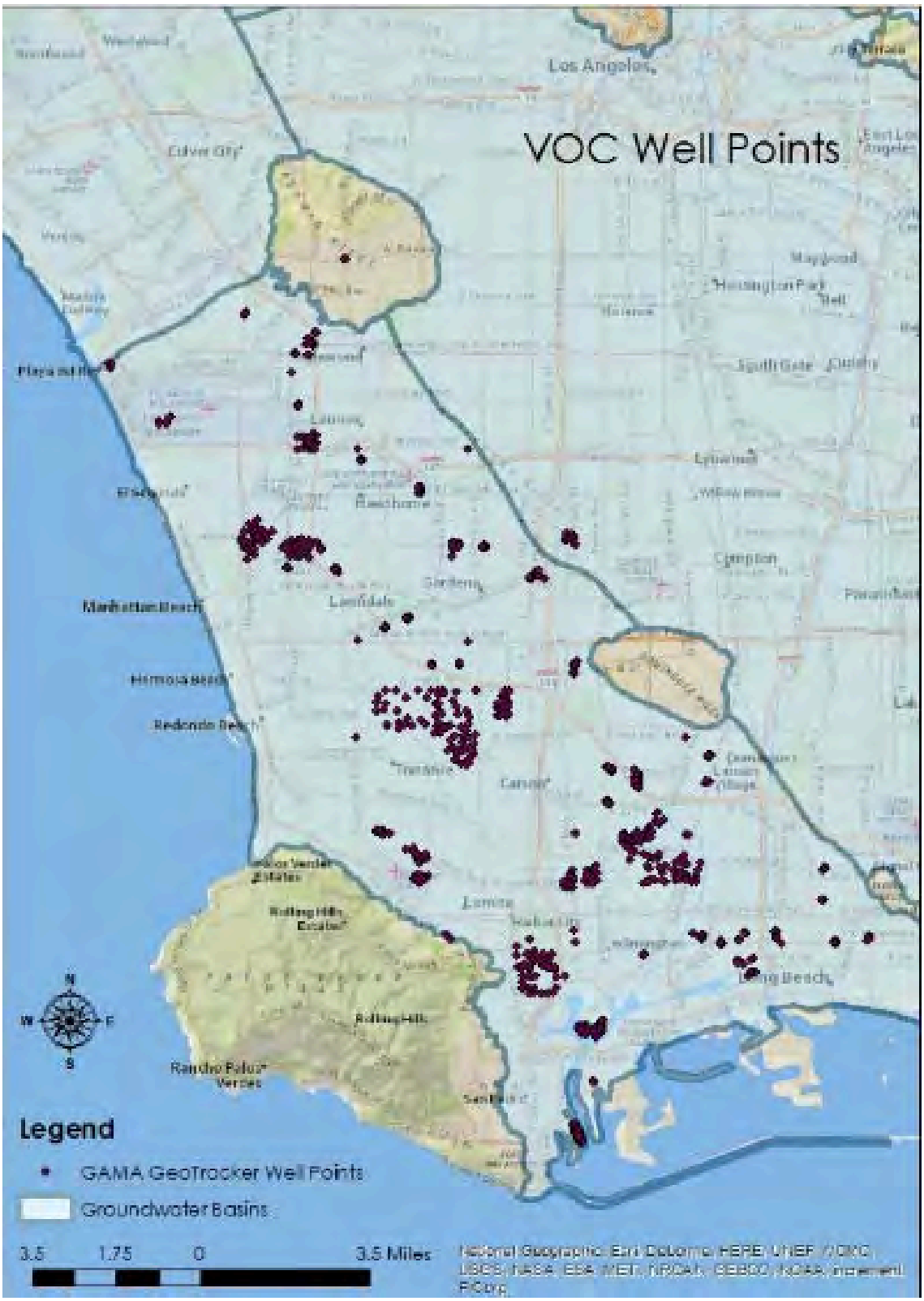


Figure 3. PCE well points from GeoTracker GAMA georeferenced onto the West Basin. The map shows a portion of the entire Los Angeles Groundwater basin.

The program ArcMap was used to create 153 polygons, each representing a cluster of joined sample well points. For identification purposes, a unique number is assigned to each cluster of well points. PCE data was used to create the one set of clusters applied to the four VOCs of interests. Since the well points sample for a number of contaminants, georeferenced well sample points should always appear in one of the 153 clusters.

Polygons were designed with two concepts in mind: proximity of the well points to each other and the purpose of the well sampling sites. Because kriging does not account for geospatial statistics, well points with data that may communicate with each other needed to be distinguished from well points that probably did not.

5.1-b Clustering of Well Points

Well points closer together influencing each other more so than do other points formed a key assumption for this clustering process. For example, wells all around a gas station were clustered together, as they were likely more related to one another than points centered on a neighborhood a mile away. Land use also informed the possible distribution of the pollutant plume: An industrial aerospace factory would be expected to produce a bigger plume than a single dry cleaning facility.

Despite the fact that clustering helps to increase the accuracy of kriging, the process also negatively impacts the model because it dilutes the data. Burke *et al.* (2013) discovered that kriging yields less error when the areas it acts upon are densely populated with well points. Not all clusters had a substantial number of points, with clusters of five or less data points being common.

Clustering would also be essential for enhanced accuracy of non-detect data treatment (Figure 4).

5.1-c Non-detect Data Treatment

To address the non-detect uncertainty, a treatment method for zero values that could more closely represent reality was sought. A brief summary of non-detect treatment methods are provided in Table 2 (Helsel and Lee, 2006) below.

Table 2. Non-detect Treatment Methods

Method	How it works	Pros	Cons	Conclusion
Substitution (Half-MDL)	Takes a fraction of the minimum detection limit	Simple	Introduces pattern external to the dataset	Does not fit because of high extent of bias
Maximum likelihood estimation (MLE)	Takes mean and variance as parameters and finds parametric values	Work best with a large (>50) sample size	A distribution must be specified	Does not fit; we have small sample size for each cluster
Kaplan-Meier (K-M) estimator	Uses survival analysis to estimate a survival function	No distribution assumed	Hardwired for right-censored data, used for less than 50% non-detect value	Does not fit because a majority of our data consists of non-detect samples
ROS	Uses a regression line in a probability plot	Can work for small population	More than 3 positive observations needed	First-choice

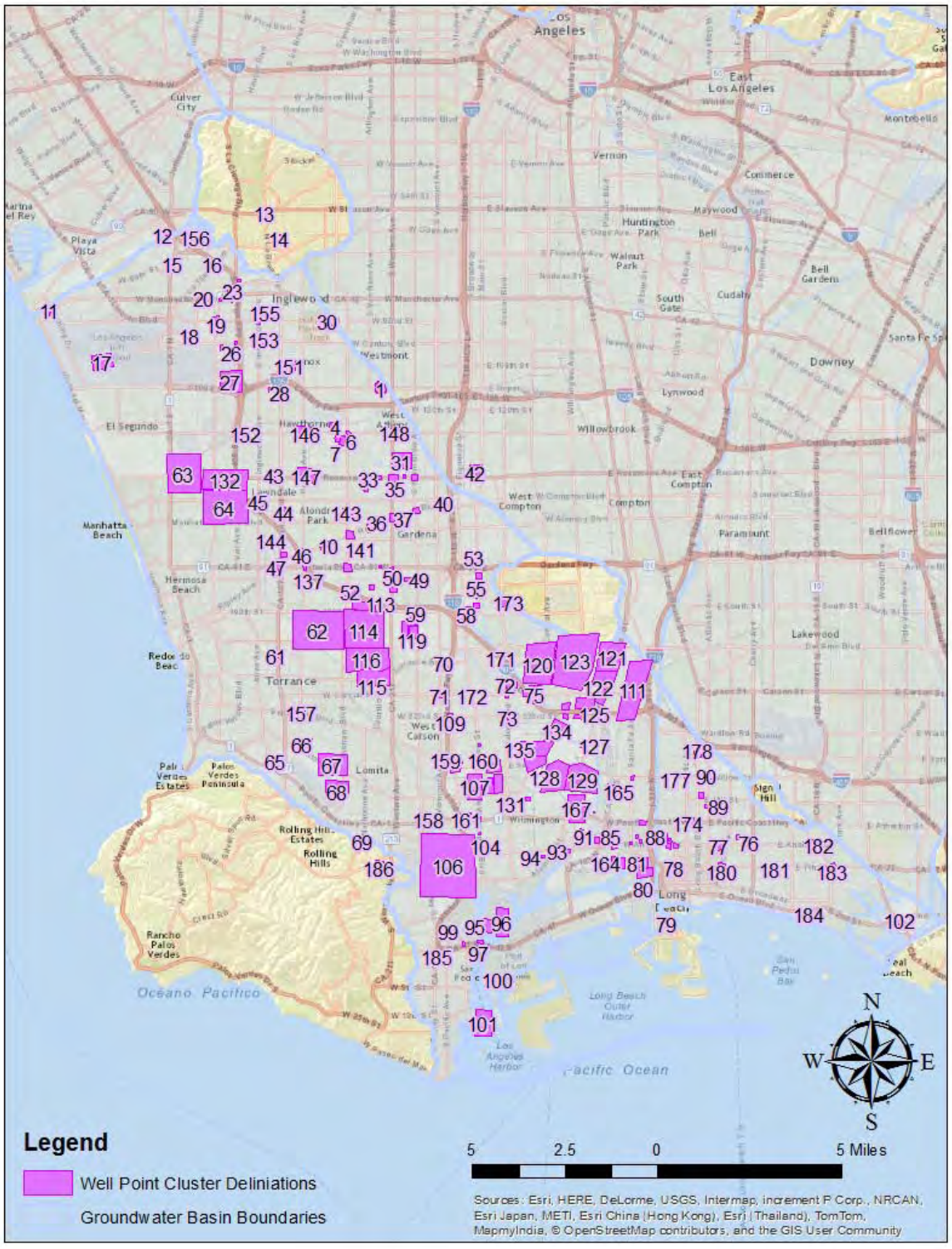


Figure 4. Delineated clusters of GeoTracker GAMA VOC well points in the Los Angeles West Coast Basin. The map displays all 153 groups of well points located in the study area.

The conclusion was made that ROS is the most suitable method for this GeoTracker dataset because of its small number of positive observations, or non-zero concentration values, per cluster. That choice was also appropriate because the dataset does not follow any distribution, a condition allowed by ROS.

ROS applies a regression line in a probability plot to estimate a non-detect value based on non-zero observations. The mean and variance of the whole population of data points can be estimated by ROS, and the non-detect value can be calculated based on the difference between the original mean and the ROS output's new population mean. The ROS substitution value would then be assigned to all zeroes within a given cluster.

Before ROS was run on the applicable clusters in dataset, data manipulation that optimized the estimations of mean and variance had to be performed.

The presence of repeated measurements was noticed within some of the same well sites. In order to avoid assigning too much weight to those repeated measurements, only one measurement was left for each well site by assigning the well the average of all repeating measurements as the value. See **Appendix B.1** for more details on the removal of repeating measurements in the programming software R.

Also, since ROS does not incorporate geographical information, concentrations from one area may affect the estimation of a non-detect value at a distant location. The clustering of data points provides the geographical component needed to increase the accuracy of the program's estimations.

According to Helsel and Lee, ROS requires at least three positive observations to draw the regression line for estimation (2006). A test was thus carried out to make sure that ROS was only run on clusters with non-detect data containing at least three non-zero concentration data points. **Appendix B.2** provides commands in R to execute such a test. **Appendices B.3 and B.4** detail additional helpful R commands for ROS analysis.

ROS yields an estimated mean and median instead of a substitution for a non-detect value. Some mathematical calculations are necessary to get the non-detect substitution value.

ROS substitution value for a cluster with non-detect data =

(ROS mean- original mean) * total # of points in a cluster / number of well points with concentrations of zeroes

ROS mean = mean group concentration when zeros are replaced with the ROS estimation;and
original mean = mean of concentration in each cluster without ROS treatment.

Figure 5 summarizes this project's approach to dealing with clusters containing non-detect data without the use of MDLs.

In addition to settling on a substitution method, another non-detect goal for this project was to determine how much of a difference the half-MDL method made for the prediction surface and concentration versus simply leaving in the non-detect data. Findings regarding this matter are shown in the Results section of this paper.

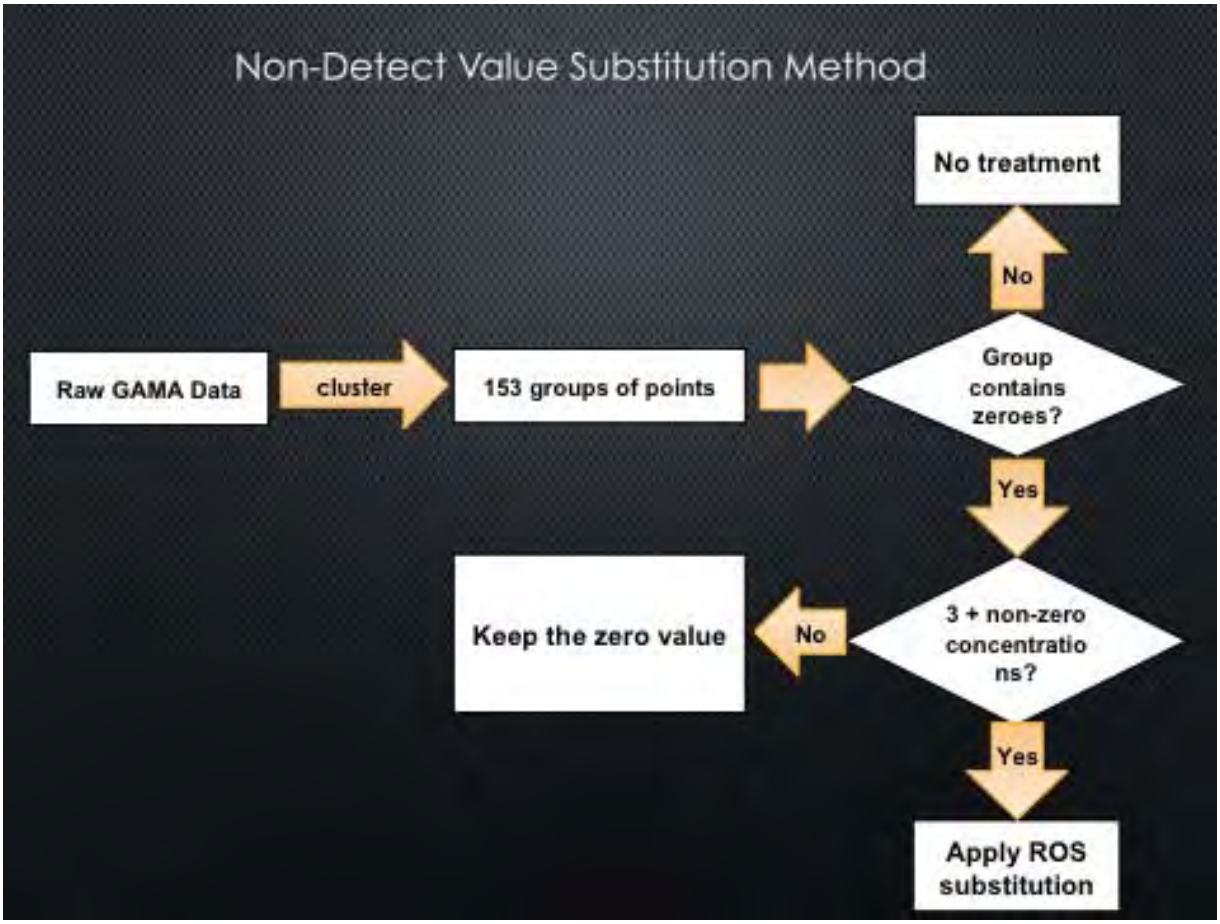


Figure 5. The project's decision-making process for treating non-detect values.

The primary method within this project's model for dealing with non-detects was to run ROS on the concentration data by cluster. ROS, however, requires that there be at least three non-zero values in the dataset to run, therefore it was necessary to devise a method for dealing with non-detect values in all clusters having fewer than three non-zero values. Use of the ½-MDL method employed by Burke *et al.* was considered, though that method would be very time-intensive.

Therefore it was decided to explore the route of leaving the non-detect values in such clusters as zeros. That was achieved by choosing three clusters and converting all non-detect values to half of the wells' respective MDL values, and running the kriging model on this data. Kriging was then run on the same clusters with the non-detects left as zeros.

5.1-d Data Normalization

The kriging model requires an input dataset that is normally distributed, but most GeoTracker GAMA information does not fulfill that condition. A large number of non-detect values caused the input data set to be heavily skewed.

The ArcMap kriging tool comes with a function to normalize input data before it is processed. Box-cox, log and sin are the transformation functions offered in the ArcMap program. A literature review was conducted to determine which transformation would produce the least erroneous model. Multiple kriging analyses were performed for the PCE data set, optimizing a

parameter and varying all other potential parameters to determine which transformation led to the most accurate model.

5.1-e Kriging to Produce a Concentration Prediction Surface

Kriging was performed with the transformed data sets to produce prediction surfaces along the X-Y, or longitudinal and latitudinal, axes. These surfaces are the result of interpolation around the well points from GeoTracker GAMA to yield estimations of pollutant concentration distribution in the north, south, east and west directions.

Ordinary kriging was used, since the method operates under an assumption applicable to this project: No true mean value exists for the data set. Like in the work of Burke *et al.* (2013), ROS was run on individual clusters comprised of localized well sites rather than on the West Coast Basin as a whole, and the technique was found to increase the accuracy and efficiency of the model.

To place the well points into clusters that allowed for a series of smaller and more accurate models, the same clusters from ROS treatment were used for input data sets in kriging.

5.1-f Assessing the Kriging Model's Accuracy

The kriging process outputs statistics for cross-validation, a means of assessing the model's accuracy. In cross-validation, one data point is removed and the whole model is run without that data point (ESRI, 2013). A predicted value for that data point will then be compared to the actual, observed value removed from the model. Every data point within an influential distance of another point is systemically removed though this cross-validation method.

Kriging uses different mathematical models to interpolate values. For each model type, ArcMap's built-in auto-optimization function was used to manipulate the rest of the variables to output the best-fit interpolation for a test site composed of a large cluster with varied concentration values.

To assess the accuracy of our model, four types of error (ESRI, 2013), which are summarized in **Table 3**, were examined.

Table 3. Measurements of Error for Optimizing Kriging Accuracy

Type of Error	Definition and Implications
Mean error	Represents the average difference between the predicted concentration values from kriging to the actual, observed values from GeoTracker GAMA. The negative sign means the predicted values are, on average, smaller than the observed values.
Mean standardized error	Illustrates variance in mean error values. Represents the average standard deviation of the difference of the predicted and observed values. More accurate models have mean standardized errors close to 0.
Root mean square error (RMSE)	Indicates the difference between predicted and observed values. This measure is the square root of the average of the square of all the differences between the predicted and observed values. Model accuracy increases as the RMSE value decreases.
Root mean square standardized effect (RMSSE)	A measure of variability among prediction values. Should this value be greater than 1, variability is underestimated in predictions, while a value less than 1 indicates overestimation of prediction variability. More accurate models have values close to 1.

5.2 USING AQUIFER TRAITS TO CALCULATE THE VOLUME OF A CONTAMINANT PLUME

Contaminant plume volume was calculated by using the equation $V = Ah\alpha_S$, where A is the area of the cell with the surface kriged concentration prediction, h is aquifer height, and α_S is the portion of the aquifer volume that is taken up by water, also known as the storage coefficient. Because kriging was conducted by cluster, the calculation of volume needed to be performed by cluster as well. To calculate a cluster's volume, the area of the kriged surface was multiplied by the height and storage coefficient value, and also by the mass coefficient.

To determine a cluster's area (A in **Eq. 1**), the cell size of the cells within the cluster's kriged concentration raster was multiplied by the number of cells in the Kriged cluster.

Height (h in **Eq. 1**) and storage coefficient (α_S in **Eq. 1**) values were calculated using raster data from the USGS that was developed in MODFLOW, a three-dimensional groundwater model, and subsequently converted to ½ mile x ½ mile GIS cells by the WRD for its groundwater basin available storage calculation (Johnson and Njuguna). These raster grids provide spatial data for the top and bottom elevations of the aquifers systems (TOL/BOL), land surface elevation (LSE), groundwater elevations (GWE), and specific yields and storage coefficients (SY/SC). That data was obtained from the WRD, and used to calculate water storage for each ½ mile by ½ mile cell in the basin.

5.2-a Obtaining Aquifer Height

The volume of water held within all four aquifer layers of the West Basin, which can extend to a depth of approximately 1000 feet, was incorporated into the project model through

use of GIS layer thickness and groundwater elevation (GWE) rasters. Storage concepts established for the WRD total available water volume calculation were used in the project to quantify the height occupied by existing groundwater. **Figure 6** on the next page shows the conceptual model assuming four different aquifer scenarios (Johnson and Njuguna**)**.

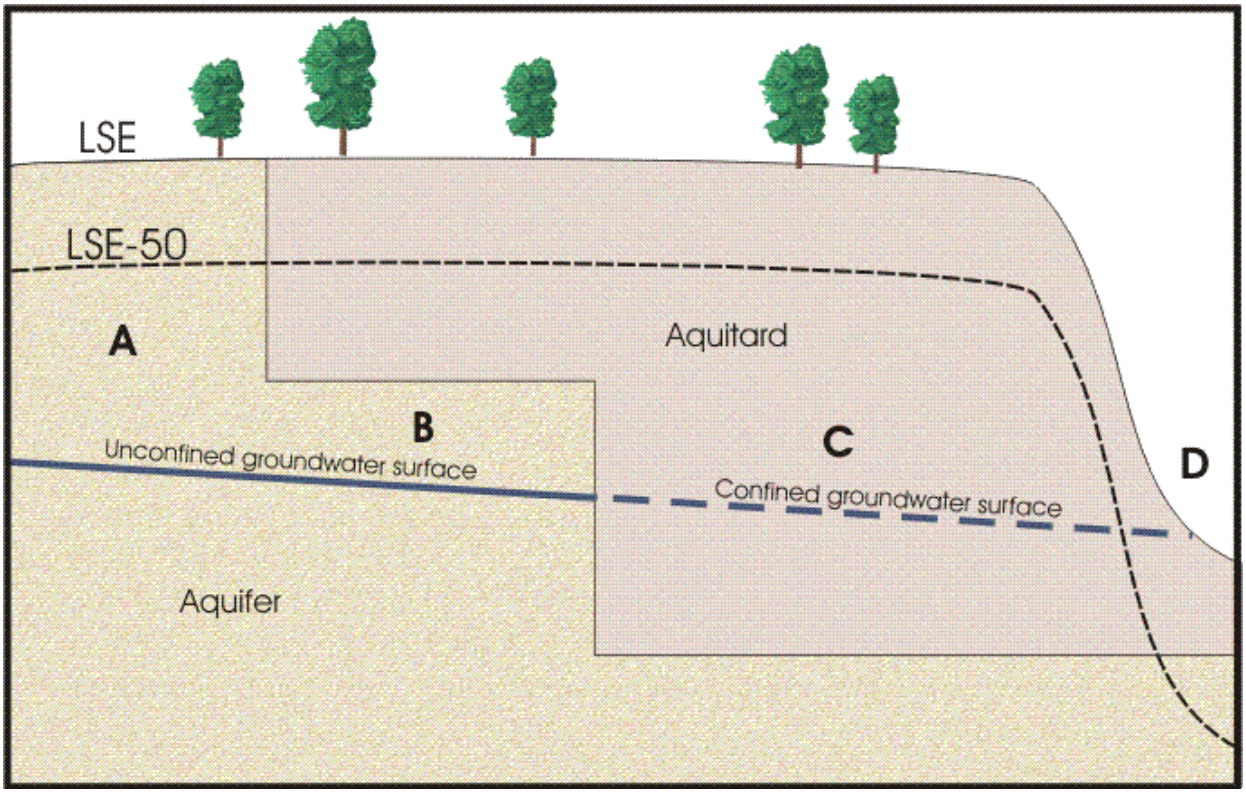


Figure 6. The four aquifer scenarios identified by Johnson and Njuguna. "A" shows a fully unconfined aquifer, while "B" shows a mixed aquifer that is both confined and unconfined. Scenario "C" depicts a fully confined aquifer and "D" displays an area in which no water storage is possible.

Table 4 lists the various layers and aquitards in the West Coast Basin, along with their aquifer type and the raster data and formula used to compute aquifer height.

Table 4. Calculating Aquifer Height

Aquifer Type	Layers	Formula for Aquifer Height Calculation
Layer 1 Unconfined	1	GWE1 - Lay1Bot*
Layer 2 Confined	Aquitard between layers 1 and 2	GWE2 - Lay2Top**
Layer 2 Unconfined	2	Lay2Top - Lay2Bot (if GWE2 - Lay2Top > 0) or GWE2 - Lay2Bot (all other conditions)
Confined	3, 4	GWE3 - Lay3Bot; GWE4 - Lay4Bot

*Lay1Bot = The depth (in feet) at which the bottom of Layer 1 occurs
**Lay2Top = The depth (in feet) at which the top of Layer 2 occurs

5.2-b Obtaining the Aquifer Storage Coefficient

Effective porosity (α_S) is represented by specific yield (SY) for unconfined aquifers, and storage coefficient (SC) for confined aquifers. Specific yield represents a ratio describing the volume of water capable of draining from a saturated area as a result of the force of gravity. SY was used to calculate volume for unconfined areas, where drainage is the predominant mechanism for groundwater flow. Similarly, SC represents the volume of water released from storage per unit change in the hydraulic head. This pertains to areas undergoing constant water compression without drainage, such as confined aquifers.

Effective porosity was assigned to each cluster based on the cell in which the centroid coordinates of each cluster wound up, with each cell of the raster grid having a porosity value designated by the USGS and WRD. The storage raster cells had ½ mile by ½ mile dimensions, while clustered cells were only 26 meters by 26 meters, and each cluster was mostly contained within a single cell.

The product of the layer height, cell unit area, and effective porosity yielded the total drainable water volume of the aquifer within each cell. In the final model, because the majority of clusters were smaller than a ½ mile by ½ mile, the cell unit area was replaced with the cluster area in to total water volume calculation.

5.3 Accounting for Contaminant Variation with Depth: the Z-Coefficient

As a contaminant travels deeper into the subsurface environment, its concentration will vary based on numerous factors, including the physical and chemical properties of that pollutant and the hydrogeological properties of the surrounding soil and rock. Not all groundwater will be saturated with contaminant, so a constant for each contaminant was introduced to address such differences in vertical distribution for a given X-Y coordinate.

The z-coefficient (β_Z) represents the fraction of a water volume that is contaminated with a VOC of interest.

A USGS report (Belitz *et al.*, 2000) that included data from a model accounting for many dynamic subsurface processes like groundwater recharge and flow, provided the information needed to construct a curve depicting contaminant concentration as a function of different depth for four contaminants: PCE, TCE, MTBE and chloroform. Each of the four were among the 12 organic pollutants that received "additional evaluation" by a Priority Basin report for the Coastal Los Angeles Basin (Belitz *et al.*, 2012).

Plotting concentration as a function of depth helps in the calculation of a z-coefficient for each of the four contaminants. Determining the z-coefficient by looking at the graphs included drawing a best-fit curve through the data points and determining a depth at which a maximum concentration occurs. The z-coefficient would be the ratio of the integral of each concentration versus depth curve divided by the integral of a pollutant's maximum concentration over the entire depth. **Figure 7** displays these curves for the four studied contaminants.

The observed maximum concentration was integrated for a given VOC under the assumption that such a value was a reasonable approximation of effective solubility for that contaminant. Intuitively, the maximum concentration and effective solubility could be expected to be similar if samples were predominantly taken from effectively saturated contamination areas, an assumption made because sampled locations were likely areas assumed to have considerable VOC pollution.

In reality, effective solubilities are difficult to generalize. True effective solubility values can lie anywhere from maximum concentration to aqueous solubility, with the latter often greater than the former by orders of magnitude. As a result, this project's model could overestimate total mass, as we would be integrating at higher solubility values.
This technique also required the generation of data points in the programming software R that were not in the original data set from the USGS. Points external to the data were generated around the 48 downloaded points to yield a more effective curve. Such interpolation was necessary considering the limited amount of data given in the USGS report. **Appendix C** shows the R commands required to generate the two integrals needed to calculate the z-coefficient.

5.4 POLLUTANT SELECTION

This project focused on providing more information about pollutants that regulators in the West Basin figured to regard as particularly concerning. An investigations report for the Coastal Los Angeles Basin Priority Basin Project informed the decisions of which VOCs would be estimated in the West Basin.

The document contained a section highlighting 12 organic contaminants that warranted "additional evaluation" (Belitz *et al.*, 2012). Those contaminants received special attention either because a notable portion of their samples within the basin showed moderate to high concentrations relative to health standards, such as MCLs, or they displayed detection frequencies greater than 10 percent throughout the basin.

These 12 organic pollutants were considered during the search for data needed to construct the z-coefficient curves. PCE, TCE, chloroform and MTBE, which each qualified for "additional evaluation" in the report, turned out to be among the few contaminants with enough data points to construct reasonable depth versus concentration curves.

With PCE, chloroform and MTBE being the three most frequently detected VOCs in domestic and public wells (USGS, 2014) and how these contaminants are representative of the most common VOC groups in the basin (Belitz and Fram, 2012), the characterization of those pollutants has a reasonable chance of being a high priority for regulatory agencies.

See **Appendix D** for an expanded explanation for the Priority Basin report's evaluation of contaminants as applied to the four contaminants of interest.

5.5 AUTOMATED CALCULATION OF VOC MASS ESTIMATE

Almost all of the project's methodology, from the importing of GeoTracker GAMA data into ArcMap through the calculation of a single VOC's mass in the West Coast basin, has been automated through the development of a tool and code (Bruguera and Tsang, 2014). These two items — the Clustered Kriging tool and Mass Output code — were developed in Python, the programming language used in kriging.

In tandem, the Clustered Kriging tool and Mass output code allows users to run pollutant data from GeoTracker GAMA through the polygon clusters, aquifer volume raster and to produce a mass-estimate for a given VOC.

Appendix E provides a comprehensive report for the creation, use and implications of this automated mass estimate technique.

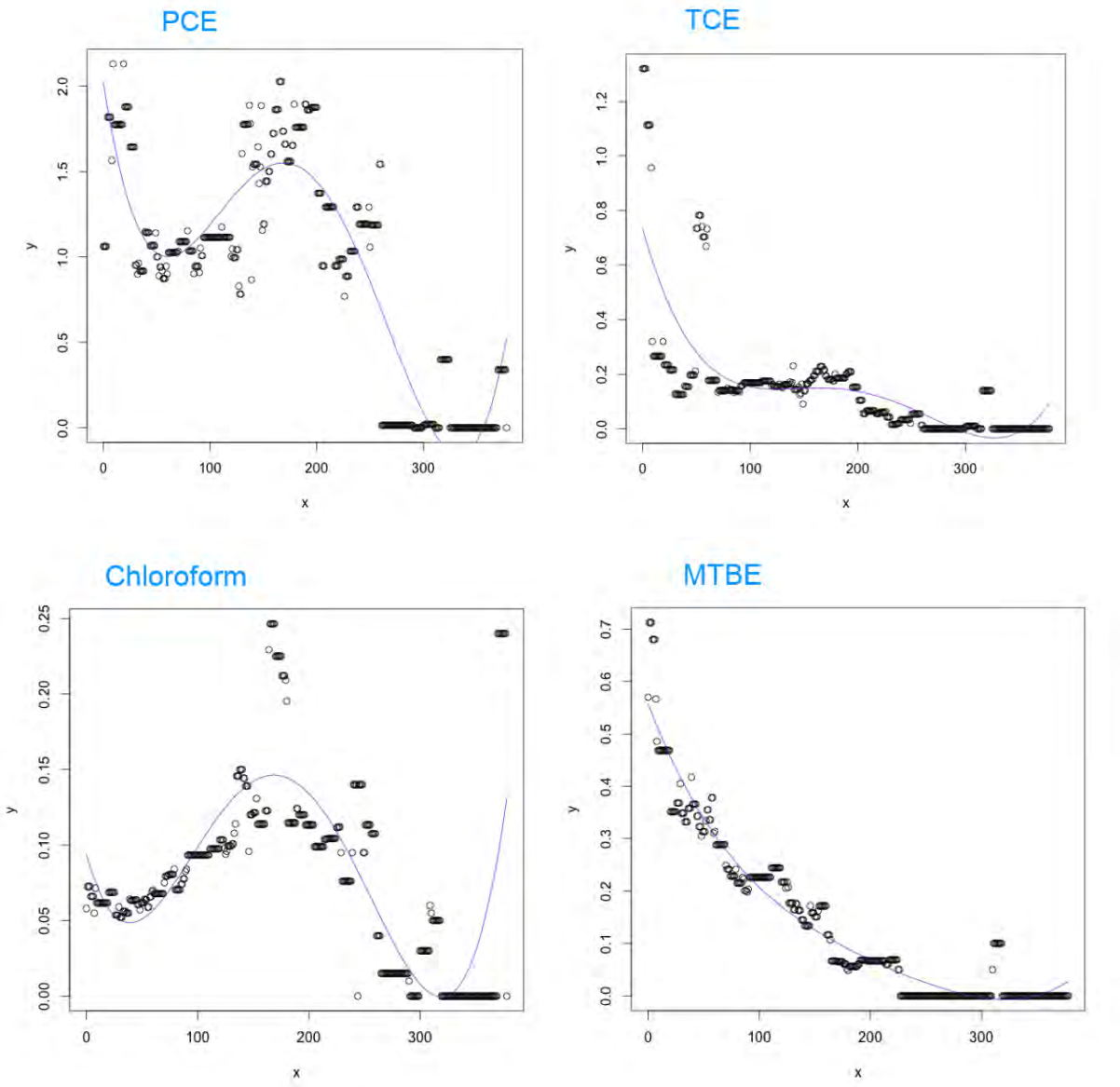


Figure 7. Graphs showing the concentrations of four contaminants over depth. The x-axis represents depth (meters) and y-axis represents concentration (micrograms/liter). Each VOC has a different distribution due to differing properties between each type of VOC.

6. RESULTS

6.1 NON-DETECT TREATMENT

Visualized in **Figure 8** below is a comparison between the effect on concentration prediction surfaces by using the ½-MDL non-detect treatment method versus that of leaving non-detect data untreated by the ROS technique. Little difference was seen between the prediction surfaces generated by kriging the two data sets.

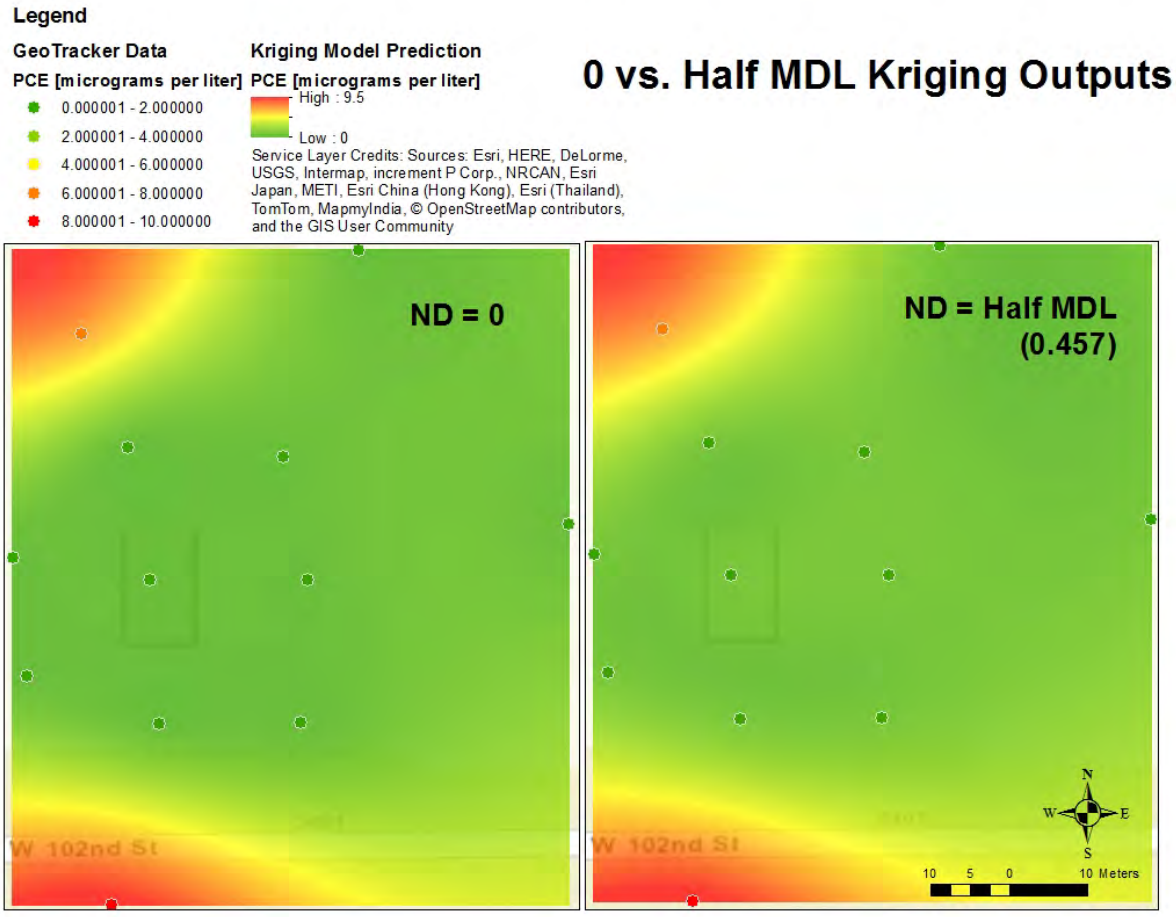


Figure 8. A comparison of a PCE kriged prediction surface between a model that keeps all non-detect concentrations at zero versus a model that substitutes half of the MDL for that contaminant at that well site.

6.2 SELECTING A DATA NORMALIZATION TECHNIQUE

The sin transformation was eliminated immediately because the GeoTracker data set was not governed by a sinusoidal function. A comparison of the box-cox and log transformations, carried out through the geostatistical wizard kriging tool, verified the finding from the literature: box-cox transformations are most accurate for modifying input data in groundwater kriging models (Varouchakis *et al.*, 2012). In all cases but one, box-cox transformed models produced less error than log transformed models. See **Table 5** for more details.

Table 5. Summary of Data Transformation Optimization Error Analysis

Transformation type	Box-cox	Box-cox	Box-cox	Box-cox	Log	Log	Box-cox	Box-cox	Box-cox	Box-cox
Power parameter	1	1	1	2	N/A	N/A	1	1	1	1
Order for trend removal	None	None	None	None	None	Second	Second	Second	Second	First
Kernel Function	N/A	N/A	N/A				Gaussian	Epanechnikov	Con-stant	Gaussian
Semivariogram Mathematical Model	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable
Smoothing	0	0.5	1	0	0	0				
RMSE	1519.57	1520.41	1522.4	4788.79	2961.31	2961.31	1521.12	1521.56	1521.92	1711.33

The results of a literature review and kriging error analysis showed the box-cox transformation to be the ideal transformation model. In the final automated model, however the empirical transformation available through ArcMap's Empirical Bayesian Kriging (EBK) tool replaced the box-cox transformation.

6.3 SELECTING MATHEMATICAL INTERPOLATION MODEL TYPES AND PARAMETERS

For each model type, kriging was executed and the cross validation statistics were recorded. **Table 6** summarizes the error analysis for the different models.

Table 6. Summary of Error Analyses for Each Model Type

Model	Mean Error	Mean Standardized (least error = close to 0)	RMSE (least error = close to 0)	RMSSE (least error = close to 1; if <1 overestimating variability, if >1 underestimating variability)	Average Standard Error
hole	-402	-0.07	3908	0.95	2637
k-bessel	-371	-0.257	4143	2.87	972
stable	-367	-0.261	4133	2.93	963
spherical	-354	-0.267	4192	3.06	946
tetraspherical	-350	-0.266	4201	3.07	945
gaussian	-351	-0.268	4201	3.1	941
exponential	-346	-0.27	4213	3.14	932
pentaspherical	-341	-0.27	4229	3.18	924
rational quadratic	-335	-0.27	4247	3.23	920
circular	-317	-0.279	4300	3.39	911
j-bessel	-308	-0.28	4327	3.42	889
empirical bayesian EBK	-55	-0.235	1990	3.86	632.58

After the input data was transformed through normalization, kriging was run to produce a prediction surface map for PCE. The final automated model executes Kriging only on clusters that have a great enough amount of mass, as determined by the criteria of having a mean well sample concentration value that exceeds 0.1 □g/L and containing at least eight well points. Because the majority of clusters do not meet these criteria, as illustrated by **Figure 9**, they are not kriged in the automated method.

As displayed in **Figure 10,** contaminant concentration was found to be low across the majority of the basin, ranging between 0 and 0.5 □g L^{-1}, with a few outlying hotspots of contamination, which the model krigs and estimates mass.

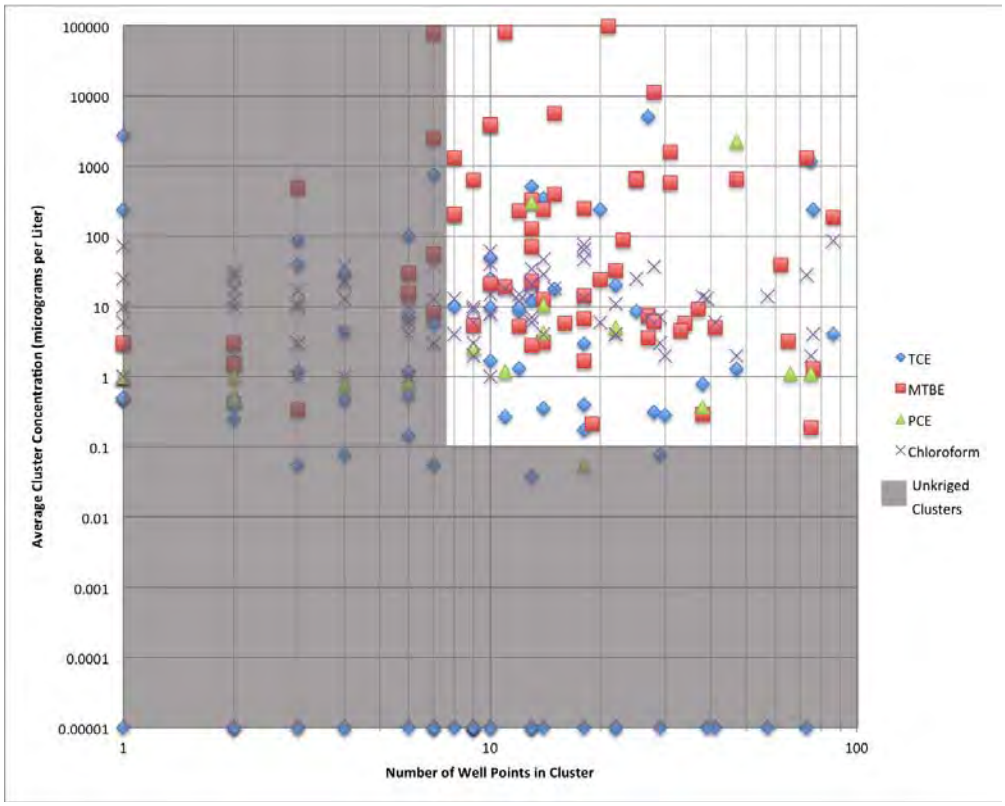


Figure 9. A plot showing the distribution of the number of well points and average concentration of each cluster. The majority of clusters which do not meet the kriging criteria have relatively low concentrations compared to clusters which do meet the criteria, with the exception of a few outliers.

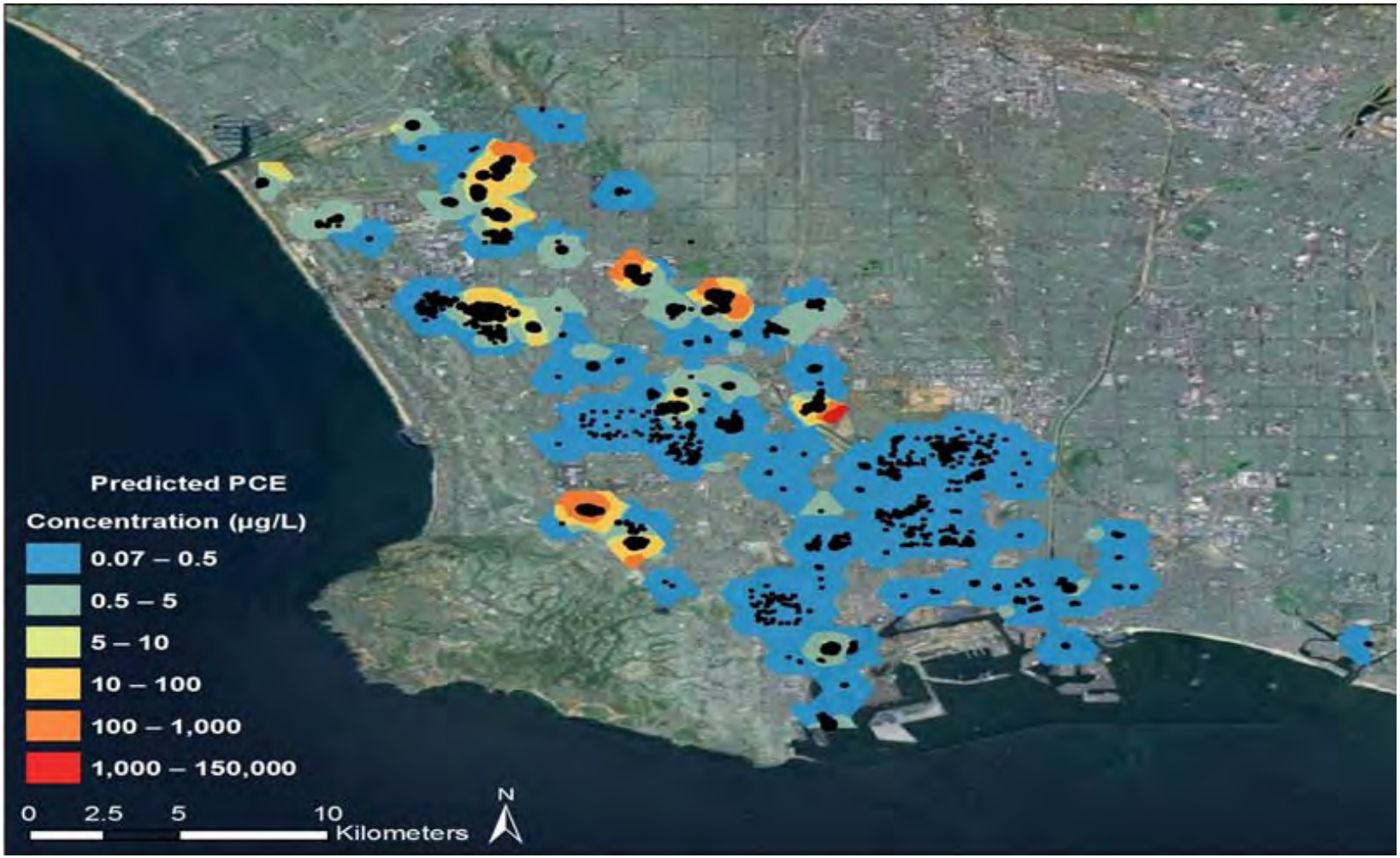


Figure 10. The predicted PCE concentration surface produced by the optimized k-bessel model, using data from all un-clustered basin well points. Note that this is representative of the concentration across the basin, but not representative of the final kriging rasters and these are only of hotspots with the high PCE concentrations. .

6.4 Calculating Aquifer Storage Volume

Pictured on the next page is **Figure 11**, a map of the West Coast Basin's aquifer storage volume. This map was generated by summing the product of:

a) the height of each aquifer layer,
b) the cell's specific yield or storage coefficient and
c) the ½ mi by ½ mi area of each cell.

Because kriged cluster cell size was 26 meters x 26 meters, and entire clusters were often less than a ½ mi x ½ mi, so clusters were often encompassed by only one or two storage cells. Since ArcMap's raster calculator sum requires that raster layers have the same cell size and extent to be summed, the planned raster calculator function of multiplying kriged cluster concentration by storage raster volume to get mass could not be conducted.

Instead, to get the aquifer storage volume value, the centroid (or coordinate center) of each cluster was calculated, and the aquifer storage value at this point was taken, and used for the entire cluster. Because the initial aquifer storage volume raster cell values included the ½ mile x ½ mile area in the volume calculation, and this area needed to be modified to be the area of the actual cluster, the aquifer storage raster cell values were divided by ½ mi x ½ mi, such that they would only include the aquifer height and storage coefficient.
In the final mass calculation, this height and storage coefficient aquifer value was multiplied by the area of each cluster, as determined by the kriged cluster cell size (26 meters x 26 meters) multiplied by the number of cells in the cluster.

6.5 Constructing Concentration Curves for the Z-Coefficient

Integral calculation in R yielded a z-coefficient of .642 for PCE, .292 to TCE, .691 for chloroform and .363 for MTBE. The four graphs depicted below in **Figure 12** show the curves and integrals used to calculate the z-coefficient for each pollutant.

6.6 Calculating VOC Mass

The code calculates pollutant mass within each of the 153 clusters and compiles that data into a single table with cluster IDs, pollutant concentration for each cluster, aquifer depth and storage value. **Figure 13** displays the part of the code that deals with calculating mass and the creation of a summary table, while **Figure 14** shows a program run's output of a summed estimated pollutant mass within the entire study area. **Appendix E** provides a comprehensive report and user manual on the code and its creation.

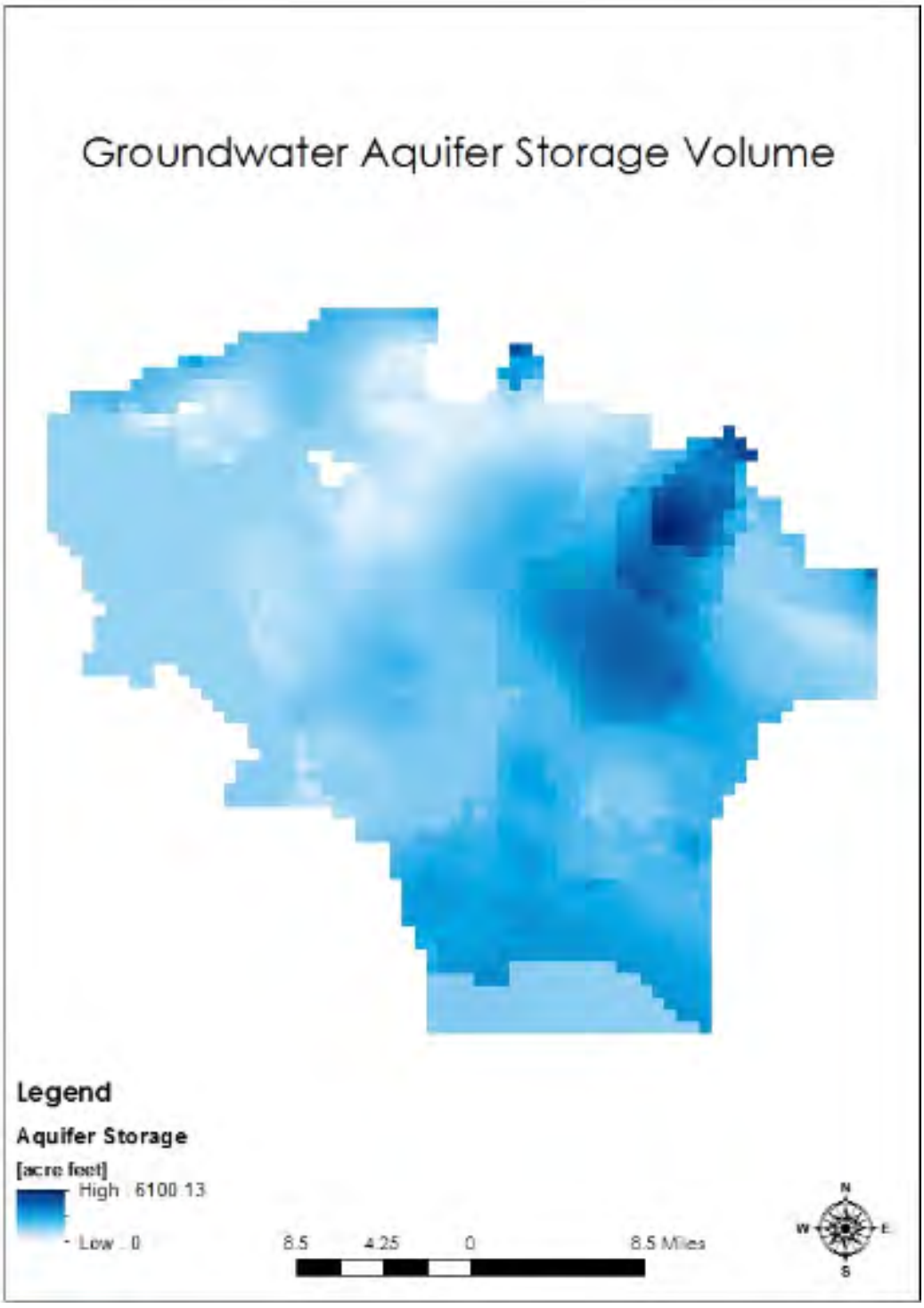


Figure 11. A map of groundwater aquifer storage throughout Los Angeles' West Coast Basin. Darker colors indicate increased ability to hold water, in acre-feet. This map was generated by summing the product the height of each aquifer layer, multiplied by its specific yield or storage coefficient, and multiplying this by the ½ mi by ½ mi area of each cell. When calculating aquifer storage for each cluster, the storage cell ½ mi by ½ mi area was replaced by the area of the cluster.

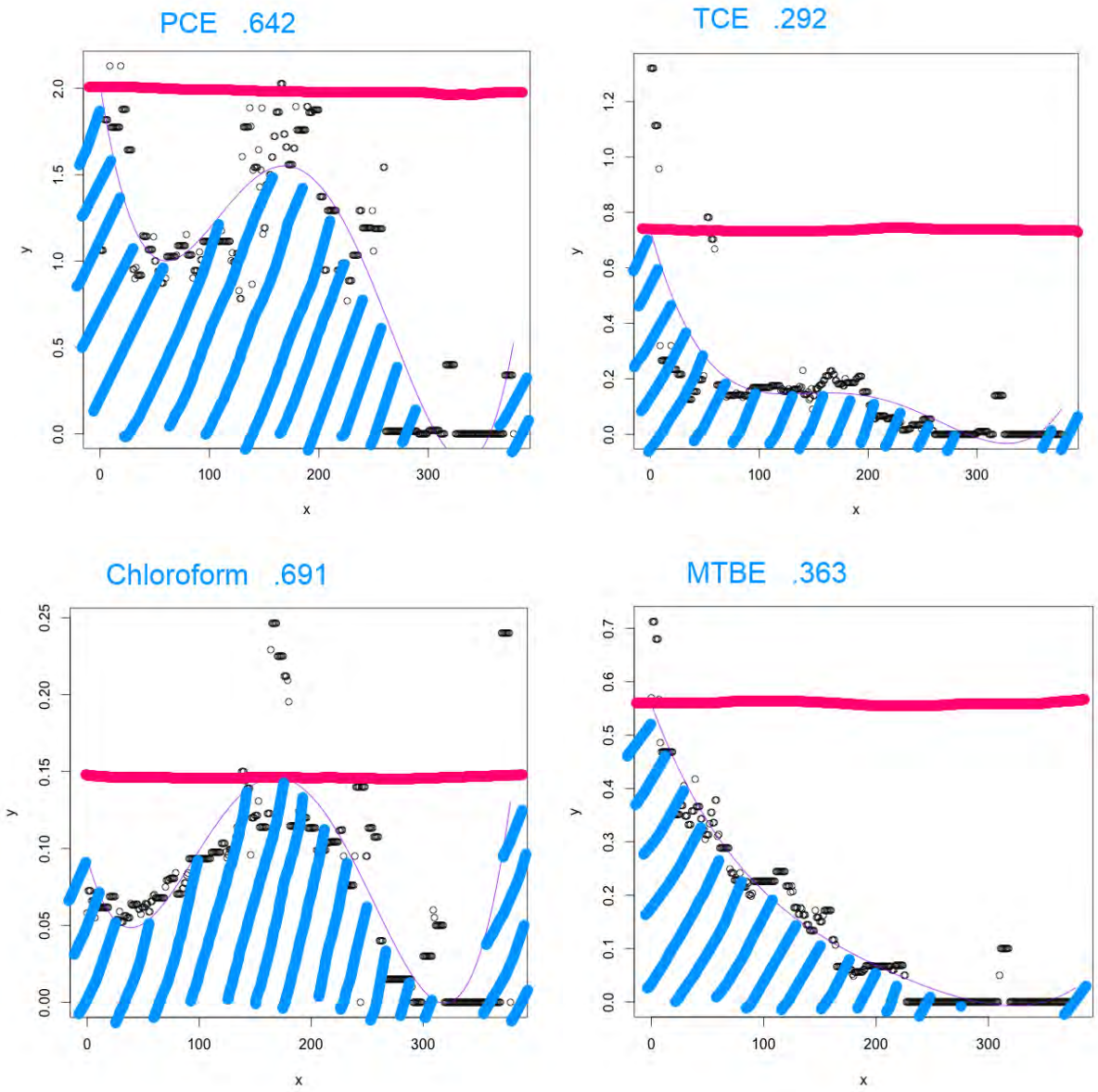


Figure 12. These are best-fit curves fitted to the concentration versus depth graphs. The z-coefficient is the ratio of the integral of the curve, shown in blue, divided by the integral of the maximum concentration over the entire depth, which is represented by the area underneath the pink line. This coefficient represents an estimation of how much a given VOC contaminates the depth of the groundwater at a particular XY location.

```
################ Multiply Kriged Rasters by Aquifer Storage Volume ###################

#For all kriged rasters, perform raster calculator
rasterlist = arcpy.ListRasters('*', 'TIF')
#Create empty list to add total mass of all clusters
mass_list = list()
#Create table for cluster concentration, cell number, and mass information
##In tool, would be USER SPECIFIED mass output table
mass_table = "Mass_Table"
arcpy.CreateTable_management(folder_path, mass_table)
#Create fields for cluster ID, total concentration, cell number, aq storage, and mass
arcpy.AddField_management(mass_table, "Clust_ID", "DOUBLE")
arcpy.AddField_management(mass_table, "Sum_Conc_ugL", "DOUBLE")
arcpy.AddField_management(mass_table, "Num_Cells", "DOUBLE")
arcpy.AddField_management(mass_table, "Aq_Storage", "DOUBLE")
arcpy.AddField_management(mass_table, "Mass_kg", "DOUBLE")

#Import numpy module
import numpy
#Create insert cursor to populate table

for krigedRas in rasterlist:
        #Determine cluster ID of kriged raster
        num = krigedRas.index('d')
        numP1 = num + 1
        cluster = int(krigedRas[numP1:-4])
        print "Cluster "+str(cluster)

        #Create array from kriged raster to get cell values
        myArray = arcpy.RasterToNumPyArray(krigedRas)
        #sum array rows to create array of total concentration value for each row
        sum1 = sum(myArray)
        #sum array of total row concentrations to get total concentration of kriged raster
        sum2 = sum(sum1)
        print "        Sum of surface concentrations: "+str(sum2)+" ug/L"

        #Calculate number of cells in kriged area
        kriged_width = len(myArray)
        kriged_length = len(sum1)
        kriged_cell_num = kriged_width*kriged_length

        #Get aquifer storage cluster index
        #Get index value of kriged raster cluster in list of cluster IDs
        clusterID_index = int(clstID_list.index(cluster))
        #print "       Cluster ID List Index: "+str(clusterID_index)
        #Get aquifer storage of indexed cluster
        aqStor = AqStor_list[clusterID_index]
        print "        Cluster aquifer storage: "+str(aqStor)

#Calculate mass of cluster
#Explanation of mass calculation: mass = concentration [ug/L] * aq storage w/ depth[m] / #
cells (kriged_cell_num from array calcs) [cell] * XY Cell area [m^2/cell] * # cells
(kriged_cell_num from array calcs) [cell]* mass coefficient (unitless) * 1000 (cubic meters
to liter conversion) [L/m^3] * 1/1000000000 (micrograms to kilograms conversion) [kg/ug]
        #####in tool, would be USER INPUT MASS COEFFICIENT, as it varies by pollutant
        mc = 0.642
        mass = sum2*aqStor*22.26*22.26*mc*1000/1000000/1000
        print "        Cluster mass: "+str(mass)+" kg"
        mass_list.append(mass)
```

```
        #Update Table Rows
        rows = arcpy.InsertCursor(mass_table)
        row = rows.newRow()
        row.Clust_ID = cluster
        row.Sum_Conc_ugL = sum2
        row.Num_Cells = kriged_cell_num
        row.Aq_Storage = aqStor
        row.Mass_kg = mass
        rows.insertRow(row)

        del rows, row, aqStor, clusterID_index, cluster, myArray
print "done"

#Sum list of mass by cluster to get total Basin Pollutant Mass
tot_mass = sum(mass_list)
print "Total Pollutant Mass in West Coast Basin: "+str(tot_mass)+" kg"
```

Figure 13. For each kriged cluster, the total number of cells, the sum of the concentration values, and the aquifer storage value are calculated, and this information is put into the table generated at the end of the code. These values are then used along with the aquifer storage and mass coefficient to produce a mass estimate.

```
Cluster 132
        Sum of surface concentrations: 392771.416015625 ug/L
        Cluster aquifer storage: 2.9927919
        Cluster mass: 373.939843786862 kg
Cluster 135
        Sum of surface concentrations: 103.31383319199 ug/L
        Cluster aquifer storage: 3.7341406
        Cluster mass: 0.122725409391036 kg
Cluster 136
        Sum of surface concentrations: 35.0153058767318 ug/L
        Cluster aquifer storage: 3.4519362
        Cluster mass: 0.0384508577960212 kg
done
Total Pollutant Mass in West Coast Basin: 2699.35 kg
```

Figure 14. **This figure illustrates the last few lines of Python Shell window output for the Mass Output code when run on PCE.** It includes properties of each cluster, which are outputted into the table as well, and a total mass estimate for the basin.

Figures 15 through 18 show the mass distribution of each of the four pollutants throughout the West Coast Basin. Not all clusters factored into the mass calculation for a given pollutant, as the code required that the cluster have at least 8 points and that the average cluster concentration be greater than 0.1 □g/L to execute kriging. The cluster shape depicted in **Figures 15 through 18** does not impact final mass calculation. The polygon file is simply used to illustrate the amount of mass determined to be within the cluster.

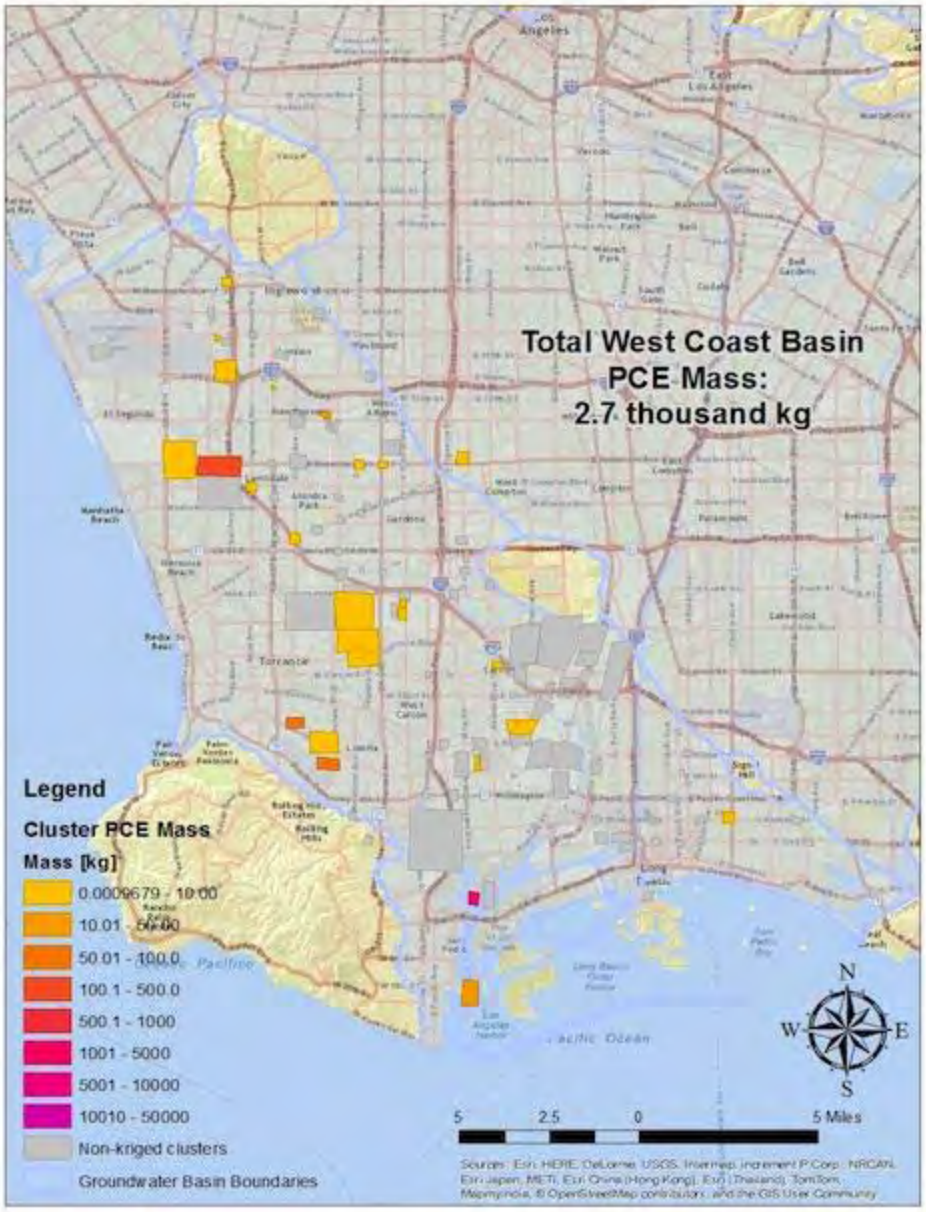


Figure 15. A mass prediction map for PCE in the West Basin. 1302 well points were contained in 88 clusters containing PCE samples, and the masses of 27 clusters were summed for a total of 2699 kilograms.

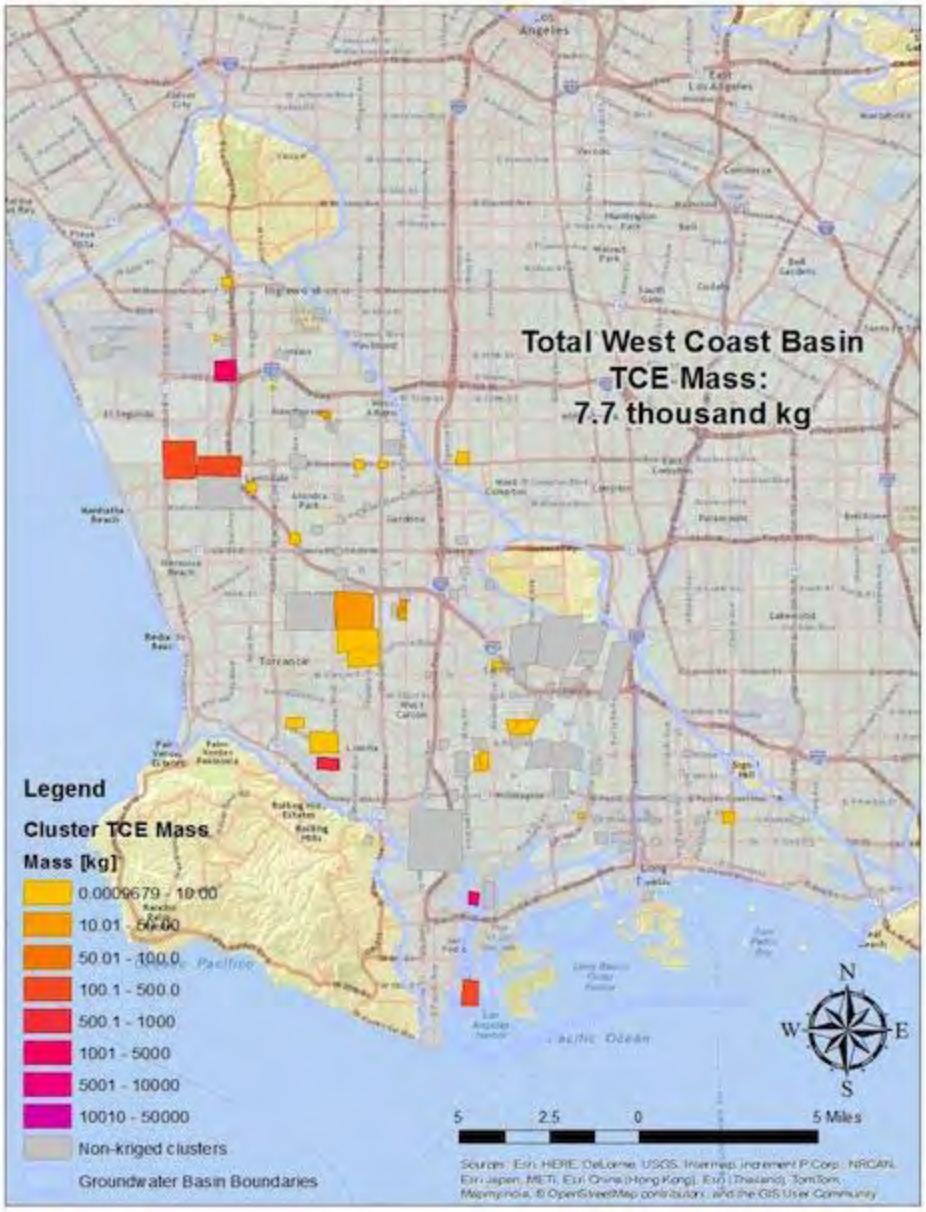


Figure 16. A mass prediction map for TCE in the West Basin. Though 1289 well points were contained in 86 clusters containing TCE samples, the masses of 29 clusters were summed for a total of 7,657 kilograms.

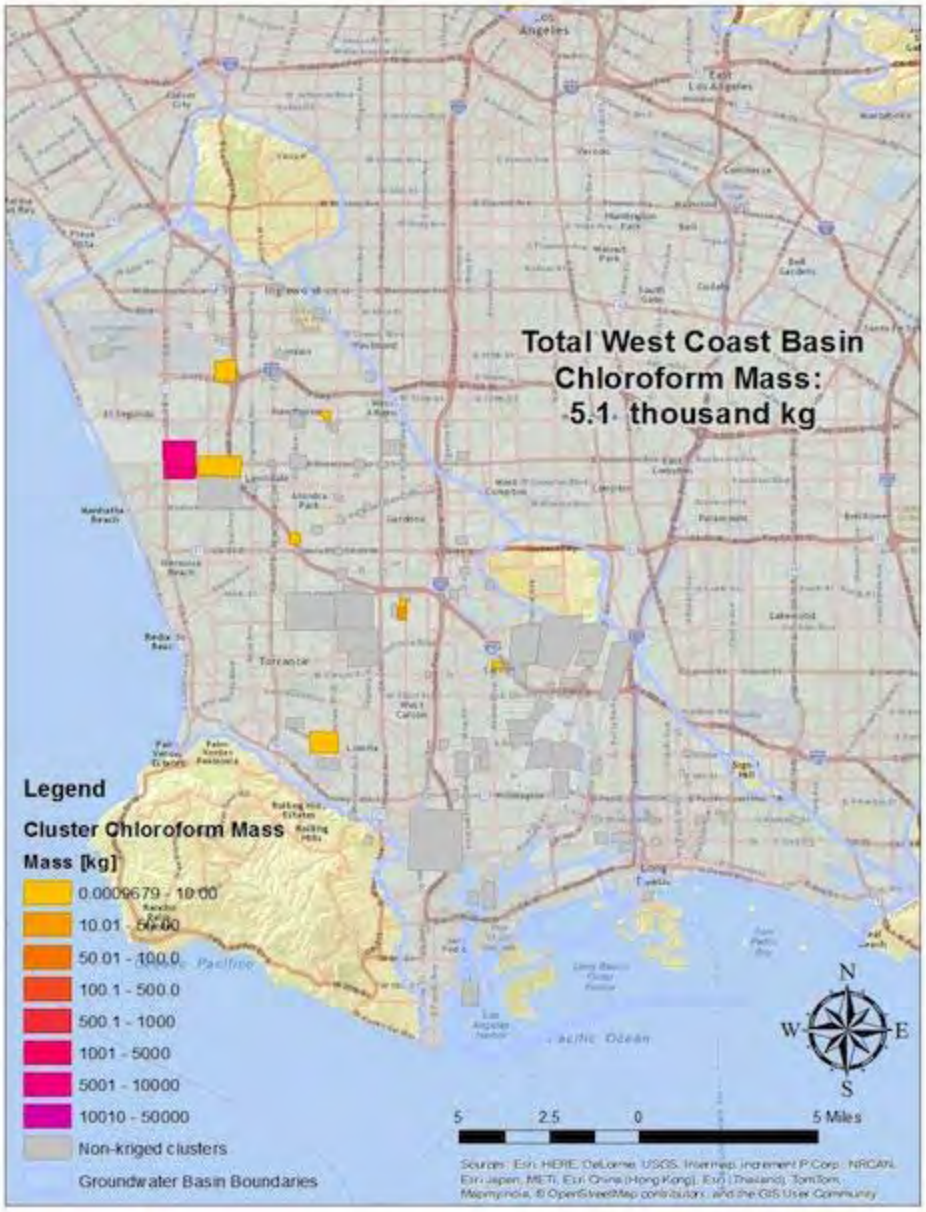


Figure 17. A mass prediction map for chloroform in the West Basin. Of the 85 clusters that sampled for chloroform, the masses of only 9 clusters were determined to be significant by the code and summed for a total of 5,102 kilograms.

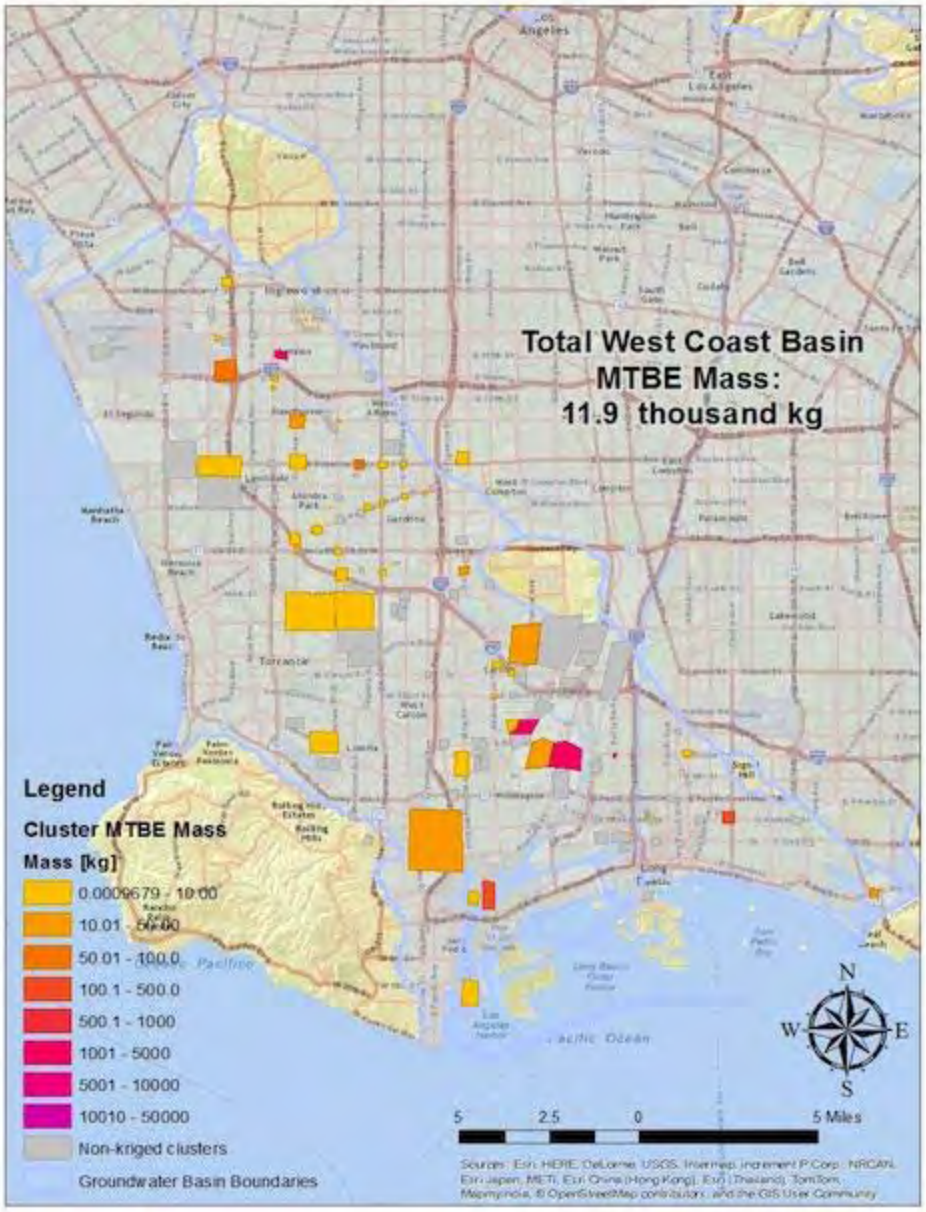


Figure 18. A mass prediction map for MTBE in the West Basin. Of the 2193 well points sampling for MTBE across 111 clusters, the masses of 47 clusters were summed for a total of 11,872 kilograms.

7. DISCUSSION

7.1 NON-DETECT TREATMENT: KEEPING NON-DETECT DATA OR USING THE ½-MDL METHOD

Figure 8 from the Results section demonstrates that little difference existed at a test site within the West Basin when switching non-detect values in the model to ½-MDL values. This result suggested that the final mass estimate might not be meaningfully impacted by leaving non-detects in ROS-ineligible clusters as zero values instead of changing them to half-MDL values. In the interest of meeting project deadlines, an executive decision was made to leave non-ROS treated zeroes as they were.

While this lack of difference is pronounced, and despite the fact that the ½-MDL method introduces external data, it was decided that MDLs should be sought out when possible in future mass estimate calculations. When ROS is not appropriate for a cluster, using half of an MDL is preferable to leaving a zero value. Given that there exist few differences between keeping the zeroes and using a fraction of the MDL, giving the well-point an actual value would probably better represent the reality of scarce, but existent, contamination. But because the process of data-mining MDL values for each well is time consuming, the option of including half-MDLs in the analysis was not pursued.

7.2 SELECTION OF A MATHEMATICAL INTERPOLATION MODEL TYPE

As shown in the Results section's **Table 6**, the k-bessel model type proved to be the most consistent and accurate model type for the test site, as it performed well in each of the three major measures of error: mean standardized error, RMSE and RMSSE. Of the 10 model types run, k-bessel boasted the RMSSE second-closest to a value of 1 and the third-lowest RMSE. It also had a mean standardized error that was the second-closest to zero. The hole effect model type was more accurate than k-bessel by every measure but average standardized error. But with an average standardized error nearly three times that of any other model type, it was decided that hole effect would not be used.

The k-bessel kriging model was run on the first contaminant, PCE, but in the process of writing a script to automate the kriging process, a more easily programmed model type called Empirical Bayesian Kriging, or EBK, was discovered. According to ESRI (2012), this model type is considered among the most accurate for spatial interpolation data, with a mean error an entire magnitude better than the other models, a high standardized error, less than half the RMSE of other models, and the lowest average standard error by far.

7.3 CALCULATED VOC MASSES AND THEIR IMPLICATIONS

While the lack of comparison to mass estimates from other West Basin studies means that the prediction values of this project mean little, the thin distribution of mass contamination among clusters suggest the utility of this model in setting remediation priorities. The total mass of each of the four contaminants can mostly be attributed to just a few clusters.

Chloroform presents the most drastic example of such a small mass distribution. With

5078 of the 5102 kilograms of total mass centered around a single cluster, over 99 percent of the predicted mass for that contaminant is contained in one group of well points.

Similarly, 92 percent of PCE mass was traced to 2 of 27 clusters sampling for that contaminant, while about 86 percent of TCE mass was accounted for by 3 of 29 clusters. A similar mass distribution among clusters is seen for MTBE. Three clusters' combined mass totals 86 percent of the West Basin's mass for that VOC. Refer to **Appendix F** for details on these calculations.

With this kind of information at their disposal, regulatory agencies may be able to address a considerable amount of pollution by targeting major inputs instead of any given cluster of contamination in the West Coast Basin.

7.4 Comparison to 2013 UCLA Model for PCE

Burke *et al.* (2013) generated a model that yielded a PCE estimate of West Coast Basin-wide 4383 kilograms. The mass prediction of this study is nearly 1700 kilograms smaller than that of Burke *et al.*, (2013) despite plume heights in this study reaching as far as 1000 feet underground, versus the Burke *et al.* volume integration over just 20 feet (2013). The conditions required for the automated kriging technique may explain some of the disparity in masses between the two projects, as many clusters were excluded from the analysis because of failure to meet the requirements for minimum average concentration requirement or well point count.

7.5 Expected Significance

A project such as this may be impactful for regulatory purposes because of how streamlined it makes the process of determining groundwater VOC mass within a given area. The automated kriging script combines GeoTracker GAMA and WRD data to make for an efficient means for regulators to determine where hot spots of pollution may be occurring within a basin. These tools could thus provide an opportunity for the better planning, coordination and priority-setting at the heart of efforts like GAMA.

This model can be run multiple times with consistent results, and, through the use of the kriging code, instantly turn concentration data from GeoTracker into mass. New datasets can be used, allowing for monitoring of progress over time, which can be difficult to do with concentration-based regulation.

Data such as that produced by this project helps to set a foundation for groundwater regulation that can examine contamination on a regional, as opposed to just a point-source, level. With agencies able to get a better idea of what inputs will cause excess pollution, mass-based calculations are more conducive to assessing the contamination of larger areas.

7.5-a Application Beyond the West Coast Basin

The WRD manages groundwater for both the West and Central sections of the Los Angeles Groundwater Basin, so that agency would likely have aquifer data of similar quality for both regions. Little may change in running the model on the Central Basin, as that basin and this project's study area mostly have unconsolidated soil. It can therefore be expected that properties of soil and its interactions with VOCs would be similar between the two regions, as well as among other locations with similar soil profiles. Differences in aquifer height may arise in that the West Coast Basin is mostly confined (with Layer 1 non-existent), while the Central Coast Basin is mostly unconfined (with Layer 1 present). But the model's calculation of aquifer height can be adjusted to accommodate that change when analyzing the Central Basin.

The assumption of unconsolidated soil probably would not hold for many other locations throughout the rest of the nation, and so major changes related to the calculation of aquifer storage and z-coefficients may be expected. Determination of aquifer depth may also change as studies get away from Los Angeles, as different basins throughout the country figure to have entirely different layer traits that cannot be accounted for in the current model. For instance, Layers 3 and 4 were entirely ignored in this study's calculations, a choice that may well not hold elsewhere. Other regions throughout the nation are also going to have different soil profiles and rock formations, as well as groundwater flow and recharge.

GeoTracker GAMA contains well data for that spans the whole of California, so this model can work within the state provided that one can find similar aquifer data for the calculation of pollutant volume. Agencies with data like that of the WRD may be easier to find in locations — like counties the Central Valley — that mostly rely on non-local water sources. Such water-starved regions could be expected to have more motivation to look into and characterize groundwater resources.

7.6 FUTURE DIRECTIONS

7.6-a Incorporating Residual Contamination into Mass Calculations

This project's model only calculates the amount of contaminants within groundwater, since the client only requested as much. But in reality, considerable sources of pollutants in the subsurface environment are created when contaminants, especially solvents like PCE and TCE, encounter layers of soil and rock that are of low permeability. Pollutant accumulates in a pool above the groundwater and over time contaminants are fed further downward into groundwater, complicating remediation efforts by requiring additional inspection of soil, as opposed to just water. Residual contamination was not included in the final calculations because of the limited knowledge concerning the amount of air in a given soil parcel and the effect of gaseous behavior on mass calculation within the soil. Improved mass estimate models would incorporate both groundwater and residual data, such as to inform the creation of more thorough remediation efforts.

7.6-b Considerations of Pollutant Fate and Transport

A thorough understanding of VOC fate and transport, including biodegradation, was not part of the project, given our client's instructions and our lack of expertise in accounting for such processes for even just a few contaminants. Considerations of the chemical environment and microbial community required to achieve differing rates of biodegradation would make for a more accurate model, and may take on the form of additional coefficients within **Eq. 1**. An early

concept idea for the model was to generalize entire VOC groups, such as chlorinated solvents, according to the behavior of an indicator contaminant or two. But generalizing to even that level proved to be beyond the scope of the literature review and project deadline, and so the mass estimation model does not include any alterations for different kinds of contaminants.

7.6-c Factoring in Dynamic Hydrogeologic Processes

The model does not account for how real-world changes in water volume over space and time alter pollutant plumes. In the original USGS model (Crawford *et al.*, 2003), the aquitards were not implicitly modeled between the layers, making it a quasi-3D model (Johnson and Njuguna). The extent and direction of groundwater flow can change the distribution of a plume and thus any mass calculations relying on that distribution. Water recharge from the likes of precipitation and runoff can also alter the amount of water and contaminant held within a specific portion of an aquifer, and consequently plume volume calculations. The incorporation of true 3D programs would help greatly in building a model that more accurately represents the behavior of contaminant plumes.

7.6-d Additions to GeoTracker

There exists the possibility that a policymaker using the tools provided by this project will experience some of the same database inefficiencies experienced during the course of this study, and be motivated to address those issues.

The database provides all inputs for the model, so any effort made to improve extraction of meaningful data from it makes the model more efficient and accurate. An especially helpful change in GeoTracker would be a column on the downloadable spreadsheet that lists the laboratory MDLs for a given contaminant and well. Such an addition could provide a better understanding of the mass-estimating model's relative error, under the assumption that a non-detect value equates to a zero concentration at that well point.

As is, many steps are required to access screen length information from the database. A more organized display of that data could be an appearance as another column on the spreadsheet for a downloaded contaminant. That feature would inform users on the distance a VOC travels underground by providing an understanding of the concentration at a particular depth relative to the depth-concentration profiles built in the model, allowing for refinement of the z-coefficient and thus improved generalization of z-axis changes.

7.6-e Visualize Model Results in a Way the Public Can Readily Understand

A suggestion made by the client at a meeting on June 12th, 2014 was to develop public outreach strategies that bring attention to groundwater contamination issues. One involved taking the massive pollutant estimations we generated, on the order of a million kilograms, and equating that mass of contaminant to a source a general audience may be familiar with, such as a battery.

Expansions of the LARWQCB's ideas include mapping out remediation costs of certain areas and delving into VOC studies that give people an idea as to how risk for different adverse health effects, such as cancer, increases with certain degrees exposure.

8. CONCLUSION

This study has produced a tool with the potential to be useful in allowing agencies to make more informed decisions about remediation and monitoring efforts of VOCs. The LARWQCB can thus better prioritize limited resources toward the most problematic areas and able to remediate more effectively.

Though much can still be learned about VOC exposure, these contaminants can negatively impact human health and so their presence in groundwater aquifers nonetheless hinders the Los Angeles Basin's ability to maximize its use of that water source.

In a region so reliant on distant water sources and susceptible to drought, the inability to expand a drinking water supply portfolio poses a major challenge to those charged with securing potable water for Los Angeles.

Most decisions regarding groundwater are still based on contaminant concentrations. Groundwater pollution management programs can become more effective by considering mass fluxes. A mass-based management considers variance caused by spatially unique hydrologic conditions. It also allows a kinematic approach and introduces dimensional aspects to plume monitoring. Mass calculation can be compared across sampling times to estimate the source and path of different contaminants. The data could then be used to deduce how chemical interactions between contaminants impact net flow and biodegradation. Regulations based on total mass allow water control programs to identify areas of high concern to prioritize the finite resources available for groundwater monitoring and risk assessment.

It should be noted that regional groundwater contamination models inherently contain a degree of uncertainty due to the reality that several complex underground phenomena influence the transport and fate of contaminants and parameter inputs into models rely upon measurements from discrete sampling points. While precise calculations of contaminant mass in smaller areas with limited data is challenging, the model presented here is well-suited for making reasonable estimates of total contaminant mass over the scale of regional water basins for the stated goal of informing the public, regulatory agencies and policy decisions.

9. WORKS CITED

Agency for Toxic Substances and Disease Registry, 2012. Toxic Substances Portal.

Asano, T., and Cotruvo, J. A. (2004). Groundwater recharge with reclaimed municipal wastewater: health and regulatory considerations. *Water Research*. doi:10.1016/j.watres.2004.01.023

Beckman, D., Blankenberg, W., Helperin, A., Inwood, D., Ledwith, V., April 2001. National Resources Defense Council. California's contaminated groundwater. New York, NY.

Belitz, K. and Fram, M.S., 2012. Groundwater quality in the Coastal Los Angeles Basin, California: U.S. Geological Survey Fact Sheet 2012-3096.

Belitz, K., Danskin, W.R., Milby Dawson, B.J., Land, M., 2000. Stable isotopes and volatile organic compounds along seven ground-water flow paths in divergent and convergent flow systems, Southern California, 2000. U.S. Department of the Interior, U.S. Geological Survey.

Belitz, K.; Fram, M.S.; Goldrath, D.; Land, M., 2006. Pollutant Selection Status of Groundwater Quality in the Coastal Los Angeles Basin, 2006: California GAMA Priority Basin Project. United States Geological Survey. Reston, Virginia.

Borkovich, J., 2012. Fact Sheet: Groundwater Ambient Monitoring and Assessment Program.

Bruguera, M. and Tsang, J., 2014. Estimating the mass of VOCs within the West Coast Los Angeles Groundwater Basin. UCLA Department of Geography, Course: GIS Programming and Development, Professor Yongwei Sheng, Teaching Assistant Evan Lyons.

Burke, S., Chu, G., Heyer, J., Lee, J., Tang, X., Tran, T., Zhang, L., 2013. Mass quantification of PCE in Los Angeles groundwater from the GeoTracker database. UCLA Institute of the Environment and Sustainability Senior Practicum. Los Angeles, CA.

Cahn, M. and Hartz, T. Load vs. concentration: implications for reaching water quality goals. University of California Cooperative Extension Monterey County, 2014.

Crawford, S., Everett, R.R., Halford, K.J., Johnson, T., Johnson, T.A., Kulshan, T.V. Land, M., Nishikawa, T., Paybins, K.S., Ponti, D.J., Reichard, E.G., 2003. Geohydrology, Geochemistry, and Ground-Water Simulation-Optimization of the Central and West Coast Basins. United States Geological Survey. Los Angeles County, CA.

Environmental Systems Research Institute, Inc. (ESRI), 2012. What is Empirical Bayesian Kriging?

Environmental Systems Research Institute, Inc. (ESRI), 2013. Cross validation (geostatistical analyst). ESRI, Redlands, CA.

Hamilton, P.A., Moran, M.J., Zogorski, J.S., 2006. United States Geological Survey. Volatile organic compounds in the nation's ground water and drinking-water supply wells – a summary.

Helsel, D. and Lee, L., August 2006. Analysis of environmental data with nondetects. American

Statistical Association. Seattle, WA.

Interstate Technology and Regulatory Council, 2002. DNAPL source reduction: face the challenge.

Johnson, T.A. and Njuguna, W.M. Aquifer Storage Calculations Using GIS and MODFLOW, Los Angeles County, California.

Legislative Analyst's Office, 1999. Supplemental Report of the 1999 Budget Act: 1999-00 Fiscal Year. Sacramento, CA.

Lidderdale, T., 2000. MTBE, oxygenates, and motor gasoline. Energy Information Administration.

Lyles J.R., 1998. Innovative technologies join the Superfund cleanup of ground water at Fort Lewis, Washington. United States Geological Survey Fact Sheet 082-98.

Matsumoto, N., 2009. Groundwater contamination prevention and cleanup in the Central and West Coast Basins. Water Replenishment District Technical Bulletin, Vol. 18. Lakewood, CA.

Moran, M. J., Price, C. V., Rowe, B. L., Toccalino, P. L., Zogorski, J. S., 2007. Occurrence and potential human-health relevance of volatile organic compounds in drinking water from domestic wells in the United States. NCBI. U.S. National Library of Medicine.

Pitt, R., February 2007. Effects of non-detected observations and selection of analytical methods. University Of Alabama. Tuscaloosa, AL.

State Water Resources Control Board, 2003. A Comprehensive Groundwater Quality Monitoring Program for California: Assembly Bill 599 Report to the Governor and Legislature. Sacramento, CA.

United States Environmental Protection Agency, 2012. Support document for the revised National Priorities List final Rule - Jervis B. Webb Co. Washington, D.C.

United States Environmental Protection Agency, 2013a. Basic information about disinfection byproducts in drinking water: total trihalomethanes, haloacetic acids, bromate, and chlorite.

United States Environmental Protection Agency, 2013b. Regulating public water systems and contaminants under the Safe Drinking Water Act.

United States Geological Survey, 2014. Volatile organic compounds in the nation's ground water and drinking-water supply wells: supporting information.

Varouchakis, E. A., Hristopulos, D. T., Karatzas, G. P., 2012. Improving kriging of groundwater level data using nonlinear normalizing transformations — a field application. *Hydrological Sciences Journal*, *57*(7), 1404-1419.

10. APPENDICES

Appendix A: Components of the GAMA Program (adapted from Borkovich, 2012)

Appendix B: Essential ROS Commands in R

Appendix C: Using R to Find Integrals for Maximum Concentration and Depth Versus Concentration in Determining a Contaminant's Z-coefficient

Appendix D: Summary of Priority Basin Report Information on Four VOCs of Interest (Adapted from Belitz *et al.*, 2012)

Appendix E: Creation, Use and Implications of Automated VOC Mass Estimate Method in Los Angeles West Coast Basin

Appendix F: Distribution of Mass Estimate Data Among Clusters

APPENDIX A
COMPONENTS OF THE GAMA PROGRAM
(ADAPTED FROM BORKOVICH, 2012)

GAMA Component	Function
1. Priority Basin Project (PBP)	Assess the public water supply quality of over 100 basins throughout California by comparing observed pollutant concentrations in untreated groundwater with drinking water benchmarks.
2. GeoTracker GAMA	Provides a free-to-access interactive map and online database of detection data from 200,000-plus California wells and over 100 million overall analytical results. Contributors include the Department of Water Resources, the Department of Pesticide Regulation, Special Studies Projects between GAMA and the Lawrence Livermore National Laboratory (LLNL). GeoTracker can be accessed here: http://geotracker.waterboards.ca.gov/gama/.
3. Domestic Well Projects	Examine the water quality of domestic, private well waters that typically serve single, home-owning families. This kind of water quality is not regulated by the California state government. Sampling effort targets include bacteria, minerals, and organic contaminants. Benzene, toluene, PCE, and MTBE are among the organic contaminants tested for by these projects. Participation is free and voluntary. Projects occur on a county-by-county basis, with Monterey, San Diego, Tulare, Tehama, El Dorado and Yuba counties studied as of 2011. The number of wells examined in each county currently ranges from 79 to 398. A description of results can be found here: http://www.waterboards.ca.gov/gama/domestic_well.shtml.
4. Special Studies Project	The Lawrence Livermore Lab conducts research on a variety of groundwater topics. Seven studies have been completed so far. Nitrate occurrence sources, management and fate and transport have been examined in the Llagas and Chico Basins, as well as in Orange County, Livermore, and Gilroy. Areas irrigated by recycled water and the effects of septic systems on shallow groundwater have also been studied. Current projects include researching groundwater recharge, developing extraction and collection tools for dissolved gases in groundwater samples, and establishing a biological assay conducted in fish to identify endocrine-disrupting chemicals. Completed and ongoing Special Studies Projects are summarized at the following website: http://www.waterboards.ca.gov/gama/special_studies.shtml.

APPENDIX B
ESSENTIAL ROS COMMANDS IN R

B.1 Removal of repeating measurements in R

```
PCE <- read.table("TCE_CLST.txt", header = TRUE)
# Import data into R
PCE$Well_ID_Name <- factor(PCE$WELLID):factor(PCE$WELLNAME)
# To combine the two categorical factors in one column
aggrPCE <- aggregate(PCE["RESULT"], by = list(PCE$Well_ID_Name), FUN = mean, na.rm = TRUE)
names(aggrPCE)[names(aggrPCE) == "Group.1"] <- "Well_ID_Name" names(aggrPCE)[names(aggrPCE) == "RESULT"] <-
"aveResult" PCE2 <- merge(PCE, aggrPCE, by = "Well_ID_Name") head(PCE2)
# Take average of repeating measurements result. Replace “result” with average and leave only one of the repeating
measurements (delete the rest)
PCE3 <- PCE2[!duplicated(PCE2$Well_ID_Name), ]
# remove the duplicated measurements
```

B.2 Running an ROS test in R

```
# IF there are less than 3 observations (>0), keep the observation and replace non-detect with half MDL or 0
# Add a new column to MD3, '1' # for positive result, '0' for Non-detect; PCE3$Calculate[PCE3$aveResult > 0] <- 1
PCE3$Calculate[PCE3$aveResult <= 0] <- 0
# Then, add the MD3$Calculate together for each group. This represents # number of positive measurements for each group
PCE3ROS1 <- aggregate(PCE3["Calculate"], by = list(PCE3$ID), FUN = sum, na.rm = TRUE)

names(PCE3ROS1)[names(PCE3ROS1) == "Group.1"] <- "ID" names(PCE3ROS1)[names(PCE3ROS1) == "Calculate"] <-
"Observation"
# Merge the sheet back to MD3
PCE3ROS2 <- merge(PCE3ROS1, PCE3, by = "ID")
# Created an indicator for whether to use MDL substitution PCE3ROS2$MDLtest[PCE3ROS2$Observation >= 3] <- FALSE
PCE3ROS2$MDLtest[PCE3ROS2$Observation < 3] <- TRUE
# Now the 'MDLtest' indicates which rows should be substituted with the Half # the MDL or 0 because there are <3 Positive
Observations with a cluster group ID PCE3ROS2$INDICATOR[PCE3ROS2$aveResult > 0] <- FALSE
PCE3ROS2$INDICATOR[PCE3ROS2$aveResult <= 0] <- TRUE
Group_with_concored <- as.data.frame(table(PCE3ROS2$ID[PCE3ROS2$MDLtest == FALSE & PCE3ROS2$INDICATOR
== TRUE])) names(Group_with_concored)[names(Group_with_concored) == "Var1"] <- "ID"
names(Group_with_concored)[names(Group_with_concored) == "Freq"] <- "no.censored" # Group ID and number of
observation that is non-detected Group_with_concored
Group_with_average <- aggregate(PCE3ROS2["aveResult"], by = list(PCE3ROS2$ID), FUN = mean) Group_with_average
names(Group_with_average)[names(Group_with_average) == "Group.1"] <- "ID" # Merge them together ROStest1 <-
merge(Group_with_average, Group_with_concored, by = "ID") ROStest <- merge(ROStest1, obs, by = "ID")
```

B.3 Running ROS in a loop in R

```
Ros = c()
x = ROStest$ID
for (i in 1:28) { Ros[i] <- mean(ros(PCE3ROS2$aveResult[PCE3ROS2$ID == x[i]], PCE3ROS2$INDICATOR[PCE3ROS2$ID
== x[i]], forwardT = "log", reverseT = "exp")) }
# Run ROS for each group that passed the test
```

B.4 Calculation of an ROS Substitution Value in R

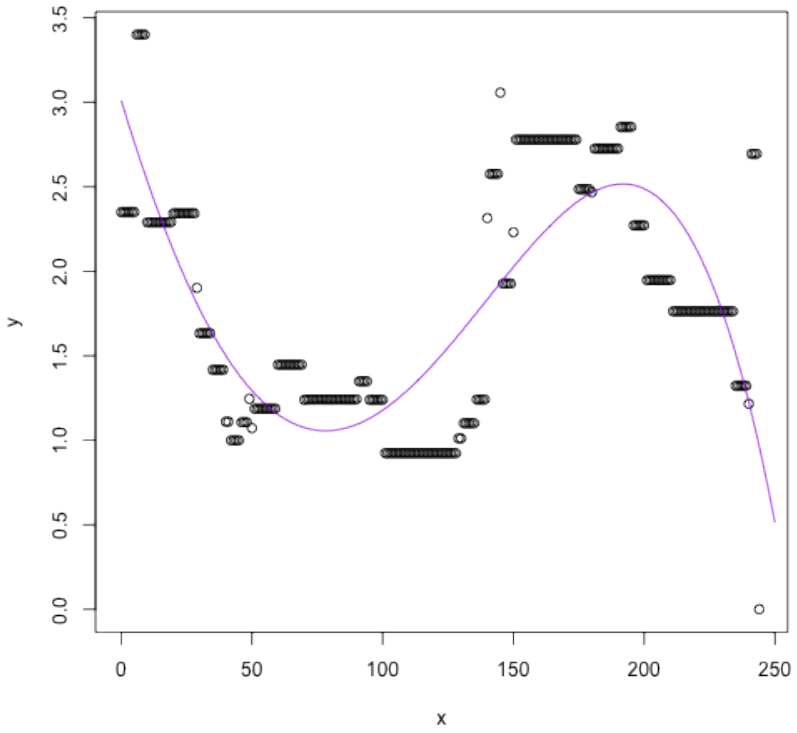
```
ROStest$ROSaverage = Ros ROStest$ROS = ROStest$ROSaverage - ROStest$aveResult ROStest$sub = (ROStest$ROS *
ROStest$Count_of_Obs)/ROStest$no.censored
```

APPENDIX C

USING R TO FIND INTEGRALS FOR MAXIMUM CONCENTRATION AND DEPTH VERSUS CONCENTRATION IN DETERMINING A CONTAMINANT'S Z-COEFFICIENT

```
Pcedata <- read.delim("~/Desktop/Practice_data/Vertical_estimation/Pcedata.dbf.txt")
# read data
Pcedata$sum = 0
data = seq(from = 0, to = 245, by = 1)
PCE = data.frame(data)
# creat a date frame for depth
PCE$sum = 0
PCE$count = 0
for (i in 1:245) {
   for (j in 1:27) {
      PCE$sum[i][PCE$data[i] <= Pcedata$min[j] & PCE$data[i] >= Pcedata$max[j]] = PCE$sum[i] +
        Pcedata$con[j]
   }
}
for (i in 1:245) {
   for (j in 1:27) {
      PCE$count[i][PCE$data[i] <= Pcedata$min[j] & PCE$data[i] >= Pcedata$max[j]] = PCE$count[i] +
        1
   }
}
# test and add overlap up together
PCE$concentration = PCE$sum/PCE$count
PCE$concentration[245] = 0
# calculate the concentration by dividing sum by number of value added
x = c(PCE$data)
y = c(PCE$concentration)
# assign to x,y for graphing and curve estimation
fit4 <- lm(y ~ poly(x, 4, raw = TRUE))
xx <- seq(0, 250, length = 100)
# 4th order estimation
plot(x, y)
lines(xx, predict(fit4, data.frame(x = xx)), col = "purple")
summary(fit4)
##
## Call:
## lm(formula = y ~ poly(x, 4, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9630 -0.3672 -0.0012  0.2494  1.6637
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.01e+00   1.35e-01   22.31  < 2e-16 ***
## poly(x, 4, raw = TRUE)1 -5.11e-02   7.70e-03   -6.64  2.1e-10 ***
## poly(x, 4, raw = TRUE)2  3.19e-04   1.29e-04    2.48   0.0139 *
## poly(x, 4, raw = TRUE)3  5.62e-07   7.94e-07    0.71   0.4796
## poly(x, 4, raw = TRUE)4 -4.72e-09   1.61e-09   -2.92   0.0038 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.433 on 240 degrees of freedom
```

```
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.615,  Adjusted R-squared:  0.609
## F-statistic: 95.8 on 4 and 240 DF,  p-value: <2e-16
library("MESS", lib.loc = "/Library/Frameworks/R.framework/Versions/3.0/Resources/library")
## Loading required package: geepack
```




```
# load the package for curve integrating
yy = predict(fit4, data.frame(x = xx))
max(yy)
## [1] 3.012
max(yy) * 250
## [1] 753.1
# maximum area
auc(xx, yy, from = 0, to = 250)
## [1] 443.4
# area according to 4th order estimation
auc(x, y, from = 0, to = 244)
## [1] 439.7
# area according to data
auc(xx, yy, from = 0, to = 250)/(max(yy) * 250)
## [1] 0.5887
# fraction
```

APPENDIX D

SUMMARY OF PRIORITY BASIN REPORT INFORMATION ON FOUR VOCS OF INTEREST

(ADAPTED FROM BELITZ *ET AL.*, 2012)

Contaminant	Report Information
1. Perchloroethylene (PCE)	Has a detection frequency of 25 percent, appearing at higher aquifer-scale proportions in 1.1 percent of the primary aquifer system, and at moderate proportions in 5.5 percent of the primary aquifer system.
2. Trichloroethylene (TCE)	Has a detection frequency of 29 percent, appearing at higher aquifer-scale proportions in 1.7 percent of the primary aquifer system, and at moderate proportions in 13 percent of the primary aquifer system.
3. Chloroform	Has a detection frequency of 45 percent. It was the most commonly detected VOC in 2001 study of LA Basin, 2003 study, national survey of groundwater VOCs.
4. Methyl tert-butyl ether (MTBE)	Found at moderate relative-concentrations in 0.4 percent of primary aquifer system. Detected in more than 10 percent of grid wells.

Groundwater quality was viewed in relationship to relative-concentration, a ratio of a sample concentration to a state, Federal or non-regulatory benchmark. Organic pollutants had three tiers of relative-concentrations:

1. Low (a relative-concentration less than or equal to 0.1),
2. Moderate (a relative-concentration greater than 0.1, but less than or equal to 1)
3. High (a relative-concentration greater than 1).

Aquifer scale-proportion, the percentage of the primary aquifer system's wells showing a high relative-concentration, was included in the analysis to help characterize the spread of a pollutant of interest.

APPENDIX E
CREATION, USE AND IMPLICATIONS OF AUTOMATED VOC MASS ESTIMATE METHOD IN LOS ANGELES WEST COAST BASIN

ESTIMATING THE MASS OF VOCS WITHIN THE WEST COAST LOS ANGELES GROUNDWATER BASIN

Maya Bruguera and Jeffrey Tsang
Evan Lyons, Dr. Yongwei Sheng
Geography 173: GIS Programming and Development

I. Problem

Over the past two centuries, industry has boomed in Los Angeles. Though this has brought the region's economy great benefits, it has also brought the region's groundwater numerous contaminants. One group of contaminants known to be harmful to human health, are volatile organic compounds, or VOCs. Because they negatively impact human health, the presence of VOCs in groundwater aquifers prevents the region from using this water resource to supply residents with potable water. In an area like Los Angeles where drought is ongoing and expected to persist throughout the next decade, the inability to take advantage of local groundwater resources to fortify the region's water supply portfolio poses a challenge to those in charge of securing potable water for Los Angeles.

II. Solution

Because of the existing contamination and need to explore local groundwater for water supply, the Los Angeles Regional Water Quality Control Board (LARWQCB), the agency which regulates groundwater pollution under the Clean Water Act, has begun looking into this contamination. In order to clean up the contamination and assess the risk to human health, the Board has set the goal of quantifying the mass of VOCs within the West Coast Los Angeles Basin. To do this, the LARWQCB has contracted out and asked a UCLA Senior Environmental Science Practicum team to conduct this mass analysis.

Over the past two years, practicum teams investigated the nature of groundwater contamination in the West Coast Basin and designed a methodology for estimating mass. This methodology involves a number of steps.

1. Downloading geospatial information on groundwater VOC concentrations from the state's Groundwater Ambient Monitoring and Assessment (GAMA) GeoTracker database for the groundwater basin. This data is downloaded as a csv.
2. Importing the csv of concentration data into ArcMap as an XY layer
3. Converting all non-detect values (aka zero-values) to the minimum positive value allowed by ArcMap (0.00001) to enable input into kriging
4. Grouping this data into clusters based on spatial proximity to increase the accuracy of the Kriging spatial interpolation

5. Performing the Kriging spatial interpolation on each cluster of the data to create a surface VOC concentration raster for the polluted regions within the basin
6. Multiplying each cluster by the aquifer depth and storage coefficient (aka the area of the cluster as determined by the latitudinal and longitudinal extent of the spatial interpolation raster pixels, multiplied by the depth of the aquifer as determined by a summation of the depth of all aquifer layers, multiplied by the storage coefficient which is representative of the volume of the aquifer actually filled with water as opposed to by sediment), and by the mass coefficient which accounts for the attenuation of concentration with depth, to give total pollutant mass in the cluster
7. Creating a table which includes data from each cluster including cluster ID, total concentration within the cluster, the number of cells within the cluster, the aquifer storage value within the cluster (depth times storage coefficient), and the total pollutant mass within the cluster

There is a large number of VOCs of interest in the West Coast Basin and a just-as-large number of steps involved in calculating the mass for each pollutant, (including conducting Kriging separately for each of the 153 clusters and multiplying each of the clusters separately by the aquifer depth and storage**).** Because of this, it is not time efficient to conduct the mass estimation by hand for each pollutant in ArcMap. Therefore, we created a tool and a code which go together to take the csv of concentrations of a given pollutant, a cluster polygon file, an aquifer storage raster, and a pollutant mass coefficient as inputs, and outputs Kriged concentration rasters for each cluster, and a table which characteristics of each cluster, including total number of well monitoring points, number of prediction cells (meaning area), a mass output value, and the sum of the concentration values of all pixels within the cluster. Together, the Clustered Kriging tool and Mass Output code enable anyone to download pollutant data from GAMA GeoTracker, use the cluster polygon file, aquifer volume raster, and data on mass coefficients as inputs as the model, and calculate the mass of any pollutant on the GAMA GeoTracker database within the Los Angeles West Coast Basin.

III. Code Design

The following sample code encompasses all steps of the mass calculation, using tetrachloroethene (PCE) data as the sample pollutant input.

The first portion of the code corresponds to Step 2 in the mass calculation methodology, where the pollutant concentration csv is imported into ArcMap as a point file. This involves taking the GAMA GeoTracker pollutant concentration csv and importing it to ArcMap as an XY layer based on the x- and y-coordinates. This step creates and saves a point layer in the WGS 1984 projection which contains data on pollutant concentration at each well point in the field *aveResult*. The code then converts this layer to a shapefile. (Figure 1) In the final code used for the tool, user can specify the pollutant concentration csv as an input, which in the code is substituted for the variable *in_Table*.

```
#Setting the workspace
import arcpy
folder_path = r"F:\Script_Data"
arcpy.env.workspace = folder_path
arcpy.env.overwriteOutput = True

#Create PCE XY Shapefile from GAMA Geotracker pollutant csv
try:
    # Set the local variables
    in_Table = "PCE_ROS.csv"
    x_coords = "LONGITUDE"
    y_coords = "LATITUDE"

    out_Layer = "PCE_layer"
    saved_Layer = "PCE.lyr"

    # Set the spatial reference
    spRef = r"Coordinate Systems\Projected Coordinate Systems\World\WGS 1984.prj"

    # Make the XY event layer...
    arcpy.MakeXYEventLayer_management(in_Table, x_coords, y_coords, out_Layer, spRef)

    # Save to a layer file
    arcpy.SaveToLayerFile_management(out_Layer, saved_Layer)

    #Convert to a shapfile
    inFeatures = ["PCE.lyr"]

    arcpy.FeatureClassToShapefile_conversion(inFeatures, folder_path)

except:
    # If an error occurred print the message to the screen
    print arcpy.GetMessages()
print "done"
```

Figure 1. Code which sets environmental variables and imports the pollutant concentration csv into ArcMap as an XY point layer, saves the layer, and exports it to a shapefile. If there are any errors, the code prints the error messages.

The following section of the code corresponds to Step 3 of the methodology, where all non-detect concentration values (aka all well concentration result values of zero) are converted to 0.00001, to allow the data to be inputted into the Kriging (Kriging does not except any zero values). The code takes the pollutant XY shapefile that was just created, creates an update cursor, and loops through each well point, replacing any concentration result values that are zeros with 0.00001 (Figure 2).

```
### The resulting PCE file
fc = "PCE_layer.shp"

###Clusters
clst = "clusters.shp"
PCE_clust = "PCE_clust.shp"

### Create update cursor
rows = arcpy.UpdateCursor(fc)
for row in rows:
    if row.aveResult ==0:
        row.aveResult = 0.00001 # Change the value in value field
        rows.updateRow(row)


del rows #delete cursor
del row #delete variable

print "Zero-values converted to 1e-6"
```

Figure 2. Code which converts pollutant concentration non-detects (zero-values) to smallest possible value (0.00001)

Next comes Step 4, where the XY point file is joined to the input polygon cluster file to assign each point a cluster ID (Figure 3). This enables Kriging to be performed on each individual cluster.

```
###Spatially join XY pollutant points to cluster polygon file to assign cluster ID

arcpy.SpatialJoin_analysis (fc, clst, PCE_clust, "", "", "", "", "", "")

print "Well points clustered"
```

Figure 3. Code that assigns a cluster ID to each point.

Step 5 follows, where each cluster of the XY pollutant point file is Kriged using the Geostatistical Analyst Empirical Bayesian Kriging function. In the code, pre-Kriging housekeeping is done, including checking out the Geostatistical Analyst extension, converting the shapefile to a feature layer, and defining the variable *clstID* as the attribute that defines the cluster ID (Figure 4). The Kriging variables are then set, where cell size is set to 10 meters, the data transformation is set to an empirical one, the maximum number of points per semivariogram is set to 20, the maximum number of semivariograms to be calculated is set to 300, the maximum radius is set to 3000 meters (based on the diameter of the largest cluster), the Kriging type is set to smooth (meaning a surface is created only for the areas surrounding the points, not for a larger area), and the output type is set to *prediction* meaning the that output raster is a prediction of concentration based on the input data points (as opposed to error, etc.) (Figure 4).

```
#Check out Geostatistical Extension Liscence
arcpy.CheckOutExtension("GeoStats")

#################### Input & Output Variables ##############################
#Inputs
saved_Layer = "PCE_clust.shp"
clstID = "ID"
#Outputs
PCE_clst_no0s = "PCE_Layer"

#################### Create Feature Layer ##################################################
#Make a feature layer
arcpy.MakeFeatureLayer_management (saved_Layer, PCE_clst_no0s)

#################### Set Kriging Variables ##################################
#Set local variables
#inPointFeatures = PCE_clst_no0s
zField = "aveResult"
cellSize = 10.0
transformation = "EMPIRICAL"
maxLocalPoints = 20
overlapFactor = 0.5
numberSemivariograms = 300

# Set variables for search neighborhood
radius = 3000
    #Note: For PCE, largest side of largest cluster polygon is 3000 m
    #Largest cluster has 86 points
    #Average cluster has 6.85 points (based on mean count in spatial join of points to
cluster polygons)
smooth = 0.0
searchNeighbourhood = arcpy.SearchNeighborhoodSmoothCircular(radius, smooth)
outputType = "PREDICTION"
quantileValue = ""
thresholdType = ""
probabilityThreshold = ""
```

Figure 4. Pre-Kriging housekeeping (check out Geostatistical Analyst and convert shapefile to layer file) and setting Kriging variables.

After setting the Kriging variables, the code runs Kriging for each cluster. Because a number of clusters contain only non-detects, and Kriging will not work when this occurs, we created a mechanism such the code will only attempt to execute Kriging on clusters which have enough non-zero values. To do this, the code selects the points within the cluster and creates a temporary table and populates it with the mean concentration result of the entire cluster. It then creates a search cursor which accesses the table and if the average concentration result exceeds 0.1, it performs kriging on the cluster. The final output of this step is a a Kriging raster of each cluster which has enough non-zero points for Kriging to be performed (Figure 5). This concludes with the portion of the code in the first tool, Clustered Kriging.

```
########################## For each cluster, run Kriging ########################

#Run kriging
for clst in range(1,153):
	arcpy.SelectLayerByAttribute_management (PCE_clst_no0s, "NEW_SELECTION", '"ID" = ' +
str(clst))
	result = int(arcpy.GetCount_management(PCE_clst_no0s).getOutput(0))
	krigedLayer = "kriged"+str(clst)+".TIF"
	krigedRaster = "kriged"+str(clst)+".TIF"
	table = "table1"
	arcpy.CreateTable_management("in_memory", "table1")
	arcpy.Statistics_analysis(PCE_clst_no0s, "table1", [["aveResult", "MEAN"]])
	search = arcpy.SearchCursor ("table1")
	row = search.next()
	if result > 0:
	print str(clst)+", "+str(result)+", "+str(row.MEAN_AVERESULT)
	clst_list.append(str(clst)+", "+str(result)+", "+str(row.MEAN_AVERESULT))
	if row.MEAN_AVERESULT > 0.1:
		if result > 8: #8 chosen because must have enough data for EBK to run
		#print "Cluster " + str(clst) + " has " + str(result) + " points"
		 arcpy.EmpiricalBayesianKriging_ga(PCE_clst_no0s, zField, krigedLayer,
		 krigedRaster, cellSize, transformation, maxLocalPoints, overlapFactor,
		 numberSemivariograms, searchNeighbourhood, outputType, quantileValue,
		 thresholdType, probabilityThreshold)
		 print "Kriged"
	del result, row, search, table
print "Done kriging all clusters"
print clst_list
```

Figure 5. Kriging is run on each cluster whose mean exceeds 0.1, outputting a raster for each cluster.

The entire second portion of the tool is dedicated to Steps 6 and 7 of the mass calculation methodology, where mass is actually calculated and information on each cluster is outputted into a table. This part of the tool is a code to be run in IDLE instead of a toolbox in ArcMap, as the functionality (in terms of amount of time to execute and magnitude of individual values which can be processed) is available in IDLE but was found to be difficult and time inefficient in ArcMap.

In this portion of the code, first, the aquifer storage value is identified. Because the aquifer storage cells are a half mile by half mile, while our kriged cluster raster cells are only 26 meters by 26 meters, and oftentimes the entire cluster doesn't even span an entire half mile, the raster cell was too large to conduct a raster calculator style multiplication of the kriged and aquifer storage rasters to calculate mass. Instead, the centroid coordinates of each cluster were determined, and the aquifer storage layer value at that specific location was found, and used as the aquifer depth and storage value for the entire cluster. Figure 6 illustrates the code that executed this process.

```
######################### Setting the workspace ##########################
import arcpy
folder_path = r"F:\Script_Data"
arcpy.env.workspace = folder_path
arcpy.env.overwriteOutput = True

######################### Get Cluster Centroid Coordinates ########################

#Define variable of cluster polygon file
##USER SPECIFIED INPUT: Cluster shapefile
#Example: "clusters.shp"
clst = "clusters.shp"

#Create search cursor to look through polygon file
cluster_cursor = arcpy.SearchCursor(clst)

#Lists
clstID_list = list()
clstCentXY_list = list()

for clstG in cluster_cursor:
        #Get geometry of cluster polygon centroids
        clstGeom = clstG.Shape
        clstCentX = clstGeom.centroid.X
        clstCentY = clstGeom.centroid.Y

        #Populate lists with cluster polygon IDs and centroid coordinates
        clstID_list.append(clstG.ID)
        clstCentXY_list.append(str(clstCentX)+" "+str(clstCentY))

del clstG

#print "cluster ID list: "
#print clstID_list
#print "cluster centroid coordinate list: "
#print clstCentXY_list

######################### Get Aquifer Storage Value at Cluster Centroid Coordinates
########################
#Create empty list for aquifer storage values
AqStor_list = list()

#USER SPECIFIED INPUT: aquifer storage raster
storage = "storage_pm"

#Populate list with values
for i in clstCentXY_list:
        #print i
        storageVal = arcpy.GetCellValue_management(storage, i)
        storageValInt = float(storageVal.getOutput(0))
        #print storageVal
        AqStor_list.append(storageValInt)
#print AqStor_list
print "done"
```

Figure 6. Cluster centroid coordinates were added to a list and accessed through the cluster polygon geometry. These coordinates were then used as inputs to the get the value of the aquifer storage at that specific point.

The subsequent portion of the tool takes the kriging concentration data, aquifer depth and storage data, and multiplies them by the area of the kriged raster and other coefficients and unit conversions to produce a mass estimate of kilograms per cluster. The code not only calculates pollutant mass for each cluster and outputs it printed in the IDLE window, but also generates a table (mass_table) which is populated with information about each cluster's ID, total concentration value, aquifer depth and storage value. Figure 7 illustrates the mass calculation and generation of the final mass table.

```
################ Multiply Kriged Rasters by Aquifer Storage Volume ###################

#For all kriged rasters, perform raster calculator
rasterlist = arcpy.ListRasters('*', 'TIF')
#Create empty list to add total mass of all clusters
mass_list = list()
#Create table for cluster concentration, cell number, and mass information
##In tool, would be USER SPECIFIED mass output table
mass_table = "Mass_Table"
arcpy.CreateTable_management(folder_path, mass_table)
#Create fields for cluster ID, total concentration, cell number, aq storage, and mass
arcpy.AddField_management(mass_table, "Clust_ID", "DOUBLE")
arcpy.AddField_management(mass_table, "Sum_Conc_ugL", "DOUBLE")
arcpy.AddField_management(mass_table, "Num_Cells", "DOUBLE")
arcpy.AddField_management(mass_table, "Aq_Storage", "DOUBLE")
arcpy.AddField_management(mass_table, "Mass_kg", "DOUBLE")

#Import numpy module
import numpy
#Create insert cursor to populate table

for krigedRas in rasterlist:
        #Determine cluster ID of kriged raster
        num = krigedRas.index('d')
        numP1 = num + 1
        cluster = int(krigedRas[numP1:-4])
        print "Cluster "+str(cluster)

        #Create array from kriged raster to get cell values
        myArray = arcpy.RasterToNumPyArray(krigedRas)
        #sum array rows to create array of total concentration value for each row
        sum1 = sum(myArray)
        #sum array of total row concentrations to get total concentration of kriged raster
        sum2 = sum(sum1)
        print "        Sum of surface concentrations: "+str(sum2)+" ug/L"

        #Calculate number of cells in kriged area
        kriged_width = len(myArray)
        kriged_length = len(sum1)
        kriged_cell_num = kriged_width*kriged_length

        #Get aquifer storage cluster index
        #Get index value of kriged raster cluster in list of cluster IDs
        clusterID_index = int(clstID_list.index(cluster))
        #print "       Cluster ID List Index: "+str(clusterID_index)
        #Get aquifer storage of indexed cluster
        aqStor = AqStor_list[clusterID_index]
        print "        Cluster aquifer storage: "+str(aqStor)
```

```
#Calculate mass of cluster
#Explanation of mass calculation: mass = concentration [ug/L] * aq storage w/ depth[m] * XY
Cell area [m^2/cell] * # cells (kriged_cell_num from array calcs) [cell]* mass coefficient
(unitless) * 1000 (cubic meters to liter conversion) [L/m^3] * 1/1000000000 (micrograms to
kilograms conversion) [kg/ug]
        #####in tool, would be USER INPUT MASS COEFFICIENT, as it varies by pollutant
        mc = 0.642
        mass = sum2*aqStor*22.26*22.26*kriged_cell_num*mc*1000/1000000/1000
        print "        Cluster mass: "+str(mass)+" kg"
        mass_list.append(mass)

        #Update Table Rows
        rows = arcpy.InsertCursor(mass_table)
        row = rows.newRow()
        row.Clust_ID = cluster
        row.Sum_Conc_ugL = sum2
        row.Num_Cells = kriged_cell_num
        row.Aq_Storage = aqStor
        row.Mass_kg = mass
        rows.insertRow(row)

        del rows, row, aqStor, clusterID_index, cluster, myArray
print "done"

#Sum list of mass by cluster to get total Basin Pollutant Mass
tot_mass = sum(mass_list)
print "Total Pollutant Mass in West Coast Basin: "+str(tot_mass)+" kg"
```

Figure 7. For each kriged cluster raster, arrays are used to calculate the total concentration per cluster and the total number of cells in each cluster. These values are then used along with the aquifer storage and mass coefficient to produce a mass estimate. A table with all of this information is created and an insert cursor is used to populate it.

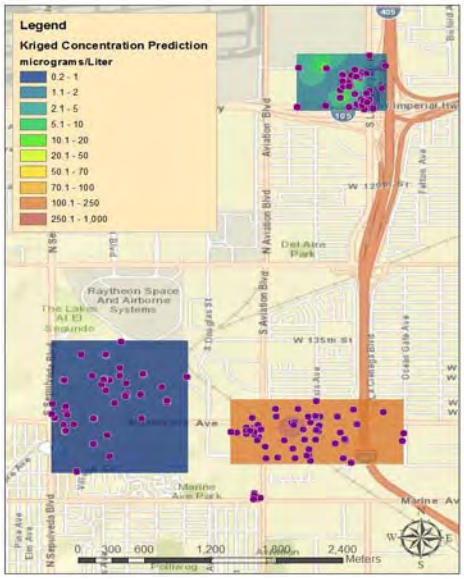
The end of the python shell window output from this step appears as follow, including a final mass estimate for the basin. (Figure 8)

```
Cluster 147
       Sum of surface concentrations: 3607.20375061 ug/L
       Cluster aquifer storage: 3.9886036
       Cluster mass: 823.8523862 kg
Cluster 149
       Sum of surface concentrations: 392771.416016 ug/L
       Cluster aquifer storage: 4.0257821
       Cluster mass: 1071408.50723 kg
Cluster 152
       Sum of surface concentrations: 103.313833192 ug/L
       Cluster aquifer storage: 2.9076476
       Cluster mass: 65.3644625966 kg
done
Total Pollutant Mass in West Coast Basin: 1670044.26144 kg
```

Figure 8. This figure illustrates the last few lines of Python Shell window output for the Mass Output code, including properties of each cluster (which are outputted into the table as well) and a total mass estimate for the basin.

IV. Tool Functionalities

Outputs



The tool set builds on itself, as the Clustered Kriging tool produces a prediction surface concentration raster, and the Mass Output tool uses these generated kriging rasters to compute total mass.

The Groundwater Pollutant Clustered Kriging outputs a shapefile of the pollutant points which have had non-detects transformed to 0.000001 and been assigned a cluster ID, and a concentration prediction raster of each cluster which was kriged. Figure 6 (to the right) illustrates an example of the output kriged concentration rasters.

When run in the IDLE Python shell, the Clustered Kriging code outputs a list, where each item includes a cluster's cluster ID, number of well points it contains, and average pollutant concentration result. This is advantageous because allows the user to see which clusters were not kriged and therefore will not be included in the total mass estimate, and why they were left out

(as detailed by the EBK code, either there was not enough variance within the data as determined by the mean concentration having a value less than 0.1) or there were not enough data points within the cluster to conduct kriging (the minimum number required for the tool to work for all clusters was 8). From the outputted kriged prediction rasters and the information on the data left out, the user can identify where surface pollutant concentration is highest and where it is monitored yet we lack enough data to accurately predict via kriging spatial interpolation. This tool could be improved by outputting the cluster ID, well point count, and average pollutant concentration result information into a table, however we did not have enough time during this project to pursue this avenue.

Regarding the functionality of the Mass Output code, it is advantageous as it provides spatial information on pollutant concentration and mass, enabling stakeholder parties to be aware of where groundwater pollution is most prevalent.

V. User Manual

Note: README.doc contains information on organization of data.

1. Groundwater Pollution Clustered Kriging Tool
   User must specify five inputs and outputs including:
   - Inputs
     - Pollutant CSV: Pollutant data from GAMA GeoTracker, csv file
     - Well Clusters: Input well cluster polygon file
   - Outputs
     - Pollutant layer: output pollutant layer, layer file
     - Shapefile output location: output location of pollutant shapefile
     - Pollutant shapefile: output pollutant shapefile

2. Mass Output Code

In this code, total mass of a given pollutant in the West Coast Basin is calculated. The inputs are the cluster polygon file and the kriging rasters generated in the Clustered Kriging Tool (therefore

it requires that this tool have been run first). The only variable the user would ever need to change or specify is the mass coefficient value, (line 111) which can easily be doing by typing it into the code. This code will output a table populated with values for each cluster of ID, number of points, number of cells, and total mass, and will print these values as text on the python shell.

APPENDIX F
DISTRIBUTION OF MASS ESTIMATE DATA AMONG CLUSTERS

Shown in each chart are the masses contained in the 10 clusters with the most mass for each pollutant. In the case of chloroform, the automated model only used 9 clusters to find the total mass within the West Coast Basin. A large proportion of each pollutant's total mass in the study area can be attributed to just a few clusters. Three clusters or less accounted for at least 85 percent of estimated basin-wide mass for each pollutant.

F.1 PCE

PCE				
CLUST_ID	MASS_KG	Total Mass of Two Clusters with Most Mass	Total Mass of All Clusters	Proportion of Total Mass In Two Clusters with Most Mass
27	2.02633188	**2482.632422**	**2699.35**	**0.919714903**
114	2.054935616			
23	2.079826271			
130	4.058466869			
101	19.32807218			
4	28.85100592			
68	71.87036106			
66	74.27833485			
132	**373.9398438**			
95	**2108.692579**			

F.2 TCE

TCE				
CLUST_ID	MASS_KG	Total Mass of Three Clusters with Most Mass	Total Mass of All Clusters	Proportion of Total in Three Clusters with Most Mass
117	17.11653	**6585.468172**	**7657.444475**	**0.860008609**
4	17.437444			
119	33.688218			
114	39.166069			
101	143.30008			
63	303.35502			
132	489.13315			
68	**798.95878**			
27	**1985.9033**			
95	**3800.6061**			

F.3 Chloroform

Chloroform				
CLUST_ID	MASS_KG	Total Mass of Cluster with Most Mass	Total Mass of All Clusters	Proportion of Total Mass in Cluster with Most Mass
72	0.0194298	**5078.420824**	**5102.634752**	**0.995254623**
117	0.0248782			
46	0.0441163			
4	0.2544996			
27	1.0324321			
132	1.5583655			
67	3.0083977			
119	18.271809			
63	**5078.4208**			

F.4 MTBE

MTBE				
CLUST_ID	MASS_KG	Total Mass of Three Clusters with Most Mass	Total Mass of All Clusters	Proportion of Total Mass in Three Clusters with Most Mass
146	49.75288	**10192.56043**	**11872.6651**	**0.858489677**
106	49.99517			
33	55.876359			
27	78.493553			
76	156.50767			
96	401.12557			
112	737.68981			
129	**1610.8778**			
151	**3918.45**			
135	**4663.2326**			