Tools to predict the hydrological response and mine pool formation in underground mines

OSM Cooperative Agreement Number: S16AC20072

Final Report

October 1, 2016 – March 31, 2019

Dr. Natalie Kruse, Dr. Dina Lopez, Jen Bowman, Nora Sullivan, Rebecca Steinberg, Lindsey Schafer, Fred Twumasi, Zachary Matthews

June 2019

OSM Award Number S16AC20072

Ohio University
1 Ohio University
Athens, Ohio 45701

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

TABLE OF CONTENTS

Abstract

Contaminated discharges can result from flooding of underground mines and acid forming chemical reactions and subsequent discharge to surface water. There is a need for more data-based prediction methods of post-mining water level for use in continued permitting of lands for coal mining. Under the Surface Mining Control and Reclamation Act (SMCRA), coal companies are required to estimate the post-mining water levels to determine if a mine pool will form and if there may be a pollutional discharge, but there lacks a consistent, data-based method for determining the hydrologic response to mining.

This project sought to address the gap in prediction by analyzing parameters of mine pool formation in post-SMCRA mines through multivariate analyses. Analyses were done in both the Unscrambler X, for multivariate statistical analysis, and Neuroshell, for artificial neural network modeling. An algorithm produced in Neuroshell, an artificial neural network program, resulted in the least amount of error and was incorporated into a tool for modeling post-mining potentiometric head elevation through ArcGIS Pro model building function. The predictive tool developed in ArcGIS Pro was made to output points of predicted post-mining water levels in the coal bearing region of Ohio. The tool only requires input of data that would be required for an underground mine permit application. This project’s final output is an empirically predictive ArcGIS tool that is publicly available for download to be used as a new approach to science-based estimation of underground mining effects on area hydrology. Methods used to develop both the algorithm and the tool in ArcGIS Pro can be used in other coal bearing regions around the world to develop a similarly useful tool for understanding connections between hydrology and underground mining.

List of Figures

List of Tables

Executive Summary

The purpose of the project was to develop empirical, multivariate relationships between water level and mining, geologic, and geographic variables in mining areas and to use that relationship to develop a GIS tool that can predict post-mining water level based on pre-mining data. This addressed the challenge of estimating probable hydrologic consequences of mining. Underground mining can have a profound effect on groundwater levels, flow rates, and directions. These complex systems do not lend themselves to deterministic prediction methods. Instead, this project undertook a multivariate empirical approach using over 2000 water level measurements in and around post-SMCRA underground coal mines in Ohio for model development.

A geodatabase was developed for the coal-bearing region of Ohio that includes pre- and post-SMCRA coal mines areas, terrain, coal elevation and thickness, overburden thickness, and composition of overburden. Additionally, precipitation and the amount of coal mined at the particular time when the head was measured served as key variables in the analysis. These data were then analyzed using two methods: multivariate statistical analysis and artificial neural networking. The multivariate statistical analysis was conducted in the Unscrambler X and included PCA, PCR, and PLSR. The three methods suggested similar weighting of variables, the error of prediction was the lowest with PLSR, however, may have still been as high as 4%. Artificial neural networking (ANN) develops complex empirical relationships between dependent and independent variables. Using ANN, a predictive algorithm with only 1% error was derived. The algorithm was selected from several options which varied in the inclusion of variables, complexity and error level. One algorithm that balanced these factors was selected as the predictive algorithm. This algorithm includes multiple variable transformations and mathematical functions. It is purely empirical and, it suggests a weighted importance of variables which was consistent with the multivariate statistical analysis.

To achieve the goal of creating a GIS tool to use the selected algorithm to predict post-mining water level, a tool was created using the Model Builder function in ArcPro with some custom Python code. Model Builder is a robust model development platform. However, it requires tight control on data format and occasionally mishandles memory, so that a reboot is required. Template input files were developed to allow a future user to input data in a format that will allow the tool to function. Within the tool, several variables are calculated or transformed. For example, the area of mines within a 4-mile buffer of the study mine is calculated to represent the interconnectedness and change in hydraulic conductivity of the area. The tool was compiled and packaged for download from the project website: http://www.watersheddata.com/MinePool_Study.aspx. In testing against known piezometric measurements in Ohio, error remained low, 1.24%.

The project website includes the model, a user’s guide, fact sheets, links to two project webinars, and links to three Masters theses developed through the project.

Introduction

Worldwide, pollutional discharges from coal mining have been and continue to be an environmental issue (Younger, 2000; Underwood et al., 2014; Lottermoser, 2015). Coal extraction has been a dominant industry providing the United States with energy since the 1800s, and with this long-term extraction comes long term environmental degradation (Crowell, 2005). Underground mining can affect surface water in the area of the mine through alteration of the local hydrology and formation of mine pools that can discharge to the surface. Coal extraction can result in a variety of chemical reactions with the minerals previously underground in anoxic conditions, now exposed to atmospheric conditions. Acid mine drainage (AMD) has been, and continues to be, a major environmental threat in the eastern U.S. in areas with a history of coal mining. Research in recent years has focused on remediation techniques (e.g. Wei et al., 2017). In addition to the focus on remediation, the high complexity of the system variables influencing AMD generation limits progress on research of such systems. Thus, research on the prediction of AMD discharge is sparse, resulting in a lack of understanding the systems and influences of this environmental hazard. Reliable prediction of the formation of mine pools and the possibility of acid generation at the permit level would prevent initial degradation and remove the need for and cost of remediation efforts. Research into understanding major variables determining the formation of underground mine pools and their discharge to the surface is needed in order to propose updated methods and tools for decreasing environmental harm from continued coal mining.

Study Area: Coal Bearing Region of Ohio

The eastern portion of Ohio sits in the Appalachian basin which, along with Pennsylvania and West Virginia, host the Appalachian coal field. The coal formed in this area of the country is high in sulfur content and thus not as pure or high quality as western coal (Crowell, 2005). The elevated sulfur and common occurrence of pyrite (FeS_2) are highly reactive when in contact with atmospheric oxygen and water, resulting in AMD. The occurrence of AMD can form naturally from mineral exposure to the surface, but also commonly forms during mining activities, both surface and underground. The effects of surface mining can be mitigated through capping waste piles, preventing water run-off, and diverting flow from passing through the mine area where minerals are exposed (Akcil and Koldas, 2006). Pollution from underground mining is difficult to mitigate, as AMD forms when a mine potentially fills with water and air, while sealing and/or collapsing mines does help reduce possible acid generation, it does not remove the possibility (Singer and Stumm, 1970).

This project focuses on post-SMCRA underground coal mines in eastern Ohio, shown in Figure 1. The Surface Mining Control and Reclamation Act (SMCRA) of 1977 made it mandatory for companies to obtain a permit prior to coal mining. The applications for mine permits require providing plans and finances for environmental protection and reclamation if pollution were to occur in the permitted area. Companies are required to determine the probable hydrologic consequences of the mine in the mine permit as part of the plans for environmental protection. If a mine pool is determined likely to form, and that the pool may form a pollutional discharge to the surface, the permit application may be denied.

Mine permit applications require coal companies to provide a characterization of area geology and hydrology. This study of hydrology includes determining where the water will rise to within the mine void post-mining, determining if the mine will form a mine pool and have the possibility of creating a pollutional discharge to surface waters. Mining companies in the Appalachian coal field do not have a strong science-based method for determining post-mining water levels and currently use the top elevation of the coal seam being mined as an estimate for post-mining water levels.

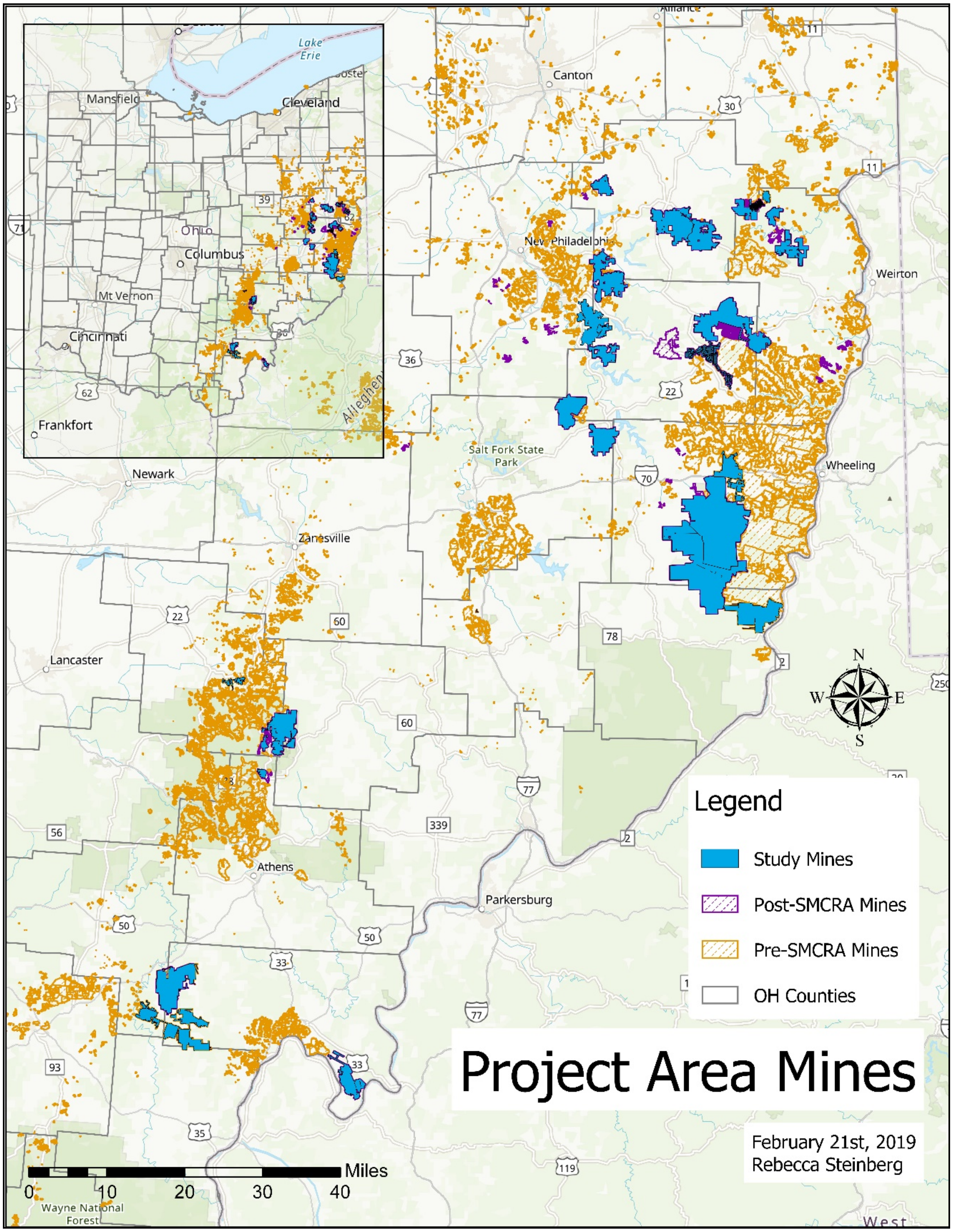


Figure 2 - Map of study area with pre-SMCRA, post-SMCRA and specific post-SMCRA study mines highlighted that are the focus of this project

Project Goals

The main research question addressed is *Can post-mining water level be predicted, within acceptable error, through multivariate analysis of hydrologic and geologic parameters and spatial interpolation?* This question is approached in several stages; first through testing approaches for multivariate analysis to determine relationships of hydrologic and geologic parameters and develop an algorithm to predict post-mining water levels using these relationships, and secondly through applying spatial interpolation methods for creating a surface of predicted post-mining water levels based on point predictions. These two approaches are finally incorporated in a user-friendly ArcGIS tool that automatically runs point predictions and risk areas as part of the on-going OSMRE funded project.

The tasks to address this question are to: 1) use individual potentiometric head measurements instead of averages to obtain larger data set for multivariate analysis, 2) determine best spatial interpolation method for expanding point predictions to area predictions, 3) develop functioning tool for ArcGIS that extracts variables, applies prediction algorithm, and runs spatial analysis to predict risk area surfaces, and 4) determine the range of acceptable error in both algorithm and interpolation surface in the empirical model. The major outcome of this project is the GIS-based tool for predicting post-mining water levels and risk for mine pool formation and surface discharge in the coal bearing region of Eastern Ohio. The method for developing the prediction tool and risk areas will be applicable to other coal producing regions with slight regional adjustments in the weighting of variables in the development of the prediction equation. The development of this empirical method for predicting post-mining water levels will thus not only address Ohio's gap of a science-based method of prediction, but also provide a greater understanding of hydrologic effects of underground mining to be applied in regions globally. This project establishes a viable methodology that can be applied to other coal regions of the world.

Experimental Approach

The response of groundwater to the perturbation of exploiting an underground mine is the object of this investigation. For the purpose of this project and the previous work, it is assumed that measures of potentiometric head in higher strata aquifers respond the same as the lower strata aquifers that contain the underground mines. While measurements in a monitoring well are clearly not the same as potentiometric head measures within the mine, area hydrology responds similarly due to the interconnectedness of groundwater hydrology (Means et al., 2018). Thus, predictions of potentiometric head measures can be extrapolated to the coal layers in the lower aquifers to gauge the hydrologic response within the mine after closure when potentiometric head data in the mined strata are lacking.

All data collected were from public data sources, consequently results can be available for public use. The majority of data analysis and compilation was performed in ArcGIS Pro, as opposed to other open source geospatial programs or ArcMap, because ArcGIS is widely used as the standard by the state agencies and mining companies that are the target users of the project outputs. ArcGIS Pro was used opposed to ArcMap due to ArcGIS Pro set to be the replacement for ArcMap as the standard within several years. The data formats are usable in ArcMap and the tool can be modified to run in ArcMap.

Due to the large amount of data and multiple variables examined for their influence on mine pool formation, complex multivariate methods were used to understand the relationship of variables and form a predictive method. Multivariate regression models were developed to further understand the variable relationships as well as form a predictive algorithm to use in the GIS model.

Initial work on this project has collected data from public sources and recent analyses have identified several key variables in determining the formation of a mine pool (Lopez and Kruse, 2015). Independent variables examined for this project include: surface elevation, bottom elevation of well, overburden thickness, thickness to mined coal, thickness of shale and clay in overburden, separate thicknesses of coal, sandstone, and limestone, total coal volume extracted, acres of underground mines within 4 miles, average precipitation, and water withdrawal over distance (Schafer, 2018). Schafer (2018) and Twumasi (2018) were instrumental in developing the approach methods and analysis for the predictive model. Multiple analytic methods were employed to verify robustness of results, and allow for selection of the best method.

Twumasi's (2018) work focused on artificial neural network (ANN) development and modeling of groundwater of the Meigs Mine complex. He used the program MODFLOW to examine the formation and sensitivity of variables causing mine pool formation in the Meigs Mine complex. For ANN work Twumasi (2018) used the Group Method of Data Handling (GMDH) simulation to run data sets with both water withdrawal data and no water withdrawal data. This final project phase also followed the methods selected for running the ANN program Neuroshell that Twumasi (2018) used in his work.

Schafer (2018) focused on multivariate statistical analysis and determining model parameters, with specific focus on the mine D-0360. The majority of the data extraction and format of data needed for analysis was determined by the work of Schafer (2018). She analyzed the data set from 11 mine permits in Unscrambler X using minimum, maximum, and averaged potentiometric head measurements over the period of record, as well as performing analyses with and without water withdrawal data. She found that using partial least squares regression using the amount of coal mined at the point of water level measurement showed a strong relationship with low error. While the results of

Twumasi's analyses correlated with Schaefer's done in Unscrambler X, the later method was selected due to findings of greater significance. Schaefer's analysis was repeated by Steinberg (2019) on an expanded data set to develop the model algorithm.

In the approach taken by Schafer (2018) and Twumasi (2018) data were collected and organized by well site and an average, minimum, or maximum potentiometric head measurement was used for multiple measurements at each site. For the final phase (Steinberg, 2019), instead of each well represented as a single measurement, each measurement in time was used, resulting in 5 times the amount of data points available for analysis and a more accurate approach.

An important assumption in these previous projects (Schafer 2018, Twumasi 2018) was the ability to extrapolate the prediction of water levels at well locations at elevations above the coal seam down to the mined coal seam. The extrapolation of the water level prediction is possible due to the interconnections of the area hydrology regardless of the discontinuous nature of some of the layers.

Collection of Data

Significant variables to mine pool formation were determined through various multivariate methods, similar to use in previous studies (Pradhan, 2010; Schafer, 2018; Twumasi, 2018). Various sources were explored for useful data for prediction of mine pool formation in the area of Eastern Ohio. Types of data sets collected have geographic references for use in spatial analysis, some were downloaded in a shapefile format and others as a data table. The mines of focus for this study are post-SMCRA underground mines in eastern Ohio, of which there is a shapefile of digitized areas from ODNR that includes information such as area of mine, type of mining, and coal seam(s) mined. Previously downloaded rasters of statewide elevation and top of coal elevation were also used in this study. Figure 2 displays data downloaded or extracted.

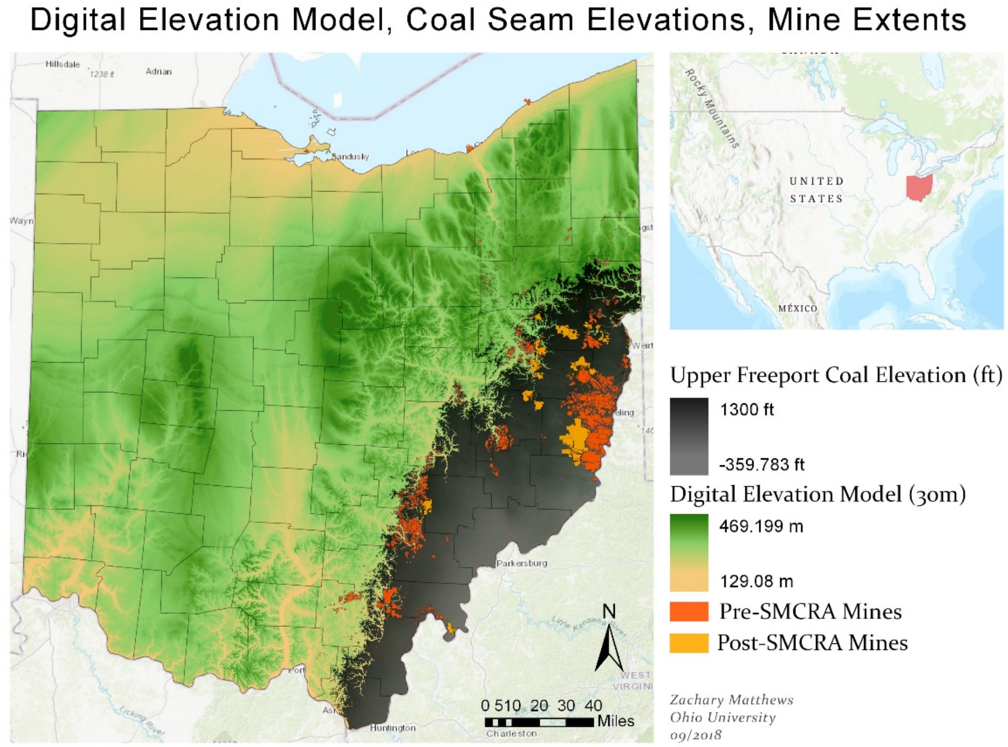


Figure 2 – A map displaying mine shape files downloaded from ODNR, as well as the raster layers of coal seams and the DEM

Scanned PDF format mine permits required manual extraction of borehole and water well data into Excel sheets. These sheets were formatted to require that only useful data were collected and assure it was recorded so it could be merged into ArcGIS for analysis. Data extracted from the mining permits for wells included: location data, projection (if recorded), surface elevation, depth of well, static water level, and aquifer type. Data for boreholes collected included: location data, projection (if recorded), surface elevation, bottom elevation, overburden thickness, thickness of coal mined, percent lithology of shale, limestone, sandstone, clay and coal. These percentages were later converted to total thicknesses.

For some mines with few points of well data in the main permit application, well data were also extracted from post-mining quarterly monitoring reports (QMRs) requested from ODNR. The same Excel format used in collecting well data from permit wells was used for QMR wells.

Precipitation data were collected for each mine, but due to the range of time the well data spans, a complete data set for local precipitation for each mine was too cumbersome to include in the analysis. Figure 3 is the map of annual average precipitation used in this analysis and in Schafer (2018) and Twumasi (2018). While the map is from a likely outdated data set from the 1990's, it was determined to be the most comprehensive and easily accessible dataset for the area of the study mines. Additionally, the precipitation across the area is not highly variable and likely would not be a significant variable between mines so averages in the area are sufficient for this analysis. This map was overlain as a tiff image and georeferenced with in ArcGIS. Precipitation values were then read off the map at each well location. Values of average annual precipitation were often the same for wells for a single mine, with the larger mines being the exception.

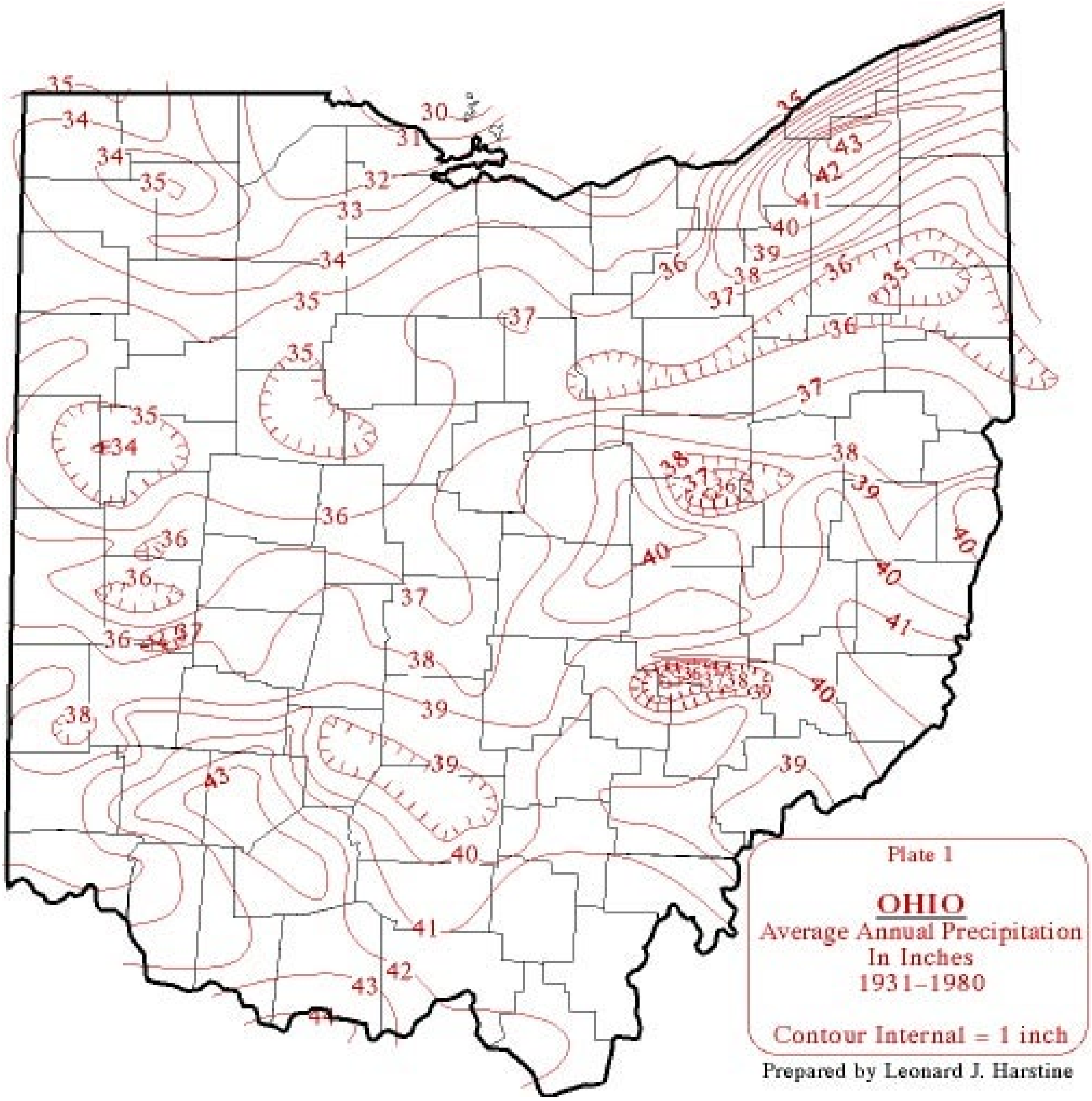


Figure 3 – Map of average annual precipitation for the state of Ohio used to extract average precipitation data for the area of mines studied, (ODNR Division of Water Resources, 1980)

Accumulative volume of coal extracted from each underground mine was also collected for use as a variable in the multivariate analyses. This variable was used to represent the amount of void space created from mining to represent how much water was pumped out of the mine. Data were downloaded for each mine permit from U.S. Department of Labor's Mine Data Retrieval System (U.S. Department of Labor, 2019). Coal volumes are recorded quarterly, so values for each quarter were copied to Excel sheets. For each well, the date of measurement was used to determine the quarter of coal extraction to calculate the total coal extracted at the point in time. The final accumulative amount of coal extracted from closed mines was also calculated. This method for extracting the accumulative coal volume extracted was developed by Schafer (2018).

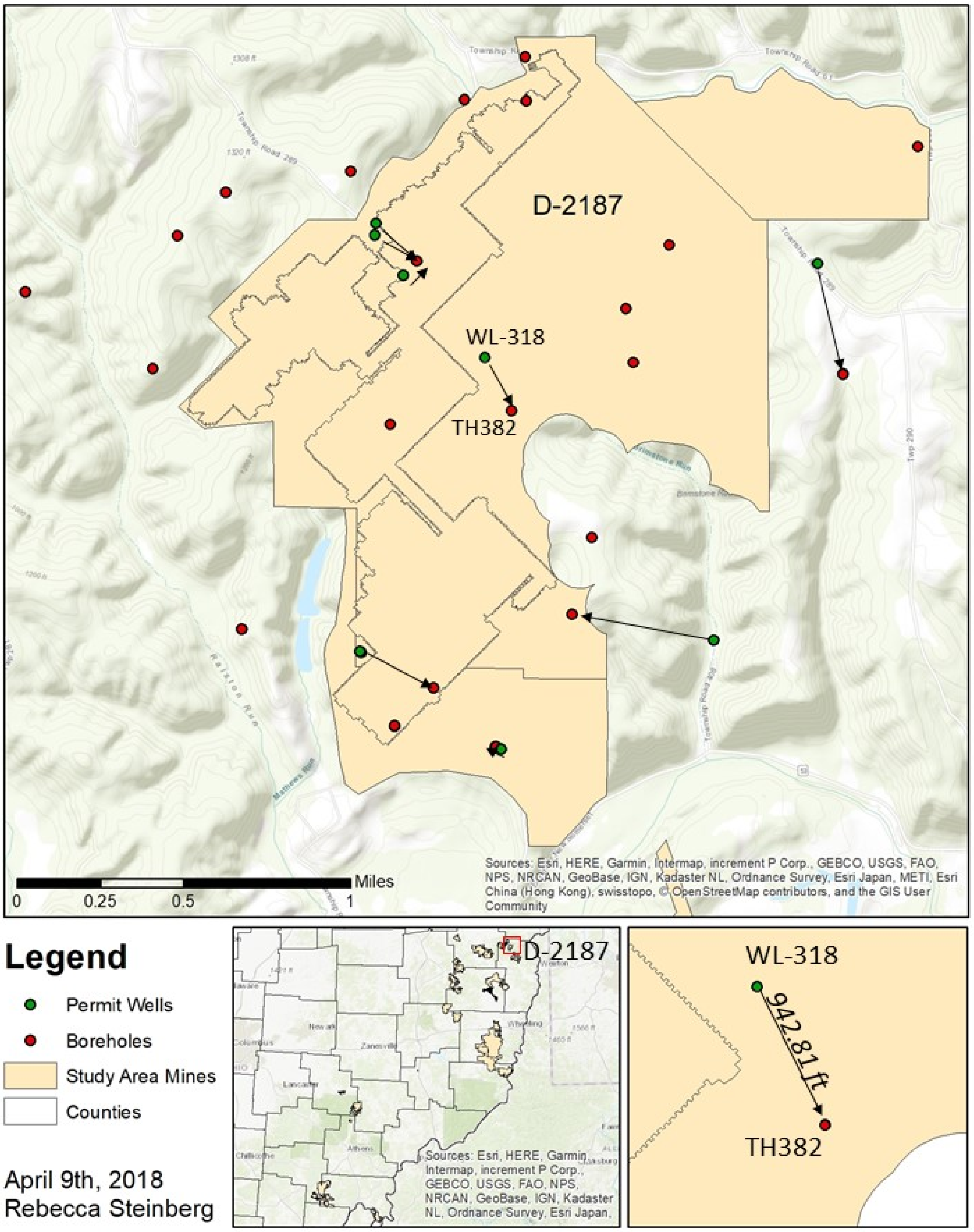
Variable Extraction from ArcGIS

Well stratigraphy was often not available in the permits. However, there are a large number of borehole data available in most mines. To relate the well stratigraphy with the borehole data, the nearest borehole to each well were extracted from maps of existing or collected data in ArcGIS Pro. Similarly, the acres of existing underground coal mines in a buffer area around the study mine was also extracted. Existing older underground coal mines must affect the hydrology of the proposed mine.

The nearest borehole to each well was used to extrapolate an approximate lithology for the area. Figure 4 displays the process in ArcGIS using the tool 'spatial join' with the parameter 'closest' used to determine which borehole was closest to each well. From this the values for borehole lithology were joined to each well, providing values for the lithology related variables (overburden thickness, coal seam mined thickness, clay/shale thickness, limestone thickness, sandstone thickness, total coal thickness).

The acreage of mined area within a buffer around each mine permit area was calculated from both the pre-SMCRA and post-SMCRA mine shapefile layers acquired from ODNR. Several buffer distances were tested to determine what distance should be used for the analysis. As displayed in Figure 5, buffers of 1, 2, and 4-miles were tested. The 4-mile buffer distance from the study mine produced the best results, likely due to the heavy influence of void space on the area hydrology (Schafer, 2018). Once a buffer was created, the pre- and post-SMCRA layers were clipped within the buffer area and those clipped shapes were used to calculate the area of void space around the study mine. This value was extracted in square feet then converted to acres for analysis.

Distance from Wells to Nearest Borehole



Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, © OpenStreetMap contributors, and the GIS User Community

Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan,

April 9th, 2018
Rebecca Steinberg

Figure 4 – Map that displays the use of the spatial join tool in ArcGIS Pro to obtain the lithology of the closest borehole and join it to the well points

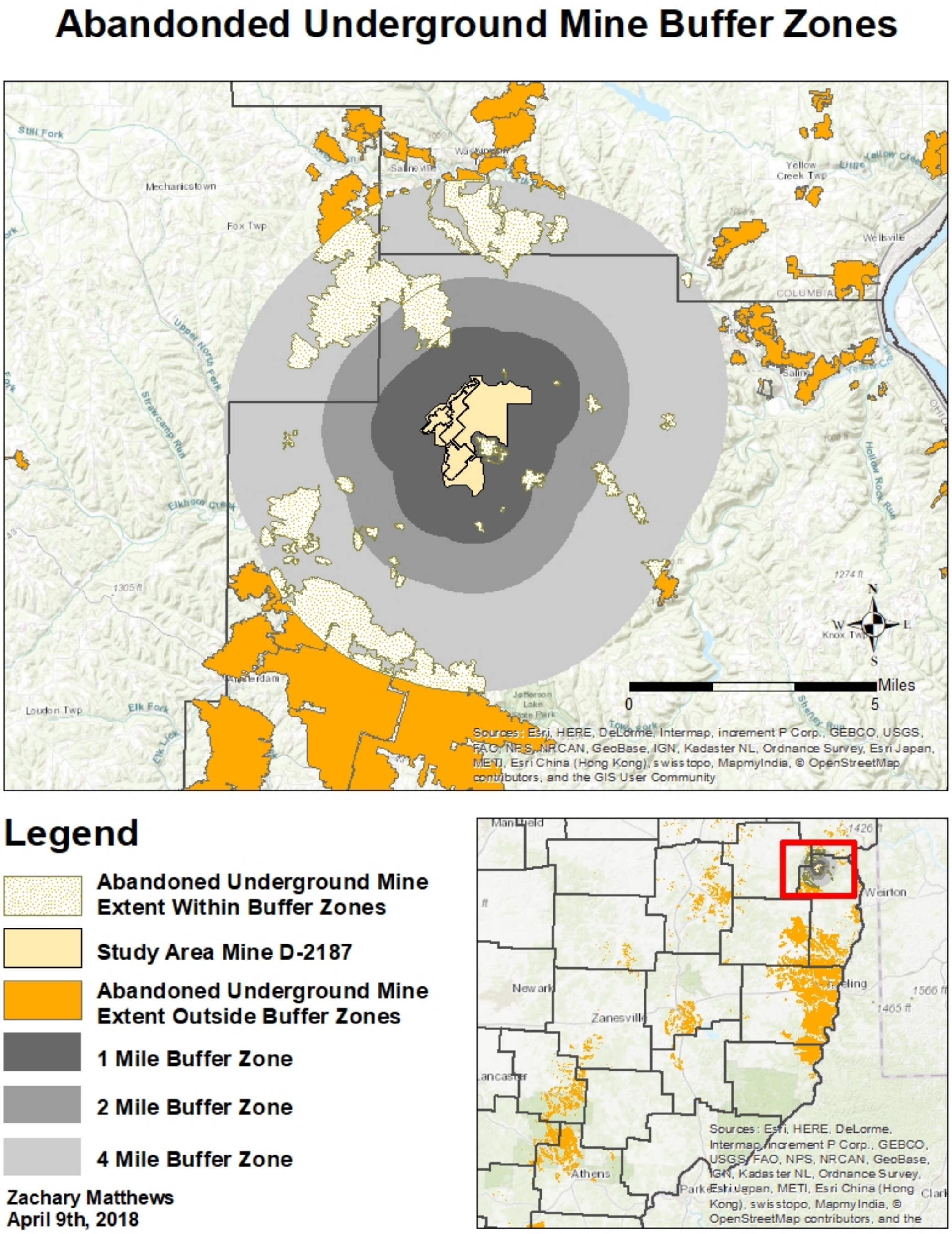

Abandoned Underground Mine Buffer Zones
Still Fork
Mechanicstown
Fox Twp
Salineville
Washington
Yellow Creek Twp
Little Yellow Creek
Wellsville
COLUMBIANA
Salineville
Upper North Fork
Strawcamp Run
Elkhorn Creek
Hollow Rock Run
1305 ft
1274 ft
Knox Twp
N
W
E
S
Miles
0
5
Amsterdam
Loudon Twp
Elk Fork
Elk Lick
Jefferson Lake State Park
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community
Legend
Abandoned Underground Mine Extent Within Buffer Zones
Study Area Mine D-2187
Abandoned Underground Mine Extent Outside Buffer Zones
1 Mile Buffer Zone
2 Mile Buffer Zone
4 Mile Buffer Zone
Zachary Matthews
April 9th, 2018
Mansfield
1426 ft
Weirton
Newark
Zanesville
Wheeling
1566 ft
1465 ft
Lancaster
Athens
Parkersburg
Clark
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the

Figure 5– Map displaying the development of buffer zones, of which ultimately the 4 mile buffer zone was used, in exctracting the area of undergound mining activity surrounding the study mine

Multivariate Analysis and Modeling

For mine pool formation, the potentiometric head is investigated as the dependent variable for determining independent variables relationships. Multivariate regression analyses were run in several programs to determine the relationships between and significance of the variables. These analyses were run first in the multivariate statistical program The Unscrambler X version 10.5, which describes the relationship of the independent variables and provides regression equations for different regression methods. Analyses of the variables were also run in a second program, Neuroshell 2.0, which uses artificial neural networks (ANN) to determine relationships of the variables and produce a complex polynomial regression equation for determining potentiometric head post-mining. These equations were compared by their complexity and root mean squared errors to determine which equation to apply in predicting post-mining water levels through the ArcGIS tool.

Multivariate Statistical Analyses

Initial statistical analysis of the variables examined were run in the program Unscrambler X, following methods developed by Schafer (2018). Methods of multivariate analysis tested were multiple linear regression (MLR), principal component regression (PCR), principal least square regression (PLS) and principal component analysis (PCA) (Schafer, 2018). These methods develop interpretations of the relationships of the variables input and produces a multivariate linear regression equation to represent those relationships. These methods were the same tested by Schafer, 2018, but re-run with the new expanded data set to compare results and accuracy with Schafer's results, and to further develop the predictive model.

MLR was not appropriate for this data set, as it requires variables be independent of one another, which is not the case with this data set. PCA was used in defining variables and determining their relationships. MLR and PCA are explained in detail in Schafer (2018) and in CAMO (2019).

PCR is a combination of PCA and MLR, where the variances of the principal components (PC) are compared in multidimensional space as in PCA, and then form a regression using the relation of the variance of the Y component to the X components as in MLR (CAMO Software AS, 2019). Figure 6 displays this method, showing the combination of PCA and MLR methods for describing the multidimensional space of the data.

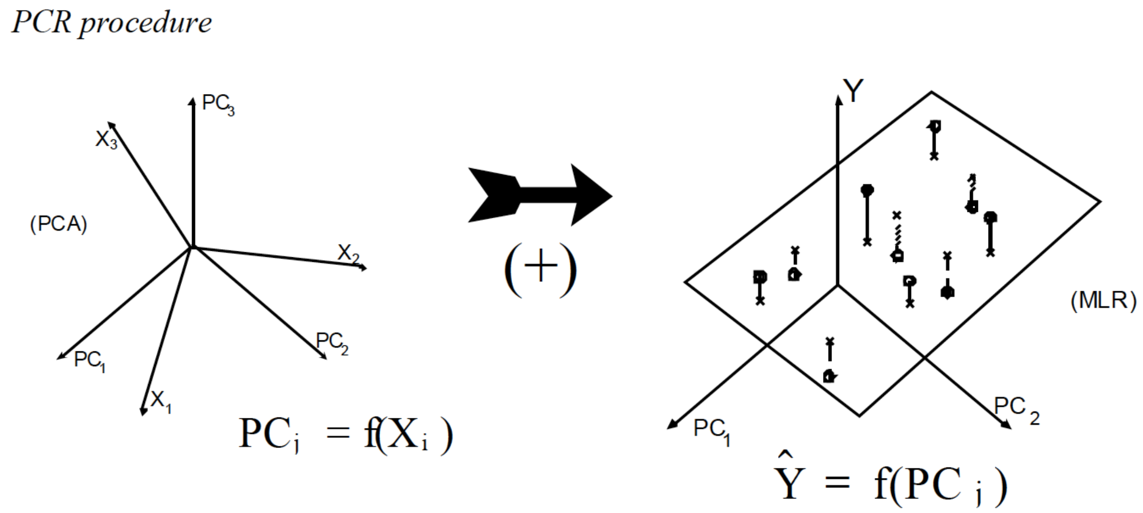


Figure 6 - Visual representation of the process of PCR, using PCs to describe the variance in the Y, (CAMO Software AS, 2006)

PLSR is a combination of PCA and MLR, but instead of comparing PCs to each other, defines the X and Y matrices as factors, which are then compared as PCs to define the X relationship to predicting Y. Figure 7 displays how these matrices define the Xs and Ys and then compare them. This data set though only has one Y variable.

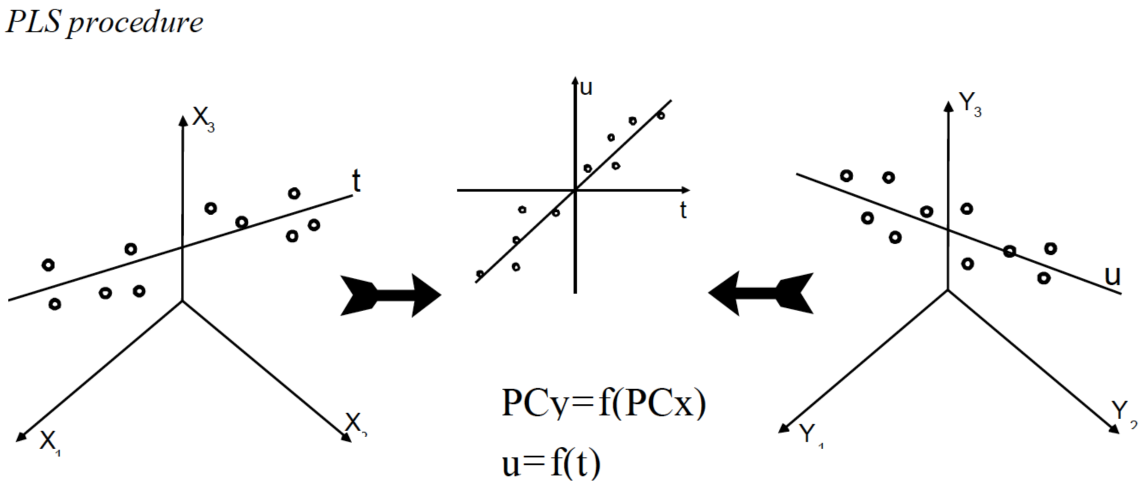


Figure 7 – Visual representation of the process of PLSR, where the X and Y variable matrices are compared as PCs, (CAMO Software AS, 2006)

PLSR and PCR previously produced the most accurate regression equations, with PLSR resulting in slightly less error, and thus were the focus for this study (Schafer, 2018). Both regression models are multivariate linear regression analyses that identify an axis in multidimensional space to represent the variance between variables and to best represent their relationships.

The PLSR and PCR analyses in the Unscrambler X also provide results that allow for identification of outliers in the data set through looking at the analysis resulting residuals. The data residuals are how far each sample is from the axis, or PC, that is defining the variable in multidimensional space (Figure 8). Samples with large residual values may be skewing results, thus can be identified as outliers and removed (CAMO Software AS, 2006, 2019). The values of the residuals are also used to determine the model error.

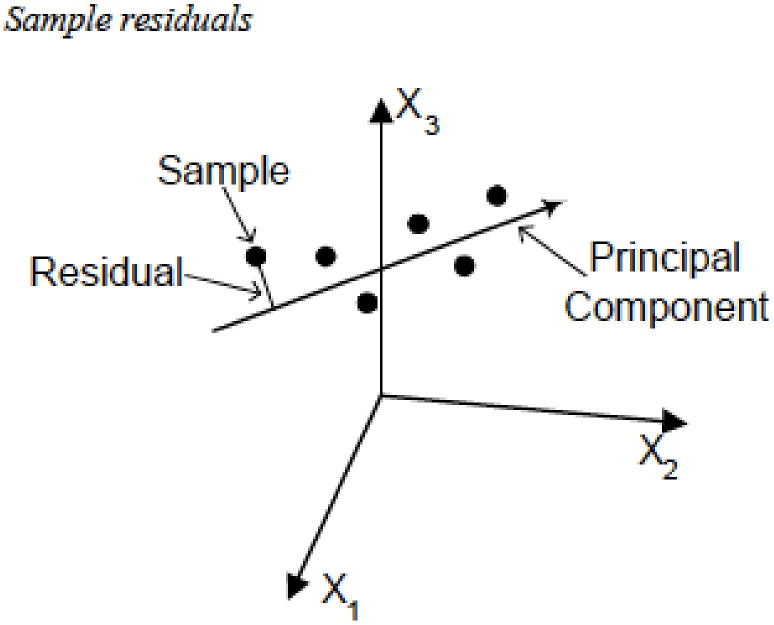


Figure 8 – Visual representation of the sample residuals along a principle component (PC) that is defining X variables in multidimensional space (CAMO Software AS, 2006)

Artificial Neural Networks

The program Neuroshell 2 version 4.0, first developed in 1993, was used as a second method for developing an algorithm to predict post-mining potentiometric head elevation. Neuroshell is a program that utilizes the construction of artificial neural networks to analyze complex non-linear relationships between input data and determine 'weights' for input variables to form a polynomial equation (Twumasi, 2018). An artificial neural network is defined as a mathematical model that runs a computational simulation that imitates the behavior patterns of neurons in the human brain to perceive patterns in data, to 'learn' from a training data set (Sánchez-Mesa et al., 2002; Twumasi, 2018). Described in 'Neural Network Overview' of Ward Systems Groups Inc.'s Neuroshell 2 help document, neural networks construct neurons to develop networks of interconnected neurons from input data (input neurons) that are able to use connections through layers of hidden neurons to produce an output network (output neurons) in which the connections or weights between neurons describe the data set relationships. Figure 9 shows the input, hidden, and output neurons, with line in between them indicating the weights of the network connections, and each type of neuron representing a layer. Multiple layers of hidden neurons are often constructed to further the learning process of the network (Ward Systems Group, Inc., 2019).

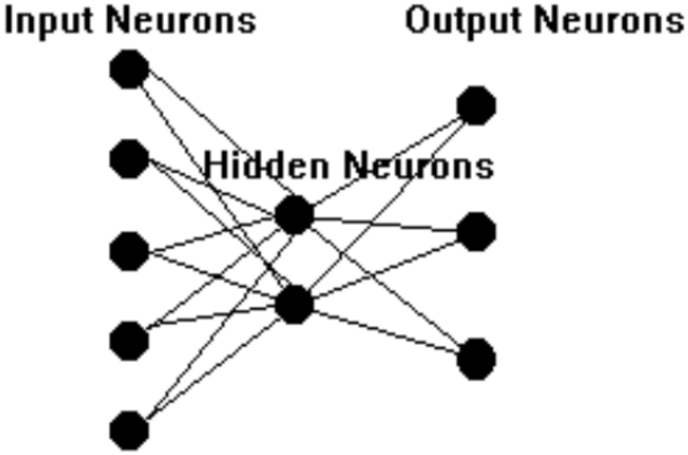


Figure 9 – Visual representation of the development of neuron layers in the creation of an artificial neural network, connected by lines representing the weighting of the network connections, (Ward Systems Group, Inc., 2019)

The learning module used for developing the equation was Group Method of Data Handling (GMDH) with the Advanced Training Criteria, the same as previously used successfully in Twumasi (2018). The advanced training option for GMDH allows the user greater freedom in selection of training criteria. These training criteria options determine how the program selects or removes 'neurons', or polynomial factors, from 'layers' in the construction of the algorithm (Ward Systems Group, Inc., 2019). Also selected from the Advanced Training Criteria were the 'schedule type' as Asymptotic with ‘decrease in maximum number of survivors’ as Gentle. For this project, only the options ‘selection criteria’ and ‘model optimization’ were varied.

Selection criteria is the most important parameter when designing a GMDH model as the options determine how neuron ‘survivors’ are selected (Ward Systems Group, Inc., 2019). For selection criterion, Prediction Squared Error (PSE), Full Complexity Prediction Squared Error (FCPSE), Minimal Description Length (MDL), Generalized Cross Validation (GCV), Final Prediction Error (FPE), and Regulatory (calibration) were all tested, with each option for model optimization as well. PSE is a combination of two terms in determining selection, the model average squared error and an overfitting penalty. FCPSE is a modified version of PSE that takes into account the model complexity instead of number of coefficients. MDL is also similar to PSE but has a greater value for the overfitting penalty. GCV is another version of applying an overfitting penalty. FPE takes into account the minimum variance of the mean-squared error of model prediction. Regulatory is different in that it looks at the average squared error of the model when applied to a test set manually selected out of the main data set.

Model optimization options tested were Smart, Thorough, and Full. The optimization options are for improving the model by removing terms deemed unnecessary, to either improve function or accuracy, and can affect how the model determines significant variables (Ward Systems Group, Inc., 2019). Smart provides a balance between calculation speed and model quality. Thorough is similar to Smart but looks closer at selecting significant variables. Full is the most complex approach in that it examines all variables combinations at each stage of model development, resulting in a highly complex but accurate model.

ArcGIS Tool Building

The tool was designed, developed and tested in ArcGIS Pro ModelBuilder. ModelBuilder allows a series of geoprocessing tools to be run in a sequence, set up as a diagram of chain connected inputs, tools, and outputs (ESRI, 2019c). Parameters required for inputs and outputs are defined in the ModelBuilder platform, which when running the resulting tool are pulled in and analyzed without further input from the user. The type and format for data needed for input into ArcGIS and to be run through the tool are defined as templates to be used with running the tool.

Python Scripting

For the application of the selected prediction algorithm, a python script was written to manually apply the calculation to variables extracted by the first part of the model. The script was written in Python 2.7 and imported as a tool in ArcGIS Pro that was then able to be added to the ModelBuilder tool flow. The manual scripting allowed for clear and correct pulling of variable values and equation application.

Tool Validation

The tool was tested using existing well and borehole data points to determine the reliability of the output of the tool as well as used for de-bugging during construction of the tool. Post-mining data from two mine complexes, the Meigs mine complex and the Corning mine complex, were explored to be used to validate the tool outputs with measured data. While the Meigs mine complex data were used in the development of the equation, the data were more complete than any other mine. The Corning mine complex was also used, but due to incomplete data, estimations were required for some variables.

Geostatistical Analysis and Spatial Interpolation

Several methods for spatial analysis of the predicted post mining water level were explored for the purpose of creating a potentiometric surface representation. Distribution of data made this impossible, however, as data points were too distant to produce connectivity within a reasonable range of error.

Results and Discussion

Multivariate Analysis

The two programs The Unscrambler X and Neuroshell 2 were used successfully in running the analysis on the post-SMCRA mine data. The analysis followed the model structure developed by the previous work of Schafer (2018), and Twumasi (2018), but with an expanded data set to further develop the model.

The Unscrambler X

The expanded data set was re-analyzed using the same statistical analyses used previously by Schafer (2018), to increase accuracy of the prediction equation and determine if additional data produced better results. Multivariate analysis in the Unscrambler X using PCA, PCR, PLSR regressions showed that PLSR still produced the best regression with the least amount of error. These runs were all done with the same expanded data set of 2872 data points, 2581 points used for prediction and 291 points used for validation (~10 percent of the data set).

PCA was previously used in determining variable relationships and was re-run here to check for consistency in variable relationships with the original analysis and this analysis. Figure 10 shows the results of the PCA correlation loading of the variables (all considered X variables in PCA) reflects the previous study variable relationships. The correlations loading chart displays the model variables in relation to each other, closer together the more related and vice versa. It also displays how much of the data variance the variables explain, with the outer ellipse representing 100% explained variance and the inner 50% variance. The variables surface elevation, bottom elevation, and potentiometric head were all displayed as closely related and near to explaining 100% of the data variance. The variable for underground mining in a 4-mile buffer and limestone thickness were also important to explaining total variance.

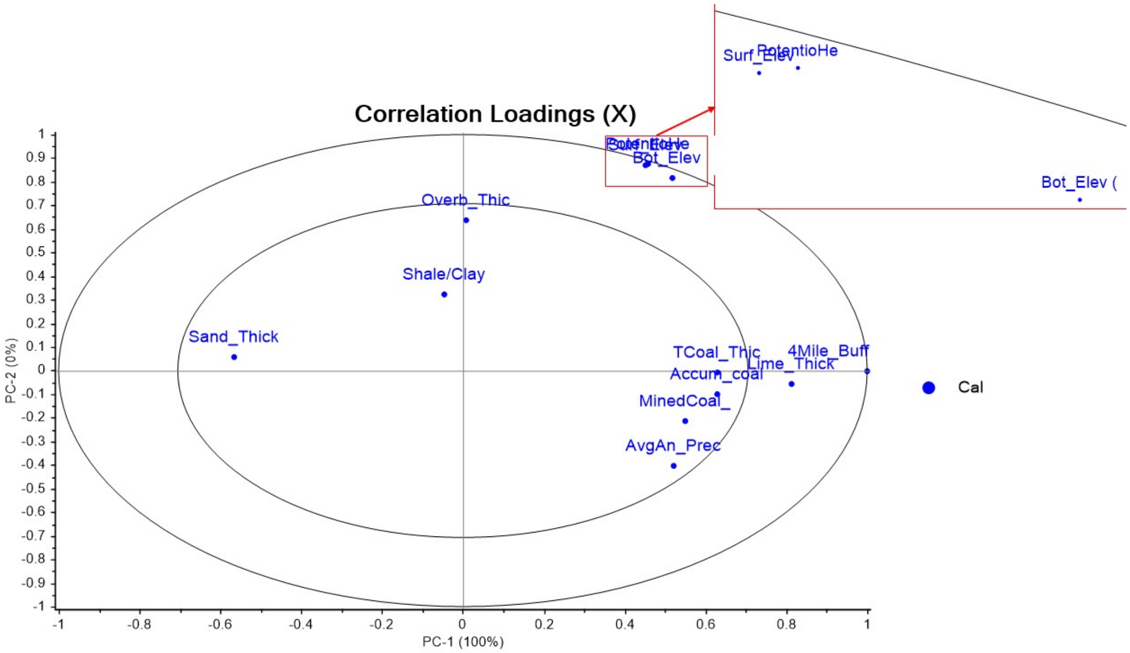


Figure 10 – Correlation loadings chart for the PCA run displaying the relationships of the variables. The outer ellipse is 100% explained variance and the inner ellipse is 50% explained variance. Variables that are closer together are more related. This displays a strong relationship between surface elevation, bottom elevation, and potentiometric head elevation.

Results from the PCR was able to explain total variance of the data by 3 PCs (Figure 11). In Figure 12 the regression's predicted values versus the actual reference values are compared, plotting both the calculation points and the 10 percent validation points, at the PC2 level where the most variance is explained. The r-squared value of 0.972 indicates an accurate regression model. The correlation's loading diagram in Figure 13 indicate a strong relationship between the X variables of surface elevation, bottom elevation, and the Y variable of potentiometric head, just as displayed in the PCA run. Figure 14 also displays the relationships of the examined variables by displaying the weights of variables on the regression, still indicating the high level of influence from surface and bottom elevations with smaller influences from the other variables. Note that even when some of the variables have a relatively low influence, we need to consider them because we are looking at reducing as much as possible the errors and eliminating some of those variables can mean increasing the errors.

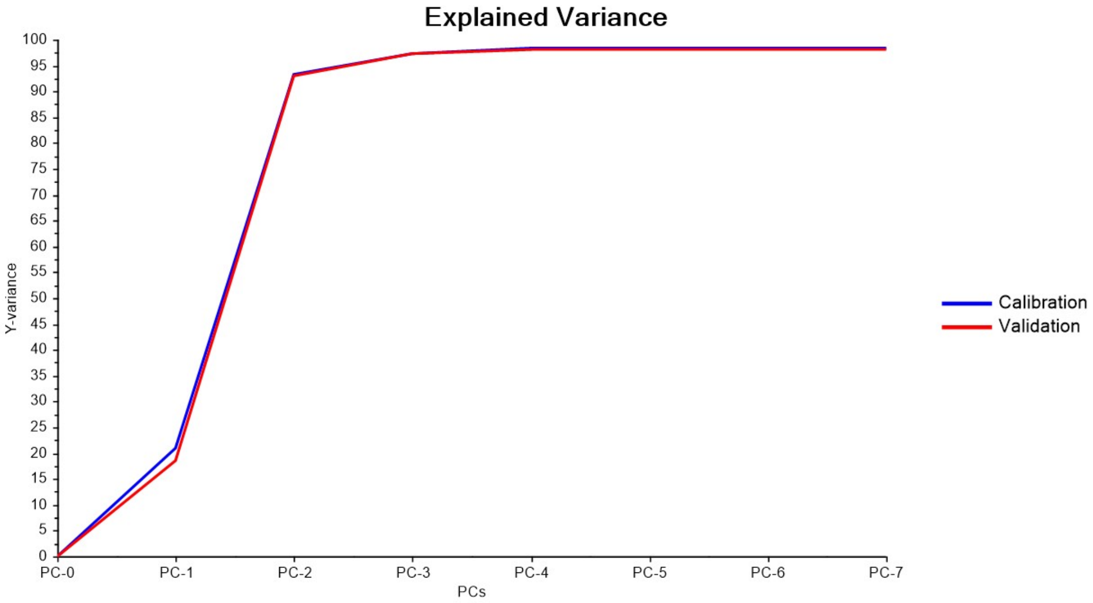
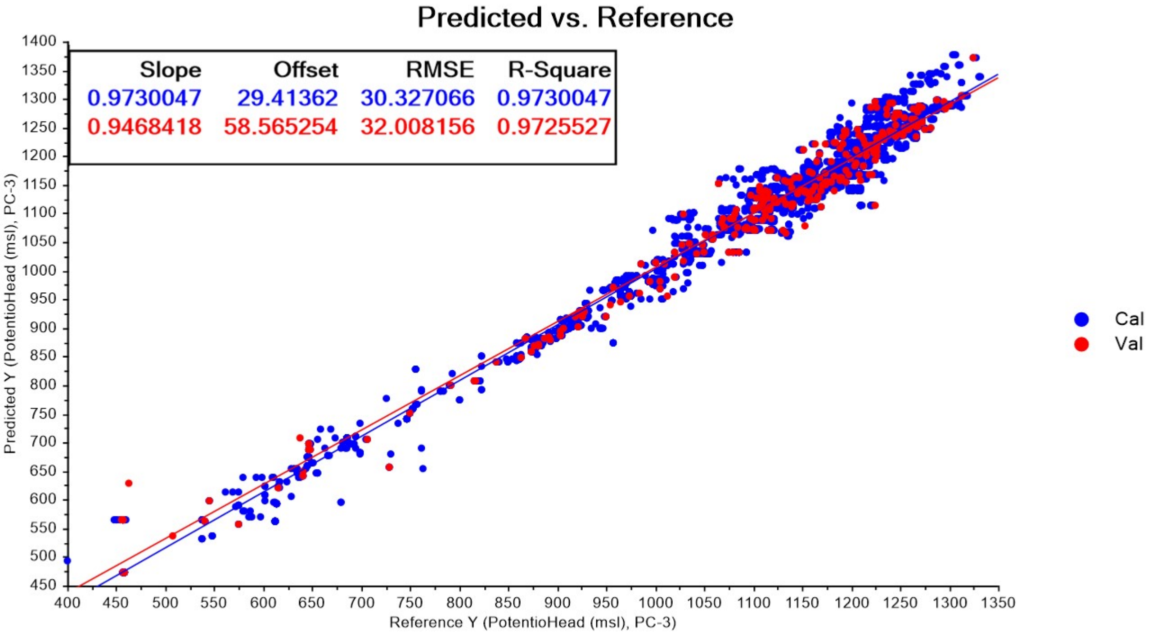


Figure 11 – Graph of explained variance in the PCR run, total explained variance required 3 PCs



Slope	Offset	RMSE	R-Square
0.9730047	29.41362	30.327066	0.9730047
0.9468418	58.565254	32.008156	0.9725527

Figure 12 – Graph of predicted versus reference values for the PCR run, displaying a decent regression with r-squared value of 0.973. Calibration data set is blue, and the 10% validation set is in red

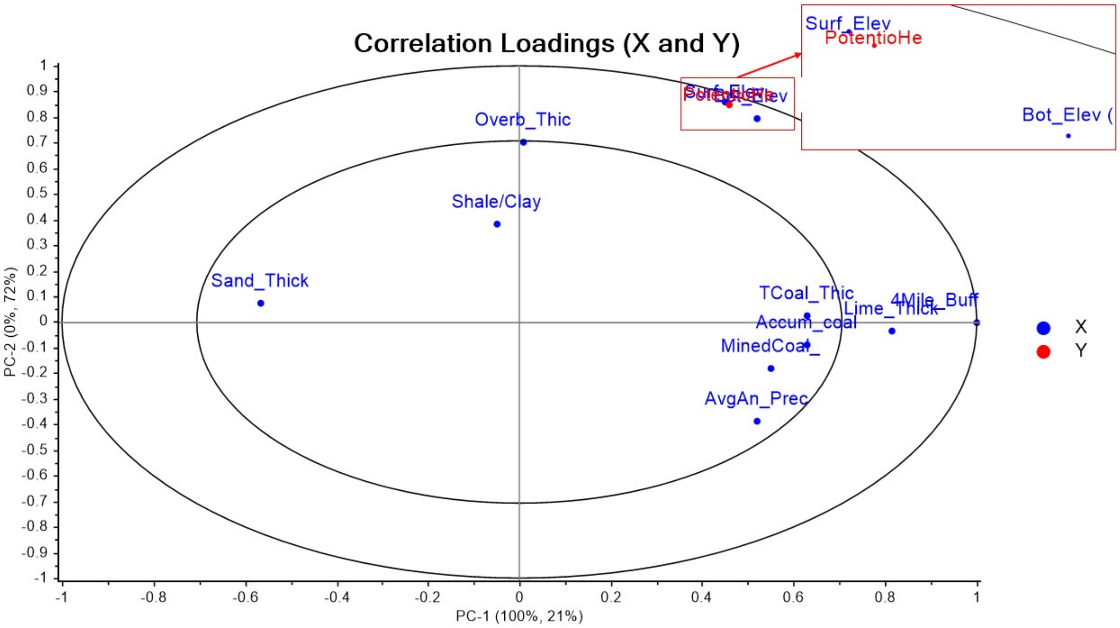


Figure 13 - Correlation loadings chart for the PCR run displaying the relationships of the variables. The outer ellipse is 100% explained variance and the inner ellipse is 50% explained variance. Variables that are closer together are more related. This again displays a strong relationship between surface elevation, bottom elevation, and potentiometric head elevation.

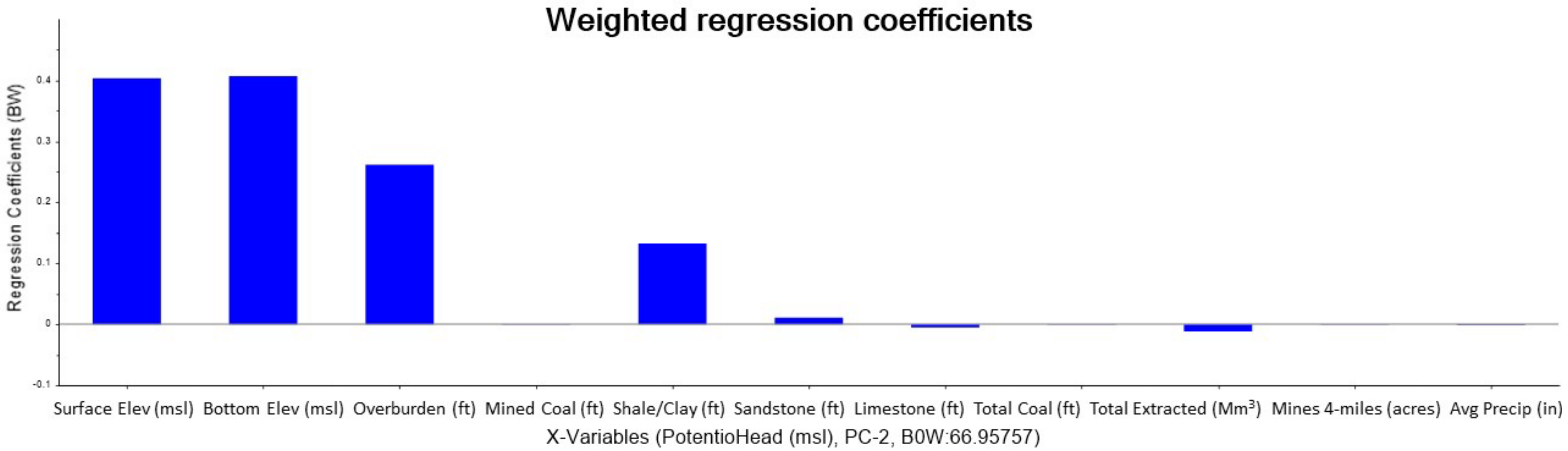


Figure 14 - Bar chart displaying the weighting of variables for the PCR run in PC2. PC1 displayed influence heavily in the variable of area of mining in the 4-mile buffer, PC2 here displays the influence from the other variables.

The results of the PLSR were similar to the PCR in that both required 3 factors/PCs to reach total explained variance (Figure 15), as well as displaying similar influences of variables (Figure 16). The correlations loadings chart from the PLSR run (Figure 17) displays the relationship of variables similar to the PCR and PCA runs in that X variables surface elevation, bottom elevation and Y variable potentiometric head are closely related and are near the outer ellipse of 100% explained variance. Also like the PCR run, the PLSR run also provided a strong regression, as seen in Figure 18 with the predicted values versus the actual reference values of the data set. Compared to the PCR run, this regression run has a slightly higher r-squared value of 0.982, and so a slightly more accurate model result. This determined PLSR as the best regression analysis in the Unscrambler X for the data set and was examined further. Table 1 displays the regression coefficients for the PLSR run with this expanded data set.

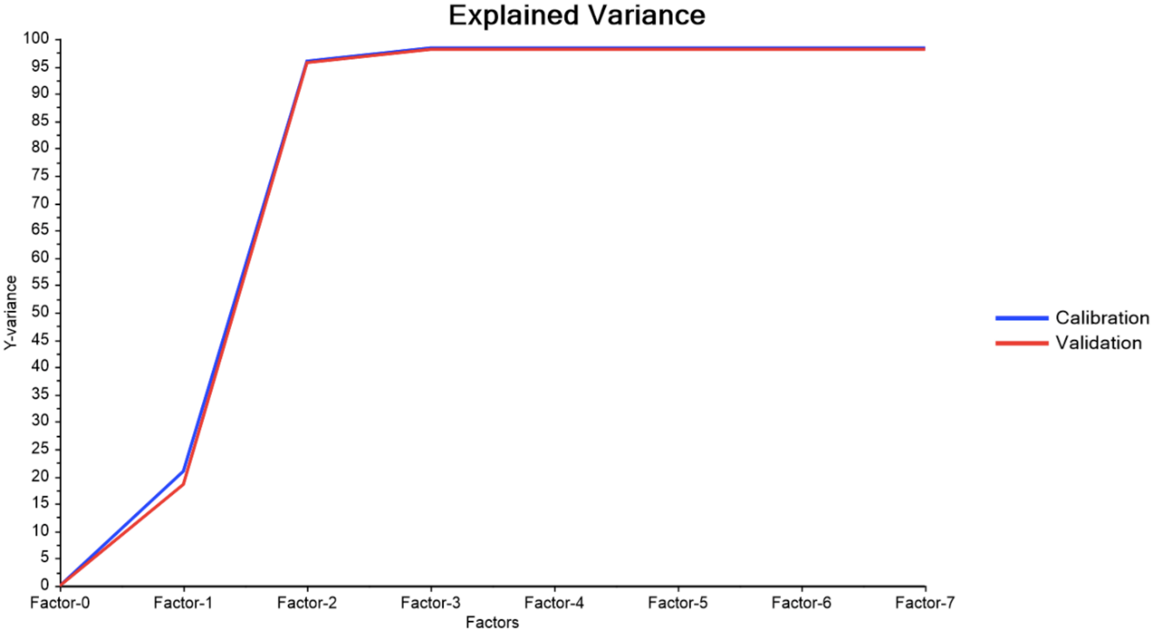


Figure 15 - Graph of explained variance in the PLSR run, total explained variance required 3 factors

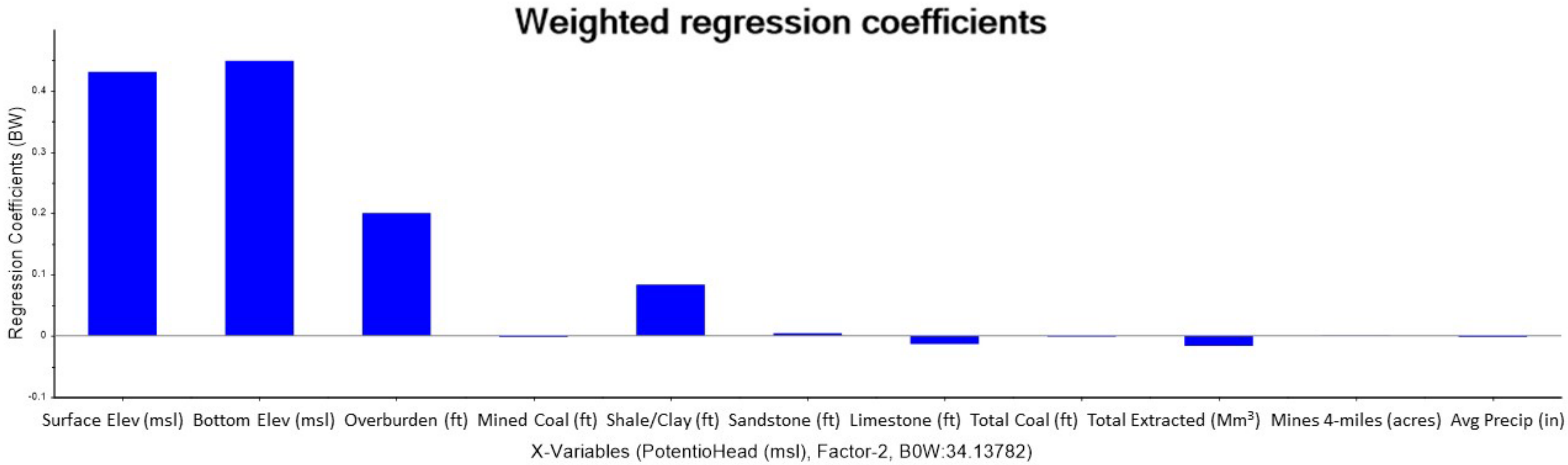


Figure 16 – Bar chart displaying the weighting of variables for the PLSR run in Factor 2. Factor 1 displayed influence heavily in the variable of area of mining in the 4-mile buffer, just as in the PCR run, Factor 2 here displays the influence from the other variable

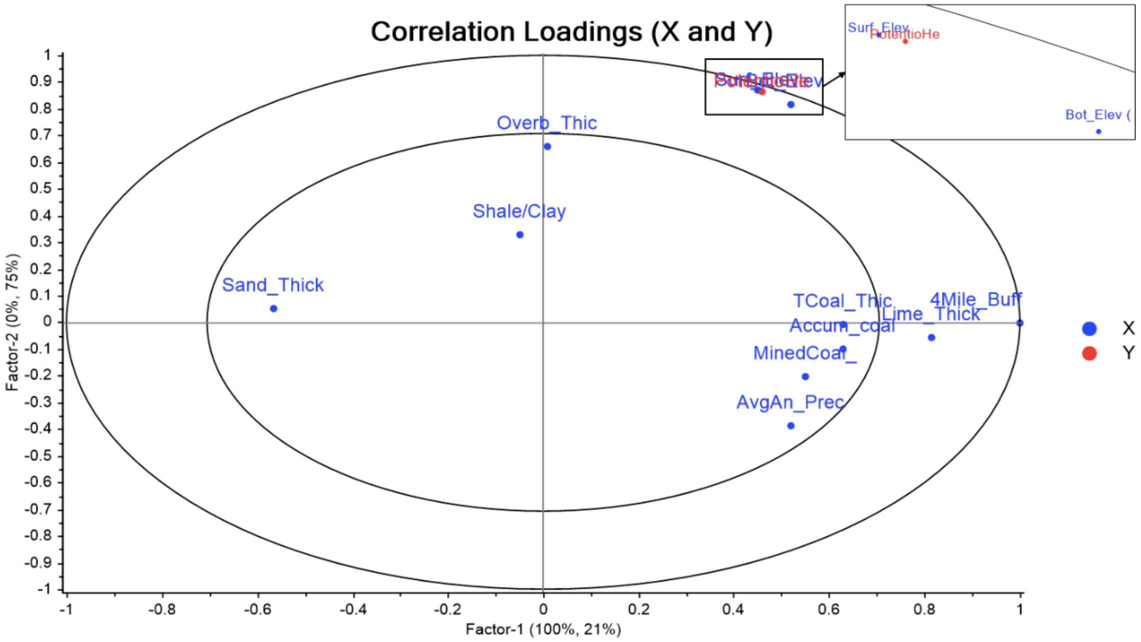


Figure 17 - Correlation loadings chart for the PLSR run displaying the relationships of the variables. The outer ellipse is 100% explained variance and the inner ellipse is 50% explained variance. Variables that are closer together are more related. This again displays a strong relationship between surface elevation, bottom elevation, and potentiometric head elevation.

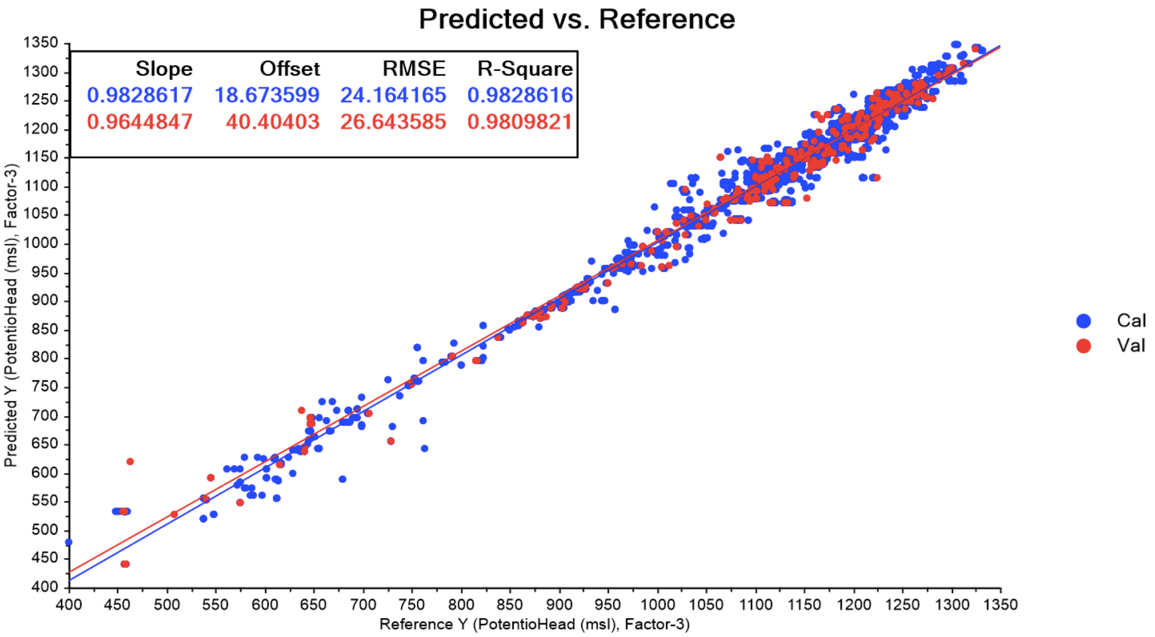


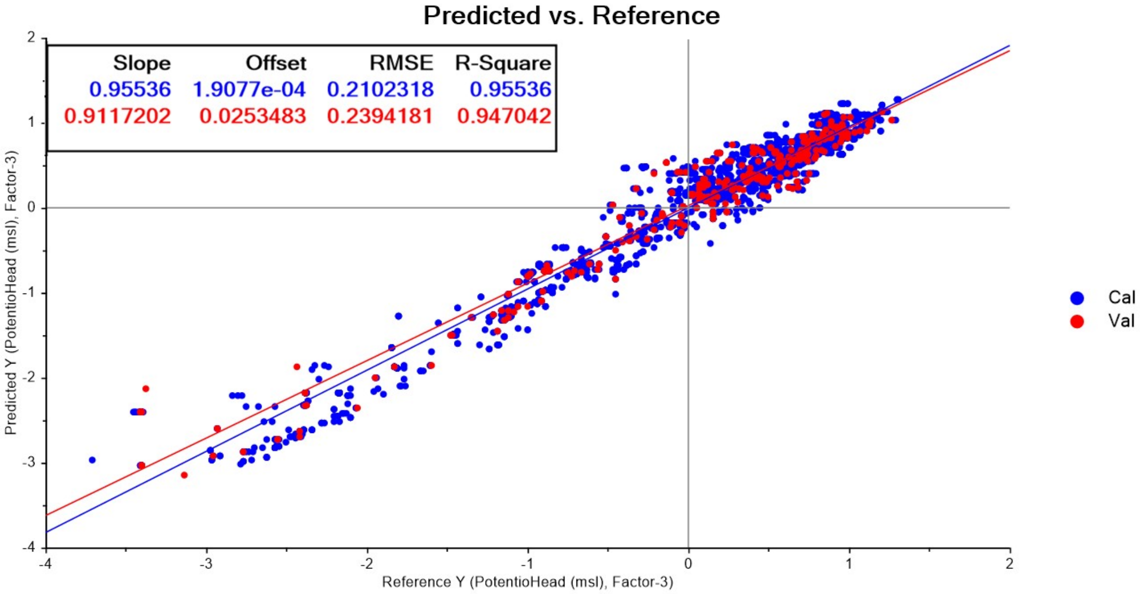
Figure 18 – Graph of predicted versus reference values for the PLSR run, displaying a decent regression with r-squared value of 0.983, better than the PCR run. Calibration data set is blue, and the 10% validation set is in red.

Table 5 - Regression variable coefficients from PLSR run

<u>Variables</u>	<u>PLS Coefficients</u>
β	-1.55728
Surface Elevation (ft msl)	0.47898
Bottom Elevation (ft msl)	0.52696
Overburden Thickness (ft)	0.03656
Mined Coal Seam Thickness (ft)	-0.00252
Shale + Clay Thickness (ft)	-0.02280
Sandstone Thickness (ft)	-0.00694
Limestone Thickness (ft)	-0.02862
Total Coal Thickness (ft)	-0.00361
Total Coal Extracted (Mm^3)	-0.02301
Underground Mines in 4 Mile Buffer (acres)	-0.00012

Average Annual Precipitation (in)	-0.00199

Also tested with the expanded dataset was the normalization of the variable values to determine if adding normalization would help decrease error any. Figure 19 displays the PLSR regression run on the normalized values with an r squared value of 0.955, less than the PLSR and PCR runs with the non-normalized values. Normalizing the dataset produced a similar resulting regression in terms of variable relationships and using 3 factors to explain total variance but produced more error than non-normalized values. From this test it was determined that non-normalized values were to be used exclusively for the remainder of the data analysis.



Slope	Offset	RMSE	R-Square
0.95536	1.9077e-04	0.2102318	0.95536
0.9117202	0.0253483	0.2394181	0.947042

Figure 19 - Graph of predicted versus reference values for the normalized data set PLSR run, displaying a regression with r-squared value of 0.955, displaying that normalized values have not produced a better regression than the non-normalized values of this data set. Calibration data set is blue, and the 10% validation set is in red.

Outliers were identified and removed from the PLSR run through use of the Leverage vs. Residual plots produced by the Unscramble X, following the method used by Schafer (2018). Figure 20 shows the plots with the outliers removed labeled by the red arrows and circles, selected out by the distinct distance from the grouping of points on the plots that represent the rest of the dataset. A total of 53 outliers were removed. With the exception of the first outliers removed in PLS run 1 from mine D-0360, the other outliers were from only two mines, D-1019 and D-2317.

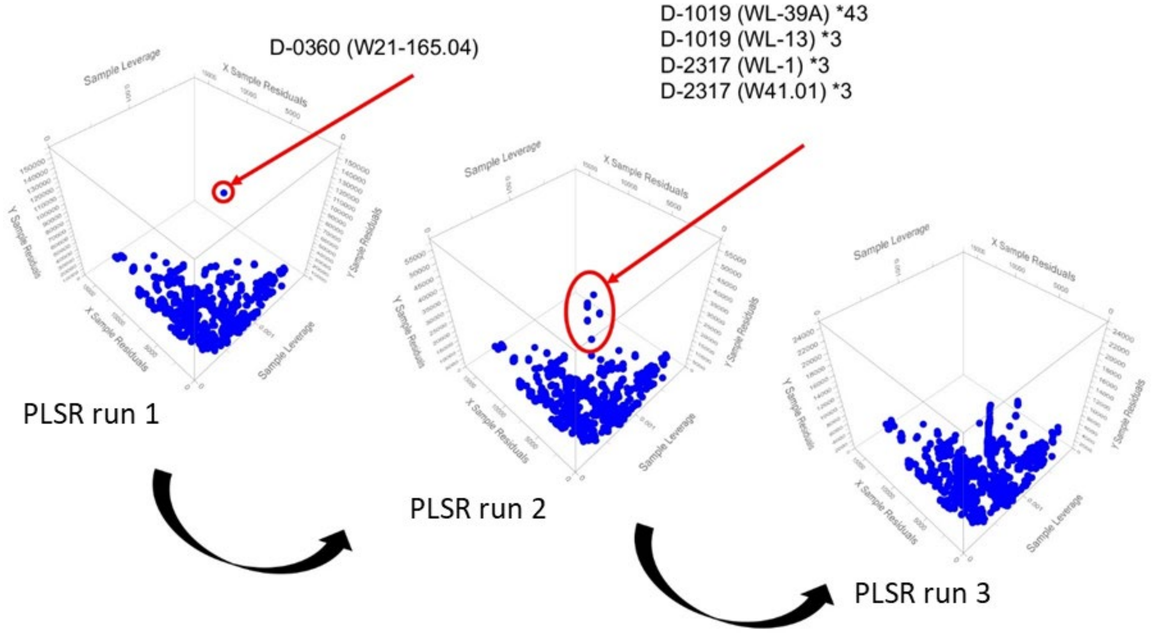
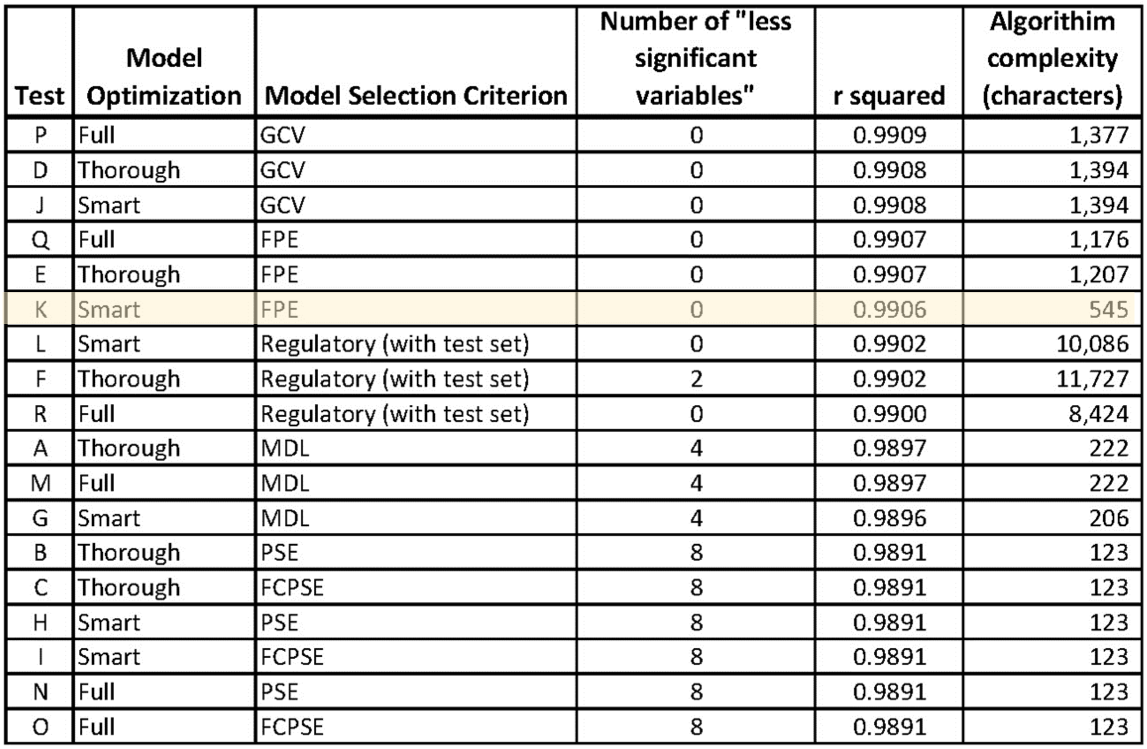


Figure 20 – Leverage versus residual 3-dimensional plots used to determine outliers from the PLSR run, re-run with the removal of the outliers to form the final regression.

Artificial Neural Network

The artificial neural network (ANN) analysis conducted by Twumasi (2018) was also re-run with the expanded data set in Neuroshell 2.0 to produce a polynomial regression equation (Table 2). As with the previous analyses by Schafer (2018) and Twumasi (2018), the ANN equation still resulted in lower error than the less complex PLSR regression produced in the Unscrambler X. Variable transformations in the previous ANN run were similar to the re-run results, indicating consistency in the analyses. The ANN equation was selected as the algorithm incorporated into the ArcGIS tool due to the increased complexity resulting in less error of post-mining potentiometric head prediction (r-squared values of 0.982 with PLSR vs. 0.9906 with ANN). Testing was done for each combination of model optimization and selection criterion parameters described in the section Artificial Neural Networks, resulting in 18 test variations, and labeled ‘A-R’, described in Table 2. The tests were compared on three model descriptors: the number of “less significant variables,” to determine which run kept the majority of input variables; r-squared values for comparing errors; and algorithm complexity (measured as number of characters) to compare how manageable the equation would be in applying to the prediction model. Table 2 was sorted by these model descriptors, starting with the lowest ‘number of “less significant variables”’, then the highest ‘r squared’, then lastly the lowest ‘algorithm complexity’. From these comparisons, equation ‘K’ was selected for further analysis to be selected as the final equation used in the ArcGIS tool, as it retains all variables determined significant to predicting post-mining water levels, has a lower complexity than other runs that retain variables and still has a high accuracy (r squared of 0.9906) like the more complex runs. The resulting equation and variable transformations for ANN run ‘K’ is displayed in Table 3.

Table 6 – Neuroshell test runs of model optimizations and model selection criterions, sorted by the lowest number of “less significant variables”, then by the highest r squared values, and lastly by the lowest algorithm complexity. The selection of test ‘K’ is highlighted



Test	Model Optimization	Model Selection Criterion	Number of "less significant variables"	r squared	Algorithim complexity (characters)
P	Full	GCV	0	0.9909	1,377
D	Thorough	GCV	0	0.9908	1,394
J	Smart	GCV	0	0.9908	1,394
Q	Full	FPE	0	0.9907	1,176
E	Thorough	FPE	0	0.9907	1,207
K	Smart	FPE	0	0.9906	545
L	Smart	Regulatory (with test set)	0	0.9902	10,086
F	Thorough	Regulatory (with test set)	2	0.9902	11,727
R	Full	Regulatory (with test set)	0	0.9900	8,424
A	Thorough	MDL	4	0.9897	222
M	Full	MDL	4	0.9897	222
G	Smart	MDL	4	0.9896	206
B	Thorough	PSE	8	0.9891	123
C	Thorough	FCPSE	8	0.9891	123
H	Smart	PSE	8	0.9891	123
I	Smart	FCPSE	8	0.9891	123
N	Full	PSE	8	0.9891	123
O	Full	FCPSE	8	0.9891	123

Table 7 – Resulting equation for test 'K' with variable transformations and error results

Polynomial Net (GMDH) Test 'K'	
Best formula:	Y=0.1*X7-4.9E-002*X11+9.2E-002-2.1E-002*X4+1.9E-002*X9+0.41*X1-1.1E-002*X3+6.5E-002*X6-0.1*X10+4.3E-002*X5+0.56*X2-0.37*X1^2-0.38*X2^2+2.5E-002*X11^2-0.14*X2^3-6.5E-002*X11^3+0.84*X1*X2-0.24*X1*X11+0.36*X2*X11+3.2E-002*X1*X2*X11-1.9E-004*X6^2+4.1E-002*X5*X6+4.3E-002*X7^2+4.E-002*X10^2-2.6E-002*X7^3+5.E-002*X10^3-0.14*X7*X10-1.1E-002*X9^2-1.6E-002*X9^3-2.5E-002*X2*X9+1.3E-002*X5^2-2.5E-002*X6^3-1.4E-002*X1^3+2.E-002*X1*X7+3.1E-002*X6*X10+2.7E-002*X1*X3+1.4E-002*X9*X11+2.9E-002*X2*X4+1.3E-002*X8^3-1.6E-002*X8*X11+6.7E-003*X4^2+4.5E-003*X1*X6
Variable Transformations:	X1=2.*(Surf_Elev (msl)-545.)/835.-1.
	X2=2.*(Bot_Elev (msl)-244.04)/1055.96-1.
	X3=2.*(Overb_Thick (ft)-65.)/638.1-1.
	X4=2.*(MinedCoal_Thick (ft)-.07)/11.69-1.
	X5=2.*(Shale/Clay_Thick (ft)-.35)/552.55-1.
	X6=2.*Sand_Thick (ft)/262.3-1.
	X7=2.*Lime_Thick (ft)/204.97-1.
	X8=2.*TCoal_Thick (ft)/33.23-1.
	X9=2.*Accum_coalextr (Mm^3)/138.61-1.
	X10=2.*(4Mile_Buffer (acres)-2061.)/108987.5-1.
	X11=2.*(AvgAn_Precip (in)-37.5)/3.7-1.
	Y=2.*(PotentioHead (msl)-400.)/932.-1.
R squared:	0.9906
Mean squared error:	324.8997
Mean absolute error:	12.3227
Min. absolute error:	0.0014
Max. absolute error:	147.93
Correlation coefficient r:	0.9953

These test runs indicate that FPE and GCV model selection criteria work best for developing an accurate algorithm with this type of data, which are very different approaches from the other selection criteria options. The other criteria, MDL, PSE and FCPSE, were quick to drop the geologic variables out of the equation while GCV and FPE kept all variables. This was likely due to the high influence of the hydrologic variables. And the selection of 'K' suggests that while full and thorough provide the most accurate model optimization options, the complexity was also high. Equation 'K' used the smart method which retained the model accuracy, r-squared of 0.9909-0.9907 to 'K's 0.9906, and halved the complexity. Due to the retention of low error and reasonable complexity, this led to the selection of equation 'K'.

The selected equation was then validated using actual measured post-mining water levels in the Meigs Mine No. 2, permit D-0354, reported in quarterly monitoring reports (QMRs) and compared with the predicted values produced by the equation. Table 4 displays the three points of comparison using the last measurement of the year for 'South Mains Shaft' in 2017 and 2018 and the last measurement of 'Roving Crew Shaft' in 2018. Publicly accessible data for recent post-mining water level monitoring is limited so this data from a well monitored closed mine complex was what existed to work with for validation at this stage of the project. The results of applying the model to these measurements, with lithology from nearby boreholes collected separately and coal extracted variable set to the final maximum value, indicated low percent errors between actual measured water level and the algorithm predicted value. Between these three points of validation, the average percent error is 1.24%. With this low error, equation 'K' was determined to be included in the final GIS prediction model.

Table 8 – Post-mining data test wells in Meigs Mine D-0354 used for validation of ANN equation 'K' with calculated percent errors. Average percent error was 1.24%.

Permit	**Well**	**Date**	**Measured Head (ft msl)**	**Predicted Head (ft msl)**	**Error (ft)**	**Percent Error**
D-0354	Roving Crew Shaft	10/22/18	456.84	443.42	13.42	2.94%
D-0354	South Mains Shaft	10/22/18	455.94	458.22	-2.28	0.50%
D-0354	South Mains Shaft	9/11/17	456.88	458.22	-1.34	0.29%

GIS Model for Algorithm Application

A tool for applying the selected prediction equation was successfully created in ModelBuilder of ArcGIS Pro version 2.2 following the structured outlined in the previous section ArcGIS Tool Building. Figure 21 is a screenshot of the final structure of the tool in ArcGIS Pro ModelBuilder. The tool successfully extracts and combines data from input mine permit data and mine extent shapefiles to form a complete table of variable data required to apply the developed prediction equation. From this constructed attribute table, the Python script is imported as a tool to run the prediction algorithm that is able to reference specific columns in the attribute table to transform variables and apply the algorithm. The attribute table then has an added column with the predicted values of post-mining water level at each point of input. These points are then compared with nearest point of the area DEM to determine how far above or below the surface the predicted water elevation may reach with a final column added to the output points.

Tool Design

Development of the design for the ArcGIS tool began in a work flow chart that indicates required inputs, GIS tools to be run, and outputs of the tool. Figure 22 is the working flow chart for the tool development that is a simplified version of the tool and was used for reference in building the structure in ModelBuilder of ArcGIS Pro. On the left side of Figure 22 the box labeled ‘Start’ indicates all the required inputs by the user for the tool to run. The model flows from left to right, arrows indicating which tools the inputs are pulled into, represented by the yellow diamonds. The orange circles indicate shapefiles output by the processes run in the tool, grey circles indicating shapefiles that are created internally but not added as an output to the user’s map.

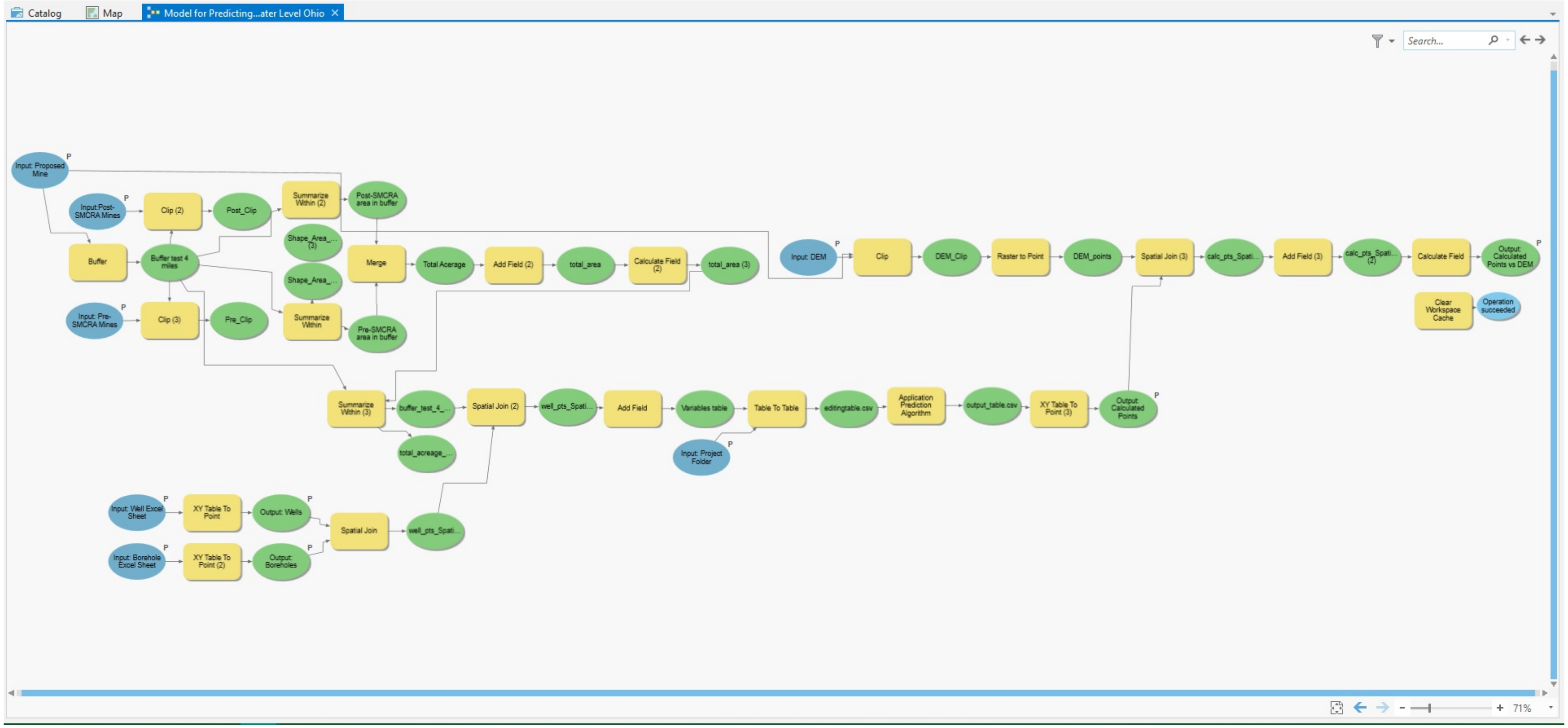


Figure 21 – A screenshot of the tool structure from within ModelBuilder in ArcGIS Pro. Inputs are blue ellipses, green ellipses are outputs, and the yellow squares are ArcGIS tools. Parameters are labeled, input and output, by the 'P' to the upper right of the shape.

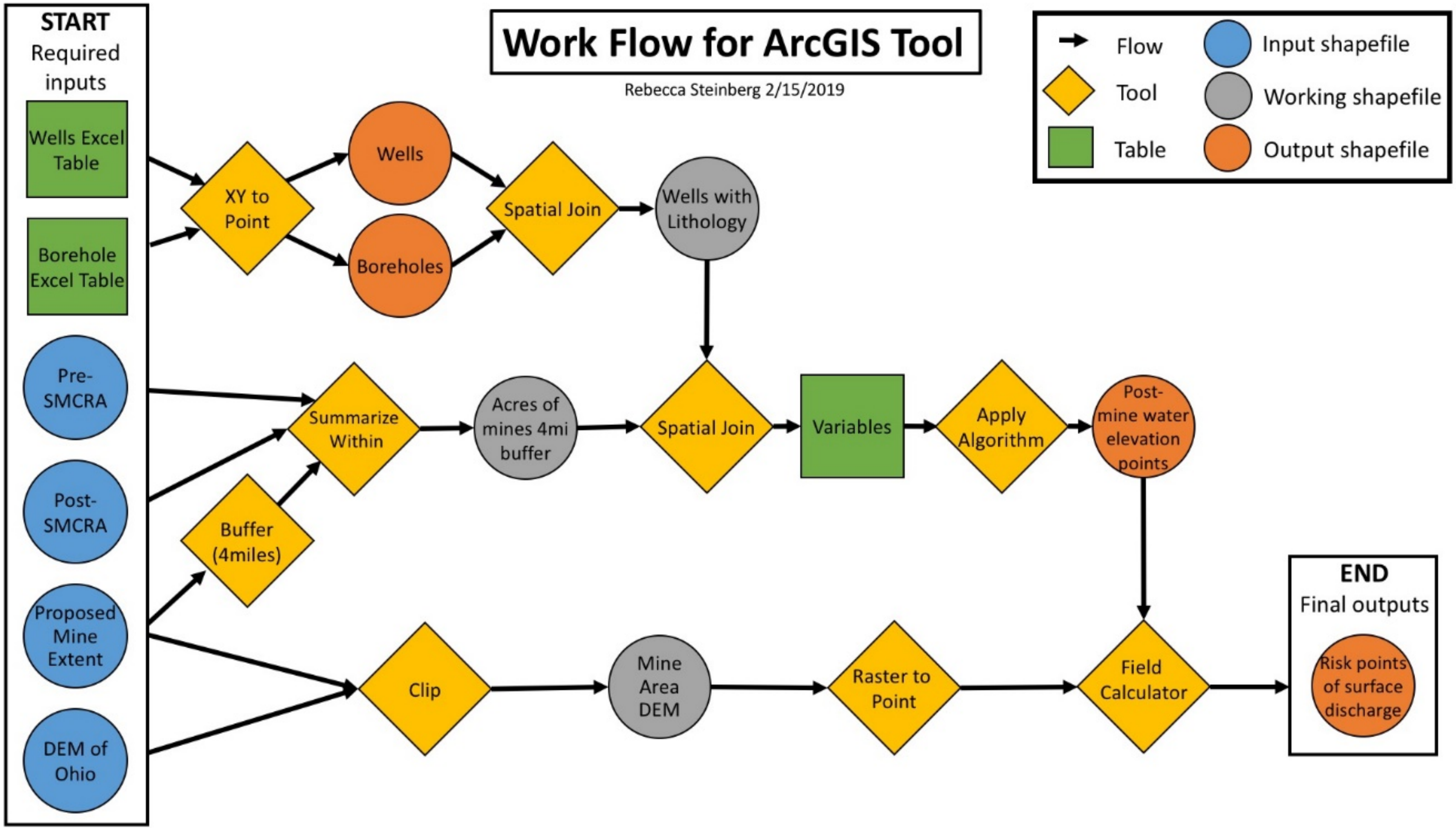


Figure 22 – Work flow diagram for the ArcGIS tool, used as a guide to develop the model in ModelBuilder of ArcGIS Pro

The required inputs are well and borehole data in standardized Excel sheets, pre- and post-SMCRA underground coal mine shapefiles, proposed coal mine shapefile, and a digital elevation model (DEM) raster of the state of Ohio. From these layers, tools in ArcGIS pull the variables needed to run the prediction equation for post-mining water level. The main table is created from the combination of the projected wells and borehole points, providing lithology to each well, as was done in the data extraction (Figure 4). The wells are the points at which the algorithm will be applied, so variables are spatially joined to the well points based on the nearest borehole. The other variable extracted is the amount of acreage mined within the 4-mile buffer of the proposed mine, determined through clipping the input shapes of pre- and post-SMCRA mines to the 4-mile buffer created around the proposed mine shape (Figure 5). The tool also calculates from the input data the bottom of coal elevation that is used when the prediction equation is applied to extrapolate the predicted post-mining water level.

Once all variables are extracted and merged into a single attribute table for the point layer, the custom Python script tool applies the prediction equation reads variables from defined columns and adds the predicted post-mining water level as another column in the table. For application of the prediction equation within the ArcGIS tool, several approaches were tested. With all variables in the same table, the possibility of using the tool 'Calculate Field' was explored. To use the ANN prediction equation in the field calculator required combining all variable transformations into a single equation. This leaves room for error in re-arranging a long complex polynomial equation. The alternative option to this approach was to develop a Python script that allows the equation to be run in steps, to avoid errors in variable transformation calculations. This custom script reads in the variable table created by the first part of the ArcGIS tool, accesses defined columns for each variable, and outputs the table with predicted values added in a new column.

Due to the importance of the format of the input data, an Excel sheet template is provided for users to organize input data in the required way. If the template is not strictly followed, variables will not be correctly labeled and result in either failure of the tool to run or inputs to the calculation of post-mining water elevation leading to an invalid result.

The final step in the tool process is the comparison of the points of predicted post-mining water level to the DEM. The DEM is converted to points of elevation so that a spatial join to the nearest elevation point can be applied to the prediction points. With the nearest elevation value added to the variable table, the final field in the attribute table is filled with the field calculator tool as the surface elevation minus the predicted head elevation, providing a measure of how far above or below the surface the water level is predicted to be at post-mining. This field calculator step also includes a conversion of units, as the DEM (as most are) is in meters and the predictions are in feet mean sea level (ft msl). This is incorporated in the ModelBuilder so that conversion of the layer is not left to the user.

Future work can be done on the creation of a spatially interpolated surface of the water table and areas of risk as a next step from the prediction points. The development of the prediction water elevation surface requires forming a continuous surface from the point data output from the algorithm. Several methods for spatial interpolation of the post-mining water level surface were tested to compare errors. Kriging methods and inverse distance weighting are being explored for methods of interpolation.

If continued work would be done on developing method for spatial interpolation, the surface of the predicted post-mining water level, a set of built in GIS tools can run to compare the DEM and the coal mine raster to the post-mining water level surface. The comparison of the coal seam raster and the predicted post-mining water level surface would show areas of possible mine pool formation (Figure 23). The difference between values of the DEM and the predicated post-mining water level surface will determine areas at risk of possible discharge to the surface. These risk areas are the main output of the tool, as well as the prediction surface and points of predicted post-mining water level.

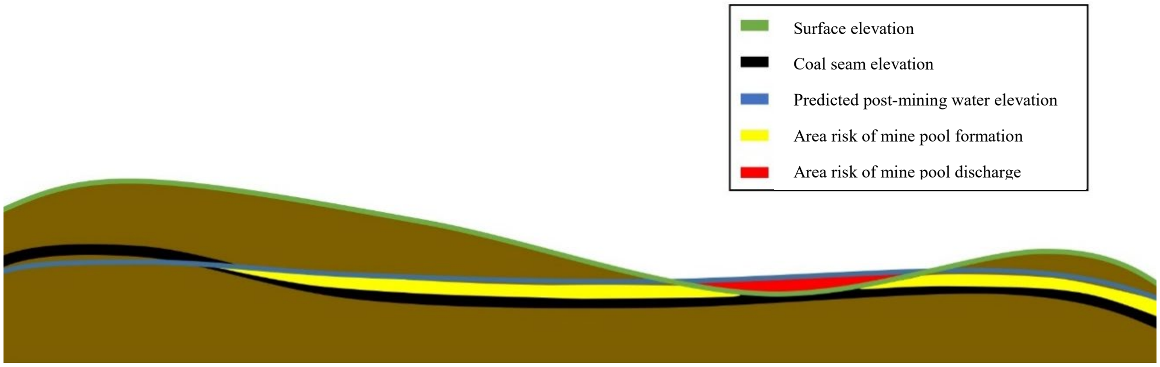

Surface elevation
Coal seam elevation
Predicted post-mining water elevation
Area risk of mine pool formation
Area risk of mine pool discharge

Figure 23 – Diagram to display the different elevation surfaces to be compared for determining areas at risk of mine pools and surface discharge

The model was then tested with a selected set of the post-SMCRA mine data for validation and trouble shooting. Once the model was running, a template map format was created that included the model for user download. As part of the packaged project with the map template, default layers are included for the required inputs, as well as templates for the Excel sheets required for inputting mine permit data, and a User's Guide developed to include in step by step instruction for running the model. Successful running and packaging of the tool required trouble shooting and discovery of bug fixes, which are also included in the User's Guide developed for the tool package (http://www.watersheddata.com/MinePool_Study.aspx).

Model Validation

Testing of the GIS model was run with existing post-SMCRA mine data previously extracted for the data analysis. Various runs were done, but the final testing was done with the shapefile of permit D-2177 with 30 well points for prediction locations. Figure 24 displays the resulting map of this analysis run with points of prediction labeled with their predicted post-mining water level values. Output by the tool are the point shapefiles of boreholes and well points of post-mining water level prediction compared to the DEM. The predictions points are symbolized displaying blue circles as greater than zero distance to surface values and red circles as negative (or less than zero) distance to the surface. These red points of negative distance to the surface are the points of predicted post-mining water level at risk of discharging to the surface (Figure 24). In this test run on D-2177, four points of predicted post-mining water level have values greater than the surface elevation that indicate a possibility of surface discharge. The user has the option of making layers, such as the coal contours or overburden thickness, visible with the results. The image shown in Figure 24 is shown with surface topography.

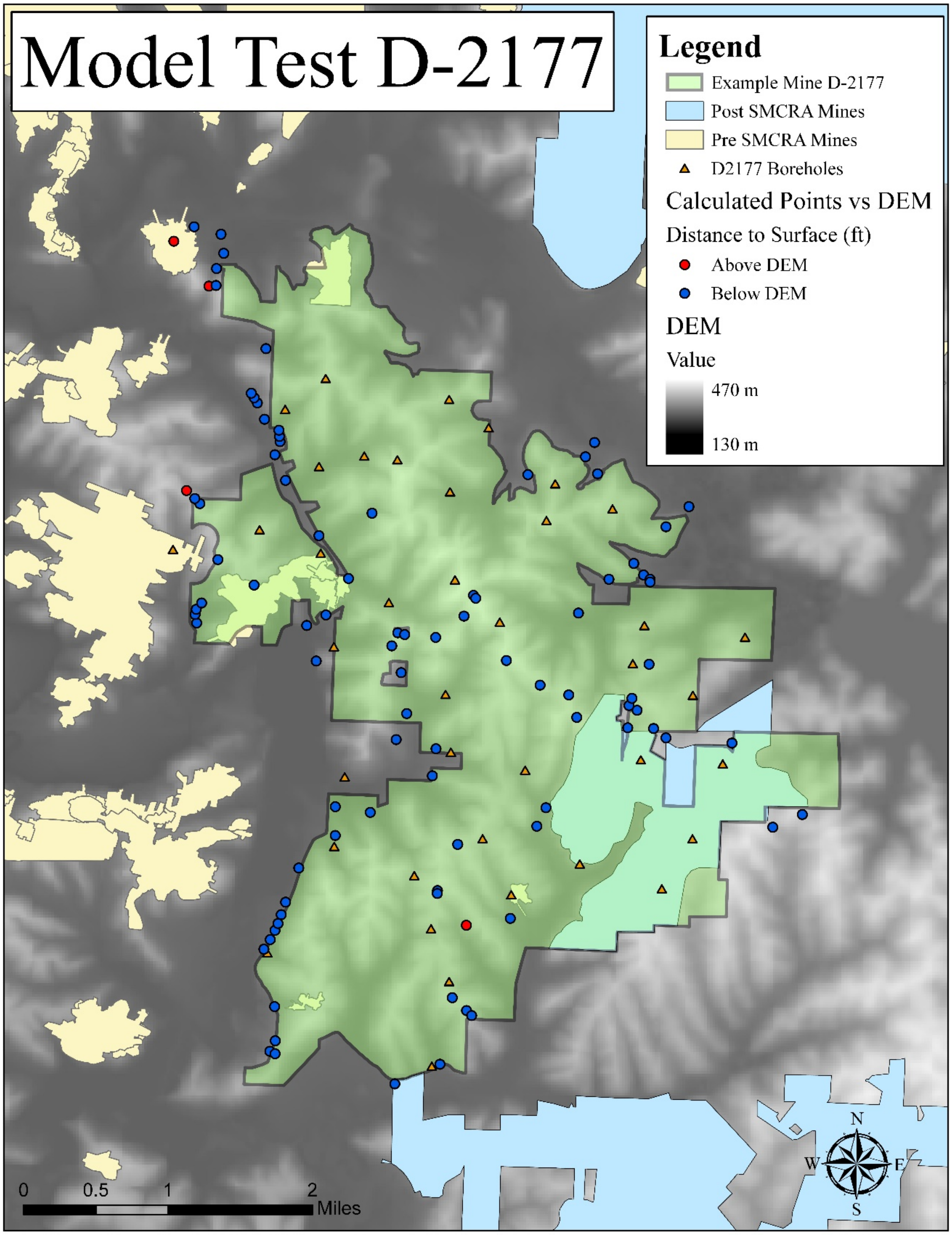


Figure 24 – Map of the final outputs of the ArcGIS model for producing points of predicted post-mining water level with a comparison to the DEM. The mine D-2177 and its permit data were used as a test for running the model.

Algorithm Application Python Script

While using existing tools, such as ‘Calculate Field’, in ArcGIS to apply the prediction equation were explored, it was ultimately determined the best way to incorporate the equation was to write a separate script to import into ArcGIS Pro ModelBuilder. Writing the script allowed for control of the exact process of extracting the correct values for each variable transformation and accurately applying the equation. Python 2.7 was used in writing the custom script.

Discussion

This project has successfully developed a multivariate statistically based empirical model for predicting post-mining water levels in underground coal mines of eastern Ohio. The methods for developing this model can be applied to develop models applicable in other regions with underground coal mines but differing geologic and hydrologic parameters.

Project Outputs

Several outputs resulted from this project. The multivariate analyses have provided an improved understanding of the relationships between the many variables examined that influence the development of mine pool. In addition to this increased understanding, the ability to develop a prediction algorithm with reasonable error is a major output of the project. Along with the algorithm itself as an output is the model developed to apply the algorithm in ArcGIS Pro. While the model is specifically an empirical model not meant to develop deterministically derived values for post-mining water level, the model is useful as a planning tool for identifying possible areas at risk for surface discharging in areas where mining is being planned. Model validation indicated a low percent error of 1.24% in output predicted post-mining water levels when compared to measured post-mining water levels, indicating that while it is an empirical model, the model can still produce predictions within reasonable error.

Model Errors

Errors in the project outputs were kept as minimal as possible through tracking percent error in the selected algorithm. The final selected algorithm from the ANN analysis had an r-squared value of 0.996, a root-mean-squared-error of 18.03, and when validated with post-mining water level data had an average error at a 1.24%.

Other areas of error possibilities are in the data itself as it is reported in the permit documents may influence the development of the model and its ability to predict post-mining water levels. There is also the aspect of human error in manual data extraction from the PDF documents into the Excel sheets that could also have influenced the model development. A source of error could also be in the availability of quality data in terms of the lack of water extraction values (where coal extraction was used in proxy), lack of borehole lithology at the exact location of the well points, and lack of detailed precipitation data instead of an outdated areal annual average. Another source of error exists in the assumption made that the empirical relationships developed from water level data from a variety of depths can be extrapolated into the mined coal layer.

Comparison of Methodologies

Figure 25 compares the PLSR results from Schafer, 2018 (A), and the later analysis considering all the head measurements (Steinberg, 2019). The re-analysis of PLSR reached 100% explained variance in 3 factors, same as with the previous run. The errors are similar, but the re-analysis with a larger data set had slightly higher error. Coefficients and relationships of the variables were also comparable to the previous run, indicating consistency in the determined relationships of variables. The correlation loading chart for the previous regression run and new regression displayed the same results in relationships of the variables (Figure 26). This re-run of the analysis validated the variable relationships with consistency between the expanded data set and the initial data set as well as a larger data set providing more reliability of the results.

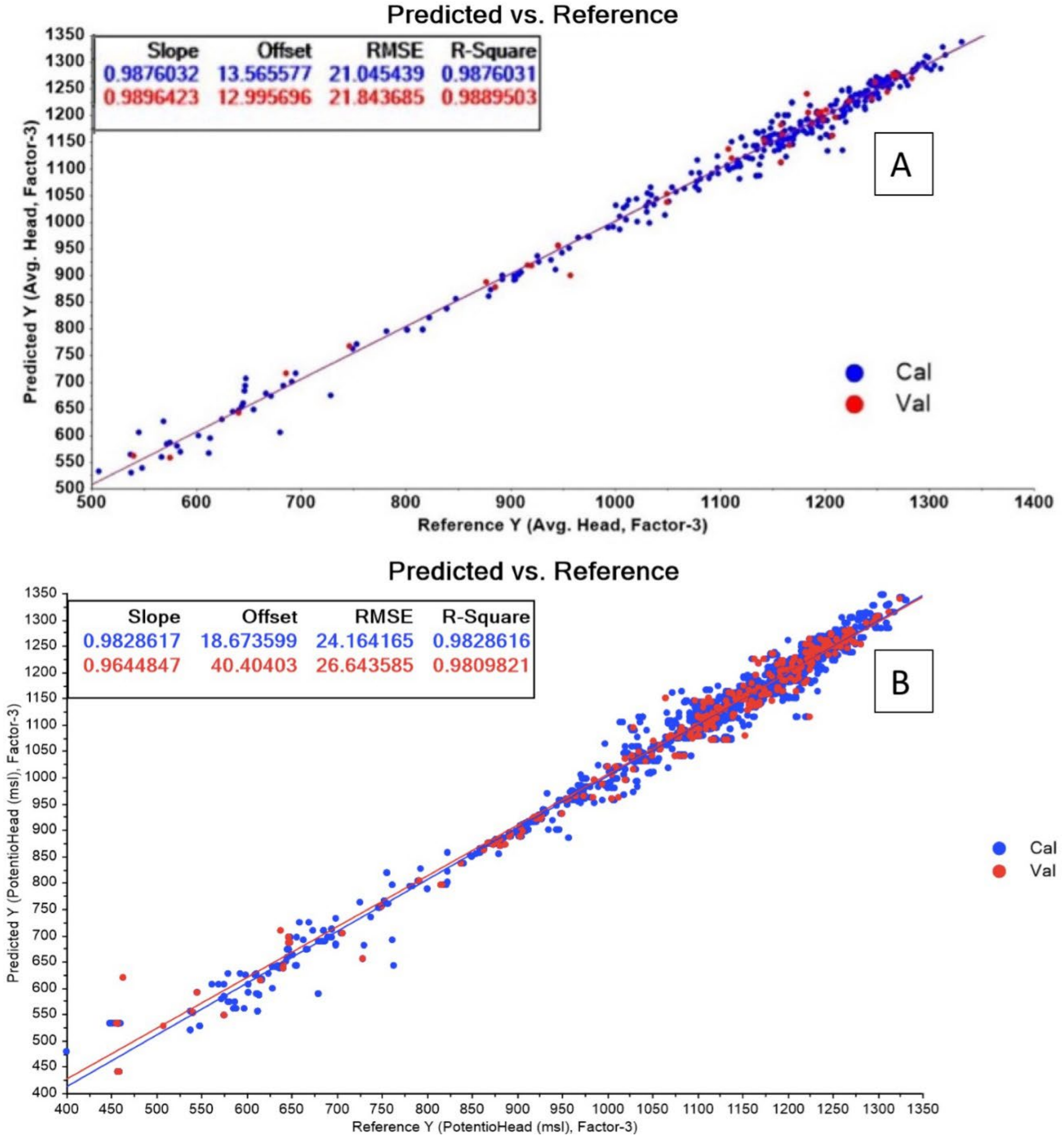


Figure 25 – Comparison of predicted versus reference regression graphs of previous PLSR analysis by Schafer, 2018 (A) and the analysis considering all the well measurements by Steinberg, 2019 (B). While r squared value was not improved, consistency was maintained with a more reliable larger data set in (B).

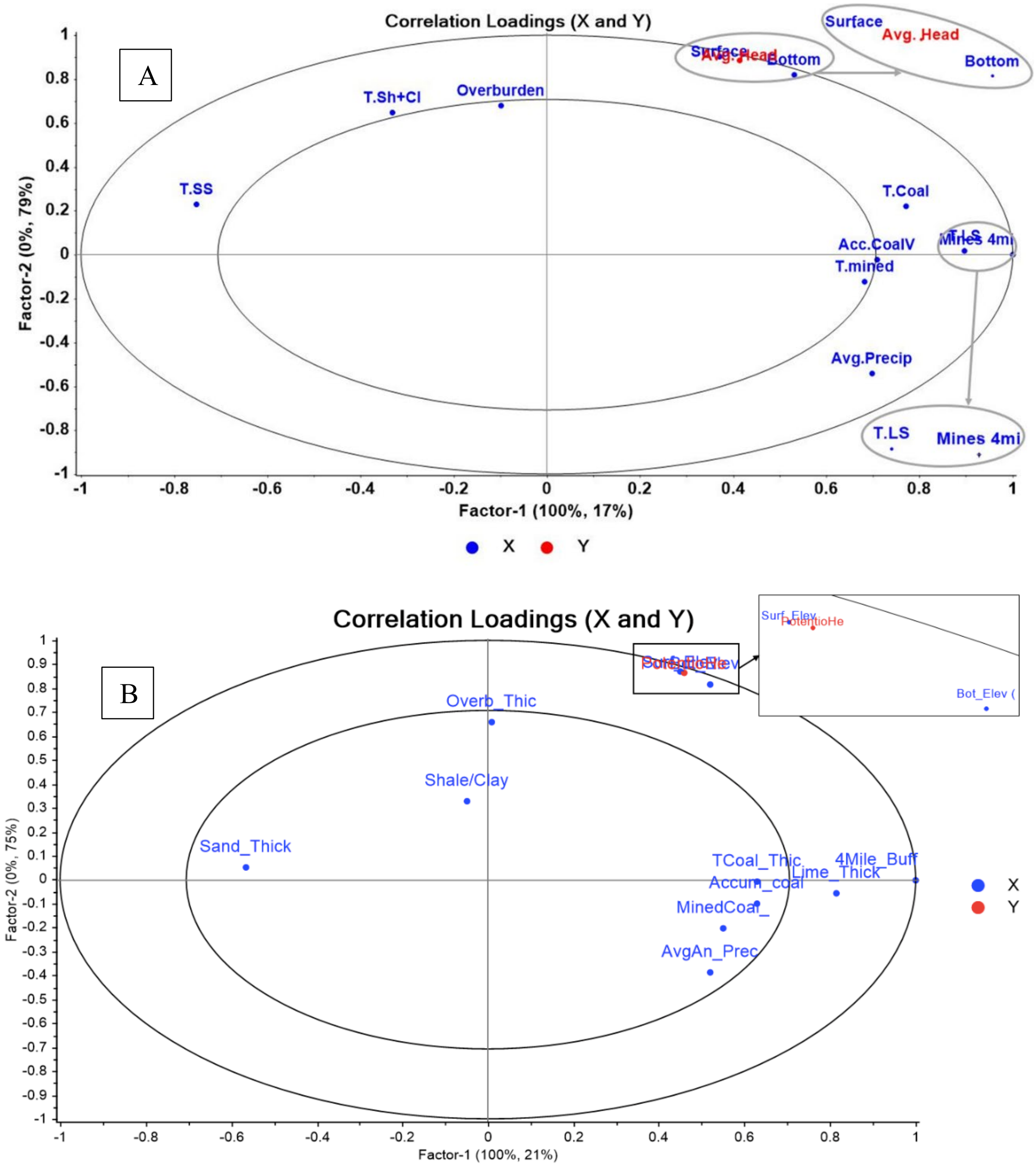


Figure 26 – Comparison of correlations loadings graphs of previous PLSR analysis run by Schafer, 2018, (A) and the analysis considering all the well measurements by Steinberg, 2019 (B). Again this comparison displays consistency in the relationships of the variables.

Conclusion

This project has addressed the need for a better method of determining post-mining water levels in undergrounds mines, using data provided in mine permit applications. The results of multivariate analyses of significant parameters guided development of an empirical model that produces estimated post-mining water levels at well locations. The final algorithm selected for use in the model was determined to estimate water levels within a reasonably usable error within 1%. Use of the model may increase the evidence for and confidence in the estimation of post-mining water level.

The methods developed during the work on this project provide the possibility of developing similar empirical models for different areas around the globe. If similar data for characterizing the area hydrology and geology can be collected, the analysis to develop the prediction algorithm can be re-run and the new area-specific algorithm can be input into the model. Results will depend on the quality and density of data collected, no matter the location.

The relationships discovered between the different hydrologic and geologic parameters have expanded on the overall understanding of how these underground mines affect the complex systems of groundwater. More research is required to determine why some of these variables are more significant than others.

Currently, requirements for hydrological data collection in the permitting process for underground coal mines in Ohio only require "*...a minimum of one test hole per one hundred sixty acres"* and do not define a requirement for well monitoring density (Ohio Administrative Code, 2016). While predictions are possible with available data, reliability of predictions would be improved, and a surface layer could be extrapolated if higher density data are collected for mine permit applications. Data depth could also be improved by the installation of piezometers in the mined layers to consistently monitor water levels, as is done in areas of coal mining in Pennsylvania. The current requirements are not sufficient for complete characterization of the area hydrology and lithology.

Recommended continued work includes improvement of the model with additional quality data, to explore the possibility of spatial interpolation methods working with a data set of higher number and higher density data. A study could be done to determine a range of necessary density of data and number of points to produce an interpolated surface with low error.

In addition to further developing the model and exploring spatial interpolation possibilities, the next step to predicting if a mine will discharge is if that discharge would be pollutional. This would require determining additional variables to the predictive model related to surface water chemistry.

This prediction model is specific to the coal fields analyzed in Ohio, but methods to develop the predictive model could be used to translate the prediction model to another area of differing geology and hydrology. In addition to applying to another area, the full extent of this model predictability would need to be determined. Continued work could be looking at how far this model can predict post-mining water levels outside of the state of Ohio but still within similar lithology.

Methods used to develop this model and approach to predicting water levels could be applied outside of underground mining as well. Other issues in understanding the multivariate relationships impacting the change in groundwater levels could adapt the approaches used in this project to address issues in other disciplines outside of mining.

Acknowledgments

The OHIO research team would like to acknowledge the Project Steering Committee for their input, ODNR for providing mine permits, Geologic Society of America for student travel grants, and the Ohio University Department of Geological Sciences Alumni Association for summer stipend support for students.

References

Akcil, A., and Koldas, S., 2006, Acid Mine Drainage (AMD): causes, treatment and case studies: Journal of Cleaner Production, v. 14, p. 1139–1145, doi:10.1016/j.jclepro.2004.09.006.

CAMO Software AS, 2006, The Unscrambler User Manual - The Unscrambler Methods (Version 9.6):, https://www.camo.com/downloads/U9.6%20pdf%20manual/The%20Unscrambler%20Methods.pdf (accessed March 2019).

CAMO Software AS, 2019, The Unscrambler X: A commercial software product for multivariate data analysis, http://www.camo.com/rt/Products/Unscrambler/unscrambler.html.

Crowell, D.L., 2005, GeoFacts No. 14: History of Coal Mining in Ohio:, http://geosurvey.ohiodnr.gov/portals/geosurvey/PDFs/GeoFacts/geof14.pdf (accessed February 2018).

ESRI, 2019c, What is ModelBuilder?—ArcGIS Pro | ArcGIS Desktop:, https://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/modelbuilder/what-is-modelbuilder-.htm (accessed March 2019).

Lottermoser, B., 2015, Predicting Acid Mine Drainage: Past, Present, Future: Leiter des Institute of Mineral Resources Engineering Mining Report 151 No. 6, 480–489 p., https://mining-report.de/english/predicting-acid-mine-drainage-past-present-future/ (accessed January 2018).

Lopez, D., and Kruse, N.A., 2015, Tools to predict the hydrological response and mine pool formation in underground mines:

Means, B., Montrella, J., Greenfield, G., and Winter, J., 2018, Mine Pool Prediction & Validation Oversight Study:

ODNR, 2019, ODNR Mines of Ohio Viewer:, https://gis.ohiodnr.gov/MapViewer/?config=OhioMines (accessed March 2018).

ODNR Division of Water Resources, 1980, Ohio DSWR Hydrologic Atlas:, http://water.ohiodnr.gov/maps/hydrologic-atlas#PRE (accessed March 2019).

ODNR Geographic Information Systems, 1997, GIS Data Search by Category:, http://geospatial.ohiodnr.gov/data-metadata/search-by-category (accessed March 2019).

Ohio Administrative Code, 2016, Underground mining permit application requirements for information on environmental resources:, http://codes.ohio.gov/oac/1501%3A13-4 (accessed February 2019).

Pradhan, B., 2010, Remote sensing and GIS-based landslide hazard analysis and cross-validation using multivariate logistic regression model on three test areas in Malaysia: Advances in Space Research, v. 45, p. 1245–1256, doi:10.1016/j.asr.2010.01.006.

Sánchez-Mesa, J.A., Galan, C., Martínez-Heras, J.A., and Hervás-Martínez, C., 2002, The use of a neural network to forecast daily grass pollen concentration in a Mediterranean region: the southern part of the Iberian Peninsula: Clinical & Experimental Allergy, v. 32, p. 1606–1612, doi:10.1046/j.1365-2222.2002.01510.x.

Schafer, L.A., 2018, Statistical Analysis of Mining Parameters to Create Empirical Models to Predict Mine Pool Formation in Underground Coal Mines: Ohio University, https://etd.ohiolink.edu/pg_10?::NO:10:P10_ETD_SUBID:165981 (accessed January 2019).

Singer, P.C., and Stumm, W., 1970, Acidic Mine Drainage: The Rate-Determining Step: Science, v. 167, p. 1121–1123.

Twumasi, F., 2018, Applying MODFLOW and Artificial Neural Networks to Model the Formation of Mine Pools in Underground Coal Mines: Ohio University, https://etd.ohiolink.edu/pg_10?::NO:10:P10_ETD_SUBID:166042 (accessed January 2019).

U.S. Department of Labor, 2019, MSHA - Mine Data Retrieval System (as developed by PEIR) Home Page: Mine Data Retrieval System, https://arlweb.msha.gov/drs/drshome.htm (accessed June 2018).

Ward Systems Group, Inc., 2019, NeuroShell 2 Help:, http://www.wardsystems.com/manuals/neuroshell2/index.html?idxhowuse.htm (accessed March 2019).

Wei, X., Zhang, S., Han, Y., and Wolfe, F.A., 2017, Mine Drainage: Research and Development: Water Environment Research, v. 89, p. 1384–1402, doi:10.2175/106143017X15023776270377.

Underwood, B.E., Kruse, N.A., and Bowman, J.R., 2014, Long-term chemical and biological improvement in an acid mine drainage-impacted watershed: Environmental Monitoring and Assessment, v. 186, p. 7539–7553, doi:10.1007/s10661-014-3946-8.

Younger, P.L., 2000, Predicting temporal changes in total iron concentrations in groundwaters flowing from abandoned deep mines: a first approximation: Journal of Contaminant Hydrology, v. 44, p. 47–69, doi:10.1016/S0169-7722(00)00090-5