



pennsylvania

DEPARTMENT OF TRANSPORTATION

Archaeological Predictive Model Set

FINAL REPORT

March 2015

By Matthew D. Harris,
Robert G. Kingsley, and
Andrew R. Sewell

URS

COMMONWEALTH OF PENNSYLVANIA
DEPARTMENT OF TRANSPORTATION

CONTRACT # 355I01
PROJECT # 120205

URS



1. Report No. FHWA-PA-2015-002-120205	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Archaeological Predictive Model Set		5. Report Date March 2015	
		6. Performing Organization Code	
7. Author(s) Matthew D. Harris, Robert G. Kingsley, and Andrew R. Sewell		8. Performing Organization Report No.	
9. Performing Organization Name and Address URS Corporation 437 High Street Burlington, NJ 08016		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No. 120205, RFQ #120205	
12. Sponsoring Agency Name and Address The Pennsylvania Department of Transportation Bureau of Planning and Research Commonwealth Keystone Building 400 North Street, 6 th Floor Harrisburg, PA 17120-0064		13. Type of Report and Period Covered Archaeological Predictive Modeling, Pre-Contact and Contact Native American	
		14. Sponsoring Agency Code	
15. Supplementary Notes This document is the final report summarizing the data, methods, findings, and conclusions of the Statewide Archaeological Predictive Model Set project. This volume builds off the content of six previous reports produced for the project and should be used in concert with them.			
16. Abstract <p>This report is the documentation for Task 7 of the Statewide Archaeological Predictive Model Set. The goal of this project is to develop a set of statewide predictive models to assist the planning of transportation projects. PennDOT is developing tools to streamline individual projects and facilitate Linking Planning and NEPA, a federal initiative requiring that NEPA activities be integrated into the planning phases for transportation projects. The purpose of Linking Planning and NEPA is to enhance the ability of planners to predict project schedules and budgets by providing better environmental and cultural resources data and analyses. To that end, PennDOT is sponsoring research to develop a statewide set of predictive models for archaeological resources to help project planners more accurately estimate the need for archaeological studies.</p> <p>The outcome of this project, contained in seven task-specific reports, documents the development of numerous statistical models created to assess the sensitivity of the landscape for the presence of Native American archaeological sites. The seven task reports of this project are organized as follows: Task 1, background literature review; Task 2, organization of study areas by physiography; Task 3, pilot model study; Task 4, models and results for Regions 1, 2, and 3; Task 5, models and results for Regions 4, 5, and 6; Task 6, models and results for Regions 7, 8, 9, and 10; and Task 7, final project synthesis. Each of Tasks 4, 5, and 6 document the bulk of this undertaking by describing the data preparation, model building process, and results for each of 10 regions that constitute the Commonwealth of Pennsylvania. The Task 7 report synthesizes the methodologies, illuminates the model building process, discusses model validation and findings, and offers possible avenues for future research.</p>			
17. Key Words Archaeological Predictive Modeling, APM		18. Distribution Statement No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 270	22. Price N/A

**PENNSYLVANIA DEPARTMENT OF TRANSPORTATION
ARCHAEOLOGICAL PREDICTIVE MODEL SET
TASK 7: FINAL REPORT**

CONTRACT #355I01

Prepared for

Pennsylvania Department of Transportation
Bureau of Planning and Research
Keystone Building
400 North Street, 6th Floor, J-East
Harrisburg, PA 17120-0064

Prepared by

Matthew D. Harris, Principal Investigator
Robert G. Kingsley,
and
Andrew R. Sewell, Hardlines Design Company

URS Corporation
437 High Street
Burlington, NJ 08016-4514

March 2015

ABSTRACT

This report is the documentation for Task 7 of the Statewide Archaeological Predictive Model Set project sponsored by the Pennsylvania Department of Transportation (PennDOT). This project was solicited under Contract #355I01, Transportation Research, Education, and Technology Transfer ITQ, Category #05 – Environmental Research. The goal of this project is to develop a set of statewide predictive models to assist the planning of transportation projects. PennDOT is developing tools to streamline individual projects and facilitate Linking Planning and NEPA, a federal initiative requiring that NEPA activities be integrated into the planning phases for transportation projects. The purpose of Linking Planning and NEPA is to enhance the ability of planners to predict project schedules and budgets by providing better environmental and cultural resources data and analyses. To that end, PennDOT is sponsoring research to develop a statewide set of predictive models for archaeological resources to help project planners more accurately estimate the need for archaeological studies.

The objective of Task 7 is to discuss the results of the project, expand on the methodology used, explore the results, and provide recommendations for model use and future directions.

TABLE OF CONTENTS

Abstract	i
List of Figures	iii
List of Tables	v
Executive Summary	1
1. Introduction	3
Data Quality	5
2. Study Areas	10
Study Region Delineation	10
Prehistoric Background	11
Archaeological Sites in Pennsylvania	17
3. Model Methodology	21
Introduction	21
Previous Studies in APM	23
Software Platform	25
Study Area Delineation and Primary Variable Creation	28
Secondary Variable Creation	30
Variable Discrimination and Prediction	31
Data Extraction and Variable Testing	34
Primary Variable Selection	37
Creation of Regression/Classification Data	39
Fitting of Statistical Models: Discussion	40
Fitting of Statistical Models: Pennsylvania Model Project	50
Final Sensitivity Layer Prediction and Thresholds	52
Final Model Selection	55
Statistical Models: Stepwise LR, MARS, and RF	56
Random Forest	60
Computational Requirements	65
4. Findings and Results	67
Classification Errors and Performance	70
Model Benchmarks	73
Model Accuracy	79
Variable Importance	81
5. Conclusions and Recommendations	89
Usage Recommendations	90
Improvements to Current Methods	96
6. References Cited	102
Appendix A. Comprehensive List of Acronyms and Glossary of Terms	
Appendix B. Variables Considered Throughout the Project	
Appendix C. Pennsylvania Model Subarea Maps	

LIST OF FIGURES

Figure 1.	Modeling regions for the Pennsylvania Model Set project.	10
Figure 2.	General organization of entire model building process.	22
Figure 3.	Occurrences of search terms in titles and abstracts of articles in the journal <i>American Antiquity</i>	23
Figure 4.	Software workflow.....	26
Figure 5.	Example of pseudo-code.	28
Figure 6.	Example of variable discrimination; red area is background and blue area is site locations.	33
Figure 7.	Boxplots of variable measurements at known sites (colored boxes) and in the background environment (white box).	34
Figure 8.	Pseudo-code for creating data frame of site-present values for environmental variables; referred to as big_df.....	35
Figure 9.	Pseudo-code for repeated K-S and MW tests for variable discrimination.	36
Figure 10.	Example of K-S test and associated D statistic.	37
Figure 11.	Pseudo-code for primary variable selection process.	38
Figure 12.	Pseudo-code for extracting a sample of point-specific background values for selected variables and joining site-present data.	40
Figure 13.	General organization of model fitting and prediction process.	41
Figure 14.	Graphical illustration of bias and variance.....	43
Figure 15.	Schematic of high bias and high variance model fits.....	44
Figure 16.	Bias and variance tradeoff for model complexity.	45
Figure 17.	Sample error and model complexity.	46
Figure 18.	Schematic diagram of model parameter optimization and fitting sequence.	48
Figure 19.	Schematic of k-folds cross validation technique.....	49
Figure 20.	Example of plotting 10-fold CV error.....	50
Figure 21.	Pseudo-code for model parameterization and error estimation.....	51
Figure 22.	Pseudo-code for preparing raster layers and predicting sensitivity.....	53
Figure 23.	Schematic example comparing linear to logistic regression.	57
Figure 24.	Schematic of first pass, over-fit linear terms and hinge functions of MARS model.....	59
Figure 25.	Schematic example of pruned linear terms and hinge functions of MARS model.	60
Figure 26.	Diagram of single decision tree.....	62
Figure 27.	Pseudo-code showing the general logic of the Random Forest algorithm.	63
Figure 28.	Schematic of prediction based on decision tree ensemble.	64
Figure 29.	Percentage of model types selected for the subareas of each region.....	70
Figure 30.	Boxplots of Kg and FNR metrics for reference models and Pennsylvania model.....	76
Figure 31.	Scatterplot of Kg and FNR metrics for reference models and Pennsylvania model.....	77
Figure 32.	Pearson's correlation r value for the 30 most important variables.....	87

Figure 33. Overview of final sensitivity layer.	90
Figure 34. Example of raster aggregation based on a 2-cell neighborhood and a maximum value function.	91
Figure 35. Example of sensitivity assessment at original 10 × 10-m resolution.	92
Figure 36. Example of sensitivity assessment at aggregate of neighborhood maximum 30 × 30-m resolution.	93
Figure 37. Model type by subarea.	94
Figure 38. Density of known archaeological sites per square mile within each subarea.	95

LIST OF TABLES

Table 1	Source of Site Location in PASS GIS Database.....	8
Table 2	Pennsylvania Site Types by Landform	18
Table 3	Quantities of Model Project Attributes	67
Table 4	Area, Archaeological Site Count, and Site Density per Region	68
Table 5	Quantification of Model Types by Landscape Position.....	69
Table 6	Confusion Matrix and Performance Metrics for Statewide Mosaicked Model	71
Table 7	Schematic of Confusion Matrix and Performance Metrics.....	72
Table 8	Kg and FNR Metrics of Models Evaluated in the Task 1 Report.....	75
Table 9	Metrics of ER Survey Model and Pennsylvania Model.....	78
Table 10	Error Rates for OOB and Hold-Out Samples for Three Model Types	80
Table 11	Variable Importance Measures for Riverine Subareas of Regions 4–10	83
Table 12	Variable Importance Measures for Upland Subareas of Regions 4–10	84
Table 13	Variable Importance Values Centered and Scaled Comparing Riverine and Upland Settings	85

EXECUTIVE SUMMARY

This summary covers the Statewide Archaeological Predictive Model Set project sponsored by the Pennsylvania Department of Transportation (PennDOT). This project was solicited under Contract #355I01, Transportation Research, Education, and Technology Transfer ITQ, Category #05 – Environmental Research. The goal of this project is to develop a set of statewide predictive models to assist the planning of transportation projects. PennDOT is developing tools to streamline individual projects and facilitate Linking Planning and NEPA, a federal initiative requiring that NEPA activities be integrated into the planning phases for transportation projects. The purpose of Linking Planning and NEPA is to enhance the ability of planners to predict project schedules and budgets by providing better environmental and cultural resources data and analyses. To that end, PennDOT is sponsoring research to develop a statewide set of predictive models for archaeological resources to help project planners more accurately estimate the need for archaeological studies.

The outcome of this project, contained in seven task-specific reports, documents the development of numerous statistical models created to assess the sensitivity of the landscape for the presence of Native American archaeological sites. The seven task reports of this project are organized as follows: Task 1, background literature review; Task 2, organization of study areas by physiography; Task 3, pilot model study; Task 4, models and results for Regions 1, 2, and 3; Task 5, models and results for Regions 4, 5, and 6; Task 6, models and results for Regions 7, 8, 9, and 10; and Task 7, final project synthesis. Each of Tasks 4, 5, and 6 document the bulk of this undertaking by describing the data preparation, model building process, and results for each of 10 regions that constitute the Commonwealth of Pennsylvania. The Task 7 report synthesizes the methodologies, illuminates the model building process, discusses model validation and findings, and offers possible avenues for future research.

To achieve the goal of this project, statistical models were developed to analyze the landscape at known Native American archaeological sites in Pennsylvania and extrapolate identified patterns to all areas of the state. The model building process included the use of three statistical machine learning algorithms: backwards stepwise logistic regression based on the Akaike Information Criterion, Multivariate Adaptive Regression Splines, and the Random Forest algorithm. These algorithms were employed in a best-practices framework that included feature selection, cross-validation for model parameterization and selection, and evaluation on independent samples. This process was repeated for each of 132 study areas that cover the extent of Pennsylvania. The final sensitivity layer derived from this process is a composite of predictions from the algorithm that best characterized that data for each area of the 132 study areas. In addition, the predictive output of each of the algorithms will be returned to PennDOT for use in future studies.

The models that resulted from this project accurately describe the pattern, to the extent one exists, of the relationship between archaeological sites and the environmental variables used to model them in

each subarea. There is, of course, room for improvement in these models through new or refined archaeological site location data, different predictor variables, and refinements of the modeling methodology. As a first best-approximation, the results of this project are successful in identifying areas of the landscape with heightened sensitivity for Native American archaeological material. They are appropriate for use on a planning scale of approximately 1:24,000 to assess the relative sensitivity for Native American archaeological material when comparing transportation alternatives. Additionally, these models are useful in the development of field survey priorities over broad areas. The models and sensitivity layers that result from this project cannot and do not replace the need for actual archaeological field survey. Field testing of these models is critical in understanding the true false-negative error rate and in evaluating the utility within different environments. Finally, these models are not static or final assessments of archaeological sensitivity. These models should serve as a starting point for field testing and future iteration based on test results and improvements in statistical techniques and our understanding of the dynamics of archaeological settlement systems.

1

INTRODUCTION

The purpose of this project is to use the existing Pennsylvania Archaeological Site Survey (PASS) file database to produce a baseline model for the sensitivity of pre-contact and contact Native American site-presence throughout the entire Commonwealth using Archaeological Predictive Modeling (APM). The resulting assessments of archaeological sensitivity will be used by transportation, planning, and other cultural resources management (CRM) practitioners to make better-informed and more consistent assessments of prehistoric archaeological sensitivity, with the ultimate goal of saving time, money, and sparing cultural resources.

Building from the previous tasks in this project—a review of APM literature (Task 1: Harris 2013a), designation of study regions (Task 2: Harris 2013b), the creation of a pilot model for central Pennsylvania (Task 3: Harris 2014), and modeling ten regions throughout Pennsylvania (Task 4: Harris et al. 2014a; Task 5: Harris et al. 2014b; Task 6: Harris et al. 2014c), this document is the final report summarizing the data, methods, findings, and conclusions of the project. This report builds off the content in the previous six reports and is designed to be a synopsis of the project and a summary of the modeling methodology that developed throughout. The previous task reports stand as individual volumes documenting more specific details of the project. Where relevant, the previous task reports are referenced in this report and, when needed, cogent concepts are reviewed for clarity. The full documentation of this project is contained within the seven-volume collection of all task reports.

The major accomplishments of this study are: 1) a complete statewide raster layer of archaeological sensitivity aggregated from 132 spatial subareas and the output from 528 statistical models; 2) a semi-automated, scalable, and parallelized model-building process capable of reproducing this analysis or repeating this analysis with a wide variety of statistical model types; and 3) a major update to the practice and practical considerations of APM using modern statistical methods.

To the first point, the final model sensitivity raster layer covers the roughly 46,000 square miles of the Commonwealth, with 1,065,669,566 (10.5×10.5 -m) cells each coded as a 1, 2, or 3 for low, moderate, and high sensitivity, respectively. This layer is the mosaic of the 132 subarea models, each of which is a selection from the candidate models derived from the Logistic Regression (LR), Multivariate Adaptive Regression Splines (MARS), or Random Forest (RF) algorithms, or a proportionally weighted additive sum model created for each subarea. This mosaic represents an expression of the relationships between 18,226 known prehistoric site locations¹ and a number of

¹ A total of 18,265 archaeological sites with prehistoric components were recorded in the PASS data base as of October 2013, but only 18,226 of these were used in the modeling process (listed in Appendix A in the Task 2 report). Sites that were either too small (less than a single raster cell) or too large (indicating that they were imprecisely drawn) were eliminated from the modeling pool because of the bias they would introduce.

environmental factors drawn from a pool of 93 total variables. The pattern in this layer is a direct reflection of the similarity of landforms at which we have recorded sites in the past and is therefore inductive. That is, the individual models that were aggregated to make the final layer were computed directly from known site locations and make few assumptions about where sites “should” be versus where we have recorded them. For the intended purpose of planning for transportation and other infrastructure projects at the appropriate scale (approximately 1:24,000), this layer is a valid tool for qualifying and quantifying the general landscape trends in Native American pre-contact settlement patterns as expressed in the Pennsylvania Archaeological Site Survey (PASS) database.

Second, the methodology that was created to carry out this project was specifically developed to be repeatable and reproducible. This is in opposition to many GIS and analytical tasks that rely on serialized manual processes with little documentation and lots of repetition with room for error. The codified process built for this project breaks the analysis into modular chunks that can be used and reused without reinvention. In this system, additions to the analysis such as new model types, variables, variable selection, etc., can be added into an existing framework. The components of this framework require minimal supervision once the correct parameters and data are set. The framework is efficient in that routines are optimized for faster data processing, and the model parameterization is parallelized to take full advantage of processor resources. Further, the code base is set up to be used on a remote server, such as Amazon AWS, with minimal adjustment.

Third, the APM created for this project (Pennsylvania model) is one of the largest and most detailed ever published. The only other statewide model of this type is for Minnesota at roughly twice the size of Pennsylvania. The first number of model iterations for the Minnesota model (Mn/Model) used 30×30 m as a base resolution as opposed to the approximately 10×10 -m cells of this model. Even with half the land area, the Pennsylvania model contains more than twice the number of raster cells as the Phase III Mn/Model. The Phase IV Mn/Model is planned to also use 10-m resolution base data and recursive portioning algorithms (Oehlert and Shea 2007). The Pennsylvania and Minnesota models stand alone in the area covered and data density. The Pennsylvania model is among the first published use of both the MARS and RF (Märker and Heydari-Guran 2009; Heilen 2013:5; Menze and Ur 2013) algorithms in archaeology, and certainly the first on this scale. Aside from using new and interesting techniques, however, this project sought to create a bridge from more traditional methods, theories, and assumptions to the use of modern data mining and Machine Learning (ML) algorithms and associated tools such as cross-validation (CV), variables selection, and bootstrapping. Through the background literature research presented in the Task 1 report, the use of the more traditional logistic regression, and the reliance on many of the theoretical boundaries established in the seminal work of Judge and Sebastian (1988), this project is not simply a use of the latest technology to do what has been done before. Additionally, building from the more recent and statistically informed work of Verhagen, Kammermans, and Oehlert and Shea this project benefits from the trials and errors encountered throughout the development of APM. It is hoped that not only does this project result in a useful product, but also serves as an object for both critique and improvement to further the discussion of modeling in archaeology.

DATA QUALITY

Of all the project components, site location data is by far the most important. Obviously, without knowing where sites are, it would be more than difficult to project where they could be. However, beyond the obvious, the characteristics of locational data for archaeological sites (drawn from the PASS database) governed the choices of methods and imposed numerous constraints on analysis. Simply, the data dictated the methods. Some of the most pressing characteristics of archaeological locational data are discussed below, in no particular order:

Bias

The majority of PASS sites (65%, $n = 14,415$) are reported voluntarily, from unsystematic survey, or from unknown sources, generally by amateur archaeologists or those with knowledge of where others have found artifacts. These reports range from highly specific to very general in terms of locational and attribute accuracy. The smaller number of PASS site locations (35%, $n = 7,727$) recorded by systematic survey, often by professional archaeologists through the environmental review process, are generally accepted to be more accurate on average, but still have quite a range of data completeness. One thing that is shared by all forms of site recordation is a general bias toward locating sites where sites are expected to be located, leading to a bias in site location. Attempting to limit this project's analysis to only those sites derived from systematic survey would greatly diminish the available data, disregard nonsystematic site records that may be very accurate, and ignore the fact that systematic survey is not necessarily significantly less biased than judgmental survey.

Neither professional nor nonprofessional archaeologists conduct archaeological surveys as pure random samples—doing such would go against many of the reasons to do what they do (e.g., recreation or project specific compliance). On the contrary, both sets of archaeologists often focus on landscapes that match their mental model for site preference. Further, environmental compliance surveys are often requested or required in areas considered to be high potential based on the same mental model. Often generalized to identify location near water, on level and dry ground, this mental model is developed through experience and training, and many archaeologists have a very keen sense of where a site may be given information about a landscape. Likely this mental model, which has been in some fashion always part of archaeological survey, often works because it can key into some portion of the true pattern of site location. However, these same archaeologists who have good “site radar” all know that sometimes a site can be found where least expected.

This anecdotal situation outlines what is perhaps the greatest issue in inductive APM, that the model will reinforce the bias found in the data. This is a valid fear and is difficult to avoid for any method that is based on true observations—inductive, deductive, and otherwise. However, it is not a fatal flaw. Similar to the archaeologist's experience, the methods used in APM seek to find the pattern in where we know sites are located and project that pattern to new areas. Further, these methods are able to analyze a much larger number of site locations and locational factors to derive an estimated

pattern based on likelihood to vast geographic areas, a task beyond human intuition. That the known pattern is a potentially biased representation of the true pattern is the reason that models are developed to be tools that aid in decision making, not to be the arbiter of decisions (Kamermans 2008). However, the acceptance of bias as a fact of the data does not license the model builder to simply mimic what is already known. As will be discussed in the methods section of this report, the use of variance reducing methods that add to the decrease in generalization error are a way of finding a broader pattern with the site location data.

Low Prevalence

Prevalence is simply a measure of how common a given condition (site-presence) is within a larger population. When considering how common archaeological sites are within a geographic population that includes the entire area of the State of Pennsylvania, they have a very low prevalence. From the point of view of this project, prevalence is measured as the number of site-present cells ($\sim 10 \times 10\text{-m}$) within the universe of cells that compose the state. To quantify this, this project considered 2,024,242 site-present cells from a population of 1,056,897,903 cells in the entire state for a prevalence of 0.0019—a very low prevalence. A more realistic view of site prevalence can be gained by calculating the same, but for only sites within areas that have been surveyed through the environmental review process. This leads to 277,649 site-present cells within 27,841,595 surveyed cells, for a prevalence of 0.01. Stated in another way, this means that 1% of surveyed areas are mapped as a prehistoric archaeological site or site component. If we are to assume that 0.01 is the true prevalence of prehistoric sites within the state, then we can model the probability of finding a site anywhere within an area requested for environmental review survey in the future as 0.045, or 4.5%.

What all this equates to is that prehistoric archaeological sites are rare occurrences within Pennsylvania. This fact has large implications in the modeling process, the basics of which are that negative data can easily overwhelm the positive data if precautions are not taken. Fortunately, many other fields of study (e.g., medical screening, fraud detection, spam filtering, etc.) encounter this same issue and have developed methods to help address issues associated with highly imbalanced data. These methods include stratified resampling, down sampling, Synthetic Minority Over-Sampling Technique (SMOTE) sampling, class probability weights, and others. The methods section of this report discusses the methods employed in this project, as does the Task 4 report (p. 78). In addition to more technical modeling issues, the fact of low prevalence also affects model interpretation. Essentially, a model predicting an occurrence of such low prevalence will tend to reveal a rather sparse estimate that may not be in line with expectations. In other words, for many purposes an archaeologist may feel uncomfortable with a model that predicts only 2% of a large area as highly sensitive for prehistoric archaeological sites. However, from a statistical perspective this may be a very generous assessment. In addition to some of the sampling solutions listed above, this project uses model probability thresholds that take prevalence into account.

Nonmechanistic and Noncontinuous

Simply put, sites can exist anywhere and everywhere, and the reasons they are there are only partially known to archaeologists. Sites are nonmechanistic in that there is no underlying reducible mechanism that leads to sites at a given location. In reality there are likely a number of definable mechanisms, each composed of countless individual human thoughts and intentions, interlaced with even more numerous and unknowable stochastic processes that would need definition to begin to formalize a site settlement system. Surely, there are mathematical and statistical ways to approximate or simulate some fundamentals of such a system, but the simulation of past cultural dynamics “*in silico*” is not the goal of this project (see Kohler and van der Leeuw [2007] for examples of this approach). Further, archaeological sites are noncontinuous beyond a local scale such that, unlike many natural phenomena, site locations cannot be reasonably assumed to be spatially continuous. In other words, based on our limited knowledge of settlement systems, knowing a site is present does not necessarily increase the probability that another site is present; and if two sites are present, the probability of there being a site between them is not necessarily increased simply by their presence. In the context of the Pennsylvania model project, archaeology sites are discrete in this sense and have definite spatial boundaries and area. This is an extension of using discretely defined site location maps as the basis of analysis and can be contrasted with spatially continuous data such as rainfall, topographic slope, bedrock depth, water table depth, and, to a degree, tree or species distributions. These are variables that can be reasonably interpolated between measured points. These qualities lead to a dataset that cannot be reasonably interpolated between data points and are created through systems that we cannot reduce to understandable mechanisms. As such, the locations of archaeological sites are essentially treated as an environmental phenomenon. While created through cultural processes, archaeological sites currently exist within, are mostly controlled by, and in this project are measured relative to the environment. The practical implication of this from the modeling perspective is that assessing sensitivity becomes a matter of binary classification—modeling the probability that a set of environmental measures from a specific point belongs to the class of site-present given what we know about site presence. Another practical implication is that many existing models applied to explicitly spatial contexts focus on continuous data and outcomes that can be interpolated. Further, other environmentally based methods focus on time-stepped mechanisms that cannot be reasonably assumed given what we know of agency and settlement system dynamics. As such, the methods to address archaeological site locations cannot be easily adopted from existing models, but the foundation of archaeological specific models can be constructed from components of its peers.

Measurement Error

The dataset of archaeological site locations is not measured directly from the environment. To model the spread of a variable such as groundwater contamination, direct measurements using some form of quantitative measuring device are taken as specific sampling locations. Conversely, under more ideal circumstances, archaeological sites are measured in reference to where artifacts are found either

through excavation, systematic testing, or surface collection, then inferred to be within a boundary measured with a GPS or GIS. In less than ideal circumstances, sites are recorded from verbal descriptions, hand-drawn maps, or memory. The large degree of variation in recordation techniques does not imply anything about the utility of the PASS files in environmental review. These voluntarily recorded locations (a sort of “crowd sourcing” in modern tech lingo) serve an invaluable purpose in recording the Commonwealth’s cultural resources and aiding in their protection. However, from a modeling perspective the life history of these locations needs to be understood.

In the past, following the submittal of a site location to the PASS files, the location on the PASS recordation form would be hand transcribed onto a 1:24,000-scale USGS quadrangle map. When the PASS files were digitized, these locations were then transcribed via heads-up digitization or geo-referencing into a GIS. More recent PASS form submissions were digitized directly into the GIS using either USGS or high resolution aerial base maps. Most recently, site locations were loaded directly from spatial data formats (e.g., a shapefile) (Table 1). Each step in this process, from the original definition in the field, to the PASS form, to the USGS quad, to the GIS multiplies measurement errors. Further, the use of circular buffers or simple ovals to define site boundaries contributes to site generalization. (It is important to note that these are just the facts of the data and do not imply anything about data collection methods.)

Table 1 - Source of Site Location in PASS GIS Database

Boundary Source	Count
Quad Sheet	12,237
GIS Point Buffer	3,352
Site Plan/Other Graphic Source	58
Text Description	15
Tax Parcel Data	2
Other	1
No Source Identified	2,600
Total	18,265

Based on the difficulties of defining site boundaries and the generalization errors associated with the recordation and digitization process, it must be assumed that many of the site boundaries used in this analysis do not contain the entirety of the original site and, conversely, may contain areas that were never part of the original site (if, that is, the concept of a “site” can be reasonably defined at all). The use of individual site cell, as opposed to site center points or averages, assumes that most of the cells falling within a site boundary are truly in the “site” area. The marginal errors from the recordation process are assumed to be a minority of the cells, but are undoubtedly present. The most practical and pressing implication of the measurement error is in establishing a minimum scale as the limit of use for the final model. While the methods used here rely on intricate statistics and high resolution environmental data, no amount of manipulation can improve the data quality of the most basic unit,

the site location. For this reason, the scale of use of the site location becomes the approximate limit of use of the models derived from them. Being that the vast majority of sites were digitized from USGS quadrangle maps published at a scale of 1:24,000 (Table 1), this scale should serve as a minimum scale of use for the final model raster layers.

The characteristics of archaeological site data described above are only a subset of the characteristics that make site data rather unique and have the greatest impact on modeling decisions. Other characteristics such as variable detectability, temporality, functionality, and fragility are other qualities that impose challenges in modeling. Theory and method within archaeology and related fields have developed to address, mitigate, or help handle these challenges. However, method and theory cannot fix everything, and these data characteristics must be extended beyond the methods and factored into our expectations and use of the results.

STUDY AREAS

STUDY REGION DELINEATION

As described in the Task 2 report, the state was divided into 10 modeling regions to ensure uniform modeling within similar landscapes and to help manage the large datasets (Figure 1). The boundaries for the 10 regions are based on grouping similar physiographic sections into regions of very roughly equal size (with the exception of Regions 3 and 10). The methods used to delineate the 10 regions are described in detail in the Task 2 report (pp. 6–8).

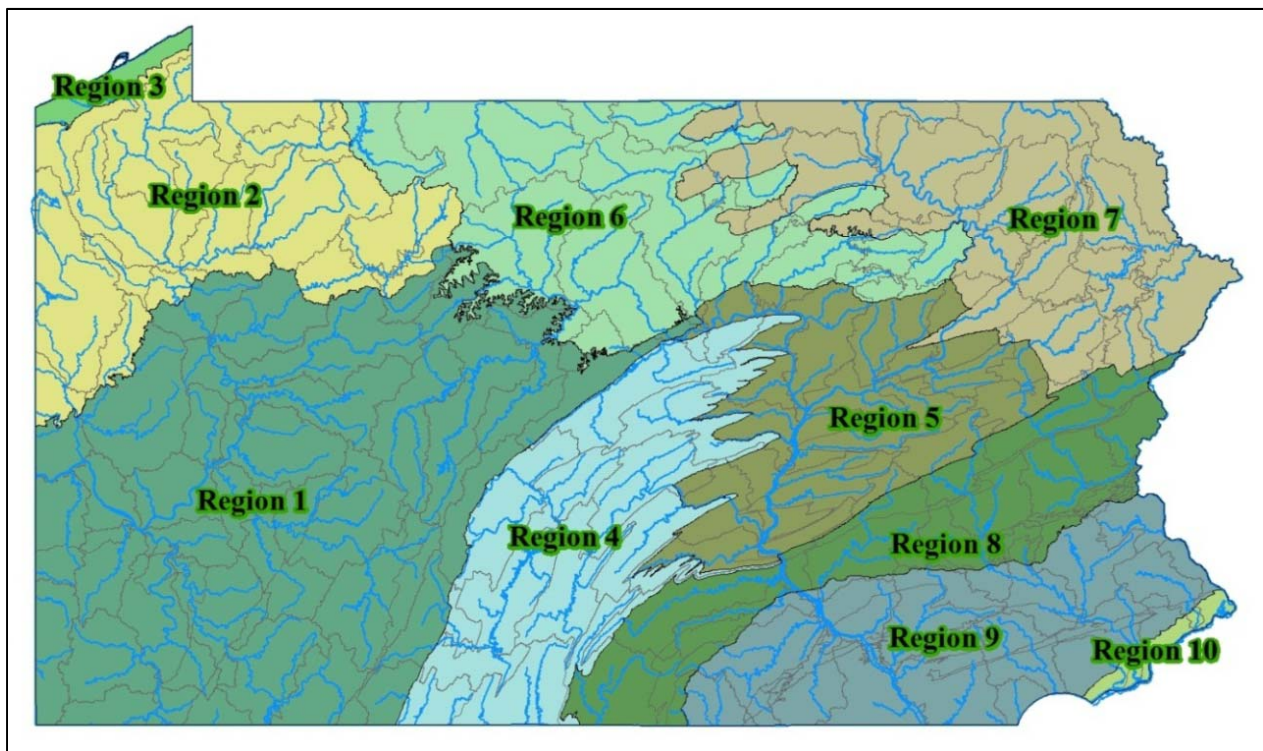


Figure 1 - Modeling regions for the Pennsylvania Model Set project.

Each of the regions was divided into smaller analytical units described in the Task 2 report as physio-sheds, which resulted from the merging of physiographic sections and Pennsylvania’s Department of Environmental Protection’s 104 State Water Plan Watersheds. These physio-sheds served as the primary modeling units. As described in the Task 4 report (p. 55), the nomenclature for dividing the state into modeling units was refined and implemented for all modeling areas. The terms used to describe the hierarchy are as follows, from largest to smallest:

Region → Zone → Section → Subarea

Regions, of which there are 10, are the largest partition of the Commonwealth. Regions may be broken down into a small number of zones based on drainage basin boundaries within physiographic provinces. The use of zones is primarily for organizing the regions into more manageable sizes for the modeling effort, and not all regions required this subdivision. Region 1 was divided into an east, north, and west zone. Regions 2 and 3 were merged to create an area comparable in size to Region 1, and the resulting Region 2/3 did not require subdivision into zones. Regions 4 and 5 were merged, but later divided into a west and east zone equivalent to the original Regions 4 and 5, respectively. Regions 9 and 10 were merged (into Region 9/10), and were not divided into zones. Regions 6, 7, and 8 also did not require subdivision into zones.

Regions or zones were further subdivided into units referred to as sections. Sections were defined based on watershed boundaries within physiographic sections. These were the same as the physiosheds described in the Task 2 report. The final division was into subareas, which is simply a section divided into riverine and upland areas. The process used to distinguish riverine from upland areas within each section is described in detail in the Task 4 report (pp. 59–64). Each subarea represents the study area for a single model, meaning that each subarea was run through the entire modeling process as an individual unit exclusive from the rest. Ultimately, the state was divided into 132 subareas, as follows:

- Region 1, 10 sections/20 subareas;
- Region 2, 4 sections/8 subareas;
- Region 3, 1 section/2 subareas;
- Region 4, 6 sections/12 subareas;
- Region 5, 7 sections/14 subareas;
- Region 6, 5 sections/10 subareas;
- Region 7, 9 sections/18 subareas;
- Region 8, 9 sections/18 subareas;
- Region 9, 14 sections/28 subareas;
- Region 10, 1 sections/2 subareas.

PREHISTORIC BACKGROUND

The three prehistoric overviews presented in the reports for Regions 1–10 (Task 4, 5, and 6 reports) were intended to be appropriately detailed and focused on three areas of the state: broadly speaking, western (Task 4), central (Task 5), and eastern (Task 6) Pennsylvania. In an effort to build upon these detailed overviews and to gain another, different perspective on the data, this summary attempts to step back and consider the prehistory of Pennsylvania from a wider, whole-state point of view. The purpose is to provide a broader, more “gestalt” perspective that might elucidate adaptational patterns and cultural variability not explicated in the previous, more focused analyses; thus, where the

previous discussions focused on the three areas individually, this broader-brush approach will consider variability and commonalities across the three analytical areas as a whole.

Viewed at large, the long prehistoric occupation of what is now Pennsylvania can be characterized as a dynamic, diverse period of time. For the most part, sociocultural development in the state generally followed that seen in surrounding regions. That said, though, prehistoric cultures in Pennsylvania did not simply move in unison with the rest of the prehistoric world. Rather, a great degree of cultural diversity may be observed through time and throughout Pennsylvania, much (or most) of which is the result of interaction, communication, and exchange of information between prehistoric groups across space; this notion is, in large measure, the sense of the term “dynamic,” as used above. Through time prehistoric people were not stagnant or immobile but moved about in their ranges and regions, formed and dissolved social and economic groups, and interacted with other groups both docile and hostile. The remainder of this discussion, then, will examine the Pennsylvania data largely from the cultural dynamic perspective explicated above.

One important question/topic that will crop up from time to time is whether the cultural diversity observed in the archaeological record is the result of in situ cultural development, or rather from imported development brought in from elsewhere.

Paleoindian through Early Archaic

During the Pennsylvania Paleoindian-through-Early Archaic period, it has been generally accepted that Paleoindians exploited the ecological niches and resources newly revealed by the slow but inexorable retreat of the glacial ice (Mason 1962; Gardner 1974). A focused subsistence system(s) is in evidence, with small groups making use of the local foodstuffs at hand. In this scenario, modes of food and resource procurement likely varied across the region, depending on what kinds of resources were available, and where (McNett 1985).

Throughout this period, it is commonly believed that Paleoindian groups continued a mobile, “wandering” lifestyle. A mobile lifestyle facilitates intersocietal relations, communication, and interaction between groups. In turn, this intergroup contact would eventually promote an overall degree of cultural consistency over space; that is, it would be expected that some or most of these formerly isolated social groups probably formed larger social units beyond the extended family or microband. Social bonding through marriage through time comes readily to mind.

Thus, this kind of social aggregation would have fostered more efficient adaptations to the late-Pleistocene/early Holocene environment, if for no other reason than by increasing the work force. Ongoing interaction and communication would have been pivotal to keep these hypothetical subsistence-settlement systems going, eventually coalescing into the larger social groups of the subsequent Middle and Late Archaic periods.

Middle to End Archaic

The period of time following the earliest cultures can be combined into an overall Middle-Late-Terminal Archaic period. That this is possible is due to the fact that sociocultural adaptations all across the state underwent dramatic, if gradual, transformations in tool manufacture and use, adaptations to the changing environment, and probably in the realm of social organization as well (Herbstritt 1980). But this is not to imply that all social groups across the region marched in lock-step with each other, somehow all evolving at the same time and across the same space.

Prehistoric cultures underwent participation in a continent-wide cultural florescence by late Middle and Late Archaic times (Custer 1996). These developments are much better known than those that preceded them. Throughout the eastern seaboard and Appalachian Piedmont, from southern New England to Georgia, the Piedmont Tradition is the predominant cultural expression (Coe 1964; Herbstritt 1980). This very large, pan-regional manifestation is marked by lanceolate and broadspear projectile points and many other diagnostic artifacts, including carved soapstone bowls. Eastern Pennsylvania is no exception, and numerous sites date to this tradition.

No one knows the precise origin and nature of the Piedmont Tradition. Some archaeologists maintain that the most likely locality would be in the neighborhood of the lower and middle Susquehanna River and upper Chesapeake Bay in Pennsylvania and Maryland. Diagnostic lithics occur in abundance in this area, as well as soapstone quarries. It is instructive to note the huge geographical spread of this tradition. Clearly, a substantial degree of social interaction was taking place along a northeast-to-southwest axis. These northeast-to-southwest connections are best indicated by the omnipresent stemmed and broadspear bifaces/projectile points found all up and down the Piedmont Province; steatite bowls have a more restrictive range, and far fewer numbers (Wittoft 1953; Stewart 1998). In addition, many archaeologists have pointed out that lanceolate points and broadspears can occasionally be found well into the interior, presumably indicating cultural connections of some kind between/among the different regions.

The Piedmont folk also contributed to the Late and Terminal Archaic cultural florescence. Sites become larger and more numerous, and large-scale (for the times) food processing begins to take place in the form of intensive fishing in many major river valleys. These collective fishing and fish-processing activities, presumably undertaken by many related social groups, would have helped solidify social and cultural relationships between groups.

A second prehistoric cultural entity can be found in Pennsylvania during the Late Archaic Period. Known as the Laurentian Tradition, the heartland of this manifestation lies to the north, principally in New York State and western New England (Kinsey 1977). Laurentian materials and sites also extend southward into Pennsylvania; Laurentian diagnostic artifacts, including notched projectile points and ground stone tools, can be found, usually sporadically, throughout the greater Mid-Atlantic region and westward across the state. Interestingly, the Laurentian culture demonstrates a probable intrusion

of people southward from a more northerly core area. This phenomenon may mirror that seen with the Piedmont Tradition, albeit writ smaller, in which social groups from a spatially extant home area moved elsewhere, whether en masse or in irregular small movements, for reasons yet unclear. The point to be made here is that, viewed from a distance, the Laurentian and, especially, Piedmont Traditions appear to represent purposeful movements of people across the landscape with specific intent; given the very large areas involved, these cases appear attributable to something more than some kind of nebulous cultural “diffusion.”

Early and Middle Woodland

For a long time, Mid-Atlantic archaeologists have lumped the Early-through-Middle Woodland periods together, which is a reflection of the considerable degree of cultural continuity evident during this lengthy interval. A clear separation between Early and Middle Woodland cultures, such as that seen further westward with the sequential Adena and Hopewell cultures, is not in evidence in Pennsylvania. Rather, the sparse data reflect continuity in subsistence practices, artifact styles, and life in general during this time. Part of the reason for this lacuna of relevant data is the simple result of fewer people being around during this period.

At the beginning of the Early Woodland, a dramatic population decline is observable all across the eastern United States and beyond (Hummer 1994; Stewart 2003). A general dearth of sites, artifacts, and populations testify to some kind of cultural contraction after the Late-Terminal Archaic period. Where the Archaic was a dynamic time, the subsequent period was not.

Not all archaeologists agree that the Early-Middle Woodland period population crash actually happened; many appeal to sampling error and uneven survey coverage to account for the apparent population slump. Alluvial burying of artifacts and sites might account for this apparent demographic decline to some extent, as well as ambiguous lithic artifact temporal assignments (e.g., presumed “Late Woodland” triangular projectile points dating to earlier periods, and stemmed “Late Archaic” points dating later, thus leading to false temporal assessments). One might also make a case that archaeologists have been working in Pennsylvania far too long to have consistently missed the remains of an entire temporal period. Whatever the case, and for reasons yet uncertain, the Early Woodland period has produced fewer artifacts and sites than the preceding or succeeding periods.

Any review of cultural dynamics in Pennsylvania cannot fail to mention the scant but undisputable evidence for various Adena-related artifacts and/or groups having been present in the state and surrounding areas during the Early Woodland time frame (e.g., Stewart 2003; Gardner 1982). Diagnostic artifacts have been recovered, typically in small quantities, in the Delaware Valley, central Delaware, and well into New England. Clearly, a wholesale population in-movement is not in evidence here; the small numbers of artifacts at the small numbers of sites suggest that small incursions of Adena people took place from time to time. The purpose of these treks is not known.

Two Middle Woodland populations can be discerned in greater northeastern and north-central Pennsylvania. These are referred to as the Fox Creek and Bushkill complexes and remain rather ephemeral and poorly known (e.g., Cavallo 1987). Bushkill is marked by distinctive ceramic wares and triangular projectile points and appears to be oriented toward the Middle and Upper Delaware valley. Distinctive Fox Creek wide-bladed, lanceolate points and shell-tempered Mockley ceramics are found over a wide portion of northern Pennsylvania, suggesting a substantial spatial extent for Fox Creek groups; further, Fox Creek artifacts are also common to the northeast, into northern New Jersey, Long Island, and adjacent New York State. Of importance here is that neither of these complexes appear to represent intensive, resident occupations in Pennsylvania; rather, these complexes appear to be ephemeral intruders originating from elsewhere.

Standing in stark contrast to the foregoing, the Abbott Farm locality on the Delaware River in New Jersey represents a vigorous, substantial occupation of this area during the Middle Woodland temporal span (e.g., Cavallo 1987). Heavily decorated, highly distinctive Abbott Farm ceramic types and other artifacts are found in considerable numbers, leading some archaeologists to suggest that population aggregations, possibly of a ceremonial or otherwise “special” nature, may have gone on here. Whatever the case, the extant data do not point to any sort of far-off “heartland” from which the Abbott Farm population might have derived. Rather, the Abbott Farm complex seems to appear abruptly and fully developed in west-central New Jersey.

Late Woodland

The Late Woodland period represents the pinnacle of prehistoric cultural development in Pennsylvania. Through the northeastern quarter and north-south center of the state, the Susquehanna River valley and upper Delaware valley were home to vigorous cultural developments on the Late Woodland time-line. In the Delaware, cultures known as Pahaquarra and Minisink succeed earlier occupants that are known as Overpeck (e.g., Custer 1989). A robust occupation of the valley floodplains is known, and large floodplain sites with storage pits and cooking facilities indicate a well-developed system in this locality.

The Clemson Island culture was an agricultural society occupying the Susquehanna River floodplains. Following Clemson Island is the Shenk’s Ferry culture. During this time Shenk’s Ferry people lived in villages and practiced maize agriculture; in the latter portions of this occupation, wooden stockades surrounded many villages, interpreted as indicating enmity of some sort between Shenk’s Ferry and other groups. Regarding origins, some archaeologists have suggested that Shenk’s Ferry entered the Susquehanna Valley from Maryland and Virginia to the south; others argue that Shenk’s Ferry clearly represents an in situ development out of an extant Clemson Island base. Whatever the case, Shenk’s Ferry was a complex society composed of principally sedentary village dwellers, agriculturalists with an eye toward social self-protection. In spite of the latter, Shenk’s Ferry was ultimately subjugated by the Susquehannock on a late prehistoric to early historic time

frame. The Susquehannock are known to have entered the state from the north, mostly following the major river valleys south.

Throughout the western portion of the state, the Monongahela culture can be said to have been the “dominant” or most complex culture in the region. Monongahela folk were agriculturalists who lived in large villages. Monongahela as a whole was loosely subdivided into different geographically extant phases. Of interest here, the McFate phase, located in the north and northwestern part of Pennsylvania, displays considerable Iroquoian influence throughout their territory near the beginning of the historic period; clearly societal contact and interaction were in play here.

Finally, the Minguannan culture is found in the southeast part of Pennsylvania and in the adjacent Delmarva Peninsula (Custer 1989). Minguannan is an imperfectly understood system and contrasts notably with most of its contemporaries. Large (or small) villages are not in evidence here. Rather, patterns seen in the preceding Middle Woodland period seem to persist, consisting of small base camps and smaller procurement camps. Also, where mature agriculture supported most other contemporary Late Woodland cultures in the region, the Minguannan seem to have never adopted the practice. The reason for this is not known, though the riverine environment of the lower Delaware would seem to be as amenable to agriculture as the middle and upper valleys, where more complex agricultural systems are known.

Summary

This brief discussion has attempted to touch on the myriad examples (and potential examples) of population movement, contact, and interaction across prehistoric Pennsylvania. As stated at the outset, an explicit dynamic perspective on the data was adopted as a focus of analysis. These few pages are not the only attempt to examine cultural dynamics in the state of course, and it is hoped that some of these ideas may be put to the test through further study. One fact stands out: pointing out cases of potential prehistoric cultural dynamics is considerably easier than explaining such potential cases.

It is believed that many more cases of potential population movement occurred in the past than are acknowledged or recognized today. It is interesting that, if one appeals to the ethnohistoric literature, one can see omnipresent examples of Native American groups moving about the landscape, sometimes impressive distances, for reasons many and varied. Are there reasons why such population movements cannot be projected backward into the prehistoric past? To this end, some archaeologists have pointed to the European fur trade as the principal disruptor of native cultures’ traditional lifeways in the East, precipitating population upheaval, movement, and sometimes confrontation. It is suggested here that many such “disruptors” can and did occur in prehistory, and the Native Americans’ reaction was likely similar.

Examples of potential intrusions into/within Pennsylvania include Paleoindians, Piedmont Tradition Archaic, Laurentian Archaic, Adena Early Woodland, Bushkill/Fox Creek Middle Woodland, and Susquehannock and Iroquois Late Woodland. Cultural developments seemingly unrelated to population movement include the Abbott Farm Middle Woodland, Clemson Island/Shenk's Ferry Late Woodland, and Monongahela, Minisink/Pahaquarra, and Minguannan Late Woodland.

In any case, a final point to be made here is that taking an explicit, dynamic approach to archaeological data can broaden the scope of an archaeological investigation in important ways. A dynamic perspective can lead to insights and conclusions previously unanticipated. The more common, complementary approach would entail a static, in situ approach that focuses down on site-specific and/or local analyses. Both approaches are fine, but the integration of site/locality and cultural dynamic perspectives can lead to new avenues of analysis and research heretofore unrecognized. One such principal avenue, as stated at the outset, is whether the cultural diversity observed in the Pennsylvania archeological record is the result of in situ cultural development, or imported development brought in from elsewhere.

ARCHAEOLOGICAL SITES IN PENNSYLVANIA

A total of 18,265 archaeological sites with prehistoric components are recorded in the PASS data base as of October 2013, 18,226 of which were used as the basis of the modeling process for Pennsylvania's Archaeological Predictive Model Set. While sites that were either too small (less than a single raster cell) or too large (indicating that they were imprecisely drawn) were eliminated from the modeling pool because of the bias they would introduce, they were included in the analyses of site types by landform and time period that appeared in each of the reports (Tasks 4, 5, and 6) for the 10 modeling regions. A statewide overview of the frequency of site types and the landforms with which they are associated is presented here (Table 2).

Table 2 - Pennsylvania Site Types by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Burial Mound	0	22	2	0	15	19	2	0	8	2	1	9	3	1	2	5	91
Cemetery	1	5	1	0	2	12	0	2	4	2	0	1	1	0	0	0	31
Earthwork	0	5	1	0	6	13	1	0	2	0	0	2	1	0	0	4	35
Isolated Find	0	29	1	0	26	47	10	9	9	2	12	4	5	11	3	2	170
Lithic Reduction	1	72	11	0	67	124	12	23	22	55	13	31	19	41	11	142	644
Open Habitation, Prehistoric	24	2693	100	46	2125	2178	822	325	250	104	83	155	193	428	75	327	9928
Open Prehistoric Site, Unknown Function	7	535	63	11	374	634	161	134	106	173	118	96	101	143	87	142	2885
Other Specialized Aboriginal Site	2	27	1	0	17	24	5	8	1	4	2	2	0	3	1	10	107
Paleontological Site	0	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	5
Path	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2
Petroglyph/Pictograph	1	5	1	5	3	5	0	6	0	3	2		1	1	2	1	36
Quarry	0	8	0	0	43	13	21	20	3	4	5	6	1	5	4	18	151
Rock Shelter/Cave	0	19	5	1	65	42	17	271	21	67	120	23	7	10	134	67	869
Shell Midden	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Unknown Function Open Site Greater than 20 m Radius	7	154	14	0	79	149	25	35	28	11	15	3	11	16	5	9	561
Unknown Function Surface Scatter Less than 20 m Radius	4	154	15	0	67	245	44	4	20	26	7	12	21	24	5	34	682
Village	1	136	11	0	13	89	26	4	39	3	2	36	28	9	2	3	402
(blank)	16	380	17	8	234	300	38	53	47	75	55	37	43	106	18	238	1665
Total	64	4248	243	71	3136	3895	1185	896	560	531	435	417	435	798	349	1002	18265

The most commonly occurring site type within the PASS data files is the Open habitation, prehistoric type, with 9,928 occurrences in the data base, representing over half of all recorded sites. The next most common site types, in descending order, are: Open prehistoric site, unknown function (2,885

sites), blank (1,665 sites—this number includes both sites where the site is identified as a type of historic site with a prehistoric component, and sites without a site type identified in the PASS data), Rockshelter/cave (869 sites), and Unknown function surface scatter less than 20m radius (682 sites). The site types in the PASS database that are the most rarely documented include Shell midden (1 site), Path (2 sites), and Paleontological site (5 sites). Open habitation, prehistoric sites are the most commonly occurring sites associated with the following landforms:

- Beach
- Floodplain
- Rise on floodplain
- Island
- Stream bench
- Terrace
- Hill ridge/toe
- Hillslope
- Hilltop
- Ridgetop
- Saddle
- Upland flats
- Blank

The most commonly occurring site on the lower slopes landform is the Open prehistoric site, unknown function. Rockshelter/cave sites are the most common site types found in middle slopes and upper slopes settings.

The landforms recorded with the most site types occurring include in descending order: floodplain (4,248 sites), terrace (3,895 sites), stream bench (3,136), hill ridge/toe (1,185), and blank (1,002). The landform types with the least numbers of sites are the beach and island types, with less than 100 sites occurring on either of those landforms. Floodplain settings are where the following site types are most commonly found:

- Burial mound
- Open prehistoric, habitation
- Other specialized aboriginal site
- Paleontological site
- Unknown function open site greater than 20 m radius
- Village
- Blank

The terrace setting was the most common setting for the following site types:

- Cemetery
- Earthwork
- Isolated find
- Lithic reduction
- Open prehistoric site, unknown function
- Shell midden
- Unknown function surface scatter less than 20 m radius

The Quarry site type was most commonly found in the stream bench setting. Hillslopes account for the highest numbers of Petroglyph/pictogram and Rockshelter/cave site types. The Path site type, with only two sites in the PASS database, was split between hill ridge/toe and hillslope settings.

In general, sites across the state of Pennsylvania are predominately associated with lowland settings. A total of 63.8% of all sites occur in lowland settings ($n = 11,657$), while 30.7% of all sites are found in upland settings. The remaining 5.5% of sites did not have landforms identified in the PASS database. Only two site types were more commonly found in upland settings than in lowland settings: Rockshelter/cave and Path.

MODEL METHODOLOGY

INTRODUCTION

The development of the Pennsylvania predictive model set followed a series of procedures that allowed for consistent and reproducible results with a bias toward transparency at all levels. The framework within which this methodology is built follows the standards and best practices of technical communities such as machine learning, data science, and statistical modeling. The framework is not specific to any one type of modeling or data set, but generalized to provide all types of modeling and analysis with guidelines for addressing sampling, variable selection, model testing, model validation, and seeking to optimize the bias/variance tradeoff. Issues associated with these topics need to be addressed for any modeling situation, be it archaeology, ecology, medical, security, or anything else. It is within this framework that specific solutions to more the explicit archaeological problems were addressed. These include addressing the bias of location and recordation of archaeological sites, the high correlation of environmentally based variables, the types of models and parameters that best address the noise inherent in archaeological locational data, differences in archaeological site types/functions, and numerous additional hurdles. Further, this methodology was tailored to deliver the best results while efficiently handling the vast quantity of data necessary to complete a series of models each covering 40,000 square miles within a project timeframe of approximately 18 months. As is the case in any time sensitive project, the consideration of development time versus benefit guided many methodological decisions. This included the use of well-documented and tested model algorithms such as LR, MARS, and RF, but also limitations on the depth of parameterization and other computationally intensive procedures. While finding the appropriate cost/benefit balance was necessary, within the confines of the overall project schedule the short-term cost of time and development were secondary to the benefit of finding justifiable and appropriate means of analysis and modeling.

The narrative below chronicles the development of the modeling framework, establishment of methods, and the intricacies of archaeological data from which methodological decisions are made. The general progression of the modeling methods is common to any field of learning that seeks to use relationships between variables to understand and predict an outcome. Figure 2 gives a generalized overview of the flow of information in this process; the specifics of each step will be explained in further detail. The process begins by gaining an understanding of the data, variables, and what correlations exist and seem worthwhile. This involves a lot of interactive manipulation of the data, cleaning, and visualization. This step is often referred to as Exploratory Data Analysis (EDA). Following the EDA, the pool of potential variables is selected and tested against numerous background samples to establish potential relationships that serve to separate site locations from the environmental background. The model building stage uses the variables to construct various statistical models that formalize the relationship between them and the locations of sites. The models,

in various ways, use these relationships to assess new data to estimate the probability of it being more like a site or more like the general background. The construction or “fitting” of the models is conducted within a framework designed to make the most use of limited data, address pitfalls such as over- or under-fitting, establish error estimates, and be repeatable. This framework incorporates k-fold CV, boot strap sampling, and parameterization based on minimizing error rates. The logic of each of these steps is documented below through the use of “pseudo-code.” This code can be used to recreate the entire framework of the model building and predicting process. In addition to the archaeological literature cited in the next section, the majority of the concepts and statistics discussed throughout are covered in detail in Hastie et al. (2009), and James et al. (2014).

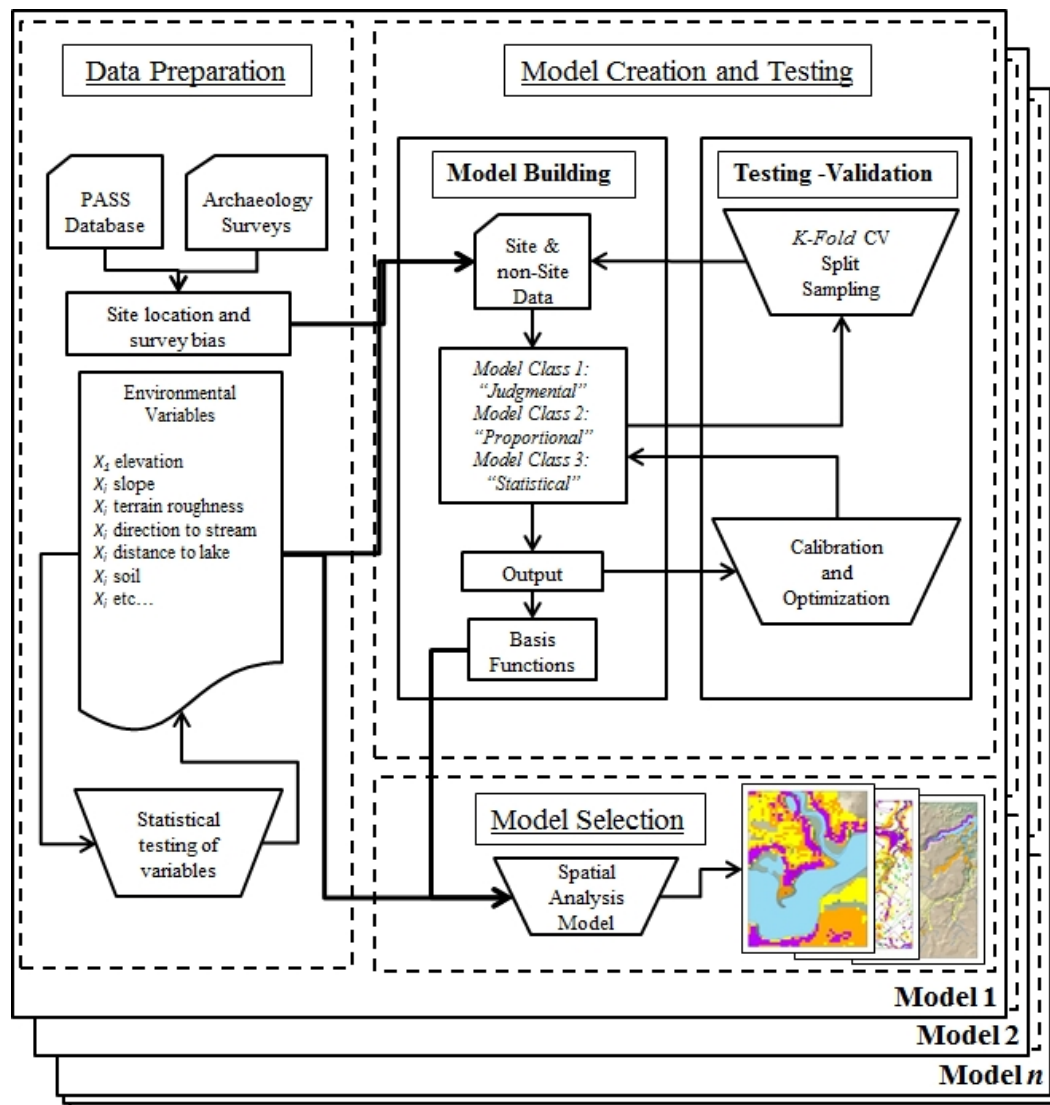


Figure 2 - General organization of entire model building process.

PREVIOUS STUDIES IN APM

The methods and models reported on here are built from the works of many people and many previous studies. The Task 1 report was dedicated to giving an overview of the scope of and examples from previous APM studies within Pennsylvania. The findings from this review are: 1) the history of APM in Pennsylvania follows the historical trends in the broader APM research field; 2) most models are ad hoc and most often based on judgmentally weighted linear combination models (e.g., they make educated guesses of where sites should be relative to a few variables, create weights, add them up for each variable, and conclude the highest value is the most sensitive); 3) reporting of methods and findings is generally poor; and 4) implementation of statistical methods is rare.

These findings are not likely to be unique to Pennsylvania. From a broader perspective, the study of APM has waxed and waned for over 40 years, but the clear heyday was in the mid-1980s during a time of overlap between the end of the quantitatively focused “New Archaeology” era and the beginning of the personal computer age. Figure 3 shows the number of occurrences for the terms “predictive” and “predictive model” within the titles and abstracts of articles in the journal *American Antiquity*, the most prestigious journal for North American archaeology. As evident in this graphic, the vast majority of major articles on the topic of APM occurred between 1980 and 1990. Following 1995, the publication of APM articles fell to a steady trickle and then dropped off to zero.

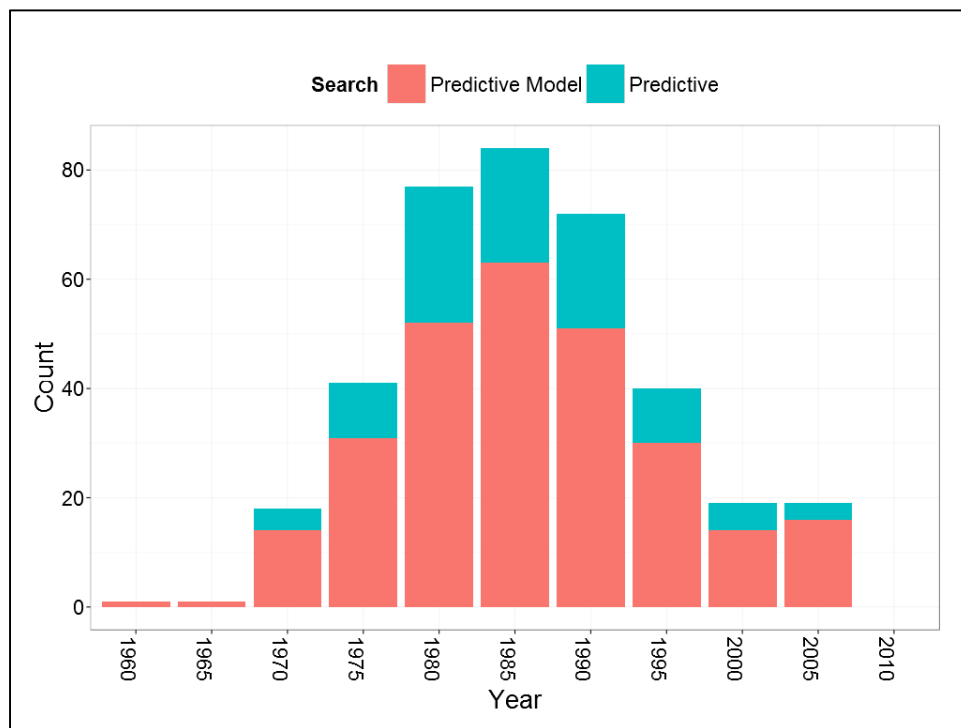


Figure 3 - Occurrences of search terms in titles and abstracts of articles in the journal *American Antiquity*.

The body of research generated from this time was full of promise and interesting ideas. The pinnacle of this early research was the 1988 volume edited by James W. Judge and Lynne Sebastian entitled *Qualifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modeling* (Judge and Sebastian 1988). Despite this, the advancement of APM was hampered by many factors including the unique challenges of archaeological data, lack of quantitative training in anthropological departments, rebuttal from the developing “Post-Processual” theoretical perspective, disillusionment from lackluster results, and, perhaps most importantly, the lack of any unified or widely successful approach. While many models were successfully applied to specific study areas, and numerous advances in theory and methods were achieved, few attempts were made by the profession to sustain a dialog or derive a generalized framework for the advancement of APM. As a corollary, other fields of study that had followed a similar trajectory such as geography, biology, ecology, and social sciences all developed strong quantitative branches that use models and develop methods to fit the character of their field’s data. While there is a notion in archaeology that the complexity of time and culture preclude quantification, there is no reason that sound and well-explored models and methods cannot be developed and add to our ability to interpret the past.

A revival of APM studies occurred in the 2000s, led by scholars at Leiden University in the Netherlands including Hans Kamermans and Philip Verhagen. With a series of publications including van Leusen and Kamermans (2005), Verhagen (2007), and Kamermans et al. (2009), this group explored new theories and introduced a number of sound statistical methods into the floundering practice of APM. These publications set a new tone for the conceptualization of predictive models and the development of best-practices—perhaps the first step to a field-wide framework. Further, the articles in these volumes looked beyond current methods and investigated new types of models and possibilities.

In the United States at approximately the same time, the Minnesota Department of Transportation (MnDOT) was making use of the third phase of a statewide APM and preparing to develop the fourth phase. Prior to the Pennsylvania model project, the MnDOT statewide model, referred to as the Mn/Model, was the only successfully completed statewide APM. Initiated in 1995, the Mn/Model was visionary in its scope, methods, and detailed reporting. Within the first five years of the project, the Mn/Model had developed from a pilot model (Phase 1) to full implementation into the State Historic Preservation Office and Department of Transportation practice (Phase 3). After nearly a decade of use, the MnDOT initiated the fourth phase of the model by reviewing the statistics and methods of the previous model and establishing a direction forward. The authors of this study Oehlert and Shea (2007) had four objectives: 1) find the best prediction methods that can be reasonably implemented with GIS; 2) produce S-Plus software (the precursor to R statistical language) to implement the predictive method; 3) provide MnDOT with advice on model evaluation; and 4) train the MnDOT users on the new methods. The report produced by the Oehlert and Shea team focuses almost solely on the first objective and is a detailed look at APM from the perspective of professional statisticians using up-to-date methods. Oehlert and Shea’s report was the first time,

perhaps arguably, that archaeological data was modeled using penalization methods such as Bayesian Information Criteria (BIC), recursive partitioning methods such as bagging and boosting trees, and the first time the results were evaluated with techniques such as Receiver Operator Characteristics (ROC) curves and k-folds CV—all techniques used or referenced throughout this project.

The cumulative effect of this history is that the study of APM soldiers on despite many challenges. This is not only because the allure of predicting archaeological site locations is strong (albeit often with impractical expectations), but because there is a substantial base of theory and practice to show that assessments of archaeological site sensitivity using these methods are useful in planning and survey applications. The way forwards is to find the right mix of assumptions, expert knowledge, and statistical methods that allow us to apply these techniques on various scales and geographies with clarity, repeatability, and management of risk. The body of publications from the 1970s and 1980s help establish the basic theories, hurdles, and methods faced when quantifying spatial archaeological data. The numerous examples of APM practiced for cultural resources management and regional survey studies since the 1980s, such as those reviewed in Task 1, demonstrate the wide range and quality of solutions found to address the issues of specific project area. Finally, the principles, techniques, and statistical models discussed by Oehlert and Shea (2007) and the Leiden University team offer rigor, substance, and direction to the study of APM. The Pennsylvania project builds on this foundation with the theoretical guidance of earlier researchers, the missteps of past projects, and the insights of modern approaches to predictive problems from within archaeology and any other field of study that deals with similar problems. It is hoped that this project helps bridge the gap between where APM faltered and where Oehlert and Shea (2007) recommended we go, and beyond to where other fields of environmental and humanities disciplines have gone—to a quantitative study of models that illuminate the systems and dynamics described by archaeological data.

SOFTWARE PLATFORM

The platform used to create the Pennsylvania statewide predictive model is a combination of software packages including ArcGIS (www.esri.com/), Python (www.python.org/), and the R Project for Statistical Computing (R) (www.r-project.org/) (Figure 4). ArcGIS is a geographic information system (GIS) created by the company ESRI. This program contains all components necessary for a GIS including map visualization, database capabilities, and tools for analysis. Python is a fully functional programming language that is integrated into ArcGIS. Through a code library called “ArcPy,” much of the functionality of ArcGIS can be accessed through the Python language allowing for a greater degree of control, flexibility, and automation of GIS tasks. Finally, R is a free and open source scripting language that is strongly geared toward statistical computation and visualization. The vast majority of the computation for this project is done in the R statistical programming language. Similar in capabilities to Python, R provides significant advantages over conducting modeling of this nature in a program such as ArcGIS. The use of R allows for the creation of scripted code to handle various aspects of the modeling process, the creation of multiuse functions that can be reused throughout the analysis, the availability of a vast array of statistical methods supported by a

large population of users, the ability to access parallel processing capabilities, and the availability of a server-based version that allows these analyses to be run across numerous cloud-based servers. Together these three programs create a platform that is capable of everything from basic GIS functions to interactive data manipulation, programmatic statistical analysis, and automation. In general, the use of each program can be characterized as follows: 1) ArcGIS for preparation of the study areas, building Digital Elevation Models (DEM), and review of PASS data; 2) Python for the creation of raster layers of environmental variables; 3) R to extract raster and PASS site data, test variables, and fit, validate, and predict models to new raster layers; and 4) ArcGIS to view prediction raster layers, select final models, and produce maps for reporting. The following discussion of project methodology will be generally organized by the three programs discussed above.

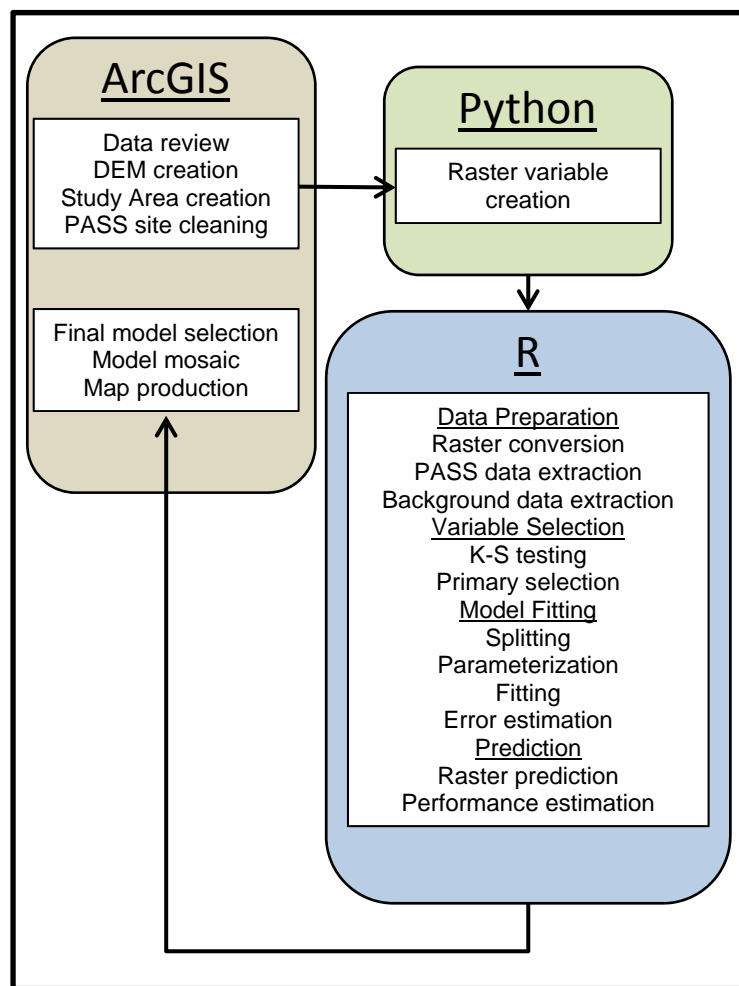


Figure 4 - Software workflow.

A Note on Pseudo-Code

In the following sections, the narrative description of modeling methods is supplemented with text boxes that contain “pseudo-code.” Pseudo-code is a simple way to present complex programming code by using plain language and distilling the major operations into concise statements. The pseudo-code provides a bridge between narrative and the actual interworkings of the scripts without requiring the reader to understand code. The component of the pseudo-code that requires a bit of new understanding is that it is arranged in the same nested logic as computer code. In computer code there are a series of control structures that guide how the code is run. These include loops, nested loops, “if” and “else” statements, and “do while” loops, as the more common examples. The pseudo-code chunks also include line numbers to help tie the narrative to the code.

Simply put, a loop repeats a chunk of code while a given condition is true. For example, a loop that prints the numbers 1–10 will cycle 10 times, until the condition “less than 10” is no longer true. Loops such as this can be nested so that a routine runs within another. “While” loops are similar, but will run while a condition is true with no set ending. For example, a “while” loop may run while a random number generator selects numbers less than 6 out of a range of 1–10. It may stop after the first random pick (i.e., if it is greater than 6), but chances are it will run for a few draws before stopping. “If” statements run only if a condition is true. For example, if the random number picker selects the number 10, then execute this code; if it selects anything else, the code is not executed. The “else” statement offers an alternative to the “if” statement if the condition is not true. A deep understanding of these control structures is not needed to understand the pseudo-code, only the knowledge that such structures exist and that the pseudo-code reflects these structures through action words and indents.

In the example below (Figure 5), the **action words** are bolded caps for emphasis, the lines are indented to show nesting, and the end of each loop is signified by an italicized *end*. Comments are prefixed by a hash sign (#). Line 1 has the action word **CREATE** to show the creation of an empty list to contain results of the following loops. These words can be a number of actions including **FOR** loops, **ASSIGN**, **EXTRACT**, **SAMPLE**, **TEST**, **PREDICT**, etc. These are plain English words that describe computational functions. The description of the action becomes apparent in the context of the pseudo-code and associated narrative. The indents in the pseudo-code show what actions happen within loops or other control structures. In the example above, there is a single **FOR** loop (lines 4–7) that cycles 10 times, each time it **SELECTS** (line 5) a random number from 1 to 100 and **INSERTS** (line 6) it in the empty list. The result of this loop is a list of 10 random numbers from 1 to 100. The **IF/ELSE** statement (lines 10–14) that follows acts on the results of the **FOR** loop if a condition is true. For example, the **IF** statement (line 10) executes the **PRINT** command (line 12) if and only if the list of 10 randomly selected numbers contains the number 67. The **ELSE** statement (line 12) executes the **PRINT** command (line 13) if and only if the **IF** condition (line 10) is false; i.e., the number 67 was not selected at random.

```
01 CREATE an empty list to hold random numbers results
02
03 # This is a single loop
04 FOR each number from 1 to 10
05     SELECT a random number from 1 to 100
06     INSERT that number into the results list
07 end
08
09 # This is an IF and ELSE statement
10 IF the list contains the number 67
11     PRINT "There is a 67 in this list!"
12 ELSE
13     PRINT "Sorry, no value of 67 selected."
14 end
15
16 # This is a nested loop
17 FOR each number from 1 to 10
18     SELECT a random number from 1 to 100 # e.g. 4
19     FOR each letter A to Z
20         PRINT the random number and letter # e.g. "4A", "4B", "4C", etc...
21     end
22 end
```

Figure 5 - Example of pseudo-code.

The nested **FOR** loop (lines 17–22) show a similar sequence, but with a loop inside of a loop. The outer **FOR** loop (lines 17–22) directs the program to **SELECT** (line 18) a random number between 1 and 100. The inner **FOR** loop (lines 19–21) takes the results of the outer loop and runs them through a loop. In this example, the randomly selected number (line 18) is sent to the **PRINT** command (line 20) for each iteration of the letters A–Z. The result is the random number joined and printed with each letter in the alphabet. Following this, the inner loop is exited (line 21) and the outer loop iterates again with the next random selection. A nest loop can have one or many sub-loops, but it gets quickly out of hand if too many loops are involved. The nested loop structure is used often throughout this project because of the nested structure of the study areas. Within a region, all the code chunks are run for each subarea. Typically, this leads to each chunk of code beginning with a **FOR** loop that iterates over each subarea in the region. The narrative description of the pseudo-code below will use the same process of line numbers and action words to tie into the code. However, each logical step of each loop will not be described in as much detail as in this illustrative example; the pseudo-code intends to clearly convey the logical hierarchy without further explanation. Further, the action words will not be bolded in the text, but will in the code.

STUDY AREA DELINEATION AND PRIMARY VARIABLE CREATION

As discussed in Chapter 2, the initial portion of the project workflow involved the creation of study areas for each model. The subarea is the smallest geographic unit of analysis for which a model is created. In total, the project divided the state into 132 separate subareas, each one being modeled as

an individual project area. Each subarea consists of either upland or riverine setting extending within a small number of watersheds all within the same physiographic section. These subareas were subsumed into a nested structure of ever-larger geographical units (i.e., region, zone, section), but it is only the boundaries of the subarea that have bearing on the model results. While the geographic division of region was divided along physiographic province boundaries, these boundaries do not have any effect on the model results. The two purposes of the larger geographic groupings are: 1) to organize the progress of the modeling effort into geographic areas for the ease and consistency of reporting environmental and cultural-historical backgrounds; and 2) group subareas into lots that form a manageable size for computation and data storage.

Building on the second purpose, since the Python and R programs are formatted to facilitate the modeling of each region, the amount of data in any one region had to be commensurate with the data storage and processing capabilities of the computer. Attempting to run the entire state as a single region of 132 subareas would require loading tremendous amounts of data into the computer and require very significant computer time to produce the results. This undertaking would require large amounts of computer memory, be prone to crashes and loss of data, and take a very long time. Alternatively, dividing the state into 132 regions, one for each subarea, would require much less computer resources at any one time, but require much more data management, a greater number of computers, produce many more files of results, and require a large amount of time for overhead management. The method used in this project sought to balance those two extremes by creating ten regions based on physiographic province boundaries and then creating subareas from those regions based on physiographic section boundaries, watershed boundaries, and site distributions. In the modeling process, the balance of data size and time was maintained by lumping regions if they were small enough or splitting regions into zones if they were too large. Regardless of the way in which they were lumped or split, the subarea boundaries based on grouping like environments was maintained and served as the unit of analysis. ArcGIS was used for this task because it allowed for interactive viewing at a range of scale, easy viewing of associated data, and the effortless overlay of additional information such as PASS site locations, drainages, geology, and physiography.

In addition to preparing the study areas for the modeling project, ArcGIS was also used to prepare the base data for the creation of environmental variables. These base data include 1/3rd Arc second DEMs, National Wetland Inventory (NWI) streams, wetlands, and water bodies, and United States Department of Agriculture (USDA) soils data. These features were then analyzed to create a number of additional features that included streams of various order, wetlands of various types, stream confluences, stream headlands, and mosaics of various soil characteristics. Python and ArcPy were used to automate the creation of the environmental variables used in the modeling process following the creation of the vector features and DEM mosaics. Since the secondary variable raster layers are generally a function of distance, cost, or mathematical manipulation of one or more inputs, automation through Python saves considerable time over manual analysis in ArcGIS. A series of scripts was created in Python integrating ArcGIS through ArcPy to process the DEMs and vector data into the final 93 secondary environmental variables used in the modeling process.

SECONDARY VARIABLE CREATION

Secondary variables are those that are created from manipulating more elemental data through statistical means. Of the 93 total variables used in this project, 91 variables are secondary derivation of elevation, hydrology, soils, or historical data. The basic DEM and a sink-filled DEM are counted in the total of 93 variables, but never used directly in model prediction. The derivation of the 91 secondary variables is done as either a function of Euclidian distance, cost distance, vertical distance, flow direction, flow accumulation, soil qualities, or a statistical manipulation of slope. The intention of this pool is to create a large number of variables representing the environmental features that may correlate to site locations. Ideally this correlation represents some environment-based component of past decision making processes, with the understanding that there are many more non-environment based decision components that cannot be correlated in this fashion. In addition to having some conceivable relevance to the environments of the past, these variables had to be available statewide at a consistent data quality and scale compatible with the scale of this analysis. Because of these restrictions, coverage such as bedrock geology (intended for a use at a scale no finer than 1:125,000), many soils attributes (inconsistent coding by county and required aggregation within map units), and features such as quarries or historically accounted villages (inconsistent coverage and data quality) were excluded from this analysis.

The table in Appendix B lists each of the variable used in this analysis, the type of measurement used, and a narrative description. The Euclidian distance function was applied primarily to hydrologic features such as streams, wetlands, water bodies, and combinations thereof, as well as stream confluences and headwaters. This function was also applied to historically mapped Native American trails (Wallace 1965). Cost distance was applied to the same set of variables. The cost distance function computed the linear distance to a feature and then applied weights based on a “cost” factor; in this case the cost is slope. Based on this, the cost distance to a stream, for example, is less along more level train, but most costly as the terrain becomes steeper. The assumption is that areas along least-cost-paths to water resources may be more preferable. The function of vertical distance is computed for streams, confluences, and drainage heads. This function simply computes the vertical difference in feet from every cell to the nearest stream, confluence, or drainage head resulting in separate variables for each. The four soils variables in Appendix B are derived from the USDA soils aggregate table as described in the Task 5 report (p. 55). These variables were chosen as they were mapped consistently from county to county, show a greater range of variability, and can be argued to have a potential correlation to site location preference. These variables are the only nominal scale variables in this pool. Another avenue to incorporating soils may be to calculate a Euclidian distance to soils with potentially favorable conditions. This would eliminate the issues associated with incorporating nominal scale data, but also requires a more heavily weighted assumption of which soils factors are more favorable to site locations.

The remaining variables covered in Appendix B are derived as various functions of slope and topography. Within this group of variables, each of the raster cells is computed as a function of the

cells within a specified neighborhood. The variables for aspect, flow direction, and slope (as degree and percentage) are all computed using the basic ArcGIS functions. Each of these functions computes the value of each cell by considering values of the eight cells that surround it (N, NW, W, SW, S, SE, E, NE)—called a Moore neighborhood. The remaining variables, prefixed in Appendix B with either *eldrop* (elevation drop), *rel* (relative topographic position), *rng* (elevation range), *splvr* (slope variation), *std* (standard deviation of slope), *tpi* (topographic position index), *tpi_cls* (classification of tpi), *tpi_sd* (standard deviation of tpi), *tri* (topographic relief index), *twi* (topographic wetness index), or *vrf* (vector roughness factor) are calculated as functions of neighborhoods of various sizes. Each of these variables uses slope and elevation in various ways to derive different measures of terrain, landforms, or landscape position. By using various neighborhood sizes (listed in the fourth column of the table in Appendix B) these measures are allowed to explore the different landscapes encountered throughout the state. For example, the range in elevation over a neighborhood composed of eight cells extending from the cell being computed for will be much different in the rugged hills of north central Pennsylvania than in the gently sloped southeastern portion of the state. The different neighborhood sizes allow for variation in the scale at which these measurements may be meaningful given different landscapes. Further, they allow for the variation in which the scale of the landscape may be meaningful to archaeological site locations. Each of the variables in Appendix B with the suffix *#c* was calculated for the range or neighborhood cell sizes listed in the “Neighborhood Cell Sizes” column and then used as a separate variable for the rest of the project. The process of looping through the calculation of each variable for each neighborhood size is automated using Python and the ArcPy library. Once the full range of variables is created and the study areas are established, these data are imported into R for the next step in the process.

VARIABLE DISCRIMINATION AND PREDICTION

The purpose of creating a large pool of variables that define numerous elements of the environment is to try and find measurements that are able to discriminate site location settings from the general environmental background. Further, the purpose of creating variables related to distance to hydrography, soils, measures of terrain, and landform definition are because we believe these are variables that, if correlated to site presence, are not spurious. However, the identification of a variable that does correlate to site location is not assumed to, in itself, be causal, but instead simply be an indicator of an unquantified variable that influenced site setting selection. While causality may be inferred from a simple variable such as distance to a stream, this analysis does not imply such a direct connection.

For the purpose of prediction, an appropriate variable is one that has a systemic relationship to the categories being predicted (site presence vs. absence) and is able to discriminate between those categories. The systemic relationship is described above and is admittedly tenuous because of the unknowable system of Native American site selection that led to known site locations and the dynamic nature of the environment throughout time. However, even though the true relationship

between these variables and site locations cannot be known, they have been used in one capacity or another throughout the history of archaeological survey with success; they form the basis of much of our understanding of settlement system dynamics; and they are the best source of continuous and consistent data to use.

The second quality of a useful variable, discrimination, is easier to quantify. A variable that discriminates well can be relatively easily split into separate ranges for each class represented. Figure 6 is a simple example of this concept. The red area is a plot of the density of background measurements for the *slpvr_16c* (slope variation within a 16-cell neighborhood) variable within a subarea. This distribution is very roughly normal with a mean around 40 and generally a consistent density on either side of that (albeit with a second small peak between zero and five). If sites were randomly distributed relative to this variable, the blue area (measures from known site locations) would follow roughly the same distribution. However, it is very clear that the blue area of site locations is distributed quite differently. The site density curve shows a bimodal distribution with a sharp main peak from zero to ten and a second peak around 20. The subarea this is generated from is a riverine area, so the low values are likely areas of flat and large floodplain. The secondary peak in the site density curve (around a value of 20) occupies the lower relief portions of the subarea landscape and at a greater proportion (measured as density) than what is present in the general environment. This figure shows the ability of the *slpvr_16c* variable to discriminate site locations from background values for areas of lower slope variation. A simple predictive model could easily be made for this single variable by drawing a decision boundary where the two densities cross around a value of 30 on the x-axis. Based on this decision boundary, an area with a *slpvr_16c* of less than 30 would be considered sensitive for sites and any value over 30 would be considered not-sensitive. This would not be a very powerful model, but indeed it would be able to tease out known site locations better than chance alone. This simple illustration underpins the general principle of what prediction is achieving: to find systemically related variables that are able to differentiate classes, understand their relationship, and extrapolate to new data.

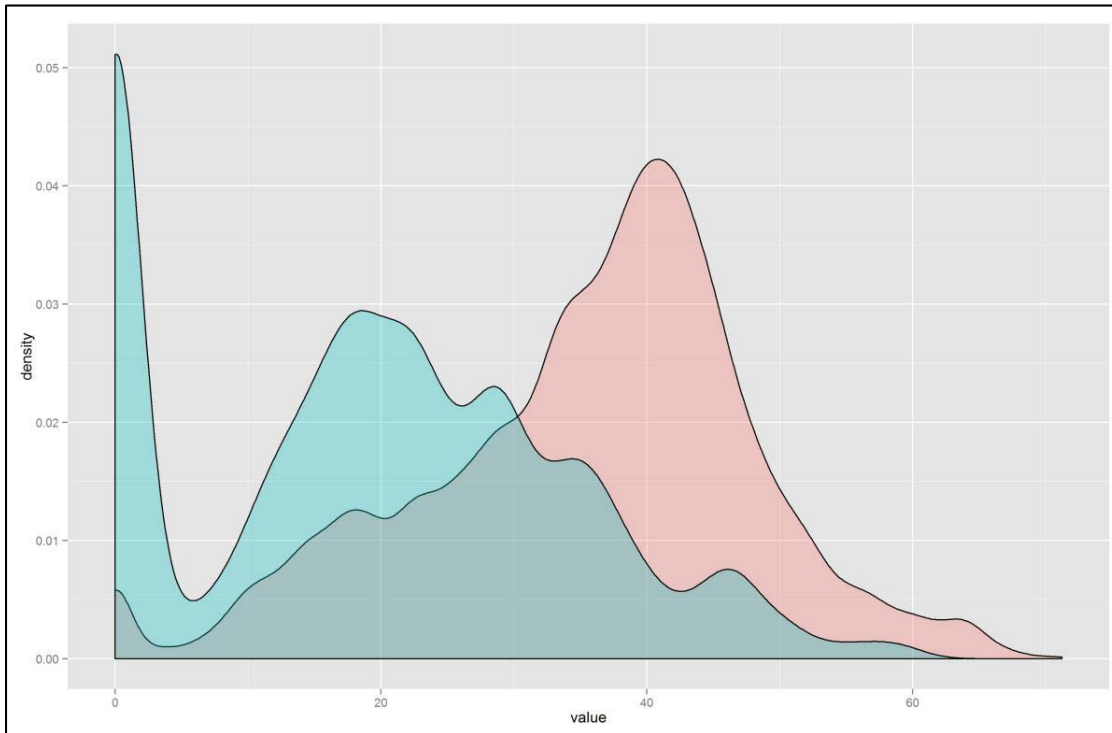


Figure 6 - Example of variable discrimination; red area is background and blue area is site locations.

A second visualization of a variable's ability to discriminate is presented in Figure 7. This example uses the measurement of the *tpi_10c* (topographic position index in a 10-cell neighborhood) variable measured for every cell within each of 50 known sites in a subarea (the colored boxplots) and the entire background population (white boxplot at right). The boxplots show the distribution of all measures of the *tpi_10c* variable for every $\sim 10 \times 10$ -m cell within that site: the taller the boxplot, the greater the range of values within that site. The stems and dots above and below the boxplot show the more extreme values and outliers. The bar across the boxplot is the distribution median and the red dot is the mean. It is evident that the majority of sites have a mean and median at approximately zero *tpi_10c*, but a few at the left have lower values and generally wider distributions. The general background has a median value of approximately -10 and a range that goes well beyond any known site location. As in the example above, a decision boundary could be drawn on this figure to help differentiate area of sensitivity. For example, if a line was drawn horizontally from the top of the white box in the background boxplot (the 75th percentile of the background distribution) it would meet the y-axis at approximately negative five. If the area above this decision boundary were considered sensitive it would contain roughly 38 of the 50 displayed sites (76%) and approximately 25% of the background area. This would not be the most accurate or sophisticated model possible, but it illustrates the concept of discriminate variables. The exercise of identifying discriminant variables is accomplished through the use of the Kolmogorov-Smirnov (K-S and Mann-Whitney (MW) U tests and is described in the section below.

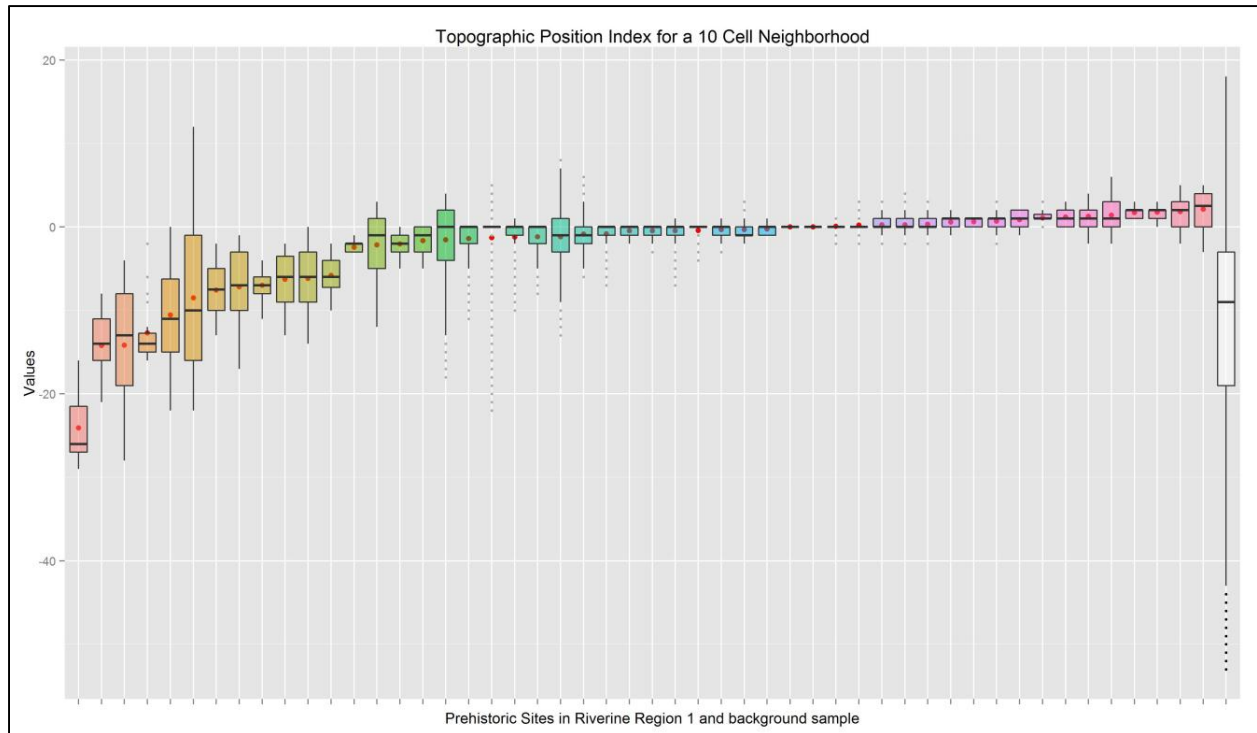


Figure 7 - Boxplots of variable measurements at known sites (colored boxes) and in the background environment (white box).

DATA EXTRACTION AND VARIABLE TESTING

Following the creation of study areas and variables, the next steps in the modeling methodology includes extracting measurements of each variable for site and background locations, and then using that data to test the ability of each variable to discriminate site locations. This process involves the steps of extracting a measurement of each variable within each ~10 x 10-m grid cell within a known prehistoric site (i.e., site-present cell), thereby creating a large database of each variable for each cell on each known site; extracting the value of each variable for up to 500,000 (or the maximum number of cells if less than 500,000) background cells in each subarea; and running a cycle of statistical tests comparing the measurements of each variable between known sites and the environmental background. The results of these procedures are a database of all variables for each site-present cell, a database of up to 500,000 background cells with each variable measured within each subarea, and the results of statistical tests indicating which variables are more likely to discriminate known site locations from the environmental background within each subarea. All of these procedures are operationalized in R.

The database of variables measured at each site is referred to during the modeling process (and in the R code) as “big_df.” The suffix of “df” is used because the particular data type for a matrix that holds

values of various types (e.g., numbers, characters, dates, nominal factors) is called a data frame. This is different from a simple matrix data type because a matrix can only contain a single class of data (e.g., only numbers or only characters). The term “big” is used because these data frames contain information on every site-present cell within an entire region and measures for all variables and descriptive information for each site-present cell, resulting in a dimension of 125 columns by upwards of 1,000,000 rows for each region. In total, the combined region-specific big_df data frames form a database of 125 columns by 2,692,082 rows representing all of the site-present cells used in this analysis.

```
01 CREATE empty dataframe: rows = site-present cells, columns = variables
02 FOR each site in Region
03     COMPUTE center coordinates for each 10x10-meter site present cell
04         FOR each background variable
05             EXTRACT variable measure at each center coordinate
06             INSERT extracted values into dataframe column
07         end
08     JOIN dataframe of extracted values for each site
09 end
10 JOIN extracted values for all sites with site descriptive data
11 SAVE as big_df
```

Figure 8 - Pseudo-code for creating data frame of site-present values for environmental variables; referred to as big_df.

The creation of the big_df is simply a process of building an empty data frame for each site that contains the number of rows for each ~10 x 10-m cell in the site and a column for each variable Figure 8. For each variable raster, the raster’s value is extracted at the center coordinates of each site-present cell and applied to the specified column in the database. Once each variable has been extracted from each site-present cell for each site, the data frames are joined into a single data frame that represents that region. While conceptually simple, this process is very time consuming because of the repeated extraction of values from each raster.

The next step in this process is to test the ability of each variable to separate site-present cells from the environmental background for each subarea (Figure 9). To do so, within each subarea (line 2) and for each variable (line 4) the entire population of measures for that variable (a value for every ~10 x 10-m cell in the subarea) is extracted (line 5) into a vector (i.e., a string of numbers in computer memory). Additionally, all of the measures for the same variable for all site-present cells in that subarea are drawn from the big_df into a vector (line 6). The statistical testing of this procedure is accomplished by comparing these two vectors numerous times. For each of 100 repeats (lines 7 to 11), a random sample of background values is drawn (line 8) from the vector. The size of this random sample is 50,000. However, since a single random sample of the background is not likely to be entirely representative of the entire background, the process is repeated 100 times. For example, if the subarea contains 10,000 site-present cells, then the first pass of the statistical tests will draw 50,000 random background samples and compute the test statistics. On the second pass, a new set of

50,000 random background samples will be drawn. This sample is done without-replacement within samples, and with-replacement between samples, meaning that a single background cell can only be pulled once per sample, but can be pulled multiple times throughout the 100 samples. In this example, after 100 passes, the total background population subjected to statistical comparison is $50,000 \times 100 = 5,000,000$. For each of the 100 passes, the statistical test scores are collected and then averaged (line 11). The reporting of the test statistics includes the mean test statistic, the mean p-value, and the standard deviation of both values to identify any large swings in values between the repeated tests. These values are joined (line 12) to the data frame (line 1) and saved for future use.

```
01 CREATE empty dataframe: rows = variables, columns = statistics results
02 FOR each subarea
03   ASSIGN variable measures for just this subarea to a new dataframe
04   FOR each background variable
05     EXTRACT total population of background values
06     EXTRACT total population of site-present values from big_df
07     REPEAT 100 times
08       SAMPLE from background value population: n = 50,000
09       TEST: Kolmogorov-Smirnov (site-present vs. background sample)
10       TEST: Mann-Whitney (site-present vs. background sample)
11       COMPUTE mean D, U, and p-value statistics for all repeats
12       JOIN mean statistic values to dataframe row
13     end
14   SAVE results dataframe for each subarea
15 end
```

Figure 9 - Pseudo-code for repeated K-S and MW tests for variable discrimination.

The actual statistical tests performed in this routine are the Kolmogorov-Smirnov (K-S) test (line 9) (reporting the D statistic) and the Mann-Whitney (MW) test (line 10) (reporting the U statistic). Generally speaking, both are non-parametric tests that measure the dissimilarity of two distributions; in this case the distributions are environmental variables measured at known site locations and those randomly picked from the background. There are specific differences in each test that contribute information valuable to understanding the way in which the two samples are different. The K-S test was the primary test used in estimating the dissimilarity between site-present and background samples. The MW test was used to support the K-S test results and show a different aspect of the distributional differences. Simply stated, the K-S test D statistic quantifies the maximum distance between two distributions. First, the K-S test computes a cumulative distribution, called the Empirical Cumulative Distribution Function (ECDF), for each background sample and the site-present sample. Second, the test compares the distance between the two ECDFs and isolates the maximum distance (measured as the maximum vertical deviation between the two curves) as the D statistic. Figure 10 visualizes the results of a K-S test using the *rng_32c* (range of elevation within a 32-cell neighborhood) variable. The distance measured by the D statistic can be seen in Figure 10 as the vertical red line annotated with “D” spanning the maximum distance between the two ECDFs. Finally, the p-value for the test is derived by estimating the probability that the D statistic would be

as large if the samples were drawn from the same parent population. Therefore, a small p-value ($p \leq 0.05$) suggests that it is unlikely that the measures of a particular environmental variable for known site locations would be drawn at random from the overall background; the use of a random value raster as a variable tests this assumption. In this sense, this is a use of a one-sample K-S test in which the background ECDF is the reference probability distribution. The result of the K-S and MW tests are computed for each variable within a subarea and saved to a spreadsheet file.

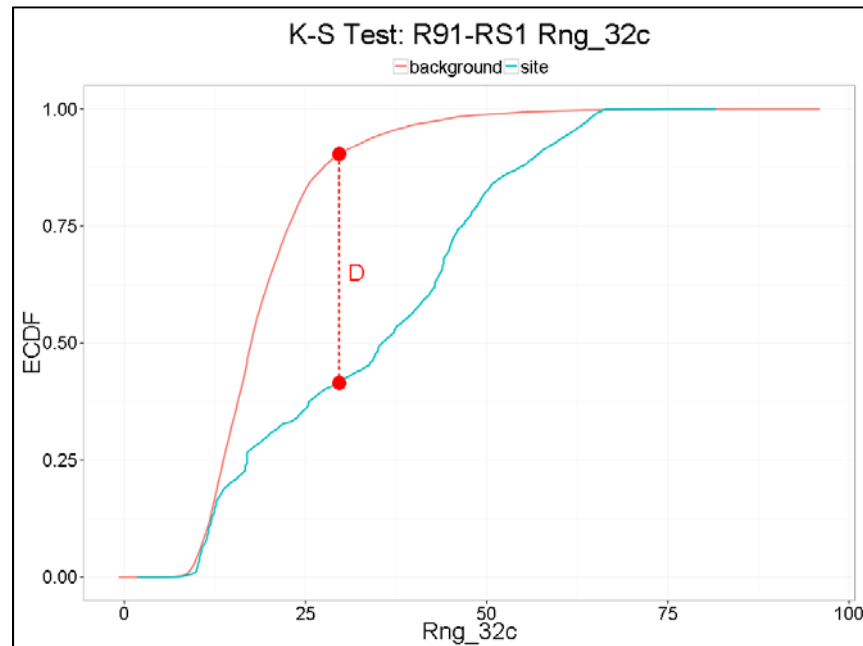


Figure 10 - Example of K-S test and associated D statistic.

PRIMARY VARIABLE SELECTION

Following the creation, value extraction, and testing of the 93 variables, a primary variable selection process is undertaken to trim the pool of variables to remove variables with known bias, high correlation, and the random variable. Figure 11 outlines this selection process. The initial stage of the selection begins by retrieving the K-S test results discussed above. For each subarea (line 1), the list of variables and K-S D statistic is loaded (line 2) and sorted (line 3) in descending order by the D statistic. The first selection (line 4) involves removing all the variables with a D statistic below the mean of all D statistics for that subarea. This removes roughly half of the variables with the lowest discrimination between site-present cells and the general background.

```

01 FOR each subarea
02   LOAD KS test results
03   SORT based on KS D statistic
04   CREATE list of variable with D statistic above mean D statistic
05   FROM the list of variables with D statistics greater than the mean
06     REMOVE variables for dem, lc neighborhood, random, buf, rel
07     SELECT first of degree or percent for slope
08     SELECT first of Euclidian or cost distance for: trail_dist, drnh,
09       h1, h2, h3, h4, h5, h6, h7, hyd_min, hyd_min_wt, conf
10     SELECT first of neighborhood size for: vrf, twi, tri, tpi_sd,
11       tpi_cls, tpi, std, slpvr, rng, rel, eldrop
12     SELECT first of h1 or h2
13     SELECT combination of distance to stream variables as:
14       IF h1, h6, and h7
15         SELECT h1 and h6
16       ELSE IF h6 and h7
17         SELECT first
18       ELSE
19         SELECT none, as there were no h1, h6, or h7
20       ELSE
21         SELECT h1, h6, or h7 as there was only one of this group
22       ELSE
23         SELECT the remaining combination of h1 and h6
24     SELECT combination of distance to wetland/water body variables as:
25       IF h3, h4, h5
26         SELECT h4 and h5
27       ELSE IF h3 and h5
28         Select first
29       ELSE
30         SELECT none, as there were no h3, h4, or h5
31       ELSE
32         SELECT h3, h4, or h5 as there was only one of this group
33       ELSE
34         SELECT the remaining combination of h3 and h4
35       ELSE
36         SELECT the remaining combination of h4 and h5
37     SELECT first of minimum distance for: hyd_min and hyd_min_wt
38     SELECT all of aws050, drcdry, drcwet, niccdcd
39   end
40   SAVE list of selected variables
41 end

```

Figure 11 - Pseudo-code for primary variable selection process.

A note of caution here is that there is danger in “cherry picking” only those variables that best differentiate the known site sample, which may lead to models with overestimated accuracy. This can result from picking only the few best variables for your dataset that may not reflect the important variables of what you are trying to predict. The initial selection described above is not seen as cherry picking in that the remaining half of variables still cover a wide variety of environmental measures and has a range of D statistics (discrimination) represented. Further, the initial selection often does not remove whole classes of variables (but rather the neighborhood sizes of variables that don’t discriminate), and finally there was a consistent lot of variables removed that simply do not help to

discriminate site locations in any environmental setting. Additionally, the stepwise logistic regression, MARS, and RF tests each have mechanisms for variables selection.

Following the removal of those variables below the mean D statistic value, the next selection (line 6) is the removal of variables that are not needed in the remainder of the modeling sequence; these include the DEM (*dem*), the sink-filled DEM (*buf*), all the variables with a 1-cell neighborhood (*lc*), and relative slope position variables (*rel*).² The DEMs are required to create many of the secondary variables, but because of the obvious trend in elevation they can greatly exaggerate survey bias. The random variable is used in the K-S testing, but is not needed in modeling. Finally, the *lc* variables and *rel* variables fail to capture any measure of the landscape that is not captured in other variables; ideally these variables would have been left out entirely, but were retained for consistency.

The set of selection procedures that follow this all compare the D statistic of a set of variables and select the variable with the highest D statistic, that which occurs first in the descending ordered list. The first of these selections (line 7) is to choose between slope measured as a percent or as degrees. Clearly, these variables measure the same thing, but with two different measurement systems. The next selection (line 8) compares between variables with both Euclidian and cost distance calculations to select which measure best discriminates each variable for that subarea. The next two selection procedures select sets of hydrology variables so as to minimize the overlap (e.g., correlation) in what they represent. The first hydrology selection (line 12) chooses between historic streams (*h1*) and NHD high-resolution streams (*h2*). Following this, various combinations of the distance to stream variables (*h6* = fourth-order stream; *h7* = third-order streams) are selected (lines 14–23) to reduce correlation. The same process is repeated (lines 24–36) for variables NHD water bodies (*h3*), NWI wetlands (*h4*), and NWI water bodies (*h5*), representing the distance to water bodies and wetlands. The final hydrology selection (line 37) chooses between two measures of the minimum distance to a hydrologic feature; one includes wetlands (*hyd_min_wt*) and the other does not (*hyd_min*). The last selection in this variable selection process (line 38) simply selects all of the soil variables. Following this, a list of the resulting selected variables is saved (line 40).

CREATION OF REGRESSION/CLASSIFICATION DATA

The K-S testing procedure for variable discrimination used above pulled large samples of background values at random. This procedure does not require the background samples to be tied to known geographic locations (provided all sampling occurs within the same subarea). Mapping coordinates were not required because the routine was testing the general background distribution against the site-present distribution. However, for the modeling of these data, the known geographical location of each background cell is required. This is because the

² Note that many other slope-related variables were used, as will be discussed in the Variable Importance section of Chapter 4, Findings and Results.

regression/classification procedures compare all of the environmental measures for selected variables at a single point on the landscape against the same measures for known site-present cells.

The pseudo-code presented in Figure 12 shows the process for sampling each subarea for up to 500,000 coordinate pairs (line 3), extracting the value for each of the selected background variables (line 6), and joining these values into a table (line 7) of point specific measures. Following the creation and saving (line 9) of the background value table, the subareas are looped over again (lines 10–14) and each big_df (site-present cell variable measures) is combined (line 11) with the background measures of the same variables. Finally, a new column is added (line 12) and coded with either “present” or “absent” denoting whether the measures in that row are from a known site or not.

```
01 FOR each subarea
02   CREATE dataframe to hold results: rows = 500000, columns = variables, x, y
03   EXTRACT 500000 random (X,Y) coordinate pairs or max cell count of subarea
04   INSERT (X,Y) coordinate pairs into dataframe
05   FOR each environmental variable selected
06     EXTRACT background value from each (X,Y) coordinate pair
07     JOIN dataframe of extracted values for each (X,Y) pair
08   end
09   SAVE results dataframe for each subarea
10 FOR each subarea
11   COMBINE random (X,Y) dataframe and big_df dataframe
12   ADD column indicating if row is a site-present or background cell
13   SAVE results dataframe as "regression_data" for each subarea
14 end
```

Figure 12 - Pseudo-code for extracting a sample of point-specific background values for selected variables and joining site-present data.

The table resulting from the primary variable selection and this procedure captures the measures of the environment for discriminatory variables for each cell on a known site and up to 500,000 background cells. The statistical procedures used to predict site locations use these tables as the reference data with the background values as the explanatory variables and the column of presence or absence as the response variable.

FITTING OF STATISTICAL MODELS: DISCUSSION

Up to this point, the focus of the methodology was on data testing and preparation. The objective of the previous steps is to result in the regression_data table described in Figure 12: a single table of background values for discriminant variables at all known site-present cells and a large sample of background locations. From these data the predictive models parameterized, trained, and tested the data and results. Figure 13 illustrates the general organization of processes that are discussed below.

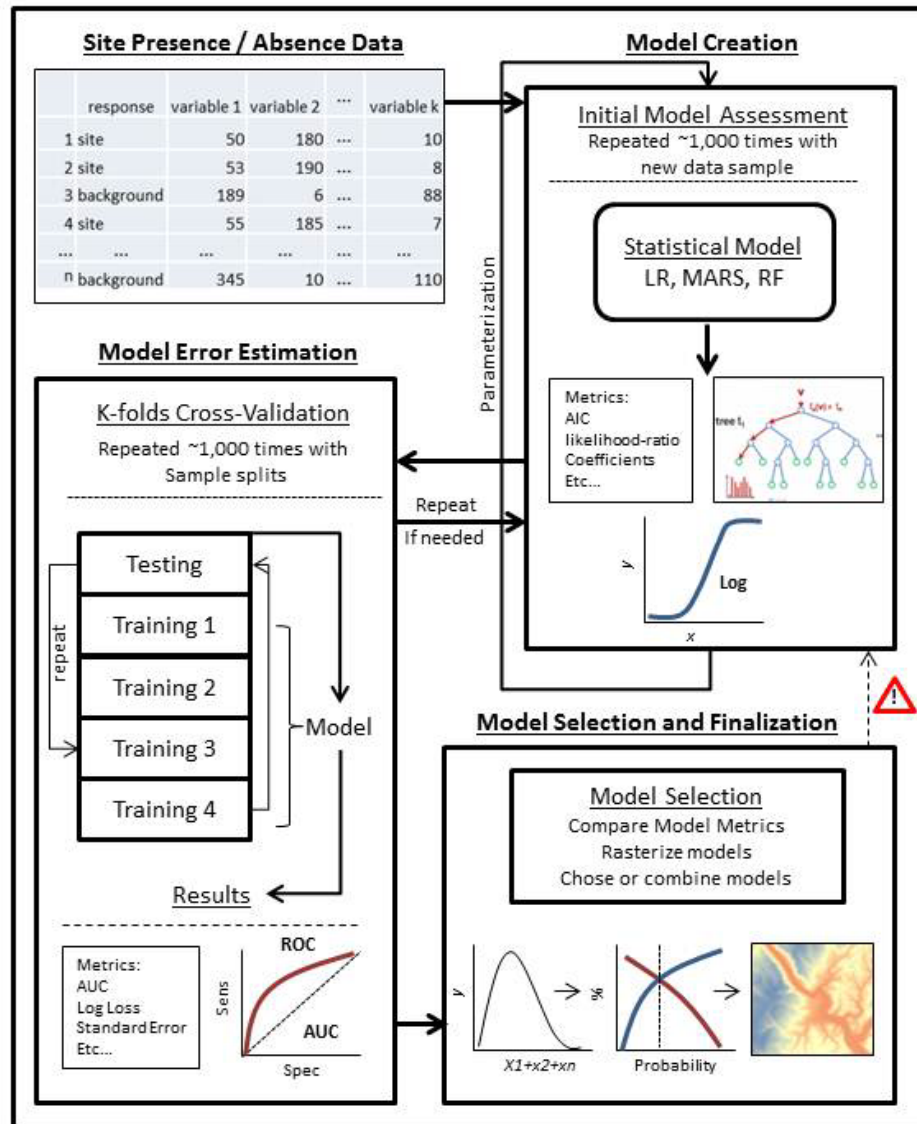


Figure 13 - General organization of model fitting and prediction process.

Model Complexity, Training Error, and the Bias/Variance Tradeoff

Creating an effective predictive model requires an adherence to a repeatable method and proper use of statistical tests, but also a great deal of subjective decision making and interpretation and an understanding of the statistical limitations of any given model. Creating an effective predictive model is not accomplished by feeding the regression data into a black-box model and mapping the output. The models for any predictive endeavor are more akin to indiscriminate machines indifferent to your intentions than they are to crystal balls or fortune tellers. While these machines will accept any properly formatted data and output a response, correct or otherwise, they offer little reassurance or

understanding when they are not properly tuned. However, the act of tuning is as much an act of art as it is science and requires consideration of the uniqueness of the data and model intention as much as it does the hard numbers of error rates and precision. For these reasons, the discussion of model tuning will require a broader discussion of model balance, described as the bias/variance tradeoff.

The bias/variance tradeoff, or dilemma as it is sometimes referred to, is an ever present issue in all forms of predictive modeling, archaeological and otherwise. Generally stated, the dilemma is that there are essentially two sources of reducible error in a prediction, the bias and the variance, and these errors are inversely proportional. As one source of error is reduced the other is often increased. This is why it is referred to as a tradeoff, and the goal is to find a model that achieves an optimum balance of the two. Consideration of the bias/variance tradeoff is critical in creating and understanding predictive models. If this tradeoff is not considered, the modeler may have no way of diagnosing an under-fit (i.e., consistently poor predictions) or over-fit model (i.e., variably poor predictions). The text below will discuss the tradeoff and techniques used in Pennsylvania's Archaeological Predictive Model Set project to find the proper balance (but see Hastie et al. [2009] for a more quantitative description of the bias/variance tradeoff).

The Components of Prediction Error.

Throughout the project the Root Mean Squared Error (RMSE) is used as an assessment of prediction accuracy. Defined below, the RMSE is simply the square-root of the average squared error where y_j is the true value of the j^{th} grid cell and \hat{y}_j is the predicted value for the same cell. In this case the squared error is simply the actual value of each cell (i.e., one for site-present or zero for site-absent) minus the predicted value (a value from zero to one) for each cell squared. Squaring the error turns any negative errors into a positive error. Add up each cell's squared errors and divide by the number of cells to get the Mean Squared Error (MSE), and then take the square root of that for the RMSE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

The RMSE can be decomposed into the sources of error with the assumptions that the data points are independent, the errors between the model and data have a mean of zero, and the variance in the irreducible error is constant (Kuhn and Johnson 2014:97). While these assumptions are idealized, they offer a situation by which to understand the nature of the bias and variance errors.

$$RMSE = \sigma^2 + (Bias)^2 + Variance$$

The equation above shows the relationship between the three forms of error contributing to the RMSE: σ^2 (pronounced sigma squared) plus the bias squared plus the variance equals the expected

value of the RMSE error. In this equation, the σ^2 is the irreducible error. This is the noise inherent in the true data that cannot be removed by the modeling process. Of interest here are the two forms of reducible error, the bias and variance. Bias is the amount by which the prediction mean differs from the true mean (how far off the prediction is, on average, from the actual values), while variance is the amount of variability in a prediction given a new set of data.

Concept of Bias and Variance

The targets in Figure 14 are a common method of depicting the schematic difference between bias and variance. The lower left target is an example of high bias and low variance: the model is not on the mark, but it is consistent in how far off it is. This is a model that generates consistent predictions given different datasets, but is not very accurate. Conversely, a high variance model such as that illustrated in the upper right-hand target is close to the mark, but varies greatly. This is a model that predicts well with some datasets, but poorly with others. The upper left-hand target is a model that predicts accurately and precisely; this is the preferred optimization of the bias/variance tradeoff. The lower right-hand target is a high bias and high variance model. This is just a poor model and is likely to be evident early in the process.

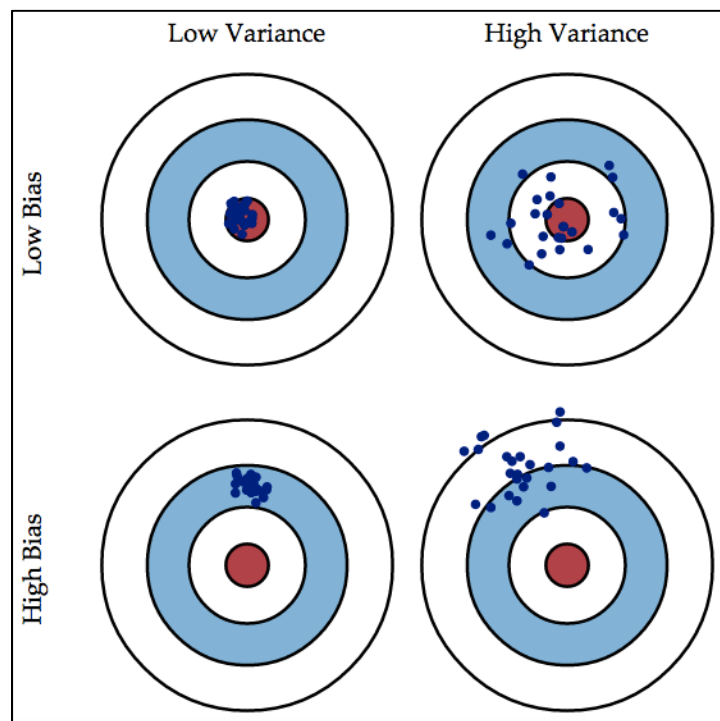


Figure 14 - Graphical illustration of bias and variance.

In terms of a model's fit, an extremely high bias model will not fit any of the data points perfectly, but will find the general trend in the data points (Figure 15, left pane). If this model was reapplied to a new set of data points generated by the same underlying process, the fit (blue line) would change

little and would maintain a consistent and relatively high degree of error—a bias. This is called an under-fit model. Conversely, an extremely high variance model will fit the data points of a particular dataset very well, perhaps perfectly (Figure 15, right pane). However, if this model was reapplied to a new set of data points generated by the same underlying process, the fit (blue line) would change drastically and maintain a low degree of error, but a high degree of variance in errors across all sets of data points. This is called an over-fit model.

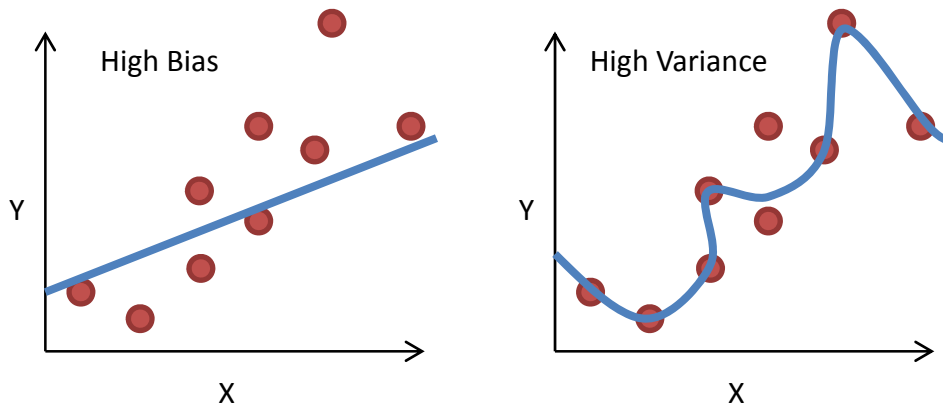


Figure 15 - Schematic of high bias and high variance model fits.

Another way to think of this is in terms of the signal and the noise around the signal. In a study, the signal is seen as the true relationship that the model is trying to approximate and the noise is the obfuscation of the signal through measurement error, imperfect predictor variables, and other processes. A model with high bias may be able to find the general shape of the signal and avoid the noise. However, since the approximation of the signal is only very general, the predictions can often be incorrect. On the other hand, a model with high variance finds the general shape of the signal by mapping very accurately to the noise around it. However, since the approximation of the signal is very specific to the noise around the signal, the predictions can often be incorrect. A model optimized to balance the errors of bias and variance will ideally achieve a close approximation of the signal while avoiding the influence of the noise. Of course, the strength of signal and amount of noise in any given dataset need to be considered in achieving the ideal balance.

Interaction of Bias and Variance over Model Complexity

The above examples are good schematics of the principals of bias and variance, but do not fully capture the tradeoff associated with bias and variance and their relationship to model complexity and prediction error. Putting this in more specific modeling terms, Figure 16 depicts how bias and variance relate to error and model complexity and overall prediction error. As shown here, the errors

of bias and variance respond in opposite directions as complexity increases (Hastie et al. 2009:223). A simple model will typically have a high bias and low variance, whereas a complex model will have high variance and low bias. Here, model complexity can be measured in a number of different ways including the number of tuning parameters, explanatory variables, or overall degrees of freedom. As the complexity of the model increases, the greater number of parameters or degrees of freedom allows the fit of a model to be more flexible and therefore more able to fit the points of a given data set. As shown back in Figure 15, the flexibility of the fit (blue line) has a direct effect on the degree of bias and variance. A simple model with an inflexible fit can only find the general trend, whereas a flexible fit can find specific noise. As shown here in Figure 16, the simple model would appear toward the left hand side of the graph and have a high bias and low variance. A complex model toward the right hand side of the graphs would have low bias and high variance. The optimum model, found by optimizing model complexity and minimizing total error is found at the balance of low bias and low variance.

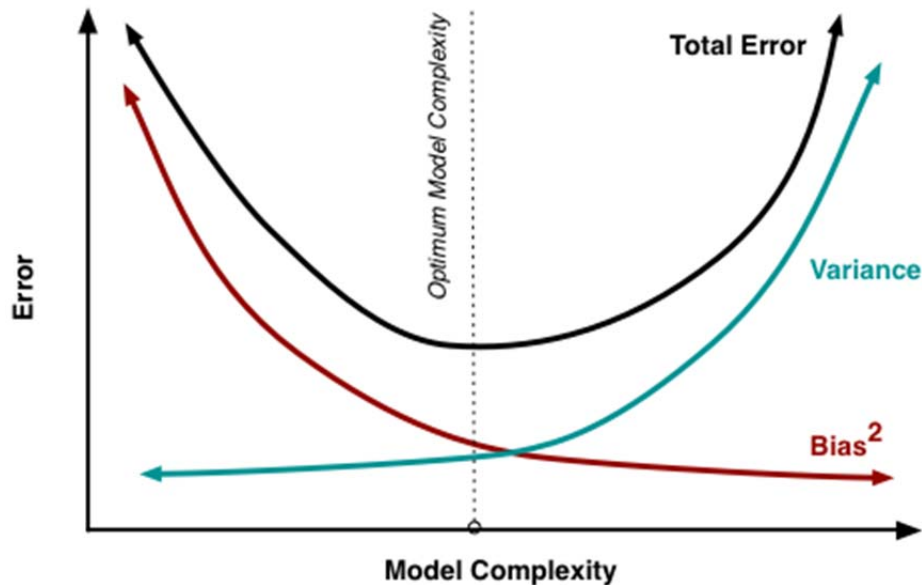


Figure 16 - Bias and variance tradeoff for model complexity.

Total Error—In-Sample and Out-of-Sample Errors

The conclusion to this train of logic is that by selecting appropriate model types and using various modeling techniques to optimize for total error, we are able to find a balance between the errors of bias and variance. The final piece to understanding this conclusion is the difference between in-sample and out-of-sample errors. Simply, in-sample errors are those calculated by predicting the data set that the model was fit on (called training sample) and out-of-sample errors are those from predicting a new data set that the model has not yet seen (called a test sample). In-sample prediction

error estimates are over optimistic because the model is fit to the same data. Out-of-sample errors are more realistic because they are independent and unbiased by the model building process. Out-of-sample error rates will always be higher than in-sample error rates.

The relationship between error rates for in- and out-of-sample predictions to model complexity, bias, and variance is shown in Figure 17. Like Figure 16, model complexity increases to the right, error increases up the y-axis, and bias/variance errors are in opposition. As model complexity increases, the fit to the training sample (blue line) will improve and lead to an ever lower in-sample prediction error. However, increased model complexity on the independent test sample (red line) will only decrease the out-of-sample prediction error to a certain point and then it will begin to increase. This is because a low complexity model will under fit the true signal with a high bias error and not predict well on independent data. A high complexity model will over fit the noise of the training data sample with high variance and not predict well on independent data. The model that predicts with the lowest error rate on independent out-of-sample data will be a balance of complexity, bias, and variance. Finding this balance requires knowledge of your data set, useful explanatory variables, and methodological considerations such as adequate variable selection, variance reduction methods, CV, sample splitting, and parameterization, to be discussed below.

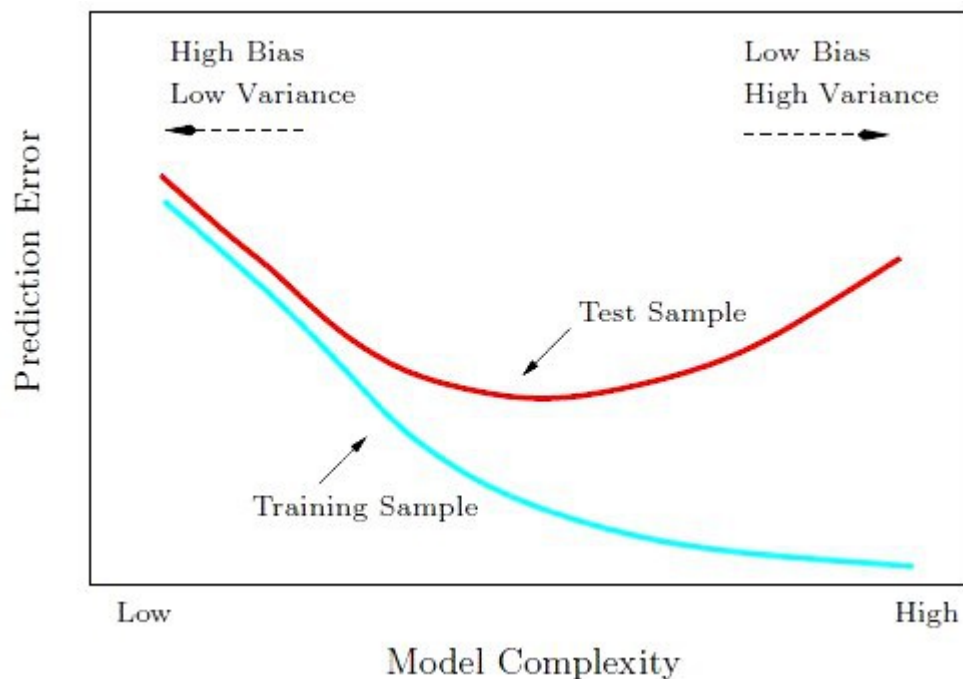


Figure 17 - Sample error and model complexity.

A final note on the topic of bias and variance tradeoff is to recognize the asymmetry of the error rates to model complexity. In Figure 16, it appears as if bias and variance errors are weighted equally and respond in kind to model complexity. However, in Figure 17 this symmetry is gone and we see that the training sample or in-sample errors steadily decrease toward model complexity and high variance. In an age when statistics and computer technology make it easy to add complexity to a model, the modeler's biggest challenge is most often to decrease variance as opposed to decreasing bias. It is relatively easy to slide down the slopes of decreasing training and test sample errors on the left side of Figure 17, but more difficult to stop before sliding down to unnecessary model complexity.

Tools for Addressing Bias and Variance

Proper model fitting requires the balance of model complexity and error rates. There are a number of techniques to achieve this goal, not the least of which being a good understanding of the characteristics of your data and an expectation of model outcomes. From a methodological point of view, tools for finding an appropriate model fit include sample splitting, k-folds CV, and parameterization. These methods were discussed in Chapter 5 (conceptually) and Chapter 6 (in practice) in the Task 3 report. The text here will not retrace all of that ground, but will provide working knowledge of these methods.

As depicted in Figure 18, the parameter optimization and fitting sequence is: 1) split data set into a training and testing sample; 2) establish a range of possible model parameters and fit a model using k-folds CV to find the parameter with the lowest average error; 3) using the optimized parameters, fit the model using the entire training sample; and 4) derive the out-of-sample prediction error rate by predicting the testing sample with the fit model. The first step of splitting the data allows for the models to be built and optimized using a training data sample and independently tested using the testing sample. The proportion of the split is relatively arbitrary and depends on the amount of data available. In most modeling applications, data are scarce and leaving any out may affect the representativeness of the sample or miss important features. However, having no testing sample is a worse outcome. Typically, splitting the training to testing samples on the order of 60/40, 75/25, or 80/20 is adequate, but the purpose of the model and characteristics of the data will influence the split.

The second step in this process is the establishment of possible parameters and the optimization of them through k-folds CV. Different types of statistical models have different types of tuning parameters. These parameters can be thought of as adjustment knobs that finely tune the predictor to best suit the given data set. In RF, the basic parameters are the number of trees in the ensemble (*ntree*) and the number of randomly selected variables to try at each split (*mtry*) (discussed in Task 3 report, pp. 52–55; Task 4 report, pp. 81–82). For the MARS algorithm the basic parameter is the number of final model terms after pruning (*nprune*) (discussed in Task 4 report, p. 61). The meanings of these parameters have been discussed in previous reports and the technical glossary, but we will use *mtry* as an example here. By default, the value of *mtry* used for the RF algorithm for

classification is \sqrt{p} ; where p is the total number of predictor variables. If we have 15 predictor variables, $mtry$ is by default set to four ($\sqrt{15} = 3.873$). If we wanted to run the model for a range of three parameter values, we could select them to be two, four, and eight. Or we could disregard the default and select a purely arbitrary range of two, six, and ten. If we were to test the model against 10 parameter values, the range could be any sequence of 10 values between one and p (the total number of variables). The selection of how wide a range of parameter values and the values themselves are arbitrary choices, but they are influenced by the data and computer resources. For every additional model parameter tested, a series of 10 additional models must be computed. Additionally, one could optimize over multiple parameters, such as $mtry$ and $ntree$. This would increase the number of models to compute exponentially. This project chose to optimize for only a single parameter per model and used a range of 3–5 different parameter values chosen by bracketing the default value of \sqrt{p} .

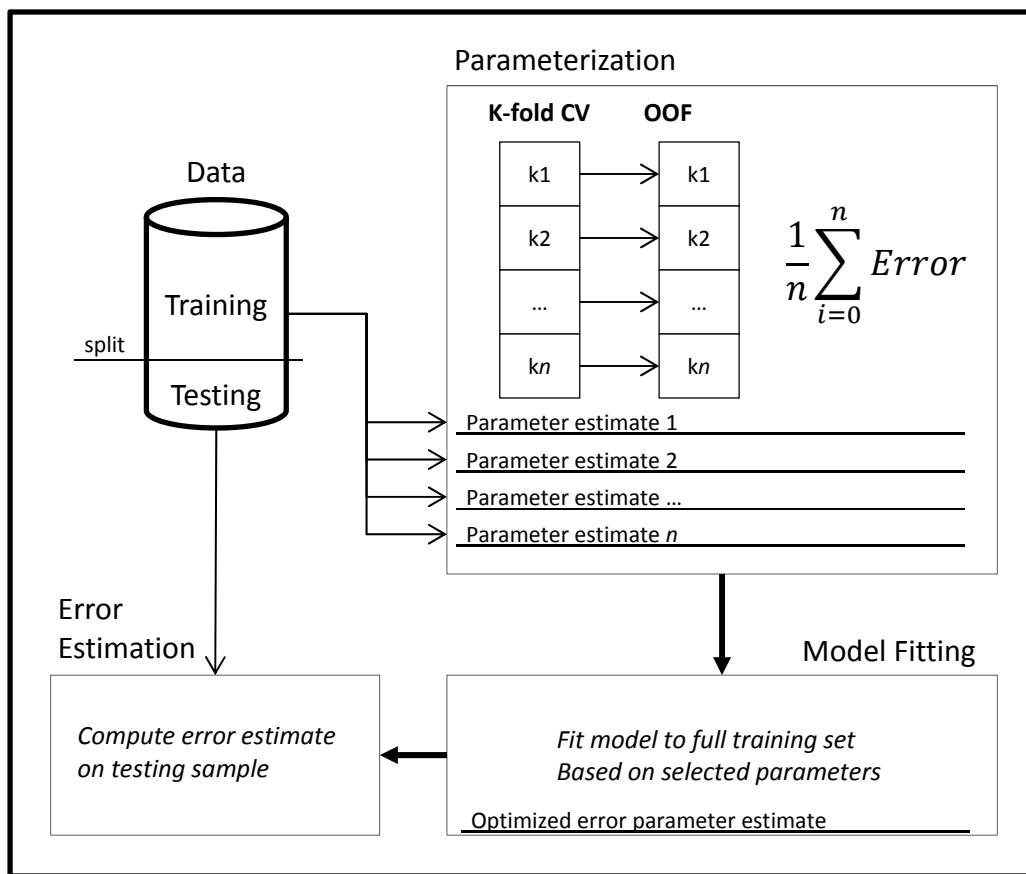


Figure 18 - Schematic diagram of model parameter optimization and fitting sequence.

The method used to select the optimized parameter value is through k-folds CV error estimation. The mechanics of CV have been covered in previous reports, but will be reviewed here because of its importance and simple concept. Basically, CV is a method by which the training data is randomly

split into a number of equally sized groups called folds. The number of folds you split the data into is referred to as “ k ,” thus k -folds. The quantity of k can be any number, but the values of 5 or 10 are most commonly selected (Hastie et al. 2009:243). Figure 19 depicts how CV works in the case of a 10-folds split. For each k folds of the data, a single model is fit to $k - 1$ groups of data, and, once fit, this model uses the remaining data as a testing set to derive an error estimation. In the example below, the data are randomly split into 10 roughly equal groups and in the first iteration, fold 1 is left out and a model is fit using folds 2–10 as a combined data set. This model is then used to predict the values in fold 1 and derive an error estimate. In the second iteration, fold 2 will be used as a testing set and the model will be fit on the combined fold 1 and folds 3–10. This process repeats until all 10 folds have been used as a testing sample, resulting in 10 separate independent error estimates that can be averaged. As such, the CV mechanism was able to utilize a single data set to derive a relatively unbiased average estimate of prediction error for that model. In our methods, this process is repeated for each parameter value in the range of parameter values to choose the value that minimizes the average CV prediction error.

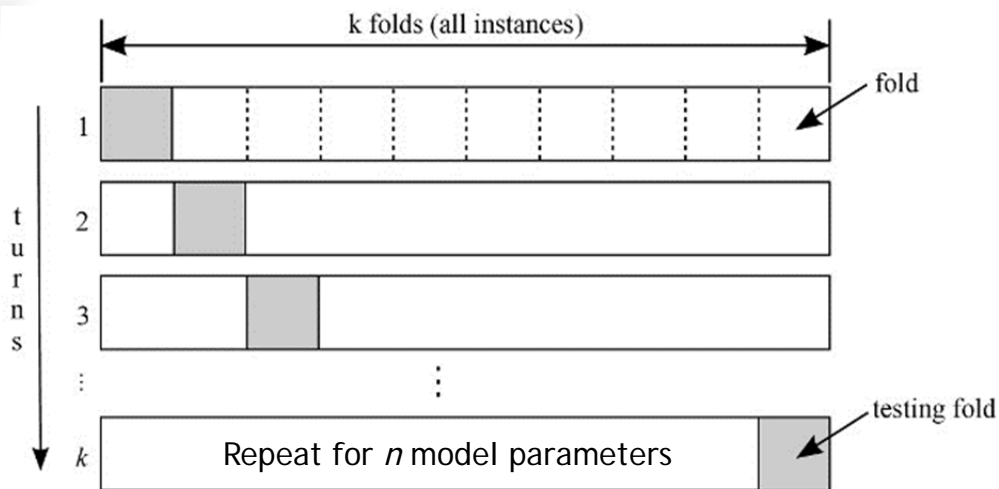


Figure 19 - Schematic of k-folds cross validation technique.

Following the repetition of the CV process we have an average error estimate for each of the range of parameter values from which we can select the parameter value that minimizes this error. Figure 20 is an example of a plot showing the average CV error for each of a range of 10 candidate parameter values. In this example the CV error is a function of RMSE and the parameter of interest is *mtry* for an RF model. From this plot it is clear that an *mtry* value of eight achieved the lowest average RMSE computed from each of 10 CV hold-out samples. Parameter values above and below *mtry* = 8 have higher, and in some cases much higher, CV error rates. These parameter values led to models that either over- or under-fit the data; decomposition of the RMSE could indicate which error (bias or variance) is most prevalent. This roughly U-shaped curve represents the black line of total error from

Figure 16 and approximates the red line for out-of-sample error from Figure 17. Based on this information, we would select an *mtry* value of eight as the optimized parameter value and use it to fit the final model.

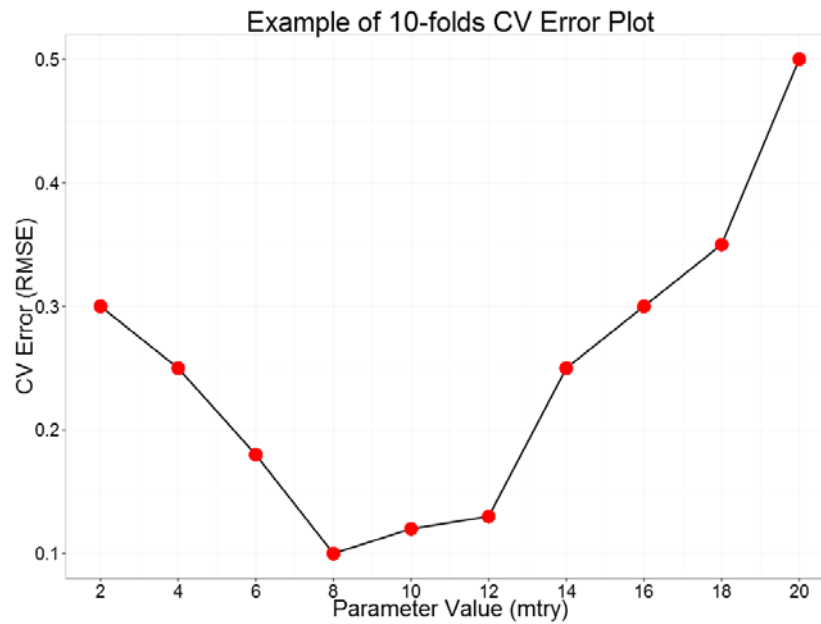


Figure 20 - Example of plotting 10-fold CV error.

The final model, chosen through the CV process above, is fit to the entire training data set to make the best use of the available training data. In the final step of the process depicted in Figure 18, the final model is used to predict the held-out testing data to derive the out-of-sample error. Provided there is no major discrepancy between CV error rate and the testing data error rate, this is the model that will be used to predict the raster layer of site sensitivity. The entire process outlined above is repeated for each of the LR, MARS, and RF models using the parameters of number of variables (*p*), number of final model terms (*nprune*), and number of variables to try at each split (*mtry*), respectively.

FITTING OF STATISTICAL MODELS: PENNSYLVANIA MODEL PROJECT

The section above discussed the theoretical and practical concerns of over- and under-fitting models and how to use data splitting, CV, and parameterization to find the optimum balance. The methodology presented above is general and could be applied to any modeling situation, although examples from the current project were included. The text and pseudo-code (Figure 21) below describe how the process of parameterization and error estimation were used in this project.


```

01 FOR each subarea
02   LOAD regression_data dataframe
03   PREPARE data: balance background to site-present as 3:1
04   PREPARE data: split data (training sample = 75%, testing sample = 25%)
05   FOR each model type: Logistic regression, MARS, Random Forest
06     FOR each parameter in range of parameter estimates (n = 5)
07       FOR each iteration in K-folds cross validation (K = 10)
08         HOLD-OUT one fold from testing data
09         FIT model on remaining folds (K - 1)
10         PREDICT response for hold-out sample data
11       end
12     COMPUTE average performance across hold-out sample predictions
13   end
14   COMPUTE optimal parameterization based on performance
15   FIT final model on all training data using optimal parameters
16   PREDICT response for testing sample data
17   COMPUTE performance for testing sample data predictions
18   SAVE model and results
19 end
20 end

```

Figure 21 - Pseudo-code for model parameterization and error estimation.

The above pseudo-code outlines the process of parameterization used here. This process is nearly identical to that described in the previous section, but structured to be run for each subarea (line 1) and for each of the three models (line 5). While looping over each subarea, the routine first loads the regression data (line 2) created by the process in Figure 12. The regression data are a data frame that contains measures of each selected variable for every site-present cell and up to 500,000 background cells. A column called “presence” contains either a 1 if the row is measured from a site-present cell or a 0 if it is a background cell. The presence column and its values serve as the response variable to be predicted for.

The first alteration to these data is to add a degree of balance to the data (line 3). This is required because the data set has a very high degree of imbalance when comparing site-present cells versus background cells. Clearly, there are multitudes more places in a subarea that do not contain sites than areas that do. The topic of imbalanced data sets and their ramifications are well covered in the Task 3 report (pp. 35–37). The take-away from that discussion is that having seriously unbalanced data makes it very easy for the algorithms to ignore the rare case (i.e., site presence) and overestimate the common case (site absence). The more imbalanced the data, the more the models will overestimate. Conversely, if the data set is balanced to 1:1, when in reality it is more like 10,000:1, the rare case will be overestimated, which in our case is not a terrible thing. For this project, the data sets were balanced at a 3:1 background to site-present ratio to address the severe class imbalance.

The second alteration to the data set is to split it into testing and training samples (line 4). From the available data, 75% is randomly split out to serve as the training data to parameterize and fit the models. The remaining 25% is held-out to be predicted by the fitted model and used to derive the

out-of-sample prediction error rate. The choice of a 75/25 split is arbitrary, but was chosen to allow for most of the data to be used for fitting given the rarity of sites in many areas. A 25% sample for testing is still adequate. The random assignment of the splitting is stratified by the “presence” column, which contains the response variable for site presence or absence.

Once rebalanced and split into training and testing samples, the data looped through each model type (lines 5–19) to be parameterized, fit, and predicted on the test set to estimate error rates. The process undertaken in this loop mirrors the parameterization, fitting, and test set predicting described in the above section (Figure 18). A range of five values for the RF *mtry* and MARS *nprune* parameter were selected for those models bracketing the default value selected by the algorithm. The LR model does not require a parameter range in the way that MARS and RF do. The optimum LR model is selected through backwards stepwise feature selection based on the AIC criteria; this form of parameterization takes place in the same loop (line 7) as the MARS and RF, but does not require a value range. The k-folds CV loop (line 7) is integrated over either the range of parameter values for RF and MARS or the combinations of different predictor variables for LR. In each iteration of the CV (lines 7–11), one of the 10 folds of data is held-out (line 8), a model is fit on the remaining nine folds (line 9), and an error rate is predicted on the one held-out fold (line 10). This is repeated and the error rate is averaged across all 10 hold-out folds and associated with that parameter value (MARS and RF) or combination of variables (LR) (line 12). After all of the parameter values or variable combinations are subjected to the CV routine, the parameter value or combination of variables that produced the lowest error rate is selected as the optimum value (line 14). The final model is fit on the entire training set (i.e., all 10 folds of data) using the optimized parameters or combination of variables for LR (line 15). This model is then used to predict the outcome of the training data set (line 16) and a final error rate is derived (line 17). Finally, the model is saved (line 18) so that it can be reused in the next step to predict the raster layers to create the final sensitivity raster for each subarea.

FINAL SENSITIVITY LAYER PREDICTION AND THRESHOLDS

The final step in this process is the creation of the raster layer for each subarea that displays the sensitivity for archaeological sites based on the model predictions, and then threshold this into low, moderate, and high sensitivity areas. The process of creating the prediction raster is relatively straightforward as all of the hard work was done in the previous steps of data preparation, variable selection, and model fitting. However, the process of raster prediction can be very time consuming and demands a great deal of computer resources. The basic approach to this step is to create a “stack” of the raster layers that represents each variable used in a given subarea’s model and then feed the values of the variables from each cell through the model created above. This process is diagrammed in Figure 22. Following this, the establishment of thresholds between the three sensitivity classes is performed.

```
01 FOR each subarea
02   FOR each environmental variable used in models
03     LOAD background raster
04     COMBINE rasters into a raster stack
05   end
06   MASK all rasters in stack to outline of subarea
07   CROP all rasters in stack to dimensions of subarea
08   SAVE all background rasters clipped to subarea
09 end
10 FOR each subarea
11   FOR each model type: Logistic regression, MARS, Random Forest
12     LOAD model
13     LOAD background rasters clipped to subarea
14     CREATE empty raster clipped to subarea
15     FOR each cell in subarea
16       FOR each background variable raster
17         EXTRACT background value
18       end
19       PREDICT response based on background variable values
20       INSERT predicted probability into empty raster at cell location
21     end
22     SAVE raster of predicted response values
23     COMPUTE square root transformation of predicted response values
24     COMPUTE confusion matrix and metrics from transformed predictions
25   end
26 end
```

Figure 22 - Pseudo-code for preparing raster layers and predicting sensitivity.

The first loop (lines 1–9) loads each raster file that represents each of the predictor variables for each subarea and crops them down to the size of the subarea. If for example, there are 10 variables used in the model for a given subarea, this loop loads each of the 15 raster files for the entire region, crops them down to the subarea, and saves them. This process is not completed earlier because of the very large memory and time commitment that would be required to preemptively crop and save all 93 raster files for each subarea.

The second loop (lines 10–26) does the bulk of the predicting work. For each subarea (line 10) and each model type (line 11), the fitted model (Figure 21, line 18) is loaded and all of the cropped background rasters from the first loop (lines 1–9) are loaded. All of the rasters are loaded (line 13) into a “stack” where each cell of each layer is perfectly aligned and overlain, and all values from the first cell of each raster (line 16) is extracted (line 17). If for example, there are 10 variables in the model, the raster stack will contain the 10 corresponding raster layers and one value from each layer will be extracted from the same cell (i.e., the same X,Y coordinate). This resulting list of 10 values will be fed into the model (line 19) and the sensitivity of that cell is predicted. Once the predicted value is obtained for that cell, it is inserted (line 20) into a blank raster layer of the same dimensions as the predictor layers. For each of the millions of cells in each subarea, this process is repeated (lines 15–21). Once all the cells are predicted, the raster layer containing the predicted values is saved (line 22). The same layer is also transformed (line 23) by taking the square root of each predicted value to

lessen the degree of right skewness. From this transformed raster, the confusion matrix and final predictive metrics (e.g., Kg, Accuracy, PPV, etc.) are calculated (line 24). At this point, the prediction process is complete.

Sensitivity Class Thresholds

The process and rationale behind model thresholding was covered in depth within the Task 4 report (pp. 72–77, 92–97). The Task 4 report should be consulted to understand the issues associated with threshold selection and how the purpose of the model is the ultimate arbiter. A number of different threshold statistics were calculated for each model, including maximizing the Kappa and Kvamme gain (Kg) statistics, balancing sensitivity and specificity, adjusting for a specific sensitivity or specificity, and prevalence based methods. From these statistics, two thresholds were selected to represent the breaks between low to moderate sensitivity and moderate to high sensitivity. Respectively, these threshold statistics are a specificity of 0.67 and predicting for a prevalence of 0.1.

For the boundary between low and moderate sensitivity, the threshold seeks to set the resulting model to a specificity of 0.67. Specificity is a statistical measure of performance that relates the True Negative Rate (TNR) of a classification, such that:

$$\text{Specificity} = \frac{\text{True Negatives}}{(\text{False Positives} + \text{True Negatives})}$$

The true negative and false positive values are derived from the confusion matrix calculated for the final model of each type; LR, MARS, and RF. As such, specificity (or TNR) is the probability of a cell being predicted a negative (i.e., site-unlikely) given that it actually is negative (background). In essence, this threshold seeks to find the sensitivity value that sets approximately 67% of the subarea to site-unlikely with the remaining 33% to be site-likely. Specificity estimates geographical area in this case because site locations make up such a small fraction of total area (i.e., low prevalence). The rationale behind this threshold is to set a bound to the maximization of specificity as suggested by Oehlert and Shea (2007). As discussed in Task 4, reducing the TNR, thereby reducing the geographical area of the model classified as moderate or high sensitivity, is not as difficult with flexible low bias models such as MARS and RF. However, ultimately this threshold selection is arbitrary and should be adjusted based on project goals.

For the boundary between moderate and high sensitivity areas, the threshold seeks to predict site-likely area equal to having an observed site prevalence of 0.1. Currently, the 18,226 known prehistoric sites used for this project occupy 2,309,463 cells (~10 × 10 m) out of the 1,065,669,566 cells that constitute the entire state, for a prevalence of 0.002 or 0.22% of the state's total area. Considering the known sites that intersect Section 106 survey areas (derived from the PHMC Environmental Review survey shapefile), the prevalence becomes 0.01 or 1% of surveyed areas. Given this very low prevalence, a threshold that predicts for a site prevalence of 0.1 is reasonable—a

10-fold increase. In many cases, this equates to a geographic area of approximately 9–11% of a subarea. As with the previous threshold, the use of a prevalence of 0.1 is subjective, but grounded in the potential use of these sensitivity layers. As discussed in the Task 4 report, there are many other potential threshold measures to use, but the choice of one should reflect the intended use and limitations of the results. This is one of the most important decisions to be made in the modeling process.

After the high and moderate threshold points are calculated for each subarea, the prediction layer is classified using these values. The raster layers, now classified into high, moderate, and low sensitivity strata are saved in the GeoTiff format for manipulation within ArcGIS.

FINAL MODEL SELECTION

Once the model fitting (Figure 21) and prediction (Figure 22) routines are complete, the results are three raster layers assessing the sensitivity for archaeological site presence; one for each LR, MARS, and RF models. For each subarea a single model from one of the three statistical models must be selected to represent the sensitivity. As will be discussed in the next section, from a functional perspective each model type relies on different assumptions, addresses the relationship between predictor variables in different ways, selects variables differently, and has different ways to fine tune the results. Because of this, the same sets of predictor values being fed into each model will derive varying predicted probabilities, but hopefully all consistently high or low. However, from a contextual perspective the three model types also work differently based on the amount of spatial autocorrelation inherent in the samples, the representativeness of the known site sample for true site locations, and the qualities of the original settlement system, if any, expressed through the documented settlement pattern. The functional differences in models types will lead to a bias that is identifiable and to a degree controllable, but differences derived from the contextual perspectives are much more qualitative. For this reason, the final model selection is not based simply on the metrics of the model fit or validation, but by a subjective consideration of the metrics combined with a visual review of each final model raster within the context of the subarea environment, predictor variables, and site sample locations. As such, future analysis of prediction error based on new sites should consider the model type used within each subarea and compare accordingly. The final model selection was accomplished using ArcGIS software for the ease of interactive panning and layers. This platform allowed for a more complete contextual understanding of the models' prediction of sensitivity than looking at the numbers alone.

The model that achieves the best balance of accuracy metrics, distribution of high and moderate sensitivity classes in respect to the quantity of sites correctly predicted (measures such as the Kg), and avoidance of obvious issues related to unrepresentative sampling and correlation was selected from the LR, MARS, and RF raster layers. The layers selected to represent each subarea are mosaicked together for an entire region. The resulting raster is the mosaic of all classified sensitivity

raster layers for all of the subareas and constitutes the final model representation. From this, the final confusion matrix of classifications is computed; this is the final graphic in each of the previous task reports for Regions 1–10. This concludes the model building and predicting process.

STATISTICAL MODELS: STEPWISE LR, MARS, AND RF

Three statistical models were chosen for use in this project; Stepwise LR (Logistic Regression), MARS (Multivariate Adaptive Regression Splines), and RF (Random Forest). While there are many hundreds of different types of statistical models that could be used to predict a response from explanatory variables, these three cover a range of complexity and flexibility and are likely to be able to capture the various patterns in our data. The Task 3 report and the glossary present a fuller discussion of the underpinnings of each model and a comparison (pp. 14–19), as well as an example of how each works (pp. 44–55). The discussion below will forgo much of the information presented in Task 3 and be more directed at the general approach, strength/weaknesses, and why each model was chosen for this project.

There is no one statistical model or algorithm that is universally better than others. There are hundreds of existing algorithms that could be applied to this project, but each model has different strengths and weaknesses with different types of data, amount of noise, types of variables, and numerous additional characteristics that set each apart. There are algorithms that work well out-of-the-box, those that require some tuning, and those that can be used in a hierarchy of models. Additionally, there is room for novel algorithms made to fit the peculiarities of archaeological data. Acknowledging that 1) archaeological data is unlike any other data set; 2) that the quality and quantity of the data are highly variable across the state; 3) that there are limited archaeological examples of algorithms beyond LR and ad hoc weighted linear combination models to study; and 4) that the project requires scalable, robust, and well-researched methods, a total of five models was chosen to represent the data. Two of the five models are weighted linear combination models referred to in previous reports as Model 1 and Model 2. Described in in the Task 3 report (pp. 9–12), these models are intended to be used when data quality is very poor. Model 1 was never employed in this project and Model 2 was used in only 5 subareas. The three remaining models are all higher-level statistical models (LR, MARS, and RF) and the subject of the following discussion. These models were chosen because LR has a long history of use in APM studies and offers a good baseline derived from a relatively noncomplex algorithm, MARS because it is a robust version of the same class as LR (Generalized Linear Model [GLM]) with dimension reduction and well-handled non-linearity, and finally RF because it requires little parameterization, is good with noisy data, includes variable selection and bootstrap aggregation (bagging), and is very well researched.

The three statistical models share the ability to conduct binary classification (e.g., site-present or site-absent), model nonparametric error distributions (e.g., binomial), and provide some form of variable selection and dimension reduction; these qualities will be explained below. Some of the main points of diversion between these models include the way in which they address non-linearity, tuning

parameters, variance reduction, and overall model complexity in terms of degrees of freedom. This is a rather simplified view of the rather complex statistics that occur within these models, but the presentation below will provide enough detail to evaluate how they fit into this project and the characteristics of their output. Further, each of these three models is well researched and documented, and numerous text and internet sources can provide information at all levels.

Logistic Regression

LR or, more specifically, backwards stepwise LR based on Akaike Information Criteria (AIC), was selected as one of the three statistical models for this project. The LR model is similar to classical linear regression, but applies a binomial error distribution to the data and uses a different loss function to fit an S-shaped curve that transforms the effects of the predictor variables on the dichotomous response. The point of this approach is to bound the predicted response to between zero and one; or in this case site presence and site absence. Figure 23 is a simple depiction of how the LR model differs from linear regression. The blue line in this figure represents a traditional linear regression line. If a linear regression is fit to a response variable (y) that contains dichotomous values (e.g., zero or one), then the blue line will predict values greater than one and less than zero. This is clearly a problem. The LR model (orange line in Figure 23) solves the problem of unbounded linear predictions by replacing the linear relationship with the logit function. This occurs by applying the logit transformation to the response variable (y) and using the predictors (x) to predict the logit, or log odds, of y . By linearizing the inherently non-linear relationship between x and y , the logit transformation function of the logistic model allows for the prediction of bounded probabilities for a dichotomous response, such as site presence vs. site absence.

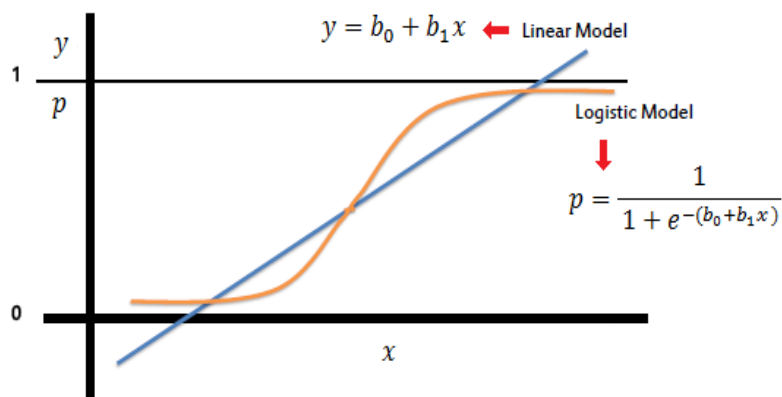


Figure 23 - Schematic example comparing linear to logistic regression.

Within this project, the basic LR model was used in a larger framework of backwards stepwise feature selection. Backwards stepwise feature is designed to make the most complicated model possible given all the variables available, which is called the saturated model. It is assumed that the

saturated model is not likely to be the best model, due to over-fitting and undue complexity. The stepwise routine begins with the saturated model, calculates the AIC for that model, and then begins to remove variables one at a time (hence the name “backwards”). The AIC is a metric that represents the relative quality of a model given the particular data set. The AIC balances the likelihood of a model against a penalty for the number of predictors used in that model (see Akaike 1974). At each step the stepwise procedure removes the variable that contributes the least to reducing the AIC metric until the model only contains variables that have the positive effect of significantly reducing the model error.

The benefits of using this model in this project are that it has a long history of use in archaeology, it is low in complexity in terms of parameters, and the coefficients of the model are relatively interpretable. With the lower complexity of this model it is likely to lead to predictions with higher bias, but lower variance. This creates a sensitivity raster that is generalized across the landscape and not likely to over-fit known site locations. At the same time, the higher bias contributes to a model that is potentially balanced more toward increasing site-likely area at the expense of accuracy. Also, the LR model is more susceptible to the error derived from spatial autocorrelation within the site-present samples and correlation between predictor variables.

Multivariate Adaptive Regression Splines

The MARS algorithm is better thought of as a model system that performs model fitting, dimension reduction, variable selection, and error estimation all in one package. An outline of this system includes 1) a first pass that fits a very high variance and over-fit model to the data; 2) a second pass that prunes that model to remove unnecessary complexity and reduced variance; 3) calculation of error rates in the second pass using the Generalized Cross-Validation (GCV) metric; and 4) variable selection performed automatically via pruning of unproductive terms. This sequence leads to a flexible and robust model that is better at adapting to noisy data and has built-in variance and dimension reduction capabilities as compared to LR.

The central function of the MARS algorithm is within the same model family as LR (GLM), but does not use a transformation to address non-linearity and interactions. Instead the MARS algorithm automatically models non-linearity and relationships by fitting a number of piecewise continuous linear splines connected by hinge functions to approximate non-linear relationships. In this way, the linear splines and hinges that connect them, referred to jointly as basis functions, can approximate a non-linearity with a series of basic linear functions applied to sub-regions of the data. Typically, to fit the same non-linearity with a polynomial would require a more complex model in terms of degrees of freedom. Figure 24 is an illustration of the first pass of the MARS model in which a high variance fit of many linear splines is built. This step purposely over-fits the data and is considered a “greedy” algorithm because at each step it tries to reduce as much error as possible at the expense of complexity and variance.

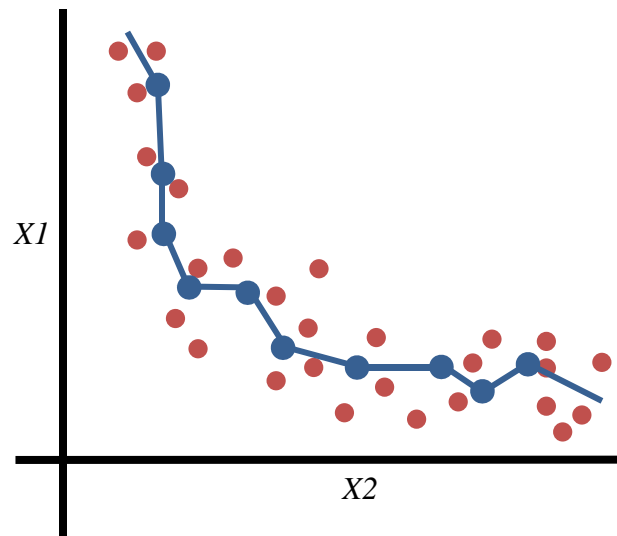


Figure 24 - Schematic of first pass, over-fit linear terms and hinge functions of MARS model.

The second pass of the MARS algorithm is where much of the work is done. This pass is referred to as “pruning” because it removes many of the over-fit linear spline terms to reduce the model to a more manageable and variance reduced fit (Figure 25). This second pass is not backwards or stepwise as it was in the LR model, but is selective across all the linear terms. The pruning process is conditional on the GCV metric and constrained by the *nprune* parameter. The GCV is similar in the AIC discussed above in that it is a regularization function that tries to balance model quality to model complexity as measured by the number of terms left after pruning. This is analogous to how AIC penalizes a model for the number of predictors. The pruning pass begins by calculating the GCV of the first pass and then removes linear terms that do not contribute significantly to the reduction of error. Note that the GCV contains the terms “cross-validation” in the name, but does not conduct k-folds CV. Instead, the GCV estimates an out-of-sample error rate by using the in-sample error rate and adding a penalty for the number of terms in the model. The pruning pass continues to reduce the GCV and remove terms until the number of terms is equal to or less than the maximum number of terms predefined by the *nprune* parameter. At this point the pruning is complete and the model is fit. From the pruning pass, an estimate of out-of-sample error is presented by the final GCV metric and variables that did not contribute to the overall goodness-of-fit are removed.

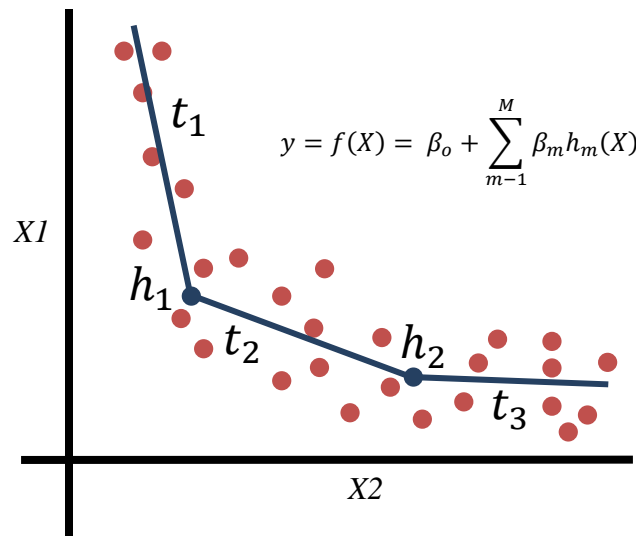


Figure 25 - Schematic example of pruned linear terms and hinge functions of MARS model.

The benefits of the MARS model are that it provides a model that is more flexible, lower bias, and regularized as compared to LR and more interpretable, lower complexity, and potentially lower variance than RF. The MARS model was selected for this project because of those benefits and because it is a very good middle ground between LR and RF. The inclusion of inherent variable selection, error estimation through GCV, dimension reduction, ability to handle data at various scales, handle both continuous and categorical data, and scalability all support the applicability of this model. Finally, while the MARS model does perform feature selection, it may still be affected by correlated variables and site-present sample locations in a similar manner as LR.

RANDOM FOREST

The final model selected for use in this project is RF. RF is unlike MARS or LR in that it is not of the GLM family, but instead is a form of recursive partitioning. The models in the GLM family result in a formula that allows for new predictions to be made based on coefficient values derived from the model fitting. Recursive partitioning models do not result in a formula, but instead result in a decision tree that is characterized by a set of rules that help predict what class a new observation falls into. More specifically, RF is an algorithm that uses many decision trees to create an ensemble of models based on randomized splits in the training data (i.e., bagging) and predicts based on agreement between all the models. Additional features of RF include variable importance, variance reduction, and out-of-sample error estimation, to be discussed below. RF was chosen for this project because it is very good at handling noisy data, can handle data measurements on a variety of scales as well as categorical data, and is a widely used and adopted model. A recent study by Fernandez-Delgado et al. (2014:3175) tested 179 machine learning classifiers to 121 different data sets and concluded that in most situations, RF will be the most accurate or among the most accurate

classifiers. It is a very robust method with many advantages, but it can over-fit data if precautions are not taken.

Additional support for the use of the RF algorithm comes from the statistical study of the Mn/Model by Oehlert and Shea (2007), described earlier in this chapter and cited throughout the task reports. Oehlert and Shea (2007) recommended that the fourth phase of the Mn/Model should use a model similar to RF. Specifically, Oehlert and Shea tested four “perturb and aggregate methods,” namely “bumped” trees, “bagged” trees, “double bagged” trees, and “boosted” trees. While the RF algorithm was developed a few years prior to Oehlert and Shea’s study, there may not have been a readily available implementation of it for them to use. However, the models they did test are very similar to RF and can be seen as precursors to RF. The description of “perturb and aggregate methods” used by Oehlert and Shea directly describes aspects of RF. “Perturb” refers to the random reshuffling of data accomplished through bagging (i.e., bootstrap aggregation) and “aggregate” refers to the combination of many trees into an ensemble, both features of RF. In addition, RF has another data permutation feature in the random selection of variables tested at each split in the tree; these features will be discussed below. From the models tested by Oehlert and Shea, the most similar to RF is likely the bagged trees, but RF is much more powerful. Of the other methods of double bagged and boosted trees, the performance noted by Oehlert and Shea was similar, with boosted trees slightly outperforming bagged and double-bagged trees. Bumped trees were poor performers. Oehlert and Shea (2007:42) recommended that the best compromise of performance and computational complexity is bagging with 10 trees. Setting aside the technicalities of double bagging and boosting, suffice it to say that RF is a more advanced implementation of the methods tested by Oehlert and Shea and most closely resembles bagging. A further recommendation was that the more trees in the ensemble, the better. Whereas Oehlert and Shea used 10 trees, Pennsylvania’s Predictive Model Set project used forests of 500 trees thanks to advances in processing speed and parallelization. Finally, the Oehlert and Shea study recommended k-fold CV, which is employed here. The Oehlert and Shea study was a well done and quite ground-breaking use of modern predictive algorithms applied to archaeological data.

As described above, the RF algorithm is an ensemble of many individual decision tree models. A decision tree operates just as the name implies, as a series of branches that guide an observation to a final decision or, in this case, classification. A very simple example would be deciding if an animal is a cat or a dog and the available predictor variable (decision criterion) is weight. The first and only branching node of this tree would divide the continuous variable of weight at a point that best divides cats and dogs, perhaps around 15 pounds. As a new observation enters the tree it will be split into greater-than 15 pounds to the right and less-than 15 pounds to the left. The leaf node (final node) to the left will contain mainly cats as it includes only observations less than 15 pounds, but a number of small dogs may be falsely categorized there. The right node for observations above 15 pounds will be almost entirely dogs, but also a small number of really big cats. If classifying dogs was more important, then this model might work sufficiently. If it was more important to classify cats or both

classes equally, then the model could use another variable that adds a layer of branching nodes and splitting points.

Moving to an archaeological example, Figure 26 illustrates a simple decision tree that could be used to classify a location as an archaeological site or not. This tree has three layers of decision as opposed to the single layer described above. The branching node labeled as “1” is where all observations enter the tree and are split according to a variable (V1). If this variable were perhaps the distance to a stream, then a split point of 200 m might be appropriate to distinguish site sensitive locations from non-sensitive locations (as will be explained, RF uses statistical measures to decide on the variable and thresholds to be used). In the example below, an observation will be split to node 2 or 3 based on the distance to water. If this split perfectly distinguished sites and non-sites in the training data, we could stop there, but there are still many sites further than 200 feet from a stream. The next layer of decisions at nodes 2 and 3 will use two new variables (V2 and V3), perhaps slope and a soil metric, to further split the observations into site-present and background classes. This process continues using different thresholds of different variables until the leaf nodes (in green) contain all the training observations. A prediction of a new observation based on this tree simply starts at node 1 and splits through each variable until it lands in a leaf node. Upon arrival at a leaf node, the observation is assigned to whichever class makes up the majority of the training data in that node.

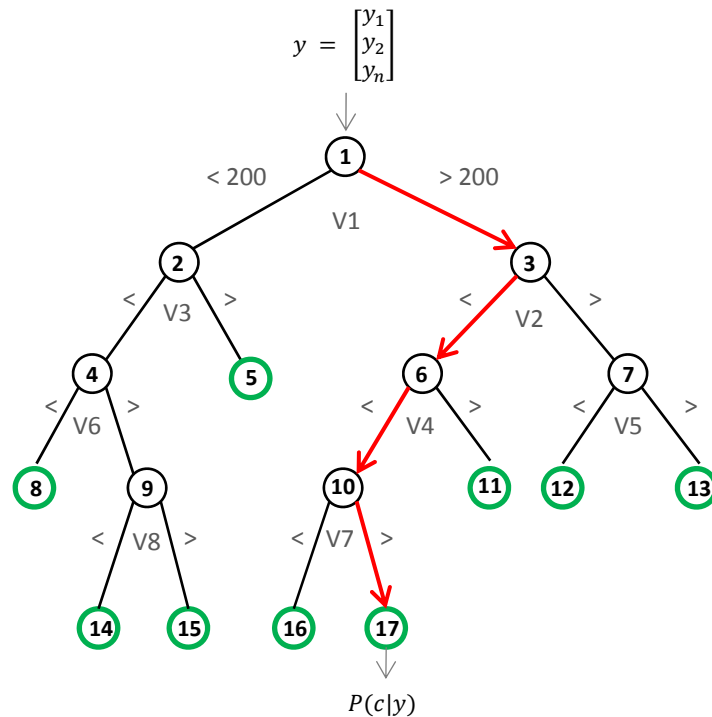


Figure 26 - Diagram of single decision tree.

The process of assigning a variable to each splitting node and establishing the value threshold to split on is automated within the RF algorithm. Figure 27 is pseudo-code for the logic behind building a single tree in RF. The creation of a tree (lines 2–16) begins with taking a bootstrapped sample (line 2) of the training data. This sample typically constitutes approximately two-thirds of the training data, while the remaining one-third is held out as a testing sample (referred to as the out-of-bag [OOB] sample [James et al. 2014:317]). With the bootstrap sample, the observations are sent to the first node for splitting. At this node, a number of the available predictor variables are selected at random (line 4). The actual number of variables to be selected at each node is governed by the *mtry* parameter. For classification problems the default *mtry* is the square-root of the number of variables available ($m_{try} = \sqrt{P}$). In this project the value of *mtry* is selected from five possible values based on 10-folds CV. The value that produces the lowest OOB error rate is used in the final model. Selecting a random set of variables to test at each split helps to perturb each tree and de-correlate variables (Kuhn and Johnson 2014:199). From the randomly selected variables, the one that creates the best split (line 5) in the data is assigned to that node. The best split can be defined in a number of ways, but typically it is the variable that leads to the biggest distinction between classes in the resulting nodes (line 6). This process occurs at each node (lines 3–7) until an ending criterion is met. This criterion could be a predefined tree depth, or that each observation rests in a single leaf node (fully grown tree), or a certain tree depth, or that each leaf node contains a certain number of observations. In this project, the stopping criterion was 15 observations per leaf node.

```
01 FOR each tree in the number of trees (ntree)
02     SAMPLE a bootstrap sample from training data
03     FOR each split in the tree
04         SELECT k variables at random (k = mtry)
05         SELECT best variable among the k variables based on criteria
06         SPLIT data based on variable and threshold
07     end
08     STOP tree at n samples per node (n = 15)
09     PREDICT OOB sample to derive error estimate
10     PERMUTE OOB sample randomly
11     FOR each node in tree
12         SPLIT data based on variable and threshold
13         RECORD decrease in node purity
14         COMPARE purity decrease to original OOB Data
15         DERIVE variable importance
16     end
17 end
```

Figure 27 - Pseudo-code showing the general logic of the Random Forest algorithm.

Figure 27 contains the logic for two additional features of the RF algorithm: OOB error estimation (line 9) and variable importance (lines 11–16). The OOB error estimation simply uses the approximately one-third of the training sample held out from the bootstrapped split (line 2) and predicts it using the tree it just built. The error rate of the prediction is recorded for each observation.

The estimation of variable importance also uses the OOB sample, but permutes it randomly (line 10) so that it does not resemble the data that were just predicted (line 10). Using the randomized OOB data, at each node (lines 11–15) they are split (line 12) based on the thresholds established for the training data (line 6). The outcome of the split is recorded (line 13) and compared (line 14) to the classification of the original OOB data. Simply put, if a variable at a given node leads to a prediction of real data that is not significantly better than random data, then the variable is not very important. On the other hand, if a variable leads to a good prediction on real data and a significantly worse prediction on random data, then it is doing its job and is important. The level of importance is tracked across all trees and reported in aggregate.

Finally, the RF algorithm is described as an ensemble because it uses a large number of individual trees (as described above) that are randomized through the bootstrap sampling (Figure 27, line 2) and random variable selection (Figure 27, line 4). In this way, the RF method computes a large number of high variance and low bias individual trees. When these high variance trees are combined into the “forest” for prediction, each tree contributes a vote, thereby reducing the variance and retaining the low bias (Hastie et al. 2009:587). Figure 28 gives a schematic representation of how the ensemble method works. This example shows three grown trees (1 through b), but the forest can have as many trees as computer resources allow. The number of trees is a tuning parameter of the RF algorithm and referred to as $ntree$. In this project this parameter was set to between 250 to 500 trees depending on the size of the data set. As trees are added to the forest (increased $ntree$) the variance and error rate are reduced. This is true to a point of dimensioning returns, but increasing the number of trees beyond this point typically does not lead to over-fitting (James et al. 2014:321).

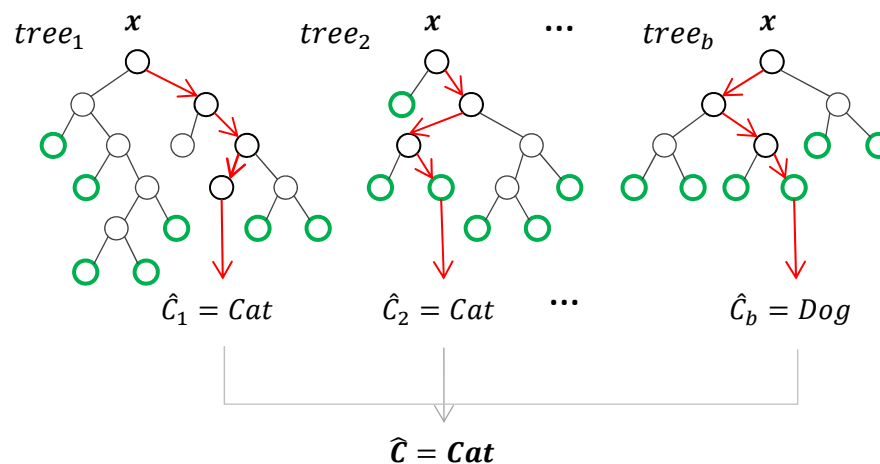


Figure 28 - Schematic of prediction based on decision tree ensemble.

The small forest represented in Figure 28 shows how a prediction can be made through majority vote. As a new observation (x) is sent through each tree, it is split at each node according to the splitting

criterion established when the tree was fit to training data. This follows the track of the bolded branches and nodes depicted in Figure 28. Even though the data of observation x is the same, each tree sends it along a different path because each tree was built with a bootstrapped sample and randomly selected variables at each node. For each tree, the observation is split until it reaches a terminal node (referred to as a leaf node) and is then predicted to the class that is represented by that node. This can be represented as $\hat{C}_b(x)$ where C -hat is the predicted class of x for the b^{th} tree. In Figure 28, we use the cat and dog example again to show that trees one and two predicted that observation x is a cat, whereas the final tree predicts it is a dog. The final prediction, represented as $\hat{C}_{rf}^B(x)$, is simply the class that the majority of trees agree on, in this case two out of three trees predicted that x is a cat. In addition, the algorithm can also provide the probability that observation x belongs to a certain class, or the probability of classification to a certain class across all trees given the observation x expressed as $P(\hat{C}_{rf}^B | x)$. The assessment of site-likely and background classes operates in the same way as demonstrated here. For the final prediction of a sensitivity raster layer, each cell is sent through each tree in the forest and the final value is derived as the probability of belonging to the class site-present.

COMPUTATIONAL REQUIREMENTS

A final note—the creation of these models is a very computationally demanding task. Each step of the process from the creation of variables, to testing variable discrimination, extracting background values, model parameterization, and predicting each incur specific demands on computer hardware. While facets of the methodology used here are shaped by computational constraints, such as the number of bootstrap samples, depth of parameter search, and model iterations, a number of efficiencies were built into the process to lessen these issues. However, it can be understated that the routines and computations described below on data of this volume take a very long time and cannot be adequately run on an office desktop without upgrades. Future implementations of these methods need to consider this as a potential limitation.

This analysis was carried out on three computers and a series of internet based or “cloud” servers. These machines represent a range of computing resources in terms of processor cores, Random Access Memory (RAM), and operating systems. The smallest office computer used was a laptop computer with 16 Gb of RAM and 4 processor cores, the other two were a desktop with 24 Gb of RAM and 4 cores, and 64 Gb of RAM with 8 cores. The internet servers that carried out much of the model parameterization are hosted by Amazon Web Services (AWS) and run on their EC2 virtual machines. These virtual machine instances ranged in size from 30 Gb of RAM and 16 cores to 60 Gb of RAM and 32 cores. These instances were configured with Ubuntu Linux operating system and R Studio Server. The laptop and smaller of the two desktops were adequate for the creation of background variables, ArcGIS interactive viewing, and data preparation, but the larger desktop and EC2 instances were needed for model parameterization, fitting, and raster prediction.

On the more powerful machines, it is not unusual for a single model fit or prediction within a single subarea to require 5 to over 20 hours of time to process; with additional steps such as raster cropping or value extraction requiring 1–5 hours per subarea. On average, each of the 10 regions required a total of approximately 350 computer hours to run. This does not include the intermittent errors, crashes, and loading of new scripts. The total project therefore required approximately 3,500 computer hours (21 weeks or 87.5 business weeks) to run if no errors occurred or models needed to be re-run. Many errors occurred and many models needed to be re-run. Fortunately, a few time-saving adaptations allowed for these models to be completed within the project time frame. First was the general optimization of the code that produces many of the model components. This required an iterative process of development and optimization that took place throughout the project. Second was the use of parallel processing that took full advantage of all available processing cores for particular tasks. While some tasks, such as computations that rely on the result from other computations, are not great candidates for parallel processing, other tasks such as ensemble modeling (e.g., RF) are very well suited for it. Where it was useful, such as in model parameterization and fitting, the code took advantage of parallel processing to reduce modeling time. Third, and perhaps most important, was the adaption of the modeling process to run on cloud-based AWS EC2 servers. Using this technology a number of high powered server instances could be recruited to take on the most arduous tasks of model parameterization and fitting. These servers are optimized to this type of computation, can be scaled up or down to fit the data needs, and are accessible from any internet connection. Therefore, broken modeling processes did not have to wait until Monday morning to be discovered.

4

FINDINGS AND RESULTS

This project resulted in a single statewide coverage demonstrating the results of a statistical sensitivity model for the presence of pre-contact archaeological material. This coverage, in the form of a raster layer, is a mosaic of four individual statewide raster layers representing each of four different statistical model algorithms. Each of these four algorithms was applied to 132 subareas that define the analytical units composing the statewide coverage. Each subarea represents either a riverine or upland subarea within the same geography. As such, there are 66 upland subareas and 66 riverine subareas. Table 3 illustrates the dimensions of this project.

Table 3 - Quantities of Model Project Attributes

Attribute	Quantity
Total model area (sq. mi)	45,293
Total model cells (~10.5 x 10.5 meters)	1,058,897,903
Site-present cells	2,024,242
Archaeological sites	18,226
Archaeological sites after being unioned with subareas	22,144
Subareas models	132
Individual models	528
Environmental variables	93
Processed cells (models + variables)	102,519,096,591

The total number of prehistoric archaeological sites or site components used in this study was 18,226. Table 4 breaks down, for each region, the number of subareas, the area in square miles, the number of sites intersecting the region, and the density of sites per square mile. For this project, these sites were unioned with the subareas, meaning that the polygon areas representing PASS archaeological sites was overlain on the polygons representing each of the subareas, and any site spanning a subarea boundary was split into two. If a site overlapped the intersection of three subareas, it would then be split into three sites along the subarea boundaries. After this process, the total number of site areas used in this study was 22,144.

Table 4 - Area, Archaeological Site Count, and Site Density per Region

Region	Subarea Count	Square Miles	Site Count	Sites per Sq Mile
1	20	12,338	7,381	0.598
2	8	5,102	2,035	0.399
3	2	240	170	0.708
4	12	4,612	1,315	0.285
5	14	4,051	1,518	0.375
6	10	5,234	422	0.081
7	18	5,265	1,239	0.235
8	18	3,669	3,336	0.909
9	28	4,553	4,701	1.033
10	2	229	27	0.118
Total	132	45,293	22,144	0.489

The four modeling algorithms were applied to each of the 132 subareas for a total of 528 individual models fitted for this study. For each subarea, one of the four model types was chosen to represent it based on internal model metrics, hold-out sample error rates, KG statistics, and a subjective assessment of fit based on the distribution of sensitivity classes. This process is discussed in Chapter 3. Table 5 documents the area, count, and number of unioned sites incorporated into each of the four model types. Additionally, Table 5 lists the same statistics broken down by upland subareas and riverine subareas. As evident in this table, the vast majority of subareas, area, and sites were modeled by the RF algorithm for both riverine and upland areas. The MARS algorithm comes in a distant second, followed by LR, and finally Model 2 (i.e., proportionally weighted linear combination model). Model 2, the most generalized model, was only used in areas that had very few known sites and therefore not enough evidence to discern a pattern. Figure 29 graphically represents the percentage of each model type that was chosen to represent the subareas of each region.

Table 5 - Quantification of Model Types by Landscape Position

Total			
Model	sq Miles	Subareas	Site Count
Model 2	3394	5	59
LR	2545	6	332
MARS	4498	24	2138
RF	34856	97	19615
total	45293	132	22144
Upland			
Model	sq Miles	Subareas	Site Count
Model 2	3362	4	50
LR	2320	3	128
MARS	3592	6	267
RF	31681	53	11958
total	40955	66	12403
Riverine			
Model	sq Miles	Subareas	Site Count
Model 2	32	1	9
LR	225	3	204
MARS	906	18	1871
RF	3175	44	7657
total	4338	66	9741

Region 6 required the most diverse selection of model types to represent the final sensitivity assessment. As documented in Table 4, the density of known sites within Region 6 is by far the lowest in the state. The only rival is Region 10, which conforms to the outline of Philadelphia County. Region 6 had numerous subareas with very few sites and therefore little pattern to identify. The use of LR and the proportionally weighted Model 2 reflect this fact. Similarly, the lower site densities of Regions 4 and 5 also required the use of less complex models such as MARS. This is not to say that areas with lower site densities cannot be modeled by RF, but that the areas of low site density often have site samples that are clearly not representative of the population.

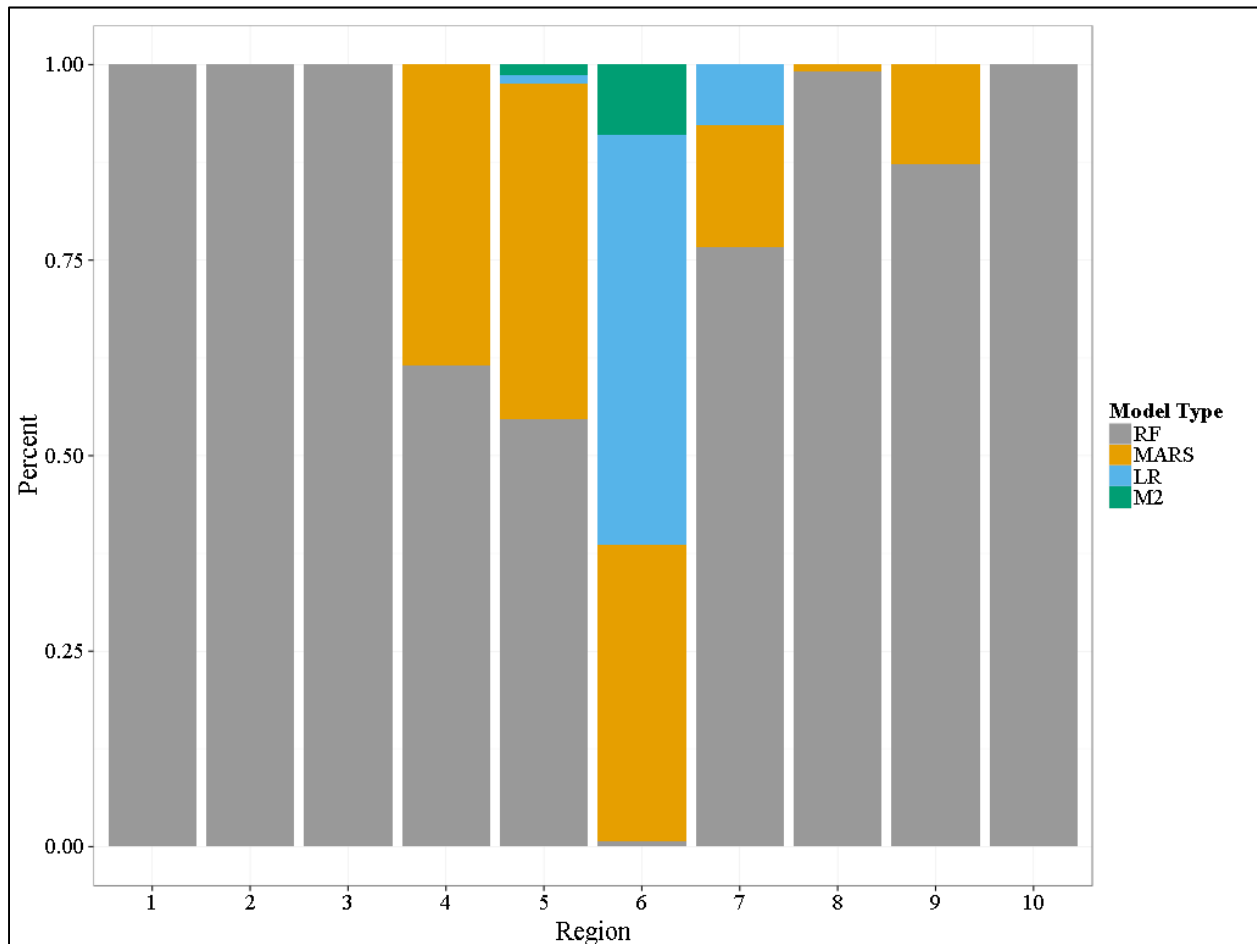


Figure 29 - Percentage of model types selected for the subareas of each region.

CLASSIFICATION ERRORS AND PERFORMANCE

A mosaic of the 132 models enumerated in Table 5 were classified based on specified thresholds into high, moderate, and low sensitivity, then mosaicked together to form the final sensitivity layer. Table 6 is a summation of the final model error rates and performance. Tables such as this, termed a “confusion matrix,” are presented at the end of each the Task 4, 5, and 6 reports. This table groups the high and moderate sensitivity classes of the sensitivity layer into a category for site presence and the low sensitivity class into site absent. This is in effect classifying high and moderate as areas where sites are likely and low as an area where sites are unlikely. Using binary classes (i.e., present and absent) allows for the construction of this confusion matrix and the derivation of the performance metrics listed below it. These metrics are discussed in the Task 3 report (pp. 67–68) and defined in the glossary, but the schematic table explaining these terms is reproduced here (Table 7).). In total, the final sensitivity model correctly classifies 98.4% of known site-present cells within an area equal to 29.2% of the state, a Kg of 0.701.

Table 6 - Confusion Matrix and Performance Metrics for Statewide Mosaicked Model

ALL REGIONS				
		Known Sites		
		Present	Absent	
Model Prediction	Present	1,992,770	309,213,157	311,205,927
	Absent	31,472	747,684,746	747,716,218
		2,024,242	1,056,897,903	1,058,922,145

Sensitivity / TPR =	0.984
Specificity / TNR =	0.707
Prevalence =	0.0019
Kvamme Gain (Kg) =	0.701
Accuracy =	0.708
Positive Prediction Value (PPV) =	0.006
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.002
Positive Prediction Gain (PPG) =	3.350
Negative Prediction Gain (NPG) =	0.022
False Negative Rate (FNR) =	0.016
Detection Prevalence =	0.294

Important observations from this table include the balance of higher sensitivity at the expense of specificity. The sensitivity, or true positive rate (TPR), is the accuracy within the class being predicted for, that is, archaeological sites. A high sensitivity demonstrates that a high proportion of known sites is included within high and moderate areas. Alternatively, specificity is the true negative rate (TNR), or one minus the false negative rate (FNR); the higher the FNR, the lower the specificity/TNR. In a model such as this, a false positive error is incurred when an area without a known site is predicted to be likely to contain a site. On the other hand, a false negative error is when a known site is predicted to be in a low/site-unlikely sensitivity area. It is clear that a false negative (i.e., misclassifying a known site) is a much more egregious error than classifying a background cell as likely to contain a site. The former stems from the model missing known site locations because of a bias error, whereas the latter is a product of the model projecting sensitivity into areas that have not been surveyed. Because the point of this model is to project into unsurveyed areas, a moderate number of false negatives, and thereby a moderate FNR, is necessary because this is the area with a higher sensitivity for finding sites. The approach taken by this project was to maximize the TPR (rate of correct predictions) while maintaining a moderate FNR so that a comfortable portion of the landscape remains as high and moderate sensitivity. In this case, it equates to approximately 30% of each subarea on average.

Table 7 - Schematic of Confusion Matrix and Performance Metrics

	Known Sites		
	Present	Absent	
Present	True Positive (A)	False Positive (B)	Total Predicted Sites (A+B)
Absent	False Negative (C)	True Negative (D)	Total Predicted Non-sites (A+B)
	Total Sites (A+C)	Total Background (B+D)	Total

$$\begin{aligned}
 \text{Sensitivity / TPR} &= A/(A+C) \\
 \text{Specificity / TNR} &= D/(B+D) \\
 \text{Prevalence} &= (A+C)/(A+B+C+D) \\
 &= 1 - ((A+B)/(A+B+C+D) / (A/(A+C))) \text{ or } 1 - \\
 \text{Kvamme Gain (Kg)} &= (\text{Detection Prev} / \text{sensitivity}) \\
 \text{Accuracy} &= (A+D)/(A+B+C+D) \\
 \text{Positive Prediction Value (PPV)} &= ((\text{Sensitivity} * \text{Prevalence}) / ((\text{Sensitivity} * \text{Prevalence}) + \\
 &((1 - \text{Specificity}) * (1 - \text{Prevalence})))) \\
 \text{Negative Prediction Value (NPV)} &= ((\text{Specificity} * (1 - \text{Prevalence})) / (((1 - \text{Sensitivity}) * \text{Prevalence}) + ((\text{Specificity}) * (1 - \text{Prevalence})))) \\
 \text{Unexpected Discovery Rate (UDR)} &= 1 - \text{NPV} \\
 \text{Detection Rate} &= A/(A+B+C+D) \\
 \text{Positive Prediction Gain (PPG)} &= \text{PPV} / \text{Prevalence} \\
 \text{Negative Prediction Gain (NPG)} &= \text{UDR} / \text{Prevalence} \\
 \text{False Negative Rate (FNR)} &= C/(A+C) \\
 \text{Detection Prevalence} &= (A+B)/(A+B+C+D)
 \end{aligned}$$

The performance metric below the confusion matrix in Table 6 depicts various dimensions of the model's classification ability and errors. While there are numerous metrics presented here, there are equally as many ways in which a simple 2×2 table of classification results can be interpreted. With the exception of the Kg statistic, the rest are commonly used in the fields practicing machine learning and classification. The terms used here are defined within this report's glossary. A few of these metrics, including PPV, NPV, and UDR, are not particularly useful in themselves because in this

context the very low prevalence of site cells skews the result. However, other metrics such as PPG and NPG build off of these by incorporating prevalence and leading to a more meaningful value. Along with sensitivity, specificity, and the Kg, the PPG metric is perhaps the most important for understanding these results. The PPG is discussed by Oehlert and Shea (2007:6–10) and is essentially a measure of how much more likely one is to find a site in a high or moderate sensitivity area given this model versus survey at random. The PPG is composed of known site prevalence, which is the probability that any randomly selected cell contains a site, and the PPV, which is the probability that a location classified as a site actually contains a site. Therefore the PPG, which is PPV divided by prevalence, divides the model's ability to correctly classify known sites with the probability of finding a site anywhere. The resulting value is an assessment of how many times more likely it is to find a site in a high or moderate area using this model than survey at random; in this case, 3.350 times more likely.

MODEL BENCHMARKS

An important way to determine the value of a model is to compare it against the standards of the field or metrics of similar models. In archaeology, it would be difficult to identify any particular benchmarks or reference models to use for comparison. This has as much to do with the regional variation and intent of different models as much as it does with lack of a consistent and agreed upon modeling approach within our discipline. However, in lieu of such a benchmark, there are other standards that can serve as a basis for comparison: 1) random chance, 2) peer models, 3) the best current model, and 4) the best possible model. The result of each of these comparisons shows that the Pennsylvania model performs better than its peers given the current data. Beyond these benchmarks, the best test of this model will be time and the recordation of new prehistoric archaeological sites.

Random Model

Comparing the Pennsylvania model against a random model is a way to assess the minimum hurdle a model would have to pass to be useful. Clearly a model that predicted on par or worse than randomly surveying the landscape is a poor choice and will not likely serve the goals of the project. With this project, we have already compared the model against random survey in the previous section. Two of the metrics provided in Table 6 provide insight into how the model functions relative to the random model. These metrics are the Kg and PPG. As discussed above, the PPG of the statewide model is 3.350. This demonstrates that the model is more than 3 times more effective than survey at random. While this number is useful at assessing the relative utility of a model, it is a little hard to interpret its true implications. This number is most effective in comparing candidate models. The Kg is discussed in detail in Appendix A of the Task 1 report and in numerous other locations in subsequent reports. The important point of comparison is that a Kg of 0.0 is essentially random survey and a Kg of 1 is a perfect model fit (and potentially over-fit). The Kg of this model is 0.701, well above random. However, being that the Kg incorporates the FPR, a necessary error within any model of this type, attempting to maximize the Kg will only reduce the area in which new sites are projected to exist. A

model that simply does better than random chance does not necessarily provide a great deal of confidence in the model, but it does show that the model is at least minimally predictive.

Peer Models

The next level of comparison is to judge a model against its peers. In the Task 1 report of this project, 32 reports were assessed to set the foundation for this model. Of these reports, 9 included one or more models that could be quantified totaling 13 individual model assessments. Table 8 lists the Kg and FNR of these 13 models to serve as a basis of comparison. The Kg is chosen because it is a common metric that is used throughout these reports and is a relative measure of a model's precision. Additionally, the FNR is chosen because it is a direct measure of a model's misclassification of archaeological sites—a measure of accuracy. As discussed above, a false negative error is considered much more costly in this context than a false positive error, and is therefore an important metric.

The final rows of Table 8 show that the average Kg of these models is 0.432 with a standard deviation of 0.170. The FNR has an average of 0.191 (or 19% misclassification of sites) and a standard deviation of 0.085. The Kg of 0.432 is not necessarily high compared to what it represents or what would generally be considered a desirable model, but it is still useful. Essentially, to achieve a Kg such as this, the percentage of sites correctly classified would have to be a little under twice the percentage of total area predicted to contain sites. For example, a model that correctly classifies 80% of sites in an area that takes up 45% of the study area would achieve a gain of 0.437. Similarly, a model that predicts 35% of sites in 20% of the study area would achieve a gain of 0.428. These models are quite different, but have the same ratio of approximately 1.7:1 of percent predicted sites to percent predicted site-likely area, an average that could use some improvement. However, with a high Kg of 0.631 to a low of 0.150, there is quite a range. Conversely, the range of FNR is relatively narrow except for an outlier of 0.40. At an average FNR of 0.191, these values are perhaps more in line with model expectations than their corresponding Kg values. An average misclassification rate of 19% of known sites is not detrimental depending on the purpose and generality of the model. The general trend of a respectable to moderate FNR paired to a moderate to poor Kg indicates that these models account for a decent number of known sites, but have to expand the site-likely area (high and moderate sensitivity levels) to a relatively large portion of the study area.

Table 8 - Kg and FNR Metrics of Models Evaluated in the Task 1 Report

Model Date	Kg	FNR
1989	0.203	0.40
1989	0.200	0.17
1994	0.536	0.27
1994	0.606	0.19
1996	0.588	0.18
1996	0.593	0.17
1996	0.532	0.25
1996	0.631	0.18
1996	0.308	0.18
1996	0.382	0.18
2002	0.484	0.13
2002	0.402	0.03
2002	0.150	0.15
Mean	0.432	0.191
Std Dev.	0.170	0.085

The Pennsylvania model has a Kg of 0.701 and a FNR of 0.016—a substantial improvement from the average of the reference models. Figure 30 shows the distribution of Kg and FNR metrics for each of the reference models (red dots) and Pennsylvania model (black triangle). Within the wide range of Kg values, the Pennsylvania model is comfortably above the reference models. At this Kg value, the model has a ratio of 3.4:1 of percent correctly predicted sites to percent predicted site-likely area, a two-fold improvement on the mean of the reference models. Similarly, the FNR of 0.016 is a much lower error rate than the average of the reference models. Only a single model was close in error rate (0.03) but was paired with a much lower Kg (0.402), indicating it had to cover a large portion of the study area to achieve the low FNR. In other words, it was quite accurate (low FNR), but not very precise (low Kg).

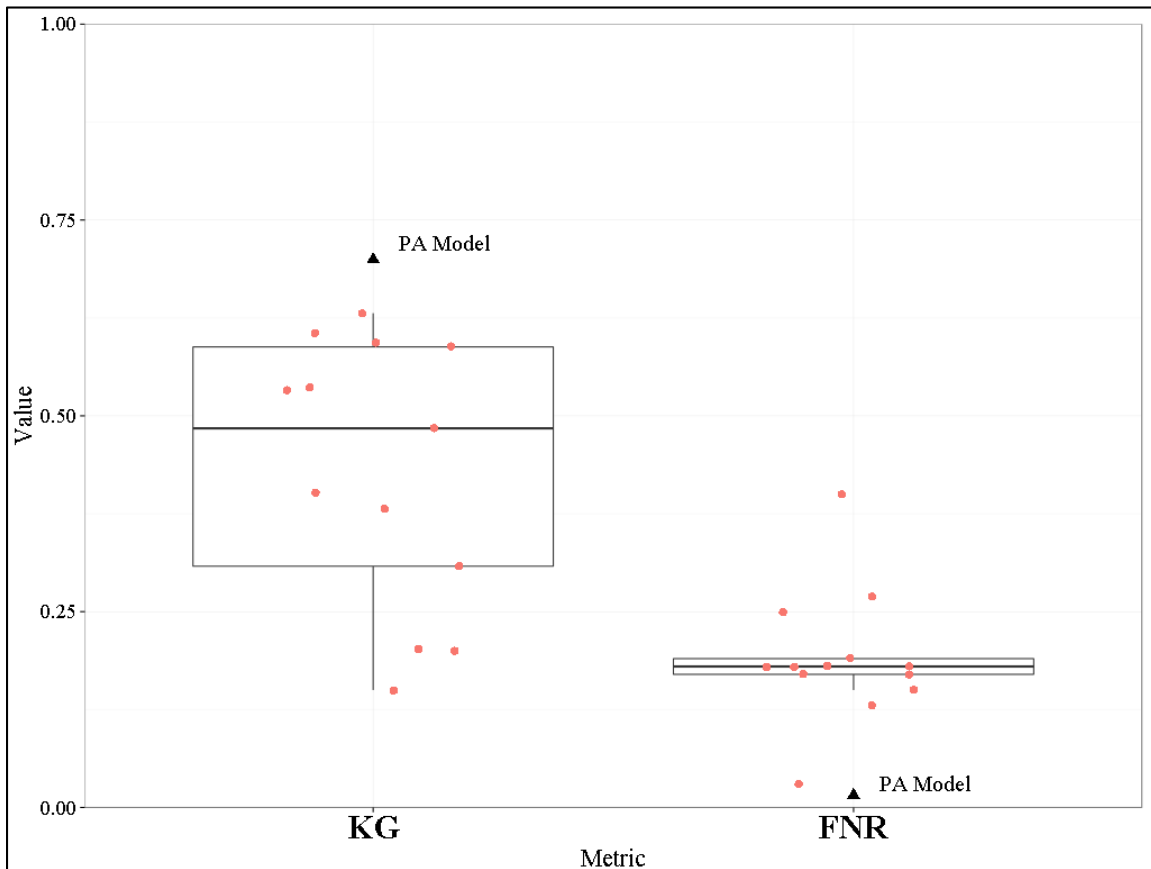


Figure 30 - Boxplots of Kg and FNR metrics for reference models and Pennsylvania model.

A bivariate scatterplot of these metrics is presented in Figure 31 for comparison. The reference models (red dots) and the Pennsylvania model (black triangle) are plotted along the axes of Kg and FNR. The Pennsylvania model achieves a separation from the reference models on both axes. To further add to the interpretation of this figure, base lines are added at subjective X-axis and Y-axis values to create quadrants corresponding to model error characteristics. Along the Y-axis of Kg, a boundary is drawn at 0.5 to separate models that are above and below a site-to-background percent ratio of 1:1. Models below this boundary would typically be poor and models above generally acceptable. On the X-axis of FNR, a line is drawn at 0.15, or a misclassification rate for sites of 15%. Models to the right of this would have a relatively poor misclassification rate and models to the left would have generally acceptable misclassification rates. Based on these boundaries, four quadrants are formed. The upper-left quadrant has acceptably high Kg and low FNR signifying models that have acceptable accuracy and precision. The upper-right quadrant has a higher Kg, but also high FNR signifying a model with precision, but little accuracy. Conversely, the lower-left quadrant has a low Kg and low FNR, signifying models with accuracy, but poor precision. Finally, the lower-right quadrant has model with low Kg and high FNR. These models are not particularly accurate or precise. The reference models are scattered about quadrants two, three, and four, but only the

Pennsylvania model is found in quadrant one (both accurate and precise). The Pennsylvania model appears to better its peers in both accuracy and precision.

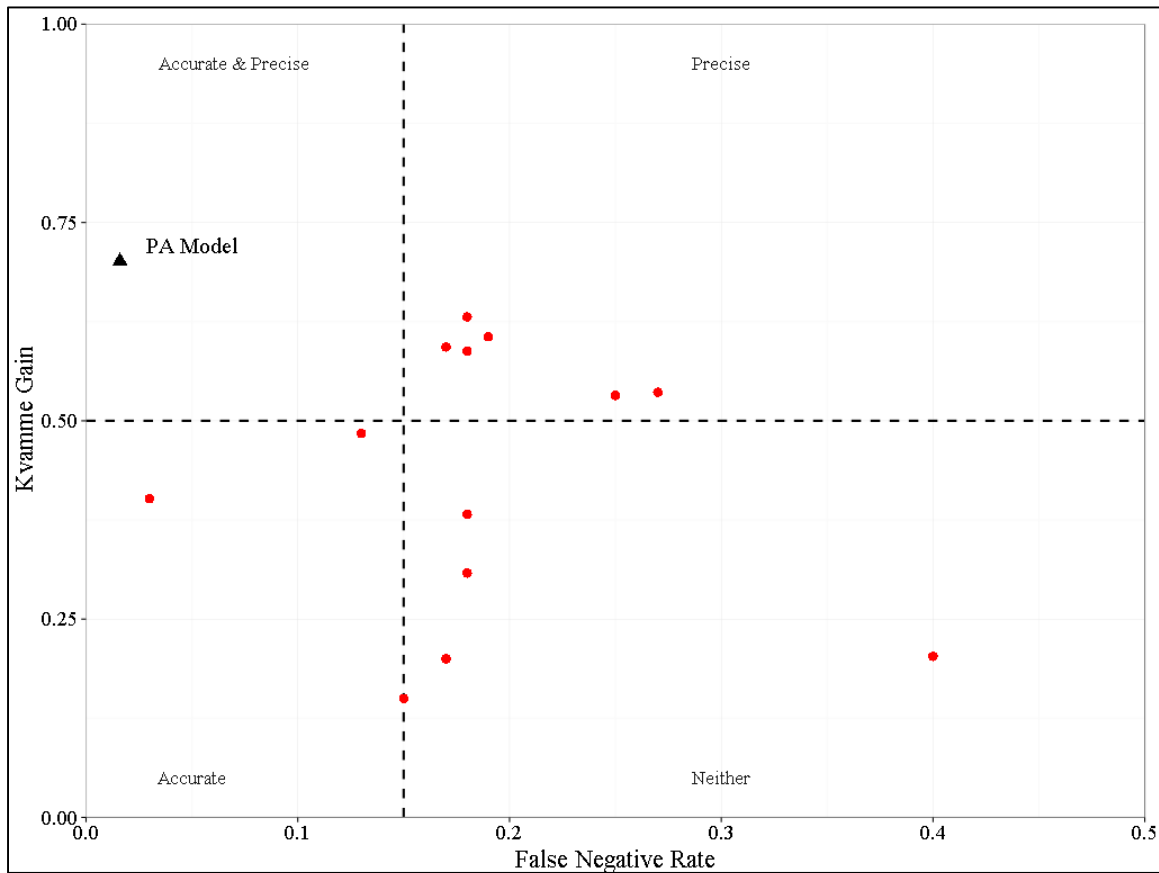


Figure 31 - Scatterplot of Kg and FNR metrics for reference models and Pennsylvania model.

Best Current Model

Identifying the best current model is problematic because no current model exists. However, the current state of archaeological survey in Pennsylvania provides a useful proxy. Given a series of assumptions, the data for site and survey locations can be combined into a confusion matrix for comparison purposes. This method of comparison is only for a general reference because of the methods by which sites are recorded and surveys are conducted. However, the values that can be reasonably justified are presented here for reference.

The following comparisons use the current environmental review (ER) survey areas as a proxy for the Pennsylvania model. Because the ER survey areas are located only where regulated undertakings occur, they do not necessarily represent the highest sensitivity areas within the state. However, it can be assumed that survey was requested by the PHMC because the project area had some degree of sensitivity for prehistoric sites and that the archaeologists who did the survey identified areas of

elevated potential for testing. This hypothesized “ER survey model” results in a precise model because it identifies 12% of the known sites within an area of 2.6% of the state, for a Kg of 0.783. As in the analysis in the section above, the FNR has no bearing here because false negatives have no meaning outside of surveyed areas. However, this relatively high Kg can be combined with the statewide site prevalence of 0.0019 to derive a PPG of 4.602. This indicates that, given the method used to call for an ER survey and the methods used to find sites (e.g., STUs, pedestrian survey, etc.), a location within an ER survey area is 4.6 times more likely to contain a site than an area outside of an ER survey area. Again, these figures are built from a number of assumptions, but will suffice for comparative purposes.

The PPG for the Pennsylvania model is 3.350, but further manipulation is required to make the ER survey model PPG and the Pennsylvania model PPG comparable. The Pennsylvania model is, by choice, much less specific than the ER survey model. While the ER survey model only covers 2.6% of the state, the Pennsylvania model considers 29.4% of the state as site-likely; these percentages are called the detection prevalence. In order to compare the PPG metrics, the detection prevalence of each model needs to be made equal. This is straightforward for non-site areas, but more complicated for site areas of the current model. The complication arises from the fact that a reduction of model area percent does not lead to an equal reduction of site percent. In the subarea models that make up the total model for this project, it is very typical for the vast majority of sites to be accounted for in the first few percent of the background area. As the amount of background area increases, the percent of sites accounted for decreases rapidly. A 50% reduction of the site-likely area will not lead to a 50% reduction of the percent of sites accounted for. Instead, a 50% reduction in site-likely area may only lead to a 10% reduction in the percent of correctly predicted sites. Therefore, for a 91% reduction in the Pennsylvania model’s site-likely area to make the detection prevalence match the ER survey model, a much smaller reduction in the site-percent is required. This percent can be found in the Kg, which is the ratio of background to site percentages. Using the Kg of the current model equates to a reduction of 70% of the correctly predicted sites to match the 91% decrease in the site-likely area needed to make the detection prevalence of the two models comparable. Finally, after the two models are made comparable, a PPG of 11.054 and Kg of 0.910 are calculated for the Pennsylvania model. As such, the adjusted version of the Pennsylvania model correctly classifies 29% of sites within an area of 2.7% of the state, whereas the ER survey model correctly classifies 12% of sites in approximately the same area (Table 9).

Table 9 - Metrics of ER Survey Model and Pennsylvania Model

Metric	ER Model	Current Model	Adjusted Current Model
Kg	0.783	0.701	0.910
Sensitivity / TPR	0.120	0.984	0.294
PPG	4.602	3.350	11.054
Detection Prevalence	0.026	0.294	0.027

Best Model

The final comparison to be made here is more of a theoretical exercise than the previous comparisons. As it is wise to compare a model to the lowest bar of the random or null model, it is also wise to compare a model against the highest bar, which is the best possible model given the data available. Using an example of a model built to predict email spam, the best model would be a model that perfectly separates spam emails from non-spam emails with no errors on either side. However, given the data available to make such a model, a perfect error-free model might not be possible. This hypothetical optimal model that makes the best prediction based on the given data is referred to as the Bayes Classifier by Hastie et al. (2009:21).

For APM, it is not as straightforward as the spam example. Typically, the point of an APM is not to perfectly predict the location of every unknown archaeological site. This is because it is unequivocally clear that the data given to build these models are not capable of creating a perfect error-free model. Therefore, the Bayes Classifier for APM will be something with a reasonable amount of error, preferably on the side of false positives. However, since it is clear that archaeological data will not lead to a perfect model no matter how optimized the hypothetical Bayes Classifier could be, the amount of error acceptable for a sub-optimal model is highly dependent on the goals of the model. The identification of how much and what kinds of error are acceptable in a model given the goals of that model are tightly coupled to a discussion of model thresholds. Such a discussion is present in the Task 4 report (pp. 86–91). Suffice to say, that without a great deal of further development in a field-wide framework of models and objectives, the “best” model given the data and project goals is an unknown quantity. Further, how this model compares to the hypothetical Bayes Classifier for planning-based APM is unknown and will only be told by time, field testing, and modeling improvements.

MODEL ACCURACY

Model accuracy for this project was established through various internal measures depending on the model type, as well as on a 25% hold-out sample for all model types. The hold-out testing sample was chosen at random from the population of site-present cells for each subarea. The values present the internal testing (Test RMSE) and external testing (CV Accuracy) of the models that compose the final mosaic. Table 10 lists the Area Under the Curve (AUC), test RMSE error from the hold-out sample, and the out-of-fold error rate from the 10-folds CV model fitting.³

³ The error rate measures were not calculated in a comparable metric for models of Regions 1, 2, and 3.

Table 10 - Error Rates for OOB and Hold-Out Samples for Three Model Types

Model	AUC	Test RMSE	CV Accuracy	Subareas
LR	0.970	0.208	0.206*	6
MARS	0.953	0.237	0.916	24
RF	0.990	0.097	0.988	67 [†]
Average	0.971	0.181	N/A	N/A

* The CV metric of LR is RMSE and not accuracy

† The total number of RF models including Regions 1, 2, and 3 is 97

The hold-out error rates of each subarea model for each region are discussed in detail under the heading of “Model Validation” in Chapter 5 of the reports for Tasks 4, 5, and 6. This section looks at the error rates in aggregate. As documented in Table 10, the error rate values for the test RMSE are quite low for the 25% hold-out sample. The RMSE for the three models ranges from a high of 0.237 for the MARS model and a low of 0.097 for the RF models. Interestingly, the LR model RMSE is lower than the MARS model. This seems a little counterintuitive given that the MARS model should decrease the bias error over LR, but it is a result of the small number of LR models selected for the final sensitivity assessment. The six LR models were selected because they were a better fit to the subarea than the MARS or RF model; therefore these six models are biased toward lower RMSE. The RMSE of the MARS models is slightly higher than the six selected LR models, but not significantly so. On the other hand, the RMSE of the selected RF models at 0.097 is much lower than the MARS or LR.

The AUC is not necessarily a measure of error, but instead a measure of how well the model balances the TPR and FPR across all thresholds. The higher the AUC the better the model fits the hold-out data. Following the same trend as the RMSE, the AUC is best (highest) for the RF model and lowest for the MARS models. Finally the CV error is the error rate calculated on the out-of-fold sample and averaged across each of the 10-folds. This error should essentially always be lower or perhaps nearly equal to the hold-out sample test. That is because the out-of-fold sample is only an estimator of true hold-out samples since the model will eventually use all of the out-of-fold data for model fitting. In this study, the CV error was measured as a percentage of accuracy for RF and MARS and as RMSE for LR.

What Table 10 shows is that the RMSE of the 25% hold-out sample is quite low for all three model types and particularly low for the RF models. This is the case for two main reasons: 1) the models are a good representation of the pattern present in the known archaeological sites and therefore good at predicting hold-out values; and 2) the spatial autocorrelation of the site-present cells being predicted for adds a degree of bias to the error rate. The first point is simply that the models are relatively accurate descriptions of the existing data. This point is elaborated on throughout this chapter and is the focus of many sections in the task reports. The second point is addressed in the methodological

sections in the previous reports and Chapters 3 and 5 in this report. The main issue associated with autocorrelation of the response variable observations (the site-present cells being predicted for) is that they are not spatially independent or identically distributed. Instead, site-present cells are dependent on the location of other site-present cells and as such are often clustered. Although the 25% hold-out sample is randomly chosen from the population of site-present cells, there is a greater probability of cells being selected from known sites that occupy large areas and from geographic areas where numerous known sites are clustered. The large area of a single site or area where sites are clustered will be more similar relative to the environmental variables than areas without clusters or large known sites. Therefore, the model predictions will be biased toward higher predicted probabilities for site-present cells near areas of large and clustered known sites. This is a known source of bias in these models, and methodological controls attempt to reduce it. However, it will only be through stratified random sampling techniques, the incorporation of random spatial effects, or the estimation of the structure of the spatial errors that the bias of autocorrelation will be greatly reduced.

VARIABLE IMPORTANCE

The degree to which the different environmental variables correlate to site presence and absence is measured by the K-S and MW U test prior to model fitting. This process is discussed in Chapter 3 and Figure 9. This routine of testing allows for the distinguishing of variables that on the univariate dimension have some power to discriminate site-present cells from the general background. However, these tests do not identify high-dimensional relationships that may affect variables differently in the presence of other variables. The RF algorithm used in this project is able to identify variable importance in the presence of other variables. The results of the variable importance measures for 102 subareas in Regions 4–10 are explored below.⁴

The RF algorithm measures variable importance by measuring the decrease in accuracy incurred by removing a variable. After the construction of each tree in the forest ensemble, the OOB data are randomly permuted and sent through each split in the tree. The variable that defines each split is essentially removed by this because of the randomized data. At each split, the accuracy is recorded on the random data and compared to the accuracy achieved at that split (by a specific variable) on the real training data. The mean decrease in classification accuracy between random and real data is a measure of the importance of each variable. The reason for this is that if a variable predicts equally as well with real and random data (a low mean decrease in accuracy), then it is not a very effective variable. On the other hand, if a variable predicts very well with real data and poorly with random data (a high mean decrease in accuracy), then that variable is taken to be important.

⁴ The variable importance measures were not calculated in a comparable metric for the models of Regions 1, 2, and 3.

The equation outlines this process with $VI(X_j)$ as the overall variable importance. This is calculated as the average variable importance of the j^{th} variable from the b^{th} tree ($VI_b(X_j)$), then averaged across all trees in the ensemble forest ($ntree$). Finally, $VI_b(X_j)$ is calculated as the difference in accuracy between the classifications of real data versus permuted OOB data.

$$VI(X_j) = \frac{\sum_{b=1}^{ntree} VI_b(X_j)}{ntree}$$

To make use of this technique, the variable importance of each RF model is calculated and then standardized to make a model-to-model comparison more appropriate. The tables below show the results of this for both riverine (Table 11) and upland (Table 12) subareas, assuming that the variables affecting settlement are likely different in these two settings. As discussed above, the VI is measured as a mean decrease in accuracy across all trees in a model: the higher the value, the more important the variable. The tables show the top 10 variables sorted by 1) the average VI for each model; 2) total summed VI from all models; and 3) the total number of times that a variable was used in these regions. The green highlighted column identifies the column for which the results are sorted in descending order. And finally, Table 13 combines the measures of total VI and average VI into a composite score to identify the variable that contributes the most to accuracy and does so in many subarea models. The composite score is the sum of the total value and average value after they are manipulated to a common scale. This table is a list of the top 30 variables from both riverine and upland settings. The variables are color-coded by general type and put side-by-side for comparison.

Table 11 - Variable Importance Measures for Riverine Subareas of Regions 4–10

Variable Importance Ranked by Average Decrease					
Rank	Variable	Average Decrease	Total Decrease	Count	Variable Description
1	e_trail_dist	134.3	2552.4	19	Distance to historic Indian trail (Wallace 1998)
2	ed_h6	127.2	3561.5	28	Distance to 4th order and up steams
3	c_trail_dist	115.5	3465.9	30	Cost distance to historic Indian trail (Wallace 1998)
4	ed_drmh	104.2	833.3	8	Distance to head of drainage
5	ed_h7	103.3	723.4	7	Distance to 3rd order and up streams
6	ed_h4	87.9	1230.4	14	Distance to NWI wetlands
7	cd_h3	78.8	157.5	2	Cost distance to NHD water bodies
8	ed_conf	73.5	514.3	7	Distance to stream confluence
9	cd_h7	72.5	435.2	6	Cost distance to 3rd order and up streams
10	cd_h6	66.8	601.6	9	Cost distance to 4th order and up steams
Variable Importance Ranked by Total Decrease					
Rank	Variable	Average Decrease	Total Decrease	Count	Variable Description
1	ed_h6	127.2	3561.5	28	Distance to 4th order and up steams
2	c_trail_dist	115.5	3465.9	30	Cost distance to historic Indian trail (Wallace 1998)
3	e_trail_dist	134.3	2552.4	19	Distance to historic Indian trail (Wallace 1998)
4	cd_drmh	64.5	1871.7	29	Cost distance to head of drainage
5	aws050	55.2	1820.2	33	Soil: available water capacity at 50 cm below surface
6	elev_2_drainh	66.2	1788.2	27	Vertical elevation to head of drainage
7	ed_h2	50.2	1605.0	32	Distance to NHD flow lines
8	cd_h4	58.6	1464.9	25	Cost distance to NWI wetlands
9	ed_h5	56.4	1410.0	25	Distance to NWI water bodies
10	std_32c	53.3	1331.9	25	Standard deviation of slope in 32 cell neighborhood
Variable Importance Ranked by Count					
Rank*	Variable	Average Decrease	Total Decrease	Count	Variable Description
9	aws050	55.2	1820.2	33	Soil: available water capacity at 50 cm below surface
10	ed_h2	50.2	1605.0	32	Distance to NHD flow lines
11	c_trail_dist	115.5	3465.9	30	Cost distance to historic Indian trail (Wallace 1998)
12	cd_drmh	64.5	1871.7	29	Cost distance to head of drainage
13	elev_2_strm	37.4	1085.5	29	Vertical elevation to stream
14	ed_h6	127.2	3561.5	28	Distance to 4th order and up steams
15	elev_2_drainh	66.2	1788.2	27	Vertical elevation to head of drainage
16	tpi_sd10c	19.9	536.7	27	Standard deviation of Topographic Position Index for 10 cell neighborhood
17	tpi_10c	19.6	529.9	27	Topographic Position Index for 10 cell neighborhood
18	cd_h4	58.6	1464.9	25	Cost distance to NWI wetlands

* Rank for variables sorted by count begin at 9 because the eight factor levels of NICCDC took up the first eight spots.

Table 12 - Variable Importance Measures for Upland Subareas of Regions 4–10

Variable Importance Ranked by Average Decrease					
Rank	Variable	Average Decrease	Total Decrease	Count	Variable Description
1	ed_h6	205.1	2051.5	10	Distance to 4th order and up steams
2	ed_h7	176.3	881.4	5	Distance to 3rd order and up streams
3	e_trail_dist	163.2	1142.3	7	Distance to historic Indian trail (Wallace 1998)
4	cd_h7	135.2	2973.9	22	Cost distance to 3rd order and up streams
5	cd_h6	112.2	1570.2	14	Cost distance to 4th order and up steams
6	ed_drnh	90.4	1175.1	13	Distance to head of drainage
7	c_trail_dist	78.7	3146.2	40	Cost distance to historic Indian trail (Wallace 1998)
8	ed_h1	74.1	148.2	2	Distance to historic streams (PSU 2004)
9	ed_h4	67.7	474.0	7	Distance to NWI wetlands
10	ed_h2	63.6	1017.4	16	Distance to NHD flow lines
Variable Importance Ranked by Total Decrease					
Rank	Variable	Average Decrease	Total Decrease	Count	Variable Description
1	c_trail_dist	78.7	3146.2	40	Cost distance to historic Indian trail (Wallace 1998)
2	cd_h7	135.2	2973.9	22	Cost distance to 3rd order and up streams
3	cd_h5	57.3	2232.9	39	Cost distance to NWI water bodies
4	ed_h6	205.1	2051.5	10	Distance to 4th order and up steams
5	cd_h4	51.1	1991.6	39	Cost distance to NWI wetlands
6	elev_2_drainh	51.1	1889.7	37	Vertical elevation to head of drainage
7	elev_2_strm	36.3	1741.4	48	Vertical elevation to stream
8	cd_conf	40.6	1703.7	42	Cost distance to stream confluence
9	cd_h6	112.2	1570.2	14	Cost distance to 4th order and up steams
10	aws050	38.1	1296.2	34	Soil: available water capacity at 50 cm below surface
Variable Importance Ranked by Count					
Rank	Variable	Average Decrease	Total Decrease	Count	Variable Description
1	elev_2_strm	36.3	1741.4	48	Vertical elevation to stream
2	elev_2_conf	26.2	1230.7	47	Vertical elevation to stream confluence
3	cd_conf	40.6	1703.7	42	Cost distance to stream confluence
4	c_trail_dist	78.7	3146.2	40	Cost distance to historic Indian trail (Wallace 1998)
5	cd_h5	57.3	2232.9	39	Cost distance to NWI water bodies
6	cd_h4	51.1	1991.6	39	Cost distance to NWI wetlands
7	elev_2_drainh	51.1	1889.7	37	Vertical elevation to head of drainage
8	aws050	38.1	1296.2	34	Soil: available water capacity at 50 cm below surface
9	cd_h2	24.1	772.0	32	Cost distance to NHD flow lines
10	tpi_sd250c	21.8	696.3	32	Standard deviation of Topographic Position Index for 250 cell neighborhood

Table 13 - Variable Importance Values Centered and Scaled Comparing Riverine and Upland Settings

Rank	Riverine Subareas				Upland Subareas			
	Variable	Total	Average	Sum	Variable	Total	Average	Sum
1	ed_h6	1.000	0.947	1.947	ed_h6	0.652	1.000	1.652
2	c_trail_dist	0.973	0.860	1.833	cd_h7	0.945	0.659	1.604
3	e_trail_dist	0.717	1.000	1.717	c_trail_dist	1.000	0.383	1.383
4	ed_drnh	0.234	0.776	1.010	e_trail_dist	0.363	0.795	1.159
5	cd_drnh	0.526	0.480	1.006	ed_h7	0.280	0.859	1.139
6	ed_h4	0.345	0.655	1.000	cd_h6	0.499	0.547	1.046
7	elev_2_drainh	0.502	0.493	0.995	cd_h5	0.710	0.279	0.989
8	ed_h7	0.203	0.769	0.972	cd_h4	0.633	0.249	0.882
9	aws050	0.511	0.411	0.922	elev_2_drainh	0.601	0.249	0.850
10	cd_h4	0.411	0.436	0.848	ed_drnh	0.373	0.441	0.814
11	ed_h2	0.451	0.374	0.824	cd_conf	0.542	0.198	0.739
12	ed_h5	0.396	0.420	0.816	elev_2_strm	0.553	0.177	0.730
13	rng_32c	0.365	0.421	0.785	ed_h2	0.323	0.310	0.633
14	std_32c	0.374	0.397	0.771	aws050	0.412	0.186	0.598
15	elev_2_conf	0.321	0.406	0.727	cd_drnh	0.337	0.199	0.536
16	vrf_32c	0.274	0.428	0.703	elev_2_conf	0.391	0.128	0.519
17	ed_conf	0.144	0.547	0.692	std_32c	0.304	0.212	0.516
18	eldrop32c	0.275	0.405	0.680	ed_h4	0.151	0.330	0.481
19	cd_h6	0.169	0.497	0.666	ed_h5	0.175	0.298	0.473
20	cd_h7	0.122	0.540	0.662	rng_32c	0.257	0.187	0.444
21	cd_h5	0.219	0.415	0.634	e_hyd_min	0.211	0.231	0.441
22	cd_h3	0.044	0.587	0.631	ed_h1	0.047	0.361	0.408
23	e_hyd_min	0.270	0.358	0.628	tri_32c	0.203	0.173	0.377
24	elev_2_strm	0.305	0.278	0.583	slpvr_32c	0.193	0.174	0.367
25	e_hyd_min_wt	0.232	0.341	0.573	eldrop32c	0.245	0.121	0.367
26	tri_10c	0.087	0.462	0.549	rng_16c	0.160	0.205	0.365
27	cd_conf	0.218	0.322	0.540	cd_h2	0.245	0.118	0.363
28	std_16c	0.053	0.466	0.519	tpi_sd250c	0.221	0.106	0.327
29	slpvr_32c	0.185	0.288	0.473	tpi_250c	0.218	0.105	0.323
30	tri_32c	0.181	0.281	0.462	ed_conf	0.096	0.209	0.305

Color coding: blue = distance to hydrology; green = measure of topography; brown = soil attribute; purple = elevation to water feature; yellow = distance to trails

From these tables, it is evident that the Euclidian distance to streams of 4th order and higher (*ed_h6*) is the most important variable in both upland and riverine settings when averaged across the RF

models for all subareas in Regions 4–10. However, for upland settings, the distance to 3rd order and higher streams (*ed_h7*) is almost equally important. In general, this finding is perhaps not surprising given the emphasis archaeologists put on sources of reliable year-round water in our mental settlement models. For the upland areas, the importance of distance to 3rd order and high streams (*ed_h7*) makes a lot of sense as these waterways originate further into the uplands than the 4th order streams. Simply, reliable flowing water is a very important factor in site location given the data we have. The next most important variables are a bit of a surprise: the distance to historically documented Indian trails (Wallace 1998) measured by Euclidian and slope-sensitive cost distance. The historic trails do have a decent degree of correlation with both 3rd and 4th order streams, but these trails also often correspond to ridge tops and other overland routes. There may also be some circularity in this correlation because people may have sought to identify sites near the trails (first published in 1965), and these trails often correlate to modern roads that people travel to find fields to surface collect.

Following these variables, there begins to be a bit of a shift in important variables between riverine and upland locations. For riverine subareas, variables tied to the location of heads of drainage (*ed_drnh*, *cd_drnh*, and *elev_2_drainh*) are important, as well as the distance to NWI wetlands (*ed_h4*). In the upland settings, the slope-sensitive cost distances to 3rd and 4th order streams and NWI wetlands (*cd_h4*) and water bodies (*cd_h5*) gain importance. The same variables focused on drainage head locations are slightly further down the list of importance. It is interesting that the variables associated with drainage heads are seemingly slightly more important in riverine areas than uplands. The heads of drainage are a decent analog to the location of spring heads and seeps, hydrology features thought to be important in upland settings. Although they have a higher rank in riverine settings, drainage heads are still important in upland settings.

The most important (and seemingly only important) soil variable is the available soil water capacity at a depth of 50 cm (*aws050*). The other soils variables appear further down the list, generally ranked in the lower 50% of all variables, with *niccdcd* appearing more important than *drcwet* or *drcdry*.

The variables that measure aspects of topography, slope variation, and landforms are generally derived from the DEM raster and appear among the more important variables in Table 11, Table 12, and Table 13. For both upland and riverine areas, the variables of range of slope (*rng_32c*) and standard deviation of slope (*std_32c*) over 32-cell neighborhoods rank high. These two measures are highly correlated (Pearson's $r = 0.956$) at known site locations and can be used by the algorithm interchangeably. Figure 32 below shows a Pearson's r correlation matrix for all of the variables in Table 13, noting that many of the correlated variables were separated during the variable selection process and are not modeled together. The lower half of Table 13 contains a number of additional DEM derived variables including the topographic relief index (*tri*), vector roughness factor (*vrf*), elevation drop (*eldrop*), and topographic position index (*tpi*). Also in the lower half of Table 13 are additional hydrology variables, most of which have analogs or cost sensitive measures in the upper half.

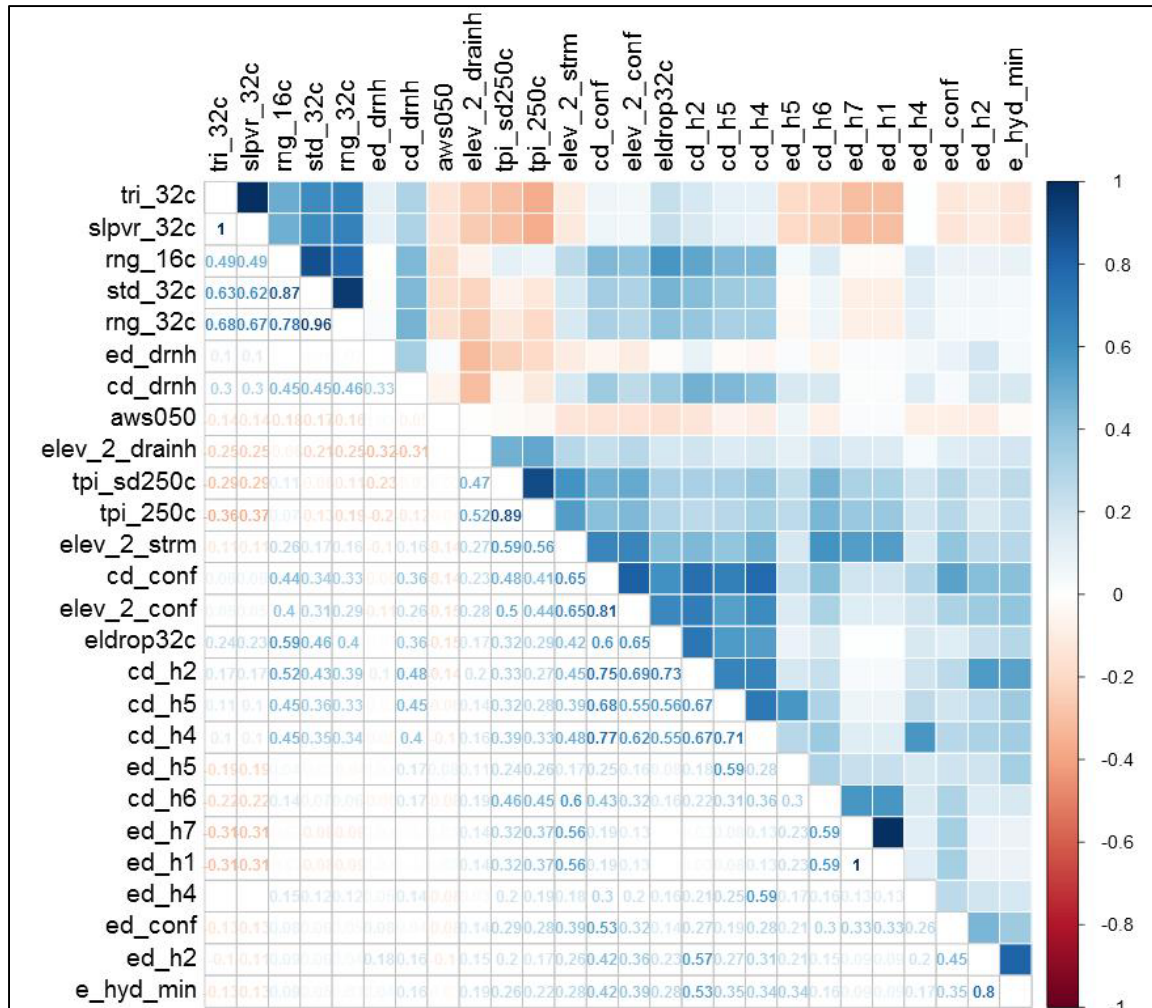


Figure 32 - Pearson's correlation r value for the 30 most important variables.

Interesting observations based on these findings include the apparent utility of a 32-cell neighborhood when calculating variables derived from the DEM. For this project, these same variables were all calculated at neighborhoods of 8, 10, 16, and 32. That intention of this range of neighborhoods was to determine at what scale the patterns became most useful. It appears that the most useful size is 32 cells, but attempting the same analysis for larger neighborhoods would be informative. The topographic position index (*tpi*) was calculated at a different range of neighbors (5, 10, 50, 100, and 250) because this variable requires a larger area to derive useful patterns. In this case again, the largest neighborhood size was the most useful. Another interesting observation is the utility of cost-sensitive distance in the upland subareas. In this project cost of distance is a function of slope, therefore it is not surprising that cost distance was not pervasive in riverine settings where slope in general is not as restrictive. However, in the uplands where slope can be a major factor in effective distance, weighting for slope apparently was effective. Also of interest is the utility of the elevation to features variables *elev_2_strm* for streams (specifically to NHD flow lines [*h2*]),

elev_2_drainh for heads of drainage, and *elev_to_conf* for the elevation above or below confluences. It is likely that these elevation variables interact with other hydrologic or topographic variables to derive this utility. Such interactions can be specified explicitly in the LR and MARS algorithms, but the RF model can detect these interactions based on node splitting parameters. New research indicates that variable interactions may be detectable from RF variable importance metrics (Kelly and Okada 2012). It is interesting that the variables associated with increasing accuracy are so similar between upland and riverine settings. Intuitively, it would seem that since upland and riverine settings are so very different in terms of hydrography and topography, the variables correlating to sites would also differ greatly. However, these results show that this is not the case. This is likely because the variables that influenced site location were not as radically different even when the environment differed. This also likely has to do with the more limited set of variables we are able to compute to serve as proxies for 10,000 plus years of environmental change. Those that work well in both settings are probably the variables that serve as solid proxies withstanding some degree of change and mapped closely to a component of previous settlement decisions, or at least what we believe them to be.

A final note to this analysis is to indicate the variables that faired very poorly. These included the classification of topographic index (*tpi_cls*) on various neighborhood sizes, topographic position index (*tpi*) of small neighborhoods, soil variables of *drcwet*, *drcdry*, and *niccdcd*, and flow direction (*flowdir*). The classified TPI likely is just a poorly constructed variable at this point, the small neighborhoods for *tpi* and other topographic variables are likely just too small to pick up on landscape trends, flow direction just has no correlation to site location, and the soils variables need some reconsideration. Soils variables would likely be improved by reworking them into continuous variables, measuring them as proximity variables, or manipulating them into composite variables that better describe the attributes of an environment resulting from different soil characteristics.

5

CONCLUSIONS AND RECOMMENDATIONS

The results of this project demonstrate that statistical techniques can identify and extrapolate patterns found within the body of known archaeological sites in Pennsylvania. Additionally, the use of split-sampling techniques and validation metrics show that these patterns can achieve a relatively low rate of error when identifying known sites not included in the pattern's development. The project further shows that the extrapolated pattern can be classified based on appropriate context-specific threshold criteria to illustrate a qualitative assessment of the sensitivity for archaeological material at a particular location given the characteristics of sites documented within environmental settings similar to that location. Finally, these assessments can be a useful tool in aiding planners and decision makers to better understand the relative potential impact of various project alternatives when considered at an appropriate scale.

The model generated by this project cannot replace archaeological field survey and, like any model derived from the abstraction of data, have strengths and weaknesses. While the model has wide geographic coverage, computational efficiency, and methodological consistency, it also uses biased and non-independent site data (as described in Chapter 1) and has limited explanatory power. These attributes must be considered in the model's implementation. Among the strengths of this approach is that the general pattern observable in known archaeological site locations can be extrapolated quickly over vast geographic areas, incorporating an array of variables. Further, this is done with statistical techniques that can identify high-dimensional relationships among variables and within a framework that helps us understand the relationships and the degree to which the patterns match the known data. This is akin to an archaeologist studying the known sites of a river valley, quantifying their relationship to features of the environment, and then impartially identifying where those features intersect throughout the valley, but on a scale that would take an infinite amount of time and brain power. The weaknesses of this approach are that the model *only* considers known archaeological sites to find a pattern, that known archaeological site distributions are not independent or necessarily representative of true site distributions, and that the variables used to describe the pattern are only a limited set of proxies for a large number of unknown variables that influenced Native American site selection processes. Building on the previous metaphor, our archaeologist finds a pattern that ignores knowledge gained from other watersheds, is aware of the faults in the data, and cannot be sure of casual relationships between the pattern and environmental variables. While these weaknesses exist, they do so as counterbalances to the strengths; in the spirit of Wolpert (1996), there is "no free lunch." Every attempt to model site sensitivity will require the balancing of strengths and weaknesses, where the "best" objective model is the one that "best" achieves the subjective goals of the project or study. To bridge the gap between model and goals requires an acceptance of strengths and weaknesses and avoidance of blind reliance on either one.

The methods used in this study were employed because they were a way to achieve the project's objectives within a given timeframe. These methods are an outgrowth of the theoretical base of APM studies old (Judge and Sebastian 1988) and new (Kamermans et al. 2009). Further, this study builds from the lessons learned from the Mn/Model's years of development (Oehlert and Shea 2007) and up-to-date research in numerous academic and industrial fields of study that employ similar models with similar goals and constraints. The end result of this process, as documented in seven task reports, is a transparently constructed and valuable tool for planning, advising, and research that balances strengths and limitations.

USAGE RECOMMENDATIONS

It is recommended that this sensitivity assessment be used for planning purposes such as comparing relative impacts from multiple project alternatives, calculating relative Phase I archaeological survey costs for multiple project alternatives, and acting as additional support for the guidance of archaeological field survey within the environmental review process (Figure 33). These uses should be implemented at a map scale no greater than approximately 1:24,000 (1 map inch = 2,000 on-the-ground feet), which is the same scale as a common USGS 7.5' series quadrangle map. As discussed in Chapter 1, this limitation is due to the scale at which the base data of archaeological site locations were originally recorded and digitized.

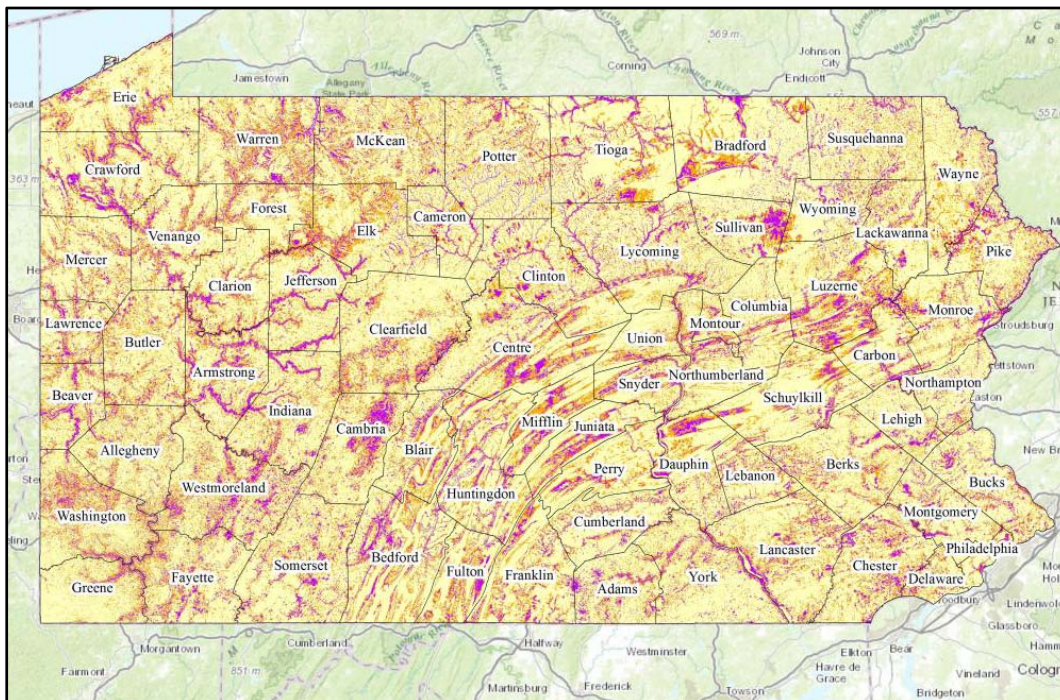


Figure 33 - Overview of final sensitivity layer.

Aggregate

Given the scale at which this model is most appropriate, roughly 1:24,000, it makes sense to have the sensitivity assessment raster reflect this scale. As generated, the final sensitivity assessment has a raster resolution of roughly 10×10 m. This resolution was established toward the beginning of the model to match the resolution of the DEM for consistency across all environmental variables. However, at the final stages of modeling, the use of this resolution is seemingly finer than can be accounted for by the base data (i.e., site locations). It therefore makes sense to aggregate the final model raster to a resolution that is more appropriate to the base data and better suited to the planning purpose of the model.

The process of aggregation involves reducing the raster resolution by grouping the sensitivity of a neighborhood of cells into a new sensitivity value. For example, if the search neighborhood is 10 cells, then for each neighborhood of 10×10 cells, a new sensitivity is calculated from the sensitivity values of that neighborhood. The resulting raster has a lower resolution than the original and a new value that is derived from the cells in the original raster. There are a number of functions to determine the new sensitivity value, but the most common are the maximum value of the original cells, their mode, minimum, or average (for continuous values). These functions simply describe how the numerous values of the original raster neighborhood are combined into the resulting lower resolution raster. The maximum function takes the highest value of the neighborhood, the minimum function takes the lowest value, the mode takes the most common value, and the average takes the mean of the original cells. Figure 34 is a graphical example of an aggregate that uses a 2×2 -cell neighborhood and takes the maximum value of the original raster and applies it to the resulting raster.

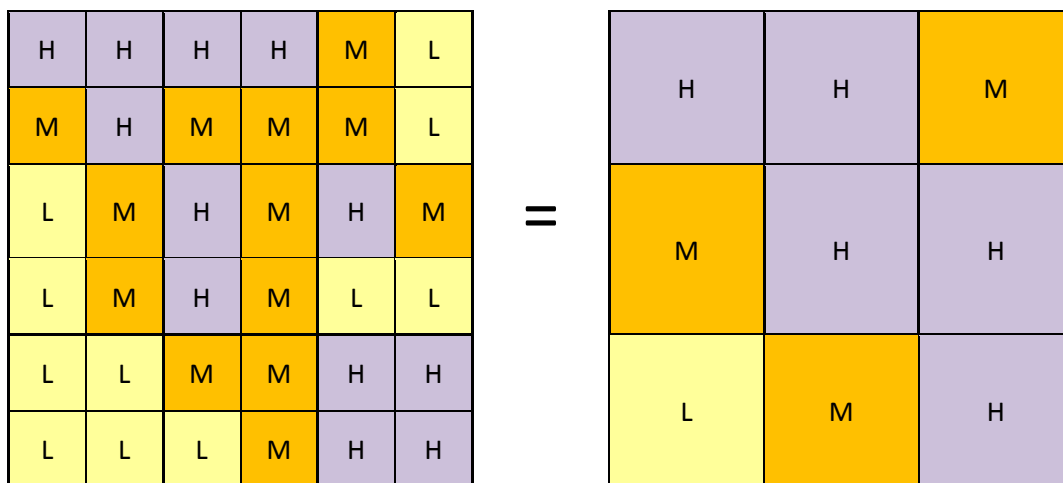


Figure 34 - Example of raster aggregation based on a 2-cell neighborhood and a maximum value function.

If the first 2×2 neighborhood of the original raster contains a high sensitivity cell, then the resulting raster will apply a value of high sensitivity to the aggregated cell. Similarly, the 2×2 neighborhood

in the lower left of the raster only contains low sensitivity, so the resulting raster will code that region as low.

For the final sensitivity raster from this model, it is recommended that a 3×3 neighborhood with a maximum function is used to derive a new aggregated raster for planning purposes. The approach would reduce the resolution down to 30×30 -m cells that more accurately reflect the base unit of analysis (i.e., archaeological sites). The maximum function is a conservative approach that uses the highest sensitivity likelihood established by the model for the resulting sensitivity. As such, if even a single cell in the nine cells that make up the 3×3 neighborhood is considered high sensitivity in the original raster, the aggregate raster will be set to high sensitivity for that neighborhood. Figure 35 is an example of the original 10×10 -m resolution sensitivity raster and Figure 36 is an example of the same area aggregated with a 3×3 neighborhood and the maximum function. The implementation of this approach has the added benefit of homogenizing geographic regions, defining landscape level trends, and removing sparse and spotty sensitivity assessments. This recommendation is designed to better align the output of this analysis with the goals of the project.

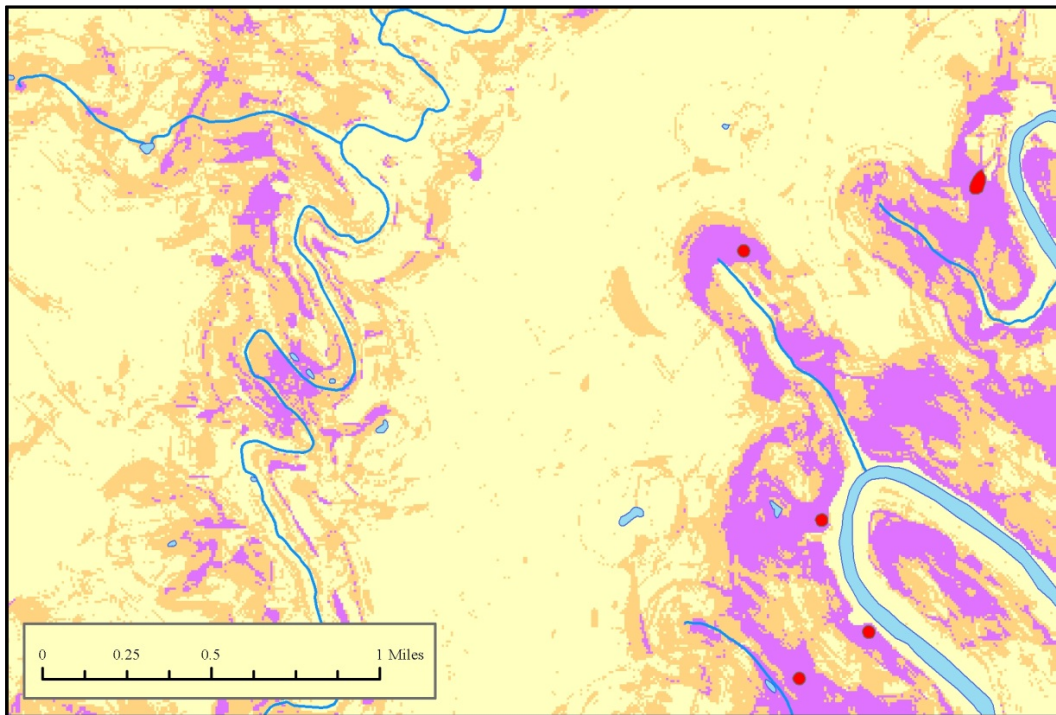


Figure 35 - Example of sensitivity assessment at original 10×10 -m resolution.

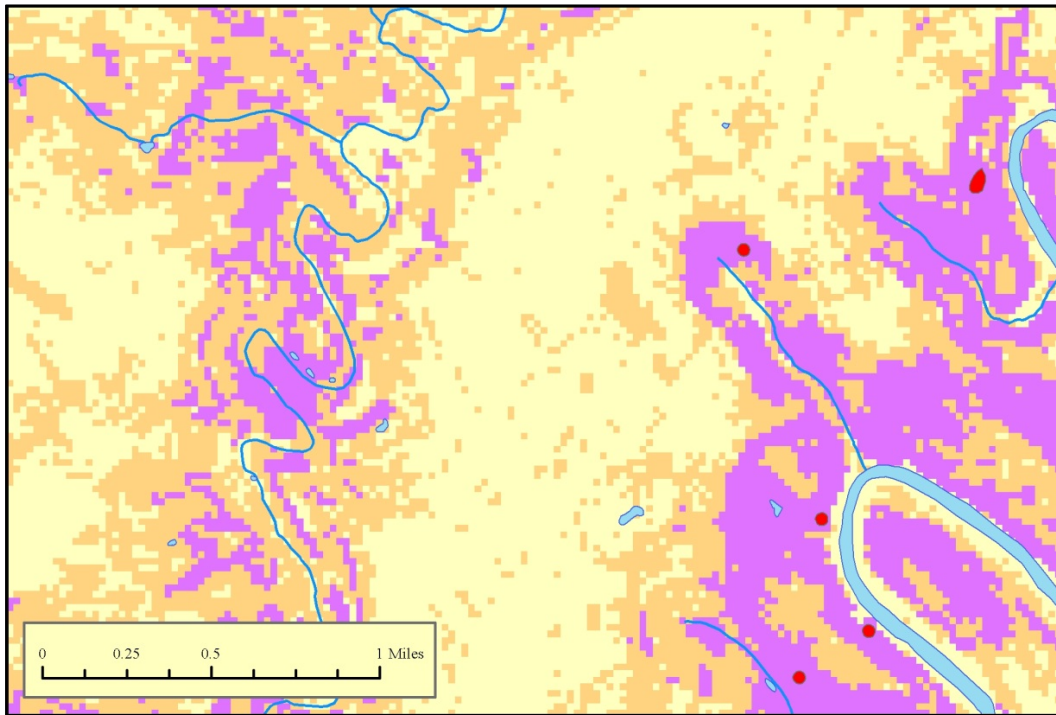


Figure 36 - Example of sensitivity assessment at aggregate of neighborhood maximum 30 × 30-m resolution.

Additional Information

The sensitivity assessment raster created through this process is a useful tool when combined with an understanding of the assumptions, models, and findings documented in this and the other task reports. However, once separated from their supporting literature, the data may be taken for granted, raising the potential for misuse. One solution to this problem may be to provide additional visual information along with the sensitivity assessment that quantifies another dimension of the analysis. Additional information pertinent to interpreting the analysis includes subarea boundaries, the type of model used in that subarea, the density of sites within the subarea, and raster layers of site and survey density.

Additional information can be incorporated by providing a separate layer or by integrating the additional information into the sensitivity assessment. This could be done mathematically or as a union of qualitative factors. As an overlay, the type of model, site, and survey densities per subarea could simply be a layer that is turned on or off by the viewer to add additional information to the decision-making process. Models in areas of very low site density are likely to be less representative of the entire site population. Figure 37 is an example of such an overlay, demonstrating the type of model used in each of the 132 subareas, and Figure 38 is the density of known sites per square mile within each subarea.

The second approach to providing additional information is to modify the sensitivity raster layer. For example, a raster layer of known site or ER survey density could be classified into high, moderate, or low density and then combined with the sensitivity assessment to make new categories that incorporate this information. The resulting layer would be symbolized to show areas of high sensitivity that are also in areas of high recorded site density; this would be an area where the data used by the model was more representative. In areas where the sensitivity is high but the survey or site density is low, the model results may be taken more skeptically.

The Mn/Model addressed this problem in a similar way. For that project, a predictive model of survey locations was created in the same manner as for the archaeological sites. Essentially, the survey area model was an extrapolation of landscape conditions similar to those that have been surveyed in the past. The result of this was a layer that showed environments similar to and dissimilar to areas that have been surveyed. The theory is that areas that are similar to previously surveyed areas will have a better representation of known sites and therefore higher confidence predictions. This is an interesting approach that could be implemented in Pennsylvania as well.

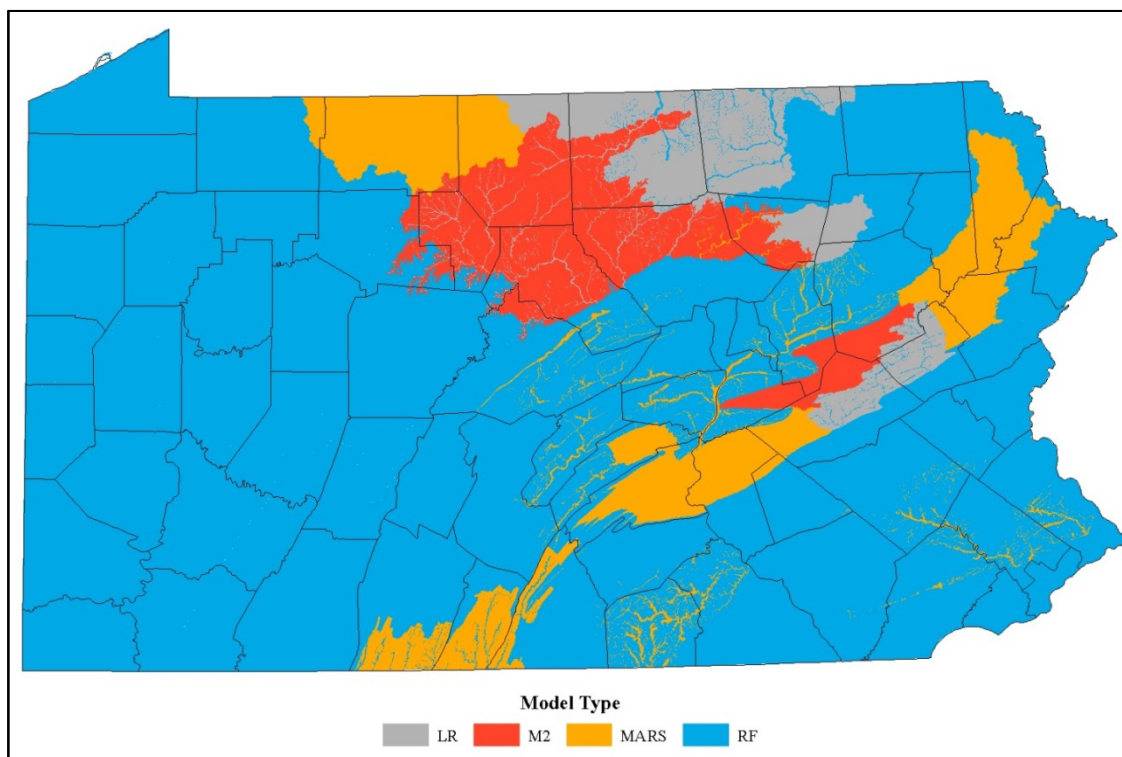


Figure 37 - Model type by subarea.

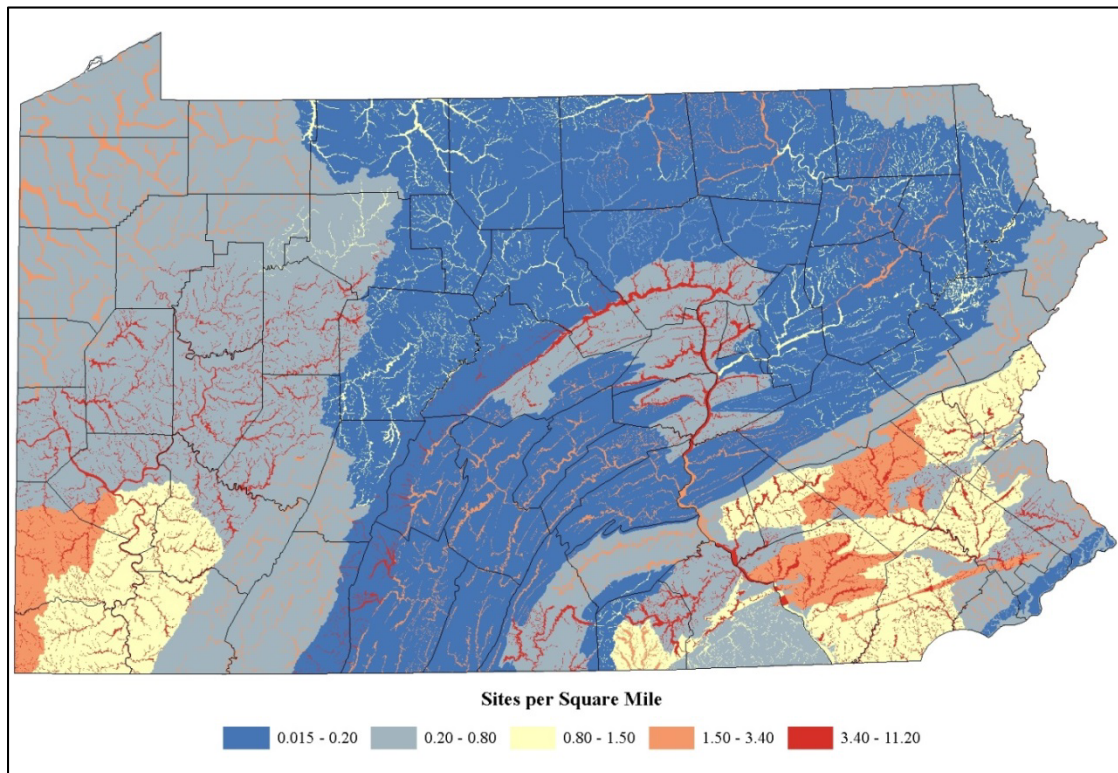


Figure 38 - Density of known archaeological sites per square mile within each subarea.

Urban Areas

The sensitivity assessment for Region 10, the entirety of Philadelphia County, is a low confidence model. Region 10 has a very low site density (0.118 sites per square mile), and the landscape has been heavily modified over the past 300 plus years. Attempting to model such an environment with the methods documented in the seven task reports introduces a whole different set of challenges beyond the rest of the state. The infilling of streams and wetlands, filling of low areas, cutting of hill tops, widespread development, and subsequent burial or destruction of site locations leads to a very unrepresentative sample of site locations and variable measurements. While Philadelphia is by far the most affected, the same can likely be said for urbanized areas such as Pittsburgh, Harrisburg, Allentown, Erie, Scranton, and so on.

The lack of apparent integrity from the view of small-scale environmental variables masks a documented truth of such settings: prehistoric sites exist in many locations and are often protected by historic-period fill. Recent work by PennDOT on the I-95 corridor has proven this (URS Corporation 2014). However, developing models for such an area requires very different methods than those employed here. A model of Philadelphia prehistoric sensitivity is a model of 300 years of land use focused on identifying the locations of remaining integrity. Researching historical documents for evidence of deep fill, shallow building construction, slab construction, stream valleys, and historic-period habitations can all contribute to successful identification of buried resources (Yamin et al.

2010). Currently, this sort of analysis only occurs on a local or site-specific scale, where historic maps and resources are available. On a larger scale, such a model could be developed using broader trends in urban development mixed with current street/alley layouts, major utilities, and historic maps. This would be a very worthwhile undertaking.

IMPROVEMENTS TO CURRENT METHODS

The methods employed in the current model are built to suit the project objectives within the constraints of time, processing power, and scalability. The numerous techniques used to work with or around these limitations are discussed throughout the task reports and are summarized here in Chapter 3. Nonetheless, there are a number of areas where improvements can be made to the current methods that may help increase the accuracy and utility of these assessments. Discussed below, many of these improvements will incur a cost relative to one or more of the previously mentioned constraints. However, that should not preclude their consideration or experimentation along these avenues.

Defining Spatial Structure

Issues associated with undefined spatial structure are potentially the most pressing that need to be addressed in this research. Spatial structure accounts for the way that site locations and variables are related and interact. This structure is multidimensional because site-likely cells are more related to neighboring cells than they are to cells in distant sites; neighboring sites are potentially related to one another on a cultural level as well as in relation to their environmental niche; environmental niches are more related to others in the same watershed; and so on to account for a vast scale of structured correlation. Many traditional and modern statistical modeling and inference methods are not designed to account for this structure as they typically are applied to controlled tests where data can be assumed to be stationary, homogeneous, and independent. Such is not the case for spatial data, a realization made 45 years ago in Cliff and Ord's (1969) paper, "The Problem of Spatial Autocorrelation."

Spatial autocorrelation can be defined as, "[a] measure of the degree to which a set of spatial features and their associated data values tend to be clustered together in space" (ESRI 2015). In the Pennsylvania model, this concept is manifested in the use of archaeological sites as the unit of analysis. Ideally, observations would be independent and identically distributed, but clearly archaeological sites are systematically related. Further, the autocorrelation is amplified through the use of individual site-present cells (as opposed to a single point representing a site) as our data points. These units are not only related on the larger systematic scale of archaeological sites, but also related as repeated observations from within a single archaeological site—that is, they are highly clustered. The effects of positive spatial autocorrelation on the models is the exaggerated influence of groups of sites and single large sites leading to a heightened estimate of probability near existing sites and on landforms with numerous or large sites.

A related issue is that of intraclass correlation. Described as the Intraclass Correlation Coefficient (ICC), this is a measure of how similar measurements are within groups. In the context of the present study, the ICC can calculate how similar the measures of each environmental variable are within sites. The greater the values of the ICC of a variable within an archaeological site, the greater the effect that variable has on the model of that site. While the Pennsylvania project does not model sites on an individual basis, an ICC that is high across many sites will have a large effect. For example, the median ICC for the distance to 3rd order or higher streams (*ed_h6*) calculated from 20 random samples of 500 individual site locations is 0.9994. The median ICC from the same sample for the topographic relief index from a 10-cell neighborhood (*tri_10c*) is 0.7749. However, with knowledge of how sites are distributed, especially riverine sites (which compose 44% of our site sample), the high correlation of distance to water values within a site should be no surprise. If a portion of the site is near water, then all of the site will be near water—and sites clearly have a positive relationship with proximity to water. The lower median ICC for the topographic variable is still relatively high, but shows that there is a greater degree of variability of this feature within the known site sample.

These issues arise from the fact that the structure and correlation of the spatial relationship between site-likely cells is not known to the statistical models. To address this issue, the spatial structure needs to be defined and accounted for or the data points need to be uncorrelated through sampling strategies. In the intervening years since Cliff and Ord (1969) drew attention to the “Problem” of spatial data, many methods have been proposed to identify and incorporate spatial structure into models, but there is still more to do. This is very much the case for the incorporation of spatial structure into archaeological site location modeling. The studies of fields such as geostatistics and spatial statistics have developed theoretical and methodological frameworks for addressing correlation, but differences between the types of data, analysis, and spatial structure commonly seen in those fields as compared to archaeology are nontrivial. Established methods such as Moran’s I, spatial lag, and variogram analysis all have a place in controlling for spatial structure in archaeological data, but are not a straight fit.

As discussed in Chapter 1, archaeological location data as measured for this study have some particular characteristics that set them apart from other spatial data, such as being nonmechanistic, discrete, heterogeneous in temporal and spatial scale, difficult to detect and measure, and derived from a biased sampling strategy. Adopting a new method of measurement such as continuous surfaces for site density or aggregating site counts into larger continuous polygons or quadrants could make them more amenable to existing spatial statistics methods, but there are tradeoffs associated with any such strategy. The solution to this problem likely lies with a mixed approach of resampling strategies, quantifying relationships, and spatially explicit models. Creating methods specific to the character of archaeological data will be challenging, but ultimately rewarding.

Variable Creation and Selection

The approach taken by this project to explanatory environmental variables is to define as many variations on environmental measures as practical, conduct univariate testing against numerous bootstrapped background samples, select approximately 30–40 variables with the best discriminatory power, remove highly correlated variables, and allow the models to use inherent variable selection techniques to choose the best subset. This is called a filtering method because the variables must pass through the univariate testing filter prior to use in the model.

Improvements on this approach would include creating additional variables to offer a larger range of potential correlations. This may include more compound variables describing landform morphology or measures of environmental attractiveness or richness; further elaborating on measures relating to soil data; and hydrologic models based on more accurate reconstructions of the past environment. Additionally, variable selection could be performed within a cross-validation framework to decrease variance. With the current univariate approach, the full body of site data is used to select variables that are then used within a CV framework to parameterize the models. In this way, the models will have “seen” or been influenced by the data that are in the hold-out sample by which they are validated (Hastie et al. 2009:345). This is a somewhat subtle point and more applicable to problems with higher ratios of variables to observations, but should be considered. The downside to using the cross-validation framework is escalated computational complexity and time. Methods such as Recursive Feature Elimination (RFE) (Kuhn and Johnson 2013) and the Boruta algorithm (Kursa and Rudnicki 2010) could be beneficial in achieving these goals. As opposed to the filter method described above, RFE and Boruta are called wrapper methods. These algorithms take a full set of variables and wrap them within a cross-validated model building sequence (commonly using RF as the base model), which results in a ranking of variables that are relevant to the classification problem.

Another approach that should be investigated is to identify a small set of variables that have widespread relevance to site locations in many different settings, as analyzed through the wrapper methods above. Restricting the model fitting to a small set of intuitive and proven relevant variables (such as distance to water) would likely lead to models with a higher bias error, but greater parsimony and interpretability across all geographies.

Model Parameterization

Chapter 3 discussed the details of the model parameterization sequence used for this project. In short, the method uses 10-fold CV to test a range of parameters and chooses the best parameters as the set that predicts the highest accuracy on the out-of-fold samples. This is a straightforward method in addressing parameter selection, but could be augmented with some additional steps to better fit the characteristics of our data. Possible additional steps include variable interaction terms and class imbalance solutions such as class weights and sampling strategies.

Variable interactions are simply the multiplicative effect of two or more variables on an outcome. This can be thought of as a synergistic effect in which the combination of two parts is greater than their sum. Interactions such as this can be included into statistical models including the LR and MARS models of the GLM family used in this study. In the parameterization of these models, no interaction terms were used, but given the complexity of variables derived from environmental settings, interactions exist. The costs of incorporating interaction terms are added model complexity, added processing time, and a potential decrease in interpretability if the interaction effects are not well understood. However, the benefits could lead to a reduction in bias. Interaction terms can be added to or searched for in LR and MARS, but RF can be used to identify interactions without additional steps.

Two common methods of dealing with highly imbalanced datasets are through weighting methods and via resampling techniques (Jeni et al. 2013). Class weighting and sampling strategies can be used in the model fitting sequence to address the highly imbalanced nature of this data set. As described in Chapter 1, given that archaeological sites make up only a very small percentage (0.22%) of the overall landscape, they have a very low prevalence (prevalence = 0.0019) within the state and even within previously surveyed areas (prevalence = 0.01). This leads to a situation where if 10×10 -m cells are selected randomly from within the state, on average a total of over 37,000 cells would be required to sample enough site-present cells to cover an average site (approximately 2-acres or 81-cells). This is an imbalance of 469 to 1. The effect of this imbalance is that the positive case of site-presence is very much in the minority and can easily be out-weighted by statistical methods; some of these implications are discussed in the Task 4 report (p. 79).

Two approaches to addressing imbalance through weighting include case weights and cost-sensitive training. Case weights can instruct the algorithm that the penalty for misclassifying a positive class observation (i.e., site-present cell) is more costly than misclassifying a negative class observation (i.e., a background cell). Cost-sensitive training is a method to alert the training algorithm that misclassification of the smaller positive class is more egregious than a misclassification of the larger negative class. Useful in partitioning methods such as RF, cost-sensitivity can set a lower threshold for the evidence required to assign an observation to a site as opposed to a higher threshold for the background. Setting specified case weights for the positive and negative classes in the algorithm and cost-sensitive approaches can produce better performance if correctly specified. Each of these approaches acknowledge that sites and background cells are not on equal terms and that sites are clearly more important than background cells in terms of correct classification. Experimentation with case weighting and cost-sensitivity was undertaken for the Pennsylvania model, but not fully implemented. Issues such as identifying proper weights and the ability to apply the costs to the raster prediction process hampered these efforts; revisiting this topic will likely help to create more generalizable models.

The other common method to address class imbalance is by resampling. This can be achieved through down-sampling the more numerous negative class or up-sampling the minority positive

class. Down-sampling is used to reduce the number of negative class observations through random sampling. Up-sampling is used to increase the number of the positive class observations through randomization or simulation. For the Pennsylvania model, the negative class was down-sampled to a ratio of 3:1 background cells to site cells. This ratio was an arbitrary choice, but through experimentation did appear to have a positive effect without overestimating the effects of small site prevalence. This down-sampling combined with bootstrap resampling and prevalence sensitive thresholds are the methods by which class imbalance was addressed in this project. Another form of resampling that should be investigated for future models is called Synthetic Minority Over-Sampling Technique or SMOTE (Chawala et al. 2002). Briefly, SMOTE sampling conducts a blend of both up- and down-sampling to achieve a more balanced dataset. The down-sampling is done through random sampling and the up-sampling is done through synthesizing new site-likely observations. The synthesis of new observations initially picks a single actual site-likely observation, picks a number of closely related observations, and then makes new observations by randomizing their values. This method should be experimented with to address class imbalance, but also to potentially address some issues related to correlation through the randomization of variable measures.

Model Averaging

The modeling process used throughout the Pennsylvania project seeks to find the “best” model to represent the archaeological site location potential of a given subarea. This is accomplished by creating numerous models through parameterization procedures and generating testing error based on CV and hold-out samples. Finally a model is chosen based on these metrics and a subjective evaluation of how the model conforms to the landscape. As discussed throughout, the different model types have difference characteristic that need to be balanced when choosing the representative model, but in the end only one prevails. However, the method of model averaging can be used to incorporate the pros and cons of each model style into a single model.

A simple form of model averaging was carried out in this project, but not introduced into the results of any task reports. This process took the results of the LR, MARS, and RF models to make two additional models: one that combined LR and RF and another that combined MARS and RF. The combination of the models was done by overlaying them and retaining the highest level of sensitivity for each cell. This is similar to the aggregation function discussed above, but the cell size does not change. The results of the model combination were interesting and seemed to pick up characteristics of both models, but interpreting the results of the individual models became unclear in the combined context. For this reason, these models were not presented in the final reports.

Additional and more sophisticated methods exist for model averaging including a method referred to as Bayesian Model Averaging (BMA). On a basic level it works on the same principle as the simple model combination described above, as well as a similar principle to the bagging method in RF: that is, to reduce variance and incorporate uncertainty by estimating the actual distribution of site sensitivity from multiple points of view (in this case models). The use of BMA for LR was

recommended for the Mn/Model by Oehlert and Shea (2007). As a general description, BMA uses the output of a number of models and blends them with a model weight based on uncertainty to produce a new model composed of the original models and the uncertainty. The choice of model weights, referred to as the model's prior probability, can be achieved in different ways to suit the purpose. A significant benefit of this approach is that not only is uncertainty accounted for to the degree it can be represented in the prior weights, but the BMA method can include models of many different types. In such a way BMA can incorporate the results of data-driven models, such as those produced here, with more deductive models that focus on hypothetical settlement mechanics and not just where we have found sites. A number of models both inductive and deductive can be appropriately weighted and blended in this way. This may not only be a beneficial approach to understanding uncertainty in the existing models, but also a way to broaden them for a more theoretically informed point of view.

Testing and Reiteration

Finally, these models should be tested with new field data and reiterated to reflect new methods and understandings. Building from a foundational understanding of past attempts to model archaeological site location sensitivity, the Pennsylvania model set project applied modern statistical techniques and algorithms, many of which have never been published in the archaeological literature, to derive an understanding of archaeological sensitivity on a statewide scale. Throughout this process best-practices were used to address bias, and a number of metrics were used to assess model fit and validity. The results of this transparent, well-documented, and reproducible process are a series of sensitivity raster layers that are an accurate representation of the pattern presented in known archaeological site locations. As presented in Chapter 4, comparisons to existing models applied to Pennsylvania show this model to have achieved a good balance between model accuracy and precision.

As with any model, prediction, or projection, there is error and there will be room for improvement. Given that this is a first attempt to model these data on this scale, there may be significant room for improvement. The information contained in the task reports of this project contains extensive documentation on error rates, performance metrics, and assumptions; however, it is through use in project planning and documentation of new sites that the true value of this model will become apparent. The identification of the strengths and weakness of these models through use and testing can be incorporated back into the model to correct blind-spots and add utility. Like the Mn/Model project, through incorporation of the suggestions above and continued iteration, the models can be adapted to take advantage of new findings and techniques, and can thereby grow in utility. The process to this point should only be the beginning of the Pennsylvania model's life, which only through continued improvement and evolution will best serve the needs of cultural resources planning and protection.

6

REFERENCES CITED

Akaike, Hirotugu

1974 A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 19(6):716–723.

Breiman, Leo

1996a Bagging Predictors. *Machine Learning* 26:123–140.

1996b Out-of-Bag Estimation. Technical Paper, Statistics Department, University of California Berkeley, Berkeley, CA.

Breiman, Leo

2001 Random Forests. *Machine Learning* 45(1):5–32.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and Richard A. Olshen

1984 *Classification and Regression Trees*. CRC press, London, England.

Cavallo, John A.

1987 Area B (28ME1-B), Archaeological Data Recovery, I-295, and Wetlands Area Interchange. Report Prepared for the Federal Highway Administration and New Jersey Department of Transportation, Bureau of Environmental Analysis. The Cultural Resource Group, Louis Berger and Associates, Inc., East Orange, NJ.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer

2002 SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16:321–357.

Coe, Joffre L.

1964 *The Formative Cultures of the Carolina Piedmont*. Transactions of the American Philosophical Society, New Series, Volume 54, Part 5. Philadelphia, PA.

Conover, W. J.

1999 *Practical Nonparametric Statistics*. 3rd ed. Wiley, New York, NY.

Custer, Jay F.

1989 *Prehistoric Cultures of the Delmarva Peninsula: An Archaeological Study*. University of Delaware Press, Newark.

- 1996 *Prehistoric Cultures of Eastern Pennsylvania*. Anthropological Series No. 7, Pennsylvania Historical and Museum Commission, Harrisburg.
- Efron, B., and R. Tibshirani
- 1993 *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Chapman and Hall, London, England.
- 1997 Improvements on Cross-Validation: The .632 + Bootstrap Method. *Journal of the American Statistical Association* 92(438):548–560.
- ESRI
- 2015 GIS Dictionary, <<http://support.esri.com>>.
- Fawcett, Tom
- 2004 ROC Graphs: Notes and Practical Considerations for Researchers. *Pattern Recognition Letters* 27(8):882–891.
- 2006 An Introduction to ROC Analysis. *Pattern Recognition Letters* 27(2006):861–874.
- Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim.
- 2014 Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *The Journal of Machine Learning Research* 15(1):3133–3181.
- Friedman, J. H.
- 1991 Multivariate Adaptive Regression Splines. *The Annals of Statistics* 19:1.
- Freund, Yoav, and Robert E. Schapire
- 1997 A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55(1):119–139.
- Gardner, William M.
- 1974 The Flint Run Paleo-Indian Complex: Pattern and Process during the Paleo-Indian to Early Archaic. In *The Flint Run Paleo-Indian Complex: A Preliminary Report, 1971-73 Seasons*, edited by W.M. Gardner. Occasional Publication No.1, Catholic University Archaeology Laboratory, Washington, D.C.
- Gardner, William M.
- 1982 Early and Middle Woodland in the Middle Atlantic: An Overview. Paper presented at the 1982 Middle Atlantic Conference, Rehoboth Beach, DE.

Harris, Matthew D.

2013a Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 1: Literature Review. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, NJ.

2013b Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 2: Designating Modeling Regions. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, NJ.

2014 Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 3: Pilot Model Study. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, NJ.

Harris, Matthew D., Susan Landis, and Andrew R. Sewell

2014a Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 4: Study Regions 1, 2, and 3. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, NJ.

2014b Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 5: Study Regions 4, 5, and 6. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, NJ.

2014c Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 6: Study Regions 7, 8, 9, and 10. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, NJ.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, T. Hastie, J. Friedman, and R. Tibshirani.

2009 *The Elements of Statistical Learning*. Vol. 2, no. 1. Springer, New York, NY.

Heilen, Michael

2013 Modeling of Archaeological Site Location in Southern New Mexico. *Newsletter of the New Mexico Archeological Council* 1.

Herbstritt, James T.

1980 *Prehistoric Archaeological Site Survey: Pennsylvania Region II, Southwestern Pennsylvania*. Prepared for the Pennsylvania Historical and Museum Commission, Harrisburg. California State College, California, PA.

Hummer, Christopher

1994 Defining Early Woodland in the Delaware Valley: The View from the Williamson Site, Hunterdon County, New Jersey. In *Recent Research into the Prehistory of the Delaware*

- Valley, edited by C.A. Bergman and J.F. Doershuk. *Journal of Middle Atlantic Archaeology* 10:141–151.
- James, Gareth, Witten, Daniela, Hastie, Trevor, and Robert Tibshirani
2014 *An Introduction to Statistical Learning*. Springer, New York, NY.
- Judge, James W., and L. Sebastian (editors)
1988 *Quantifying the Present and Predicting the Past: Theory, Method and Application of Archaeological Predictive Modeling*. U.S. Department of the Interior, Bureau of Land Management, Denver, CO.
- Jeni, Laszlo, Jeffrey F. Cohn, and Fernando De la Torre
2013 *Facing Imbalanced Data Recommendations for the Use of Performance Metrics*. Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland.
- Kamermans, Hans
2008 Smashing the Crystal Ball. A Critical Evaluation of the Dutch National Archaeological Predictive Model (IKAW). *International Journal of Humanities and Arts Computing* 1(1):71–84.
- Kamermans H., M. van Leusen, and P. Verhagen (editors)
2009 *Archaeological Prediction and Risk Management. Alternatives to Current Practice*. Leiden University Press, Leiden, The Netherlands.
- Kelly, Cassidy, and Kazuroi Okada
2012 Variable Interaction Measures with Random Forest Classifiers. *29th IEEE International Symposium on Biomedical Imaging (ISBI), Proceedings* 2012:154–157.
- Kinsey, Fred W.
1977 Patterning in the Piedmont Late Archaic: A Preliminary View. *Annals of the New York Academy of Sciences* 28:375–391.
- Kohler, Tim A., and Sander van der Leeuw
2007 *Model-Based Archaeology of Socionatural Systems*. SAR Press, Santa Fe, NM.
- Kuhn, Max, and Kjell Johnson
2013 *Applied Predictive Modeling*. Springer, New York, NY.
- Kursa, Miron B., and Witold R. Rudnicki
2010 Feature Selection with the Boruta Package. *Journal of Statistical Software* 36(11).

Kvamme, Kenneth L.

- 1988 Development and Testing of Quantitative Models. In *Quantifying the Present and Predicting the Past*, edited by W. Judge and L. Sebastian, pp. 325–428. U.S. Government Printing Office, Washington, D.C.

Lehmann, Erich L

- 1986 *Testing Statistical Hypothesis*. 2nd ed. Wiley, New York, NY.

Lehmann, Erich Leo, and George Casella

- 1998 *Theory of Point Estimation*. 2nd ed. Springer-Verlag, New York, NY.

Liaw, Andy, and Matthew Wiener

- 2002 Classification and Regression by randomForest. *R News* 2(3):18–22.

Madsen, Henrik, and Poul Thyregod

- 2011 *Introduction to General and Generalized Linear Models*. CRC Press, London, England.

Märker, Michael, and Saman Heydari-Guran

- 2009 Application of Datamining Technologies to Predict Paleolithicsite Locations in the Zagros Mountains of Iran. Paper presented at Computer Applications to Archaeology, Williamsburg, VA.

Mason, Ronald I.

- 1962 The Paleo-Indian Tradition in Eastern North America. *Current Anthropology* 3(3):277-283.

McNett, Charles W., Jr. (editor)

- 1985 *Shawnee-Minisink: A Stratified Paleoindian-Archaic Site in the Upper Delaware Valley of Pennsylvania*. Academic Press, Orlando, FL.

Menze Bjoern H., and Jason A. Ur

- 2013 Multi-Temporal Classification of Multi-Spectral Images for Settlement Survey in Northeastern Syria. In *Mapping Archaeological Landscapes from Space*, edited by Douglas C. Comer and Michael J. Harrower, pp. 219–228. Springer-Verlag, New York, NY.

Milborrow, Stephen

- 2011 earth: Multivariate Adaptive Regression Spline Models R package: R package.
Electronic document: <<http://CRAN.R-project.org/package=earth>>.

- 2014 Notes on the Earth Package. Electronic document: <<http://cran.r-project.org/web/packages/earth/vignettes/earth-notes.pdf>>.

Oehlert, Gary W., and Brian Shea

2007 Statistical Methods for Mn/Model Phase 4. Research Services Section of Minnesota
Department of Transportation, St. Paul.

Pampel, F. C. (editor)

2000 *Logistic Regression: A Primer*. Sage, Thousand Oaks, CA.

Salkind, Neil J. (editor)

2007 *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA.

Stewart, R. Michael

1998 Ceramics and Delaware Valley Prehistory: Insights from the Abbott Farm. Trenton Complex
Archeology, Report 14. Report submitted to the Federal Highway Administration and the
New Jersey Department of Transportation. Louis Berger & Associates, Inc., East Orange, NJ.

Stewart, R. Michael

2003 A Regional Perspective on Early and Middle Woodland Prehistory in Pennsylvania. In
Foragers and Farmers of Early and Middle Woodland Periods in Pennsylvania, edited by
Paul A. Raber and Verna L. Cowin, pp. 1–33. Recent Research in Pennsylvania Archaeology
No. 3., Pennsylvania Historical and Museum Commission, Harrisburg.

URS Corporation

2014 Digging I-95: The Archaeology of Northern Liberties, Kensington-Fishtown, and Port
Richmond. Report prepared for the Pennsylvania Department of Transportation and the
Federal Highway Administration. URS Corporation, Burlington, NJ.

van Leusen, P.M., and Hans Kamermans (editors)

2005 *Predictive Modelling for Archaeological Heritage Management: A Research Agenda*.
Nederlands Archeologische Rapporten 29. Rijksdienst voor het Oudheidkundig
Bodemonderzoek, Amersfoort.

Verhagen, Phillip

2007 *Case Studies in Archaeological Predictive Modelling*. Leiden University Press, Leiden, the
Netherlands.

Verhagen, Philip

2009 Testing Archaeological Predictive Models: A Rough Guide. In *Archaeological Prediction
and Risk Management. Alternatives to Current Practice*, edited by H. Kamermans, M. van
Leusen, and Ph. Verhagen, pp. 63–70. Archaeological Studies Leiden University 17. Leiden
University Press, The Netherlands.

Viera, Anthony J., and Joanne M. Garrett

2005 Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine* 37(5):360–363.

Wallace, Paul A.W.

1965 *Indian Paths of Pennsylvania*. Pennsylvania Historical and Museum Commission, Harrisburg.

Wasserman, Larry

2000 Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology* 44(1): 92–107.

Wheatley, David, and Mark Gillings

2002 *Spatial Technology and Archaeology. The Archaeological Application of GIS*. Taylor & Francis, London, England.

Witthoft, John

1953 Broad Spearpoints and the Transitional Period Cultures. *Pennsylvania Archaeologist* 23(1):4–31.

Wolpert, David

1996 The Lack of A Priori Distinctions between Learning Algorithms. *Neural Computation* 1996: 1341–1390.

Yamin, Rebecca, Matthew D. Harris, Douglas C. McVarish, and Grace H. Ziesing

2010 Independence National Historical Park Archaeological Sensitivity Study (Phase IA Archeological Assessment, Independent Living History Center, North Lot). Prepared for Independence National Historical Park, Philadelphia, Pennsylvania. John Milner Associates, Philadelphia, PA.

APPENDIX A

COMPREHENSIVE LIST OF ACRONYMS AND GLOSSARY OF TERMS

ACRONYMS/ABBREVIATIONS

AIC	Akaike Information Criterion
AMS	Amazon Web Services
ANOVA	Analysis of Variance
APM	Archaeological Predictive Modeling
AUC	Area Under Curve
AUROC	Area Under Receiver Operating Characteristics Curve
CART	Classification and Regression Trees
BIC	Bayesian Information Criterion
BMA	Bayesian Model Averaging
CoV	Coefficient of Variation
CRGIS	Cultural Resources Geographic Information System
CV	Cross-Validation
DEM	Digital Elevation Model
ECDF	Empirical Cumulative Distribution Function
EDA	Exploratory Data Analysis
FNR	False Negative Rate
FPR	False Positive Rate
GCV	Generalized Cross-Validation
GIS	Geographic Information Systems
GLM	Generalized Linear Model
GRSQ	Generalized R-Square
ICC	Intraclass Correlation Coefficient
Kg	Kvamme Gain
K-S	Kolmogorov–Smirnov
LR	Logistic Regression
MARS	Multivariate Adaptive Regression Splines
MLE	Maximum Likelihood Estimate
MnDOT	Minnesota Department of Transportation
Mn/Model	Minnesota Statewide APM

MSE	Mean Squared Error
MW	Mann-Whitney
NPG	Negative Prediction Gain
NPV	Negative Prediction Value
NWI	National Wetland Inventory
OLS	Ordinary Least Squares
OOB	Out-of-Bag Sample
OOS	Out-of-Sample
PASS	Pennsylvania Archaeological Site Survey
PPG	Positive Predictive Gain
PPV	Positive Prediction Value
R ²	R-Squared
RF	Random Forests/randomForest
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristics
RSS	Residual Sum-of-Squares
SD	Standard Deviation
SMOTE	Synthetic Minority Over-Sampling Technique
SSE	Sum of Squares Error
TNR	True-Negative Rate
TPR	True-Positive Rate
UDR	Unexpected Discovery Rate
USDA	United States Department of Agriculture

TERMS

first use:

- Accuracy (in error estimates for MARS and RF models) Task 5, p. 59
The measurement of accuracy is used in many classification methods. This measure is simply the percent of observations (site-present or site-absent) that are correctly classified by the algorithm. As used in this report, the accuracy is the percentage of observations from the out-of-bag sample that were correctly classified by the model. This is an internal metric that assess the model's ability to correctly predict data that were not used in the fitting of the model.
- Adaptive Regression Splines (see Multivariate Adaptive Regression Splines) Task 3, p. 16
- Akaike Information Criterion (AIC) Task 3, p. 62
A measure of relative model quality that balances goodness of fit and model complexity. This measure is used in model selection to choose the model that has the best fit relative to complexity for a given data set. Within a series of nested candidate models, the one with the lowest AIC will likely represent the model with the best goodness of fit without being over-fit or over-parameterized (see Akaike 1974).
- Analysis of Variance (ANOVA)..... Task 3, p. 47
ANOVA is a suite of statistical models used to test the difference in variation between groups. In linear model creation, ANOVA can be used to estimate the variance explained by each variable or whether there is a significant difference in variance explained by each model (see Freedman 2005).
- Archaeological Predictive Modeling (APM) Task 1, p. 1
The field of study concerning the use of existing archaeological data or theory to predict the sensitivity of locations for the presence of archaeological material.
- Area Under Curve (AUC) (see also Receiver Operating Characteristics) Task 3, p. 21
Also referred to as Area Under Receiver Operating Characteristics Curve (AUROC), AUC is a measure of the balance between a model's Sensitivity and Specificity across the full range of cut-off points. The AUC is a single measure that captures a model's ability to balance True Positive Rate and False Positive Rate across the full range of the model's output. The higher the AUC, the higher the Sensitivity and Specificity across the full range of the model, and the more likely the model is to correctly classify a randomly chosen positive instance. AUC is used in model selection to assess a model's ability to correctly classify observations (see Fawcett 2006).

- Bagging (see Bootstrap Aggregating)..... Task 3, p. 18
- Bayesian Information Criterion (BIC) Task 7, p. 25
BIC is a measure of relative model quality that balances goodness of fit and model complexity, similar to the Akaike Information Criteria (AIC). This measure is used in model selection to choose the model that has the best fit relative to complexity for a given data set. Within a series of nested candidate models, the one with the lowest BIC will likely represent the model with the best goodness of fit without being over-fit or over-parameterized. BIC penalizes model complexity more than AIC for models with greater than seven predictors (James et al. 2014).
- Bayesian Model Averaging (BMA)..... Task 7, p. 100
BMA is an ensemble technique that uses random sampling and Bayes theorem to account for uncertainty in the model selection process. While there are a number of ways to apply BMA, the general principle is to randomly sample the model space of numerous models to derive a posterior probability based on all models, weighted by some criteria. The end result attempts to decrease the risk of choosing a single over-fit model, but comes at the cost of computational complexity (see Wasserman 2000 [not in refs]).
- Boosting Task 3, p. 20
Boosting is a term that defines a family of statistical regression and classification methods that use random subset selection and weighting to minimize variance and lessen the potential for over-fitting. The concept of boosting is used in a number of regression and classification models with the general commonality of providing a means to achieve an ensemble of models. The final model is often selected through the weighted average of sub-models (see Freund and Schapire 1997).
- Bootstrap Aggregating Task 3, p. 18
Bootstrap Aggregating (or Bagging) is a term that defines a method of statistical regression and classification often applied to tree-based machine learning algorithms. Simply, bagging uses the regression or classification of numerous bootstrapped samples to create an ensemble. Taking the average output of this ensemble generally reduces model variance and lessens the potential for over-fitting (Breiman 1996a).
- Bootstrapping Task 3, p. 14
Bootstrapping is a statistical method of resampling that draws numerous samples from a sample or population with replacement. This means that each time a sample is chosen, its value is returned to the sampling population so that it may be drawn again. Bootstrapping offers a method of estimating population parameters from small samples or complicated distributions (see Efron and Tibshirani 1993).

Classification and Regression Trees (CART).....	Task 3, p. 18
<p>CART is a statistical learning algorithm. In a simple form, the CART is used to classify training observations based on the nested splitting of input variables. Called nodes, the split point of each variable creates a branch-like structure that begins with all of the training observations at the base of the tree and ends with the classification of each training observation at the tips of each branch. A predictive model can be drawn from the ability of the tree to correctly classify training and test observations based on the splits in each variable. The general structure of the CART is used as the basis for a number of algorithms such as Bagging, Boosting, and Random Forest (see Breiman et al. (1984).</p>	
Confusion Matrix.....	Task 3, p. 35
<p>A classification table in the form of a 2-cell \times 2-cell contingency table that shows how many sites were correctly predicted as sites and how much of the non-site area was correctly predicted as such. This method is frequently used as a means to assess the ability of a model to classify observations (see Fawcett 2006).</p>	
Cost Variable	Task 3, p. 7
<p>A Cost Variable is a predictive variable derived through a cost analysis. The cost associated with a cost variable may be anything that is thought to introduce a difficulty or impediment to movement. For example, the linear distance for any point to the nearest stream only considers the straight line distance between those two points. A cost distance to the nearest stream will consider an impediment or set of impediments between any given point and the nearest stream. If crossing a wetland is considered costly, the least cost path from a given point may not be the shortest linear path, but may be a non-linear path that avoids traveling over wetlands.</p>	
Coefficient of Variation (CoV).....	Task 3, p. 65
<p>The CoV is a statistic that measures the normalized dispersion within a frequency distribution. The acronym CoV is used in this study to avoid confusion with the acronym used for Cross-Validation (CV). The CoV is calculated as the ratio of the standard deviation to the mean and is also referred to as Relative Standard Deviation (RSD). The CoV represents the percentage of standard deviation from the sample mean (see Lehmann 1986).</p>	
Cohen's Kappa Coefficient (see Kappa).....	Task 4, p. 61
Cross-Validation (CV) (see Generalized Cross Validation and K-folds Cross-Validation)	Task 3, p. 14

Cultural Resources Geographic Information System (CRGIS)	Task 2, p. 14
Computerized database and mapping tool for the visualization and analysis of cultural resources data within the Commonwealth of Pennsylvania. This tool is developed and administered through a join agreement between the Pennsylvania Historical and Museum Commission and the Pennsylvania Department of Transportation. (This tool is available at: www.portal.state.pa.us/portal/server.pt/community/crgis/3802 .)	
Digital Elevation Model (DEM)	Task 1, p. 8
A digital elevation model is a computer based representation of the topography at earth's surface. DEMs are stored as a raster format composed of square cells representing a single elevation measure for a given resolution. DEMs are available in a range of resolutions and are created and curated by the United States Geologic Survey. (Information and data sets are available at: http://ned.usgs.gov .)	
Earth (see also Multivariate Adaptive Regression Splines)	Task 3, p. 16
Earth is an implementation of the Multivariate Adaptive Regression Splines algorithm written in the R Statistical Language (see Milborrow 2011).	
Empirical Cumulative Distribution Function (ECDF)	Task 7, p. 36
A cumulative distribution function (CDF) is a statistical function that estimates the probability of a random variable or value being equal to or less than a point in a given distribution. The term "empirical" signifies that the CDF is derived from real data and not a hypothetical distribution. The term ECDF is used in the context of this report to describe the way in which the K-S test compares two sequences of numbers.	
Euclidian Distance	Task 3, p. 7
The simple, or straight-line, distance between two points, colloquially described as "as the crow flies."	
Exploratory Data Analysis (EDA)	Task 7, p. 21
Exploratory Data Analysis is an approach to modeling and understanding data that uses many techniques to visualize different dimensions of the data outside of the formal modeling process. EDA is often an early step in the modeling process to better understand the data and identify preliminary patterns, bias, and distributions.	
Factor & Factor Level	Task 5, p. 51
A factor is the data type used by the R statistical language to code data that are categorical (nominal), as opposed to quantitative data such as continuous integers. The factor data type is composed of the qualitative categories represented as levels (e.g., "high," "moderate," "low") and a string of integers to represent the categories (e.g., 1, 2, 3). The categorical data are actually stored as a string of representative integers, but referenced back to the levels so that	

the data can be converted to its original category when needed. Among other reasons, this allows the program to work very efficiently with integers as opposed to storing and computing a long list of category labels.

False Negative Rate (FNR) Task 3, p. 70

The fraction of the positive observation (site locations) that are incorrectly classified as a negative observation (site not-likely). The FNR is derived from the Confusion Matrix and calculated by dividing the number of false negatives by total number of observed positive observations. This number is also interpreted as the Type-II error rate, or beta (β).

False Positive Rate (FPR) Task 3, p. 36

The fraction of the negative observations (background locations) that are incorrectly classified as a positive observations (site likely). The FPR is derived from the Confusion Matrix and calculated by dividing the number of false positives by total number of observed negative observations. This number is also interpreted as the Type-I error rate.

Generalized Cross-Validation (GCV) Task 3, p. 18

GCV is a statistical method that estimates performance or prediction error from within a model based on weight assigned to model complexity. GCV approximates the measure of performance that would be derived through leave-one-out Cross-Validation. In this project, the GCV relates to the internal performance measure derived from the Multivariate Adaptive Regression Splines model (see Milborrow 2014).

In more formal terms, the cross-validation (CV) prediction error can be defined by the equation below. Where \hat{f} (pronounced f-hat) is the approximated function of data set x , what we call the model. The term $L(y_i, \hat{f}^{-k(i)}(x_i))$ is the error or loss function that defines the error—RMSE in our case. The term $\hat{f}^{-k(i)}(x_i)$ represents the model fit to the remaining training data x_i minus the i^{th} k -fold. So this reads that the CV error of the model is the average error computed over each of i iterations of the CV given the loss function of the left-out fold (y_i) and the model fit to the data of the remaining folds ($\hat{f}^{-k(i)}(x_i)$)

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}^{-k(i)}(x_i))$$

A slight alteration as used in the method of parameterization is the addition of the selected parameter value (α) to the outcome and loss function.

$$CV(\hat{f}, \alpha) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}^{-k(i)}(x_i, \alpha))$$

Generalized Linear Models (GLM) Task 3, p. 18

GLMs are a family of models that extend linear regression by allowing for error distributions other than the normal distribution. This is achieved by using link functions to relate the response variable to the appropriate error distribution. Logistic Regression and Multivariate Adaptive Regression Splines are examples of GLM regression (see Madsen and Thyregod 2011).

Generalized R-Square (GRSQ)..... Task 3, p. 53

This metric is used in the Multivariate Adaptive Regression Splines model to normalize Generalized Cross-Validation and estimate a model's R-Squared when predicting for independent data. This measure is a ratio of 1 – GCV divided by the GCV of the Null Model or intercept only model (see Milborrow 2014).

Geographic Information Systems (GIS) Task 1, p. 4

A GIS is a computer application that stores, manages, displays, and manipulates information with a spatial component (see Wheatley and Gillings 2002).

Gini Importance Criterion or Gini Impurity Task 4, p. 81

The Gini Importance criterion is a metric used within the random forest algorithm for both branch splitting and variable importance. For the former, the Gini criterion is used to measure the purity, or how well segregated the representation of sites versus background values, of the node following its split on one of p variables. The split is made using the variable the leads to the largest increase in node purity from the parent node to the two descendent nodes. For the latter, the Gini Importance criterion is used to assess the value of each variable's contribution to the model. For each instance that a variable is chosen to split a node, the decrease in Gini is added up and compared to the other variables. Those variables that contributed to a greater decrease in the Gini are considered to be more important to the model's ability to correctly classify (see Breiman 2001).

Intraclass Correlation Coefficient (ICC)..... Task 7, p. 97

The ICC is a measure of how similar measurements are between groups. In the context of the Pennsylvania model project, the ICC is used to describe the relationship between the measurement of a given environmental variable between known site locations. A high ICC

would indicate that the variable is strongly correlated to individual site locations and therefore may lead to difficulties in model building based on site-present cells. A low ICC for a variable would indicate a lower correlation between site locations and that variable, which is preferable.

K-folds Cross-Validation (CV)..... Task 3, p. 14

Cross-Validation is the method by which a sample of observations is split into a number of different but equal-sized classes. The number of classes is referred to as K and the classes themselves are referred to as folds, hence “K-folds Cross-Validation.” This is a method by which models can be validated on test sets that were not part of the training set, while at the same time, using the entire data set for modeling (see Efron and Tibshirani 1997).

Kappa coefficient..... Task 4, p. 61

The Kappa coefficient, or Cohen’s Kappa coefficient, is a statistical measure of a predictions agreement with real observations after accounting for chance agreement. In this project, the Kappa is used in a similar fashion as the Kvamme Gain statistic. However, the Kappa’s calculation of by-chance observation is more inclusive than the Kvamme Gain. The Kappa statistic is derived from the confusion matrix and is used to compare model results of similar prevalence (see Viera and Garrett 2005).

Kolmogorov–Smirnov (K-S) Test Task 3, p. 8

A non-parametric statistical test that measures the equality of continuous unpaired probability distributions to each other (two-sample test) or a reference distribution (one-sample test). In this study, the K-S test is used to test whether the distribution of an environmental variable is significantly different between known site locations and the overall environmental background (see Conover 1999).

Kvamme Gain (Kg)..... Task 1, p. 27

The Kg is a metric used to assess the ability of a model to correctly classify positive observations (site present) given the area in which positive observations are predicted to occur (site-likely area). The higher the gain, the greater the ratio of percent sites present to percent of the modeled area considered site-likely. This measure does not take into account model precision or True Positive Rate (Sensitivity), meaning that an equivalent Kg statistic can be reached by correctly predicting 16% of known sites in 5% of the area or 95% of known sites in 30% of the area (see Kvamme 1988).

Likelihood Ratio Test Task 3, p. 50

This is a statistical test used to compare the fit of two models. In this project, the Likelihood Ratio Test is used to compare Logistic Regression models by testing the likelihood ratios of the Null Model and alternative models. This test uses a p-value to accept or reject the null model based on the likelihood ratio.

- Logistic Regression (LR)..... Task 1, p. 17
Logistic Regression is a statistical model used to predict for a binary response (0 or 1) or to classify a categorical response (“dead” or “alive”) based on one or more predictors. This method uses a S-shaped logistic transformation to model the binary response probability as the log odds of the linear function of the predictor variables. Simply, the model fits the linear model to the S-shaped curve so that the prediction is kept between 0 and 1 (see Pampel 2000).
- Mann-Whitney (MW) U Test Task 3, p. 8
The Mann-Whitney U Test is a non-parametric statistical test that evaluates the dissimilarity of unpaired distributions by ranking the observations and comparing the mean ranks. This test is similar in concept to the Kolmogorov–Smirnov Test, but uses a ranked approach as opposed to a distance approach. The MW U Test is more sensitive to changes in the median of two distributions (see Lehman 1975).
- Maximum Likelihood Estimate (MLE) Task 3, p. 50
The MLE is a statistical procedure used to estimate parameters within Logistic Regression. This function uses an iterative approach to identify a set of parameters for which the probability of the observed data is the greatest (see Pampel 2000).
- Mean Squared Error (MSE)..... Task 3, p. 56
The MSE is a statistic, or loss function, used to quantify the difference between an estimate and a true value. In this project, the MSE is used to quantify the difference between the predicted values (\hat{y}) and the observed test values (y). MSE is calculated as the Sum of Squared Errors divided by the number of observations (see Lehman and Casella 1998).
- Model Formula Task 5, p. 51
As used in this project, the model formula is a symbolic representation of the *a priori* relationship between the model predictors ($x_1, x_2, x_3, \dots, x_n$) and the outcome (y). Typically, the tilde symbol (\sim) is used to specify that the response is a function of one or more predictors. For example, the formula ($y \sim x_1$) specifies to the statistical model that y as the response variable is a function of the linear predictor x_1 . Further, the formula symbols specify the relationship between the predictor variables. For example, the formula ($y \sim x_1 + x_2 + x_3$) specifies that y is an additive function of the linear predictors x_1, x_2 , and x_3 . Additional symbols can be used in the formula to represent interactions between predictors, non-additive relationship, and polynomials. However, this project uses only linear and additive formulae.

<i>mtry</i>	Task 3, p. 54
<p>This is the name of a key parameter in the RF model. One of the key features of RF is the random selection of a subset of the predictor variables to test at each node in the tree building process. The number of randomly selected variables to try is called “<i>mtry</i>.. By default, <i>mtry</i> is set to \sqrt{p} for classification problems and $p/3$ in regression problems. In this project, <i>mtry</i> is optimized through cross-validation to the lowest error rate of the out-of-fold sample.</p>	
Multivariate Adaptive Regression Splines (MARS)	Task 3, p. 14
<p>A statistical model that is an extension of the Generalized Linear Model. This method approximates a non-linear model by fitting piecewise linear segments that are connected at nodes referred to as hinge functions. The hinge functions provide the point at which the two straight lines join. A sequence of lines and hinges approximates a non-linear Spline. The MARS model uses a forward pass to find the best fit that minimizes the Sum of Squared Error. This first pass is referred to as “greedy” because it seeks the best fit regardless of how many terms, or line and hinge segments, it creates. To avoid over-fitting, the MARS method has a second pass that prunes the terms created in the first path to assess which can be removed without having large negative effects on the model’s performance; this lowers the model’s complexity and variance. The MARS method uses Generalized Cross-Validation to assess how pruning affects performance. This method was introduced by Friedman (1991).</p>	
Negative Prediction Gain (NPG)	Task 3, p. 70
<p>The NPG is a statistic that is derived from the confusion matrix to assess a model’s ability to correctly classify site-unlikely areas. The NPG quantifies how much less likely a site discovery is at a location labeled site-unlikely using the model than if surveying at random. Ideally, a model would have a low NPG and a high Positive Predictive Gain (see Oehlert and Shea 2007).</p>	
Negative Prediction Value (NPV)	Task 3, p. 70
<p>The NPV is a measure that is derived from the confusion matrix. This measures the probability that a non-site cell is correctly labeled as a background cell (see Oehlert and Shea 2007).</p>	
<i>nprune</i>	Task 5, p. 56
<p>This is the name of a key parameter in the MARS model. This algorithm includes a backwards pass that prunes the model down to reduce variance and eliminate unneeded model terms. The <i>nprune</i> parameter is used to set the <i>maximum</i> number of terms that are allowed to remain in the model; the fewer terms, the more simple the model. Through this parameter, models can be trimmed for the purpose of model size, complexity, or generality of the fit. By default, <i>nprune</i> is set to NULL so that the model is unrestrained in the number of terms. For this project, the <i>nprune</i> parameter is set through cross-validation to the lowest error rate of the out-of-fold sample.</p>	

- Null Model Task 3, p. 50
The term Null Model refers to a Logistic Regression that only contains the response variable (in this case, site prediction), with no predictor. It is essentially a flat-line regression that uses the average of all values, thereby providing a baseline against which the saturated model (that is, the model that incorporates the predictors) can be tested. If the saturated model is not better than the null model by a statistically significant amount (in this case as measured using the Likelihood Ratio Test), then new predictors need to be chosen.
- Ordinary Least Squares (OLS)..... Task 3, p. 50
OLS is the statistical method most commonly used to estimate unknown parameters within linear regression. OLS seeks to fit a line that minimizes the Sum of Squared Errors between the predictor (in this case, the environmental variable) and the response (site-present vs. site-absent).
- Out-of-Bag (OOB) Sample Task 3, p. 19
The term OOB is used to describe the internal testing data set within predictive algorithms such as Bagging and Random Forests. Within these algorithms, the training data set is sampled via the bootstrap with roughly two-thirds of the data used for model training and the remaining one-third used for testing and variable selection. This remaining one-third is referred to as the OOB Sample. OOB Sample error rates are calculated from this hold-out set (see Breiman 1996b).
- Out-of-Sample (OOS)..... Task 3, p. 37
The term OOS refers to the portion of data within the hold-out sample from K-folds Cross Validation. For example, in $k = 10$ fold CV, the first pass will use folds 1–9 to train the model and fold 10 to test the model. The tenth fold is the OOS fold and the error estimate derived from this is called the OOS estimate (see Efron and Tibshirani 1997).
- Pennsylvania Archaeological Site Survey (PASS) Task 1, p. 65
The PASS files are a collection of paper forms, maps, reports, and photographs that document the location and attributes of known archaeological sites within the Commonwealth of Pennsylvania. These files have been digitized and can be accessed through the Cultural Resources Geographic Information System.
- Positive Predictive Gain (PPG)..... Task 3, p. 36
The PPG is a statistic that is derived from the Confusion Matrix to assess a model's ability to correctly classify site-likely areas. The PPG quantifies how much more likely a site discovery is at a location labeled site-likely using the model than if surveying at random. Ideally, a model would have a high PPG and a low Negative Prediction Value (see Oehlert and Shea 2007).

Positive Prediction Value (PPV).....	Task 3, p. 70
The PPV is a measure that is derived from the Confusion Matrix. This measures the probability that a site cell is correctly labeled as a site-likely cell (see Oehlert and Shea 2007).	
Prevalence	Task 5, p. 77
Prevalence is the proportion of a population found to have a particular condition. In this case, the population is the total number of $\sim 10 \times 10$ -m raster cells that make up each subarea and the condition is that a cell be within a known archaeological site. Determining prevalence is important in these models because the low number of cells within known archaeological sites is very small compared to the overall area being predicted, leading to highly imbalanced data in terms of site-presence versus site-absence.	
Pseudo R-Squared (Pseudo R^2).....	Task 3, p. 55
Pseudo R-Squared describes a statistic that is intended to mimic the qualities of the R-Squared, but is applicable to models that do not use Ordinary Least Squares, such as Logistic Regression. In general, Pseudo R-Squared is similar to R-Squared in that the numerous variations of the measure range approximately from 0 to 1 and a higher number indicates a generally better fit. However, Pseudo R-Squared should not be compared directly to R-Squared because they are derived quite differently. A number of Pseudo R-Squared variations have similarities to the Likelihood Ratio Test (see Pampel 2000).	
Python Language	Task 3, p. 38
Python is a widely used high-level programming language. (Information available at: http://www.python.org/ .)	
R-Squared (R^2).....	Task 3, p. 50
R^2 , also referred to as the Coefficient of Determination, is a metric used in the evaluation of variance and goodness-of-fit for primarily linear models using Ordinary Least Squares. The R^2 is calculated as a one minus the residual sum of squares divided by the total sum of squares. The most common interpretation of R^2 is for the fit of a linear model. An R^2 of 1 indicates a perfect fit between the regression line and data points.	
R Statistical Language	Task 3, p. 38
R Statistical Language is a widely used statistical computing environment. (Information available at: http://www.r-project.org/ .)	
Random Forests	Task 3, p. 14
Random Forests is trademarked statistical classification algorithm created by Leo Breiman and Adele Cutler. Random Forests is a tree based ensemble method that builds off the ideas	

of Classification and Regression Trees and Bagging. The primary features of Random Forests include internal testing through Bootstrap Aggregating and variable importance via random subset selection (see Breiman 2001).

randomForest (RF) (see also Random Forests)..... Task 3, p. 18
RF is an implementation of the Random Forests classification algorithm written in the R Statistical Language (see Liaw and Wiener 2002).

Receiver Operating Characteristics (ROC)..... Task 3, p. 21
The ROC is a graphical representation of statistical classification model results. The ROC graph typically takes on a curved shape and is therefore often referred to as the ROC curve. The x-axis of the ROC graph is a model's False Positive Rate and the y-axis is the True Positive Rate; both are scaled from 0 to 1. The quantities on the x- and y-axes are also referred to as 1 – Specificity and Sensitivity, respectively. The actual curve in the graphic is generated by calculating the True Positive Rate and False Positive Rate for each cut-point of the model's prediction. The graphic also contains a line (often dashed) that originates at point 0,0 and goes at a 45-degree angle to point 1,1. This line represents a model that has no predictive power. The closer the ROC curve is to the upper left corner of the graph (which is point $x = 0$, $y = 1$), the greater the predictive power. Put another way, the best classification has the largest area under the curve. A line of this description will have a high True Positive Rate for the entire range of False Positive Rates. The ROC curve can be used to estimate the total predictive power of the model, often enumerated as the Area Under Curve, to compare similar models across all cut-points, or select an optimal cut-point to use for classification, resulting in a Confusion Matrix (see Fawcett 2004).

Residual Sum-of-Squares (RSS)..... Task 3, p. 53
The RSS is a measure of the model fit. The RSS is calculated as the sum of squared differences between the estimated and actual observations (see Salkind 2007).

Root Mean Square Error (RMSE)..... Task 3, p. 21

The RMSE is a statistic, or loss function, used to quantify the difference between an estimate and a true value. The RMSE is calculated as the square root of the Mean Squared Error. When calculated on Out-of-Sample predictions, such as in this project, the RMSE represents the sample standard deviation of the prediction errors. The formula below is how RMSE is calculated, where n = the number of data values, y_j is the observed j^{th} value and \hat{y}_j is the predicted j^{th} value for all j values from 1 to n . Therefore the RMSE is the square root of the average of all squared errors.

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

A benefit of RMSE over Mean Squared Error is that it is scaled to the dependent variable and is therefore directly interpretable. With a binary dependent variable (0 to 1), the RMSE is taken as the distance on average between the predicted probability and the true value (see Salkind 2007).

Sensitivity (see also True Positive Rate)..... Task 3, p. 21

Sensitivity is a term used for a classification's True Positive Rate; this value is also referred to as Recall. Sensitivity is the total fraction of sites that are classified by the model to be in the site-likely area. This measure is akin to the concept of precision and Type II errors. Sensitivity is calculated for a cut-point within a classification model as the number of correctly predicted positive observations (correctly classified sites) divided by the total number of actual positive observations (known sites) (see Oehlert and Shea 2007).

Synthetic Minority Over-Sampling Technique (SMOTE)..... Task 7, p. 6

This is a technique developed to deal with highly imbalanced class data in machine learning. Highly imbalanced data occur when one of the two classes (e.g., positive and negative) is represented at a much greater rate than the other; typically the positive class or class of interest is much smaller than the negative class. In the context of the Pennsylvania model project, the data are highly imbalanced because the number of site-present cells is much smaller than the number of background cells in the environment. SMOTE tries to create a more equal class distribution by down-sampling the negative class to make it smaller and up sampling the positive class by creating new synthetic observations that are random permutations of existing observations (see Chawala et al. 2002).

Specificity (see also True Negative Rate)..... Task 2, p. 21

Specificity is a termed used for a classification's True Negative Rate. Specificity is the fraction of background that is classified as site-unlikely by the model. This measure is akin to

the concept of accuracy and Type I errors. Specificity is calculated for a cut-point within a classification model as the number of correctly predicted negative observations (correctly classified non-sites) divided by the total number of actual negative observations (background cells) (see Oehlert and Shea 2007).

Spline Task 3, p. 18
A curve that connects two or more points. The shape of the Spline is determined by a mathematical function that interpolates the space between the points into a smooth curve.

Sum of Squared Error (SSE)..... Task 3, p. 50
The SSE is a measure of prediction accuracy. This measure is calculated nearly the same as the Residual Sum-of-Squares, but more commonly used on prediction errors as opposed to model fit. Within this project, the SSE is used as a part of the Mean Squared Error and Root Mean Square Error statistics to assess the accuracy of prediction results (see Lehman and Casella 1998).

True Negative Rate (TNR) (see also Specificity) Task 3, p. 67
The TNR is a measure of a model's classification at a given cut-point. Often referred to as a model's Specificity, the TNR is calculated as the percent of negative observations correctly classified as such. In this project, this would be the rate at which background cells are correctly classified as site un-likely cells (see Oehlert and Shea 2007).

True Positive Rate (TPR) (see also Sensitivity)..... Task 3, p. 67
The TPR is a measure of a model's classification at a given cut-point. Often referred to as a models Sensitivity, the TPR is calculated as the percent of positive observations correctly classified as such. In this project, this would be the rate at which known site-present cells are correctly classified as site-likely cells (see Oehlert and Shea 2007).

Unexpected Discovery Rate (UDR)..... Task 3, p. 70
The UDR is a measurement of a model's classification ability at a given cut-point. The UDR is defined as the probability of a cell containing a site given that the model predicted it as site-unlikely. That can be thought of as the rate of unintentional discovery, or "oops" rate (see Oehlert and Shea 2007).

APPENDIX B

VARIABLES CONSIDERED THROUGHOUT PROJECT

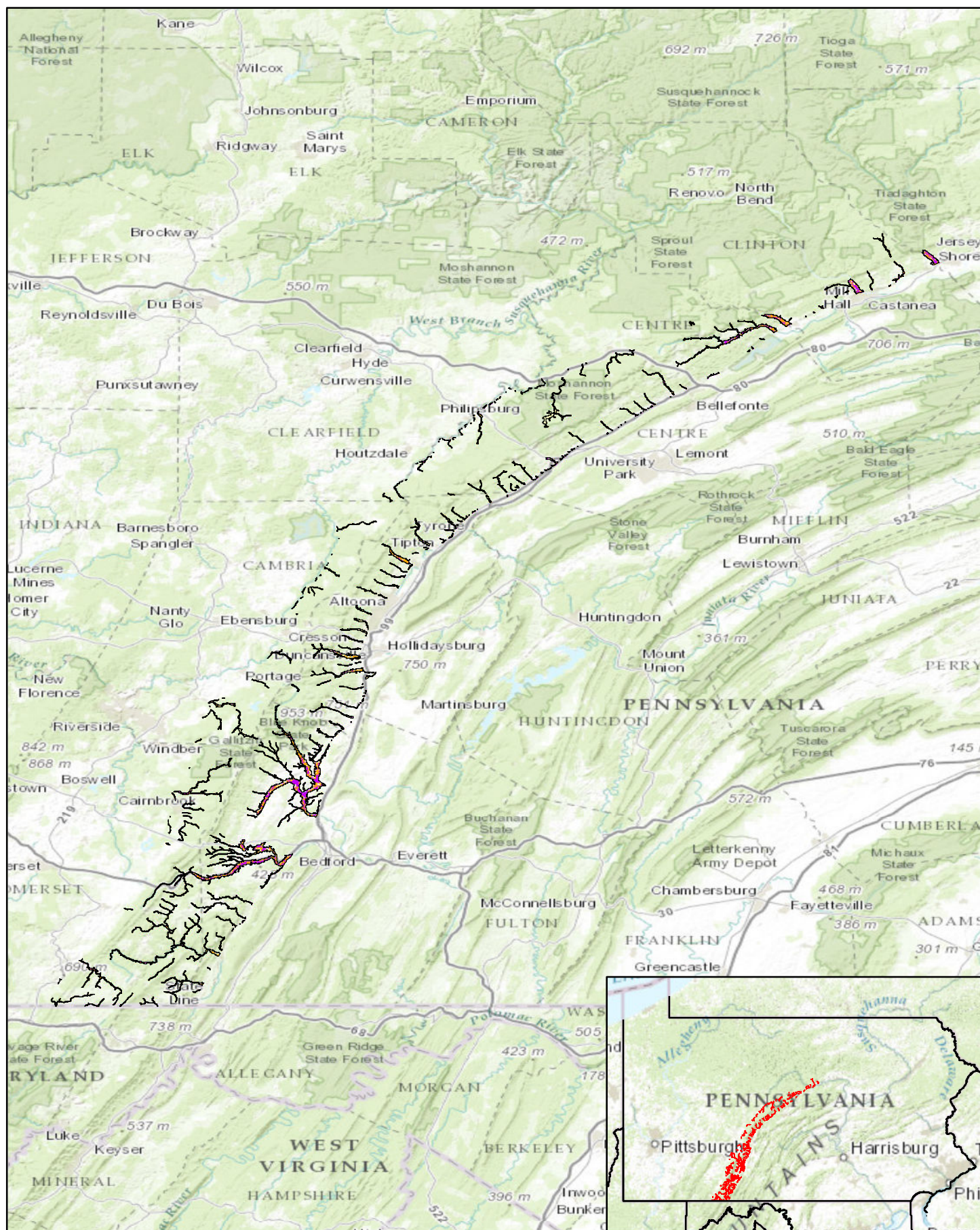
Predictor	Family	Measure	Neighborhood Sizes	Description
aspect	Topography	bearing	n/a	Orientation of slope relative to north
aws050	Soils - aggregate	water storage - integer	n/a	Water that is available to plants in the top 50 cm of soil. AWS is expressed as centimeters of water, reported as the average of all components in the map unit.
c_hyd_min	Hydrology	cost-distance	n/a	Minimum distance to stream or water body
c_hyd_min_wt	Hydrology	cost-distance	n/a	Minimum distance to stream, water body, or wetland
c_trail_dist	Topography - Cultural	cost-distance	n/a	Cost-distance to historically documented Native American trails (Wallace 1965).
cd_conf	Hydrology	cost-distance	n/a	Cost-Distance to stream confluence (NHD flow lines)
cd_drnh	Hydrology	cost-distance	n/a	Cost-Distance to stream heads (NHD flow lines)
cd_h1	Hydrology	cost-distance	n/a	Cost-distance to historic streams
cd_h2	Hydrology	cost-distance	n/a	Cost-distance to NHD flow lines
cd_h3	Hydrology	cost-distance	n/a	Cost-distance to NHD water bodies
cd_h4	Hydrology	cost-distance	n/a	Cost-distance to NWI wetlands
cd_h5	Hydrology	cost-distance	n/a	Cost-distance to NWI water bodies
cd_h6	Hydrology	cost-distance	n/a	Cost-distance to 4th order and higher streams
cd_h7	Hydrology	cost-distance	n/a	Cost-distance to 3rd order and higher streams
dem_fl1	Topography	elevation, meters (float)	n/a	1/3rd Arc-second digital elevation model as float, with sinks filled
drcdry	Soils - aggregate	classification, nominal	n/a	Drainage class (dominant condition) - the NRCS describes natural soil drainage classes that represent the moisture condition of the soil in its natural condition throughout the year
drcwet	Soils - aggregate	classification, nominal	n/a	Drainage class (wet conditions) - the NRCS describes natural soil drainage classes that represent the moisture condition of the wettest soil component in its natural condition throughout the year

Predictor	Family	Measure	Neighborhood Sizes	Description
e_hyd_min	Hydrology	Euclidian-distance, meters	n/a	Minimum distance to stream or water body
e_hyd_min_wt	Hydrology	Euclidian-distance, meters	n/a	Minimum distance to stream, water body, or wetland
e_trail_dist	Topography - Cultural	Euclidian-distance, meters	n/a	Euclidian distance to historically documented Native American trails (Wallace 1965).
ed_conflu	Hydrology	Euclidian-distance, meters	n/a	Euclidian distance to stream confluence (NHD flow lines)
ed_drnh	Hydrology	Euclidian-distance, meters	n/a	Euclidian distance to stream heads (NHD flow lines)
ed_h1	Hydrology	Euclidian-distance, meters	n/a	Euclidian distance to historic streams
ed_h2	Hydrology	Euclidian-distance, meters	n/a	Euclidian distance to NHD flow lines
ed_h3	Hydrology	Euclidian-distance, meters	n/a	Euclidian distance to NHD water bodies
ed_h4	Hydrology	Euclidian-distance, meters	n/a	Euclidian distance to NWI wetlands
ed_h5	Hydrology	Euclidian-distance, meters	n/a	Euclidian distance to NWI water bodies
ed_h6	Hydrology	Euclidian-distance, meters	n/a	Euclidian distance to 4th order and higher streams
ed_h7	Hydrology	Euclidian-distance, meters	n/a	Euclidian distance to 3rd order and higher streams
eldrop#c	Topography	elevation, meters	1,8,10,16,32 cells	Drop in elevation over # cell neighborhood
elev_2_conf	Topography - Hydrology	vertical-distance, meters	na	Elevation to stream confluence (NHD flow lines)
elev_2_drainh	Topography - Hydrology	vertical-distance, meters	na	Elevation to stream head (NHD flow lines)
elev_2_stream	Topography - Hydrology	vertical-distance, meters	na	Elevation to stream (NHD flow lines)
flowdir	Hydrology	direction, bearing	na	Flow direction based on DEM
flw_acum	Hydrology	accumulation, cells	na	Flow accumulation based on DEM

Predictor	Family	Measure	Neighborhood Sizes	Description
niccdcd	Soils - aggregate	classification, nominal	n/a	The broadest category in the land capability classification system for soils; the dominant capability class, under nonirrigated conditions, for the map unit based on composition percentage of all components in the map unit.
random	Random	random float (0 to 1)	na	Randomly selected number between 1 and 0
rel_#c	Topography	index, 0 to 1	1,8,10,16,32 cells	Relative topographic position
rng_#c	Topography	elevation range, integer	1,8,10,16,32 cells	Range of elevation in # cell neighborhood
slope_deg	Topography	slope, degrees	n/a	Topographic slope measured in degrees
slope_pct	Topography	slope, percent	n/a	Topographic slope measured in percent rise over run
slpvr_#c	Topography	slope range, integer	1,8,10,16,32 cells	Slope variability within # cell neighborhood
std_#c	Topography	standard deviation	1,8,10,16,32 cells	Standard deviation of elevation range within # cell neighborhood
tpi_#c	Topography	index, integer	5,10,50,100,250 cells	Topographic Position Index. Position of cell relative to surrounding landscape within # cell neighborhood
tpi_cls#c	Topography	classification, nominal	5,10,50,100,250 cells	TPI standardized and classified into 1 standard deviation groups within # cell neighborhood
tpi_sd#c	Topography	standard deviation	5,10,50,100,250 cells	Standard deviation of TPI within # cell neighborhood
tri_#c	Topography	index, integer	1,8,10,16,32 cells	Topographic Ruggedness Index. Measure of terrain roughness within # cell neighborhood
twi#c	Topography - Hydrology	index, integer	1,8,10,16,32 cells	Topographic Wetness Index. Measure of upslope accumulation within # cell neighborhood
vrf_#c	Topography	index, integer	1,8,10,16,32 cells	Vector Roughness Factor. Measure of three-dimensional variation in slope within # cell neighborhood

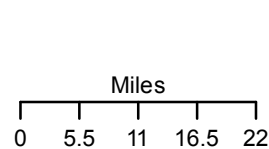
APPENDIX C

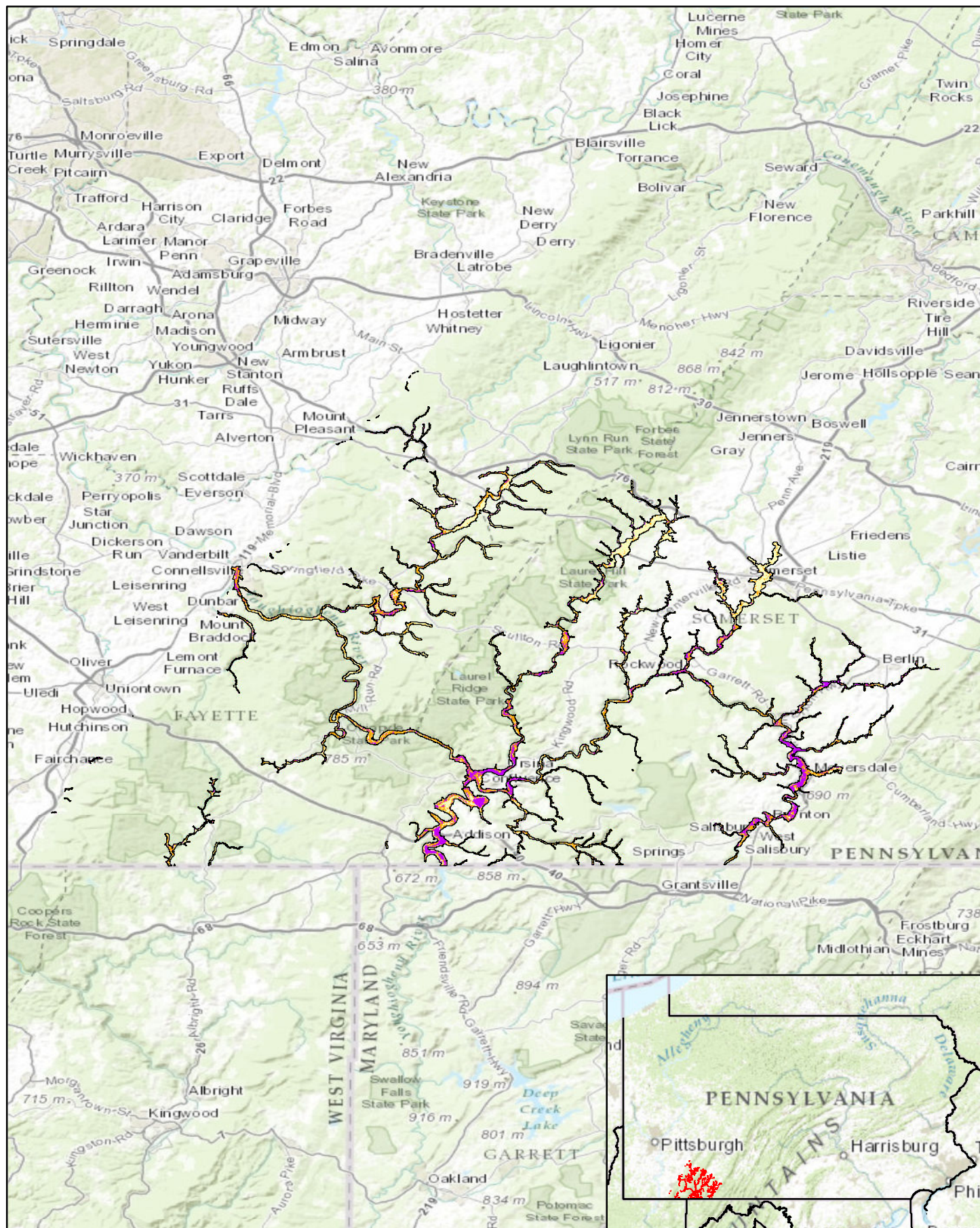
PENNSYLVANIA MODEL SUBAREA MAPS



Pennsylvania Predictive Model Set
 Region: 1, Zone: east, Subarea: riverine section 1

Sensitivity
 High
 Moderate
 Low





Pennsylvania Predictive Model Set

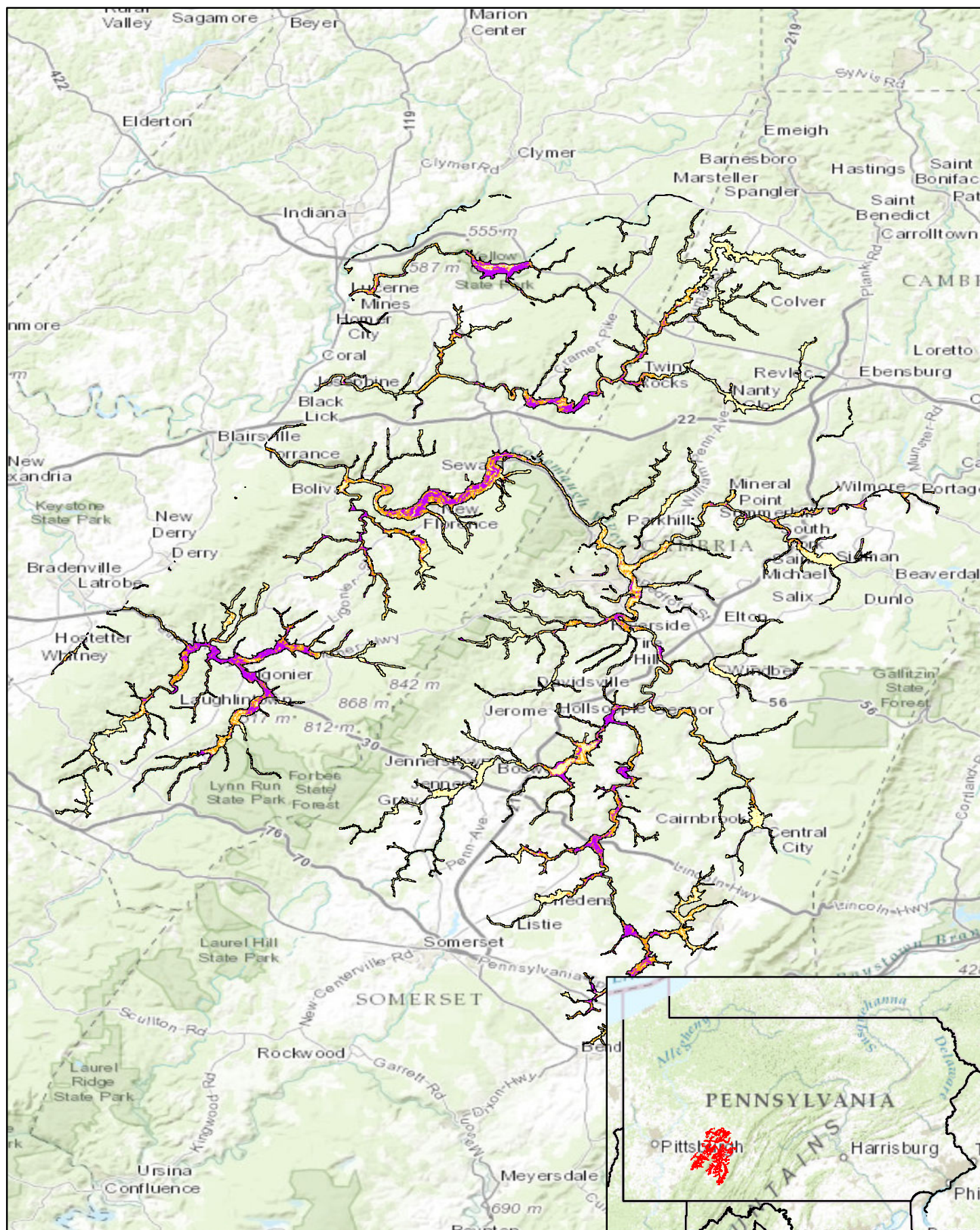
Region: 1, Zone: east, Subarea: riverine section 2

Sensitivity

- High
- Moderate
- Low

Miles
0 2.5 5 7.5 10

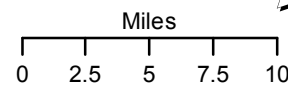
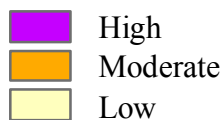


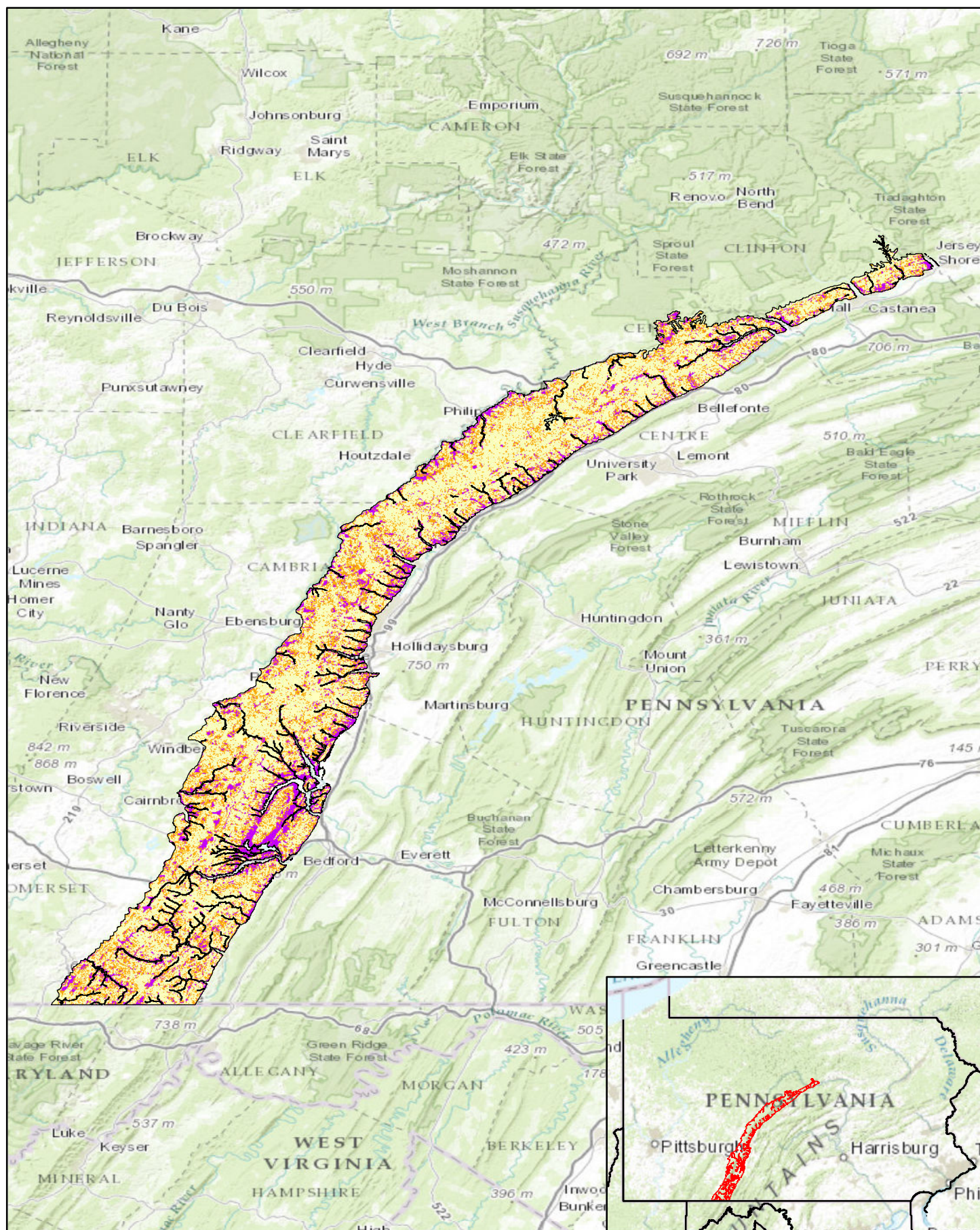


Pennsylvania Predictive Model Set

Region: 1, Zone: east, Subarea: riverine section 3

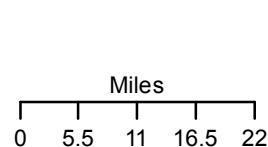
Sensitivity

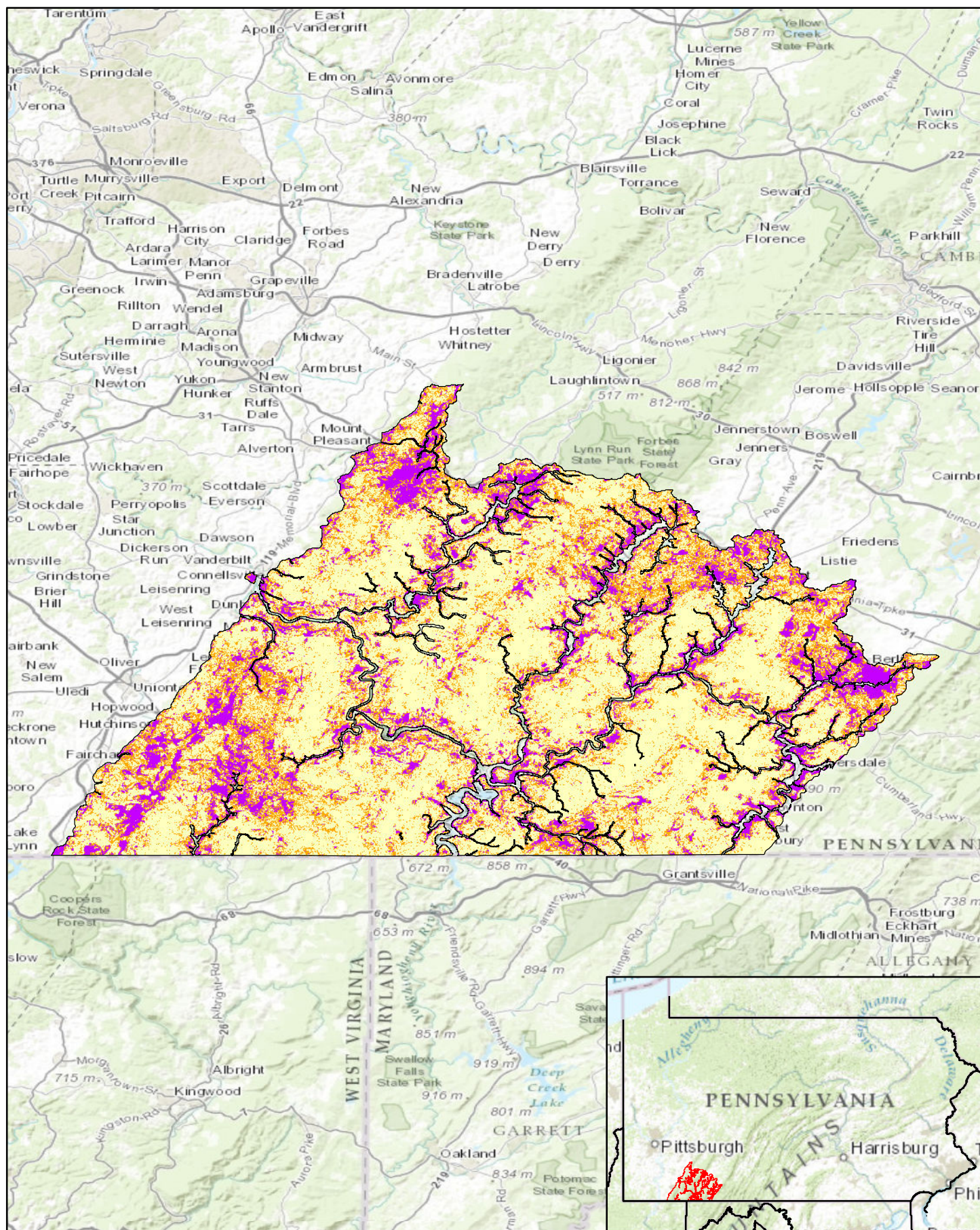




Pennsylvania Predictive Model Set
 Region: 1, Zone: east, Subarea: upland section 1

Sensitivity
 High
 Moderate
 Low

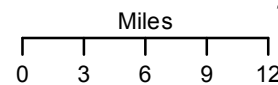


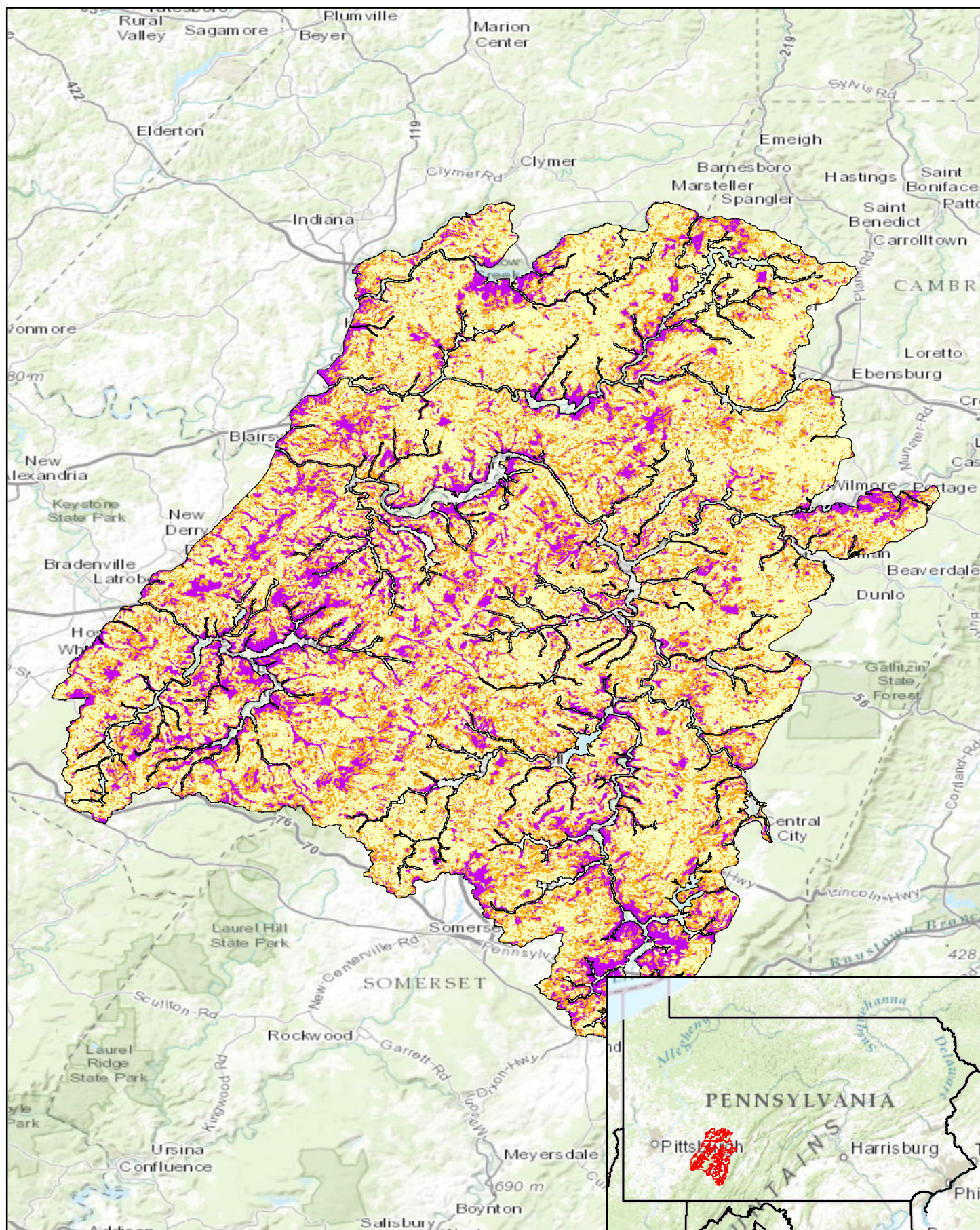


Pennsylvania Predictive Model Set
 Region: 1, Zone: east, Subarea: upland section 2

Sensitivity

- High
- Moderate
- Low



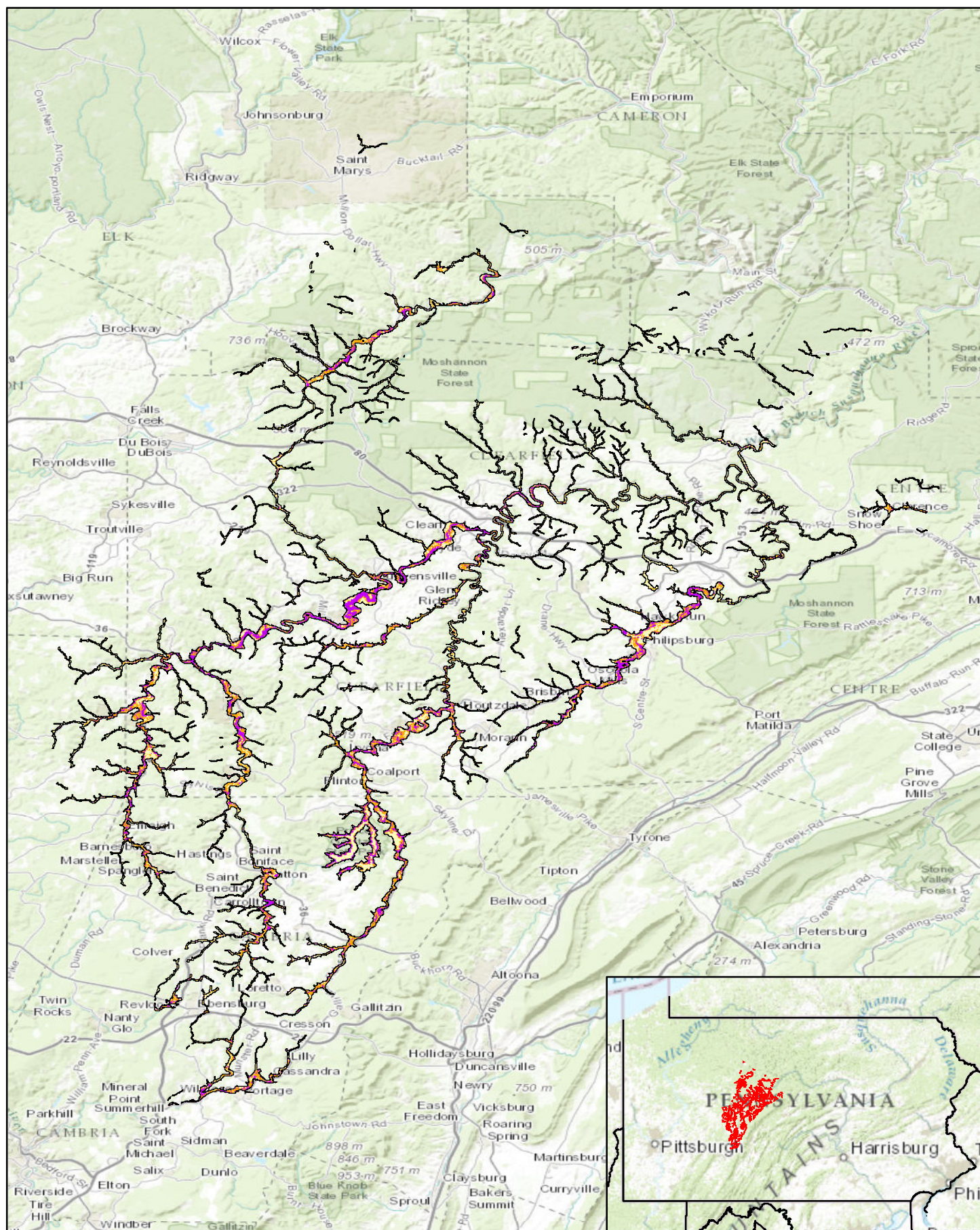


Pennsylvania Predictive Model Set
 Region: 1, Zone: east, Subarea: upland section 3

Sensitivity
 High
 Moderate
 Low

Miles
 0 2.5 5 7.5 10



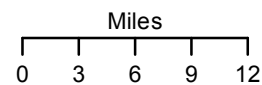


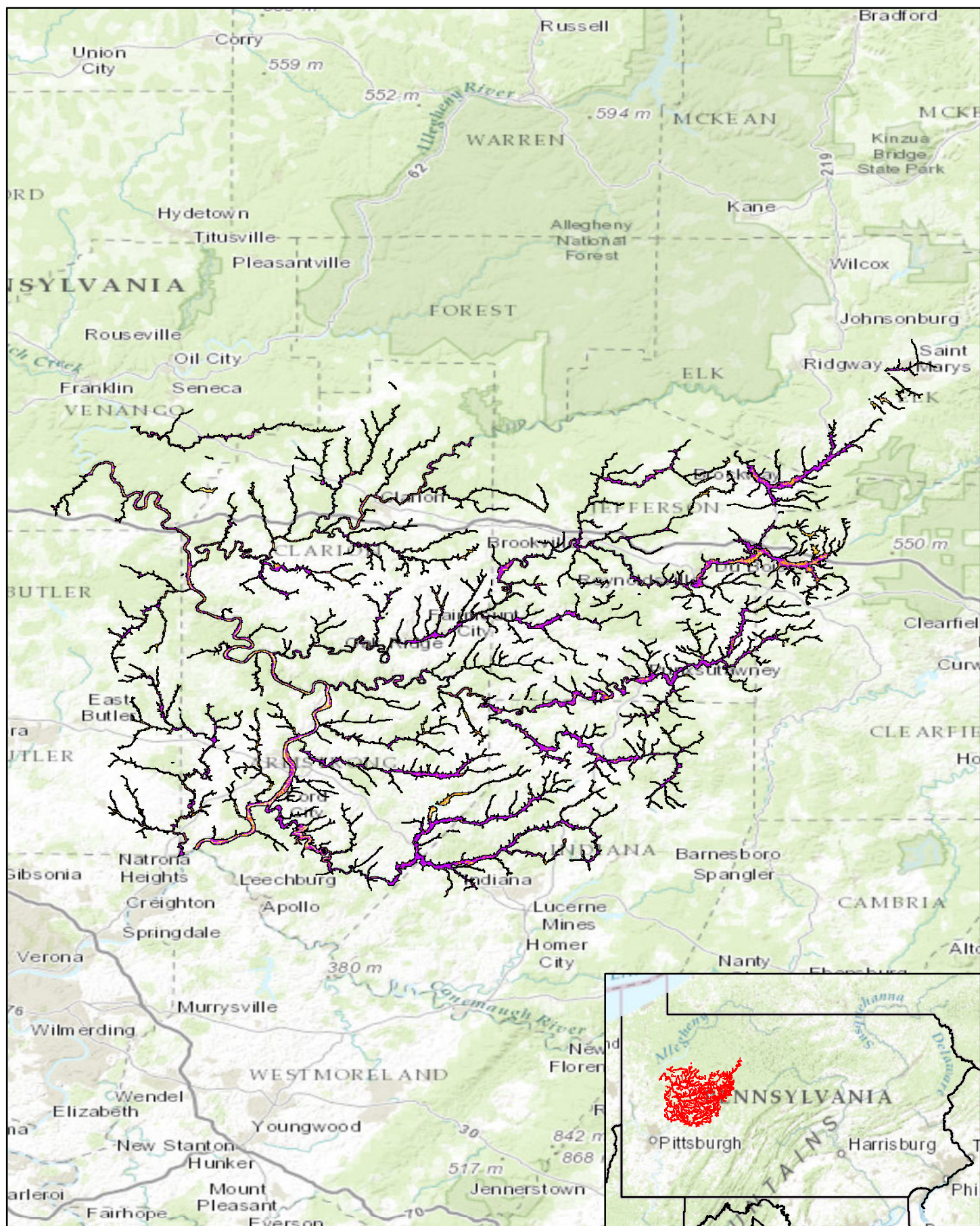
Pennsylvania Predictive Model Set

Region: 1, Zone: north, Subarea: riverine section 1

Sensitivity

- High
- Moderate
- Low



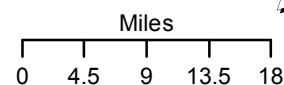


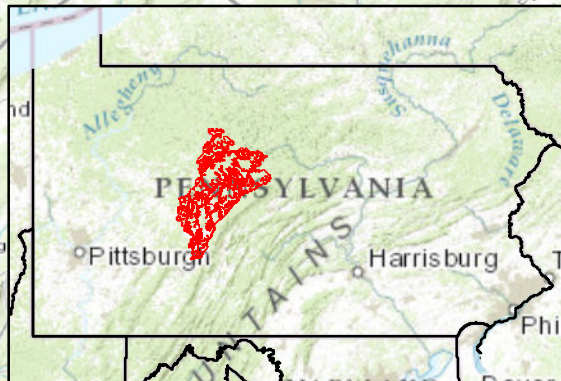
Pennsylvania Predictive Model Set

Region: 1, Zone: north, Subarea: riverine section 2




Sensitivity

- High
- Moderate
- Low

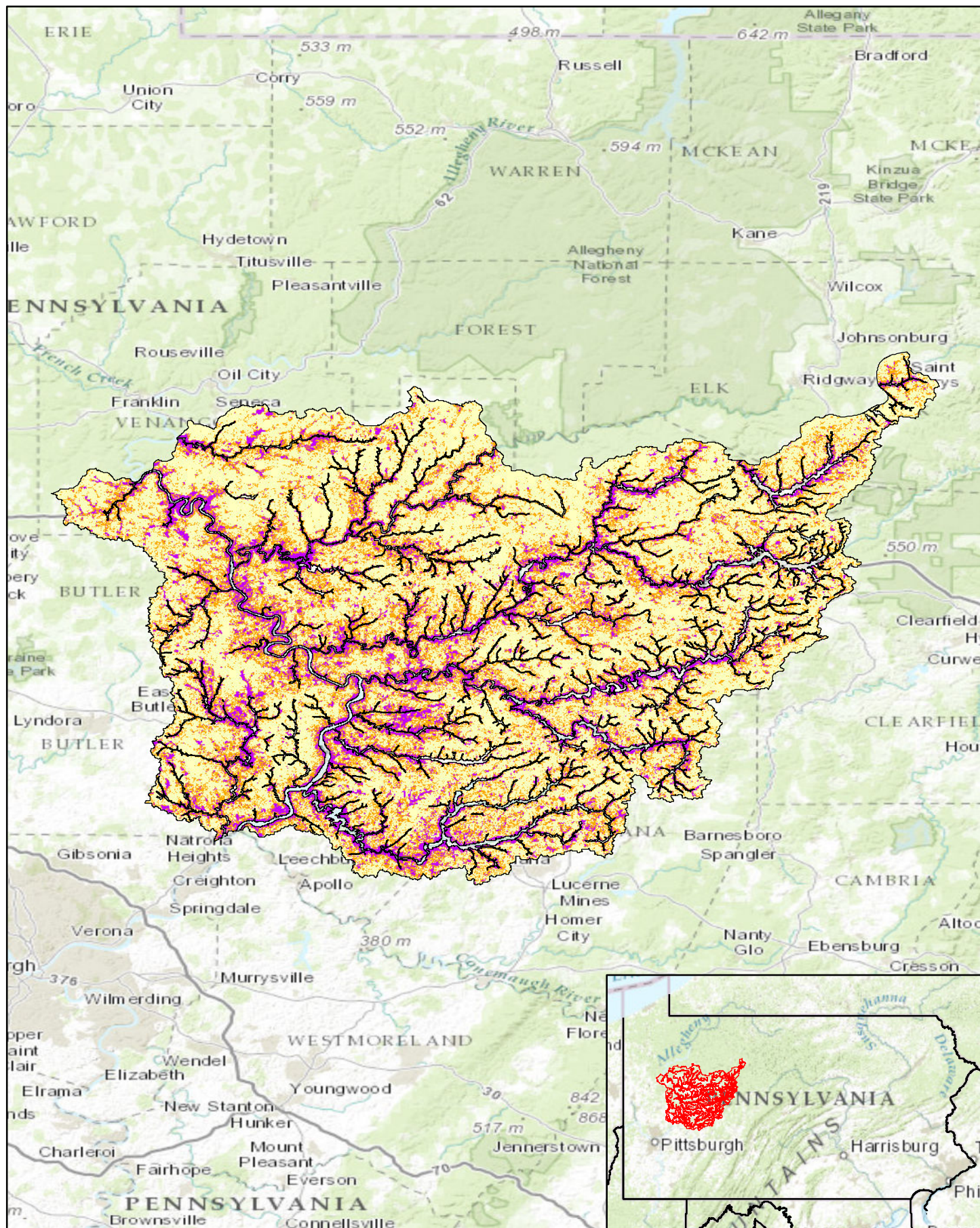




Region: 1, Zone: north, Subarea: upland section 1

	High
	Moderate
	Low

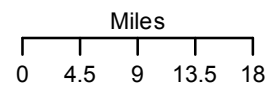
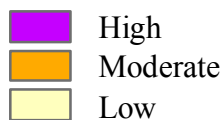


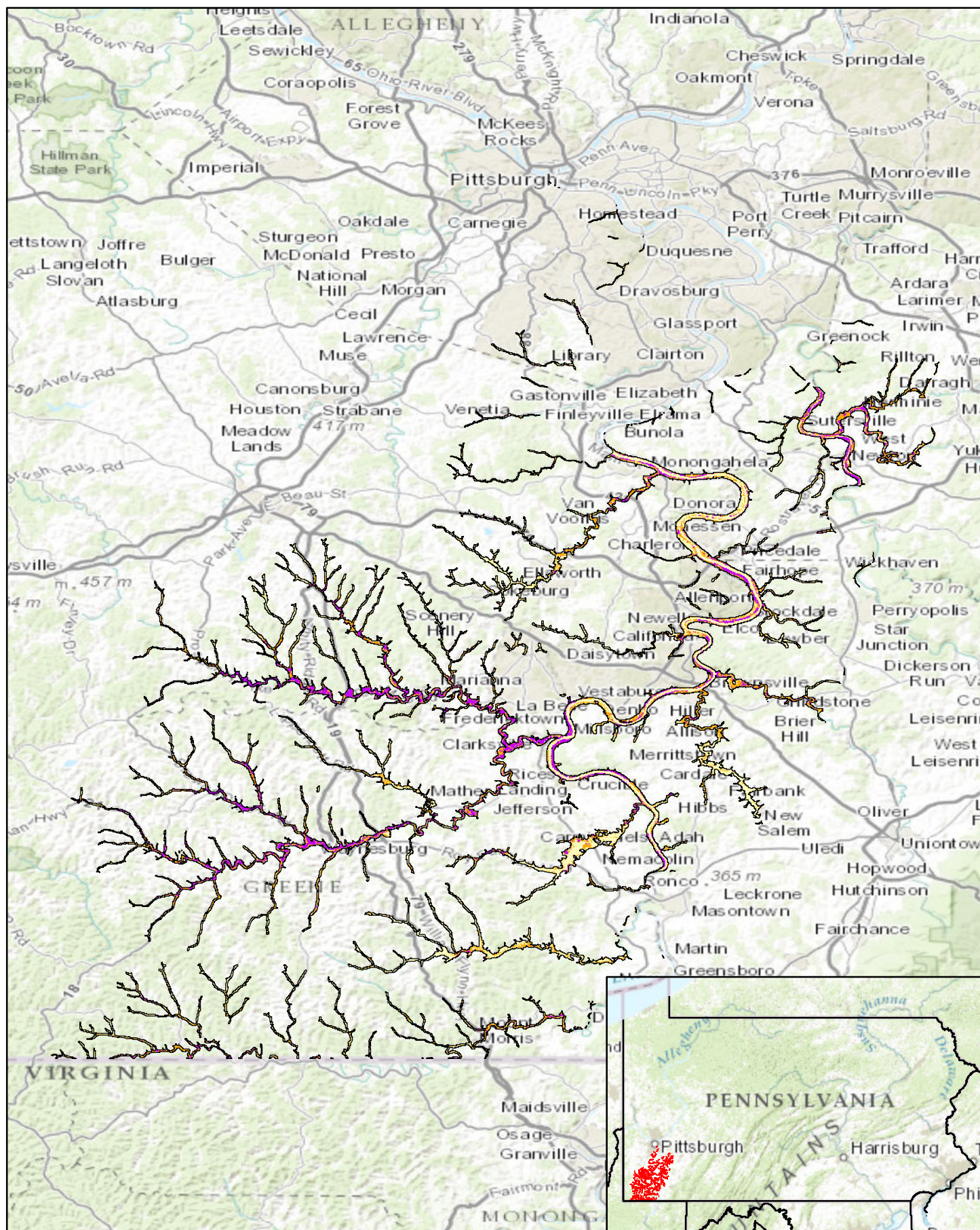


Pennsylvania Predictive Model Set

Region: 1, Zone: north, Subarea: upland section 2

Sensitivity

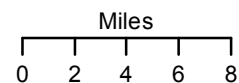


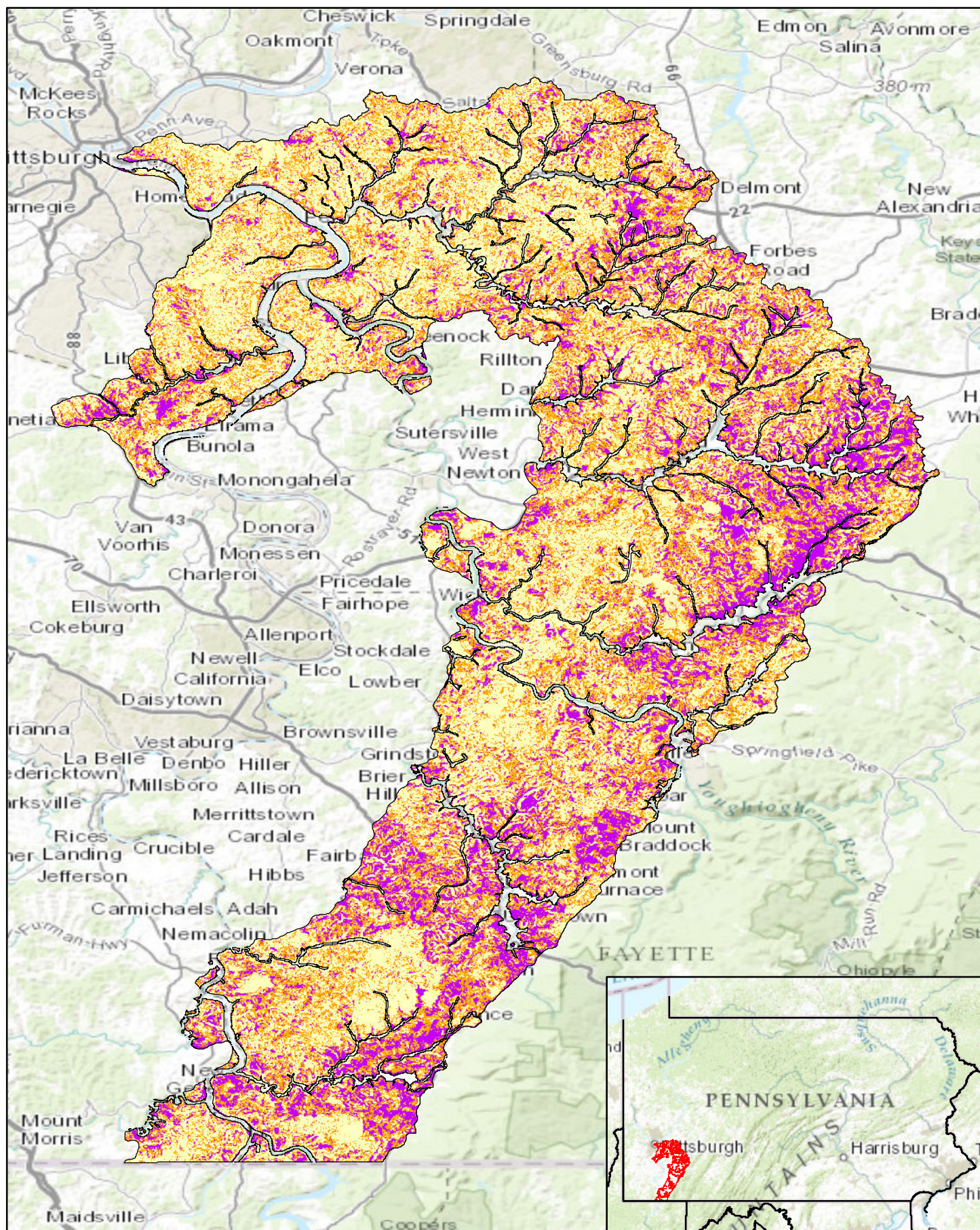


Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: riverine section 1

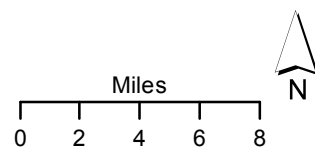
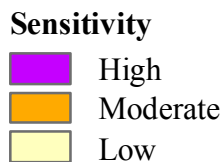
Sensitivity

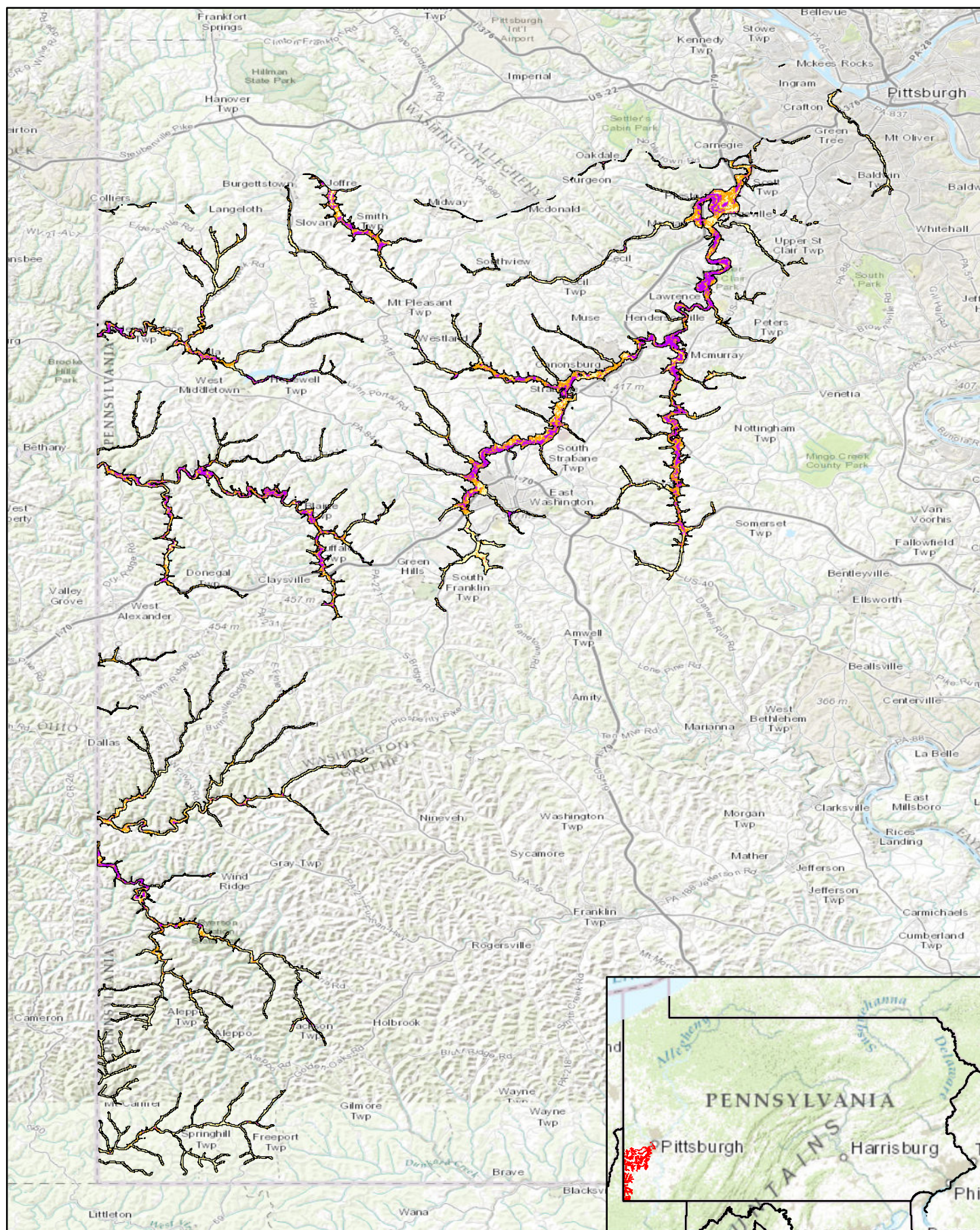
- High
- Moderate
- Low





Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: upland section 5





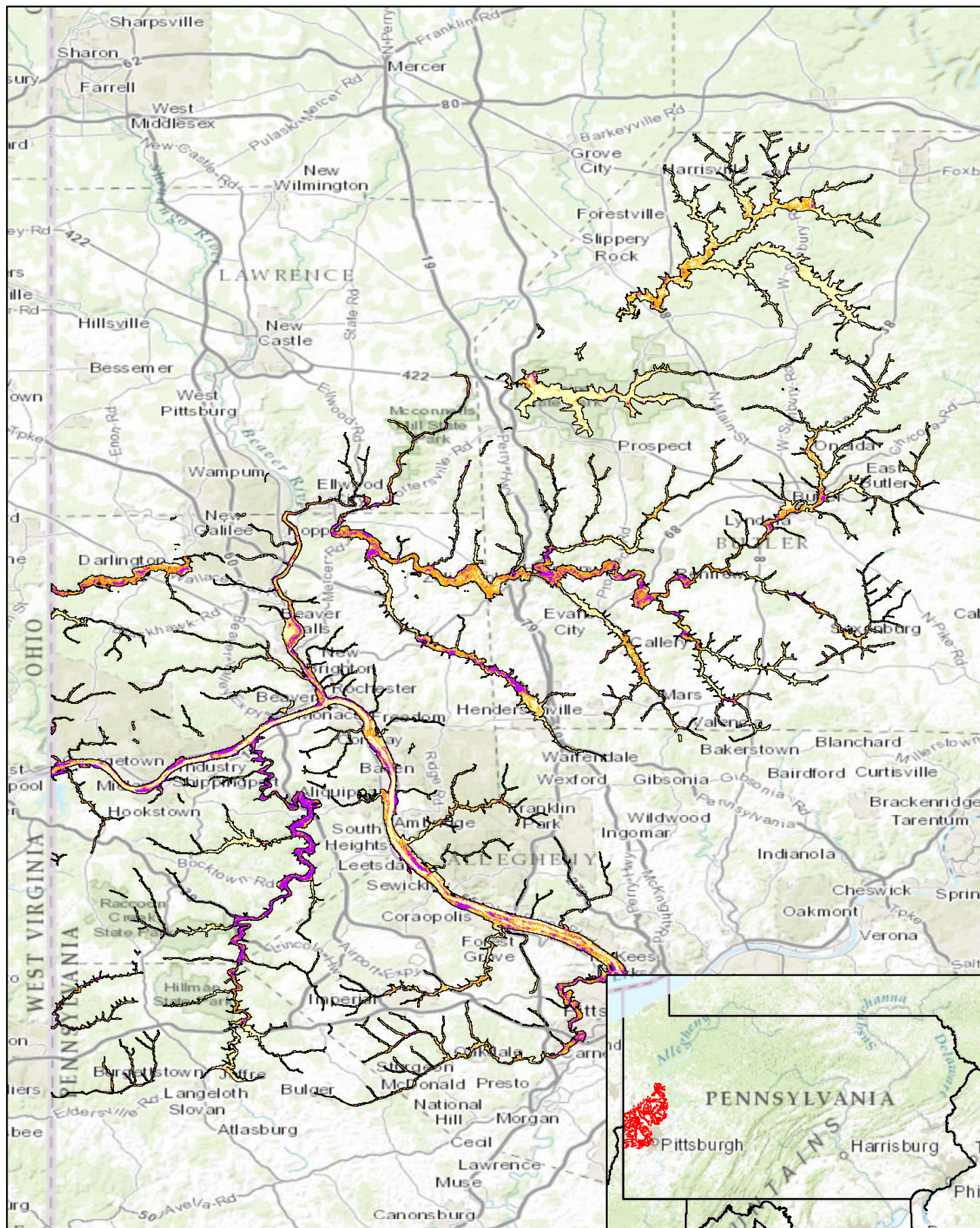
Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: riverine section 2

Sensitivity

- High
- Moderate
- Low

Miles
 0 2 4 6 8

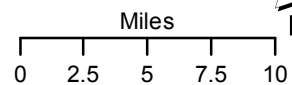


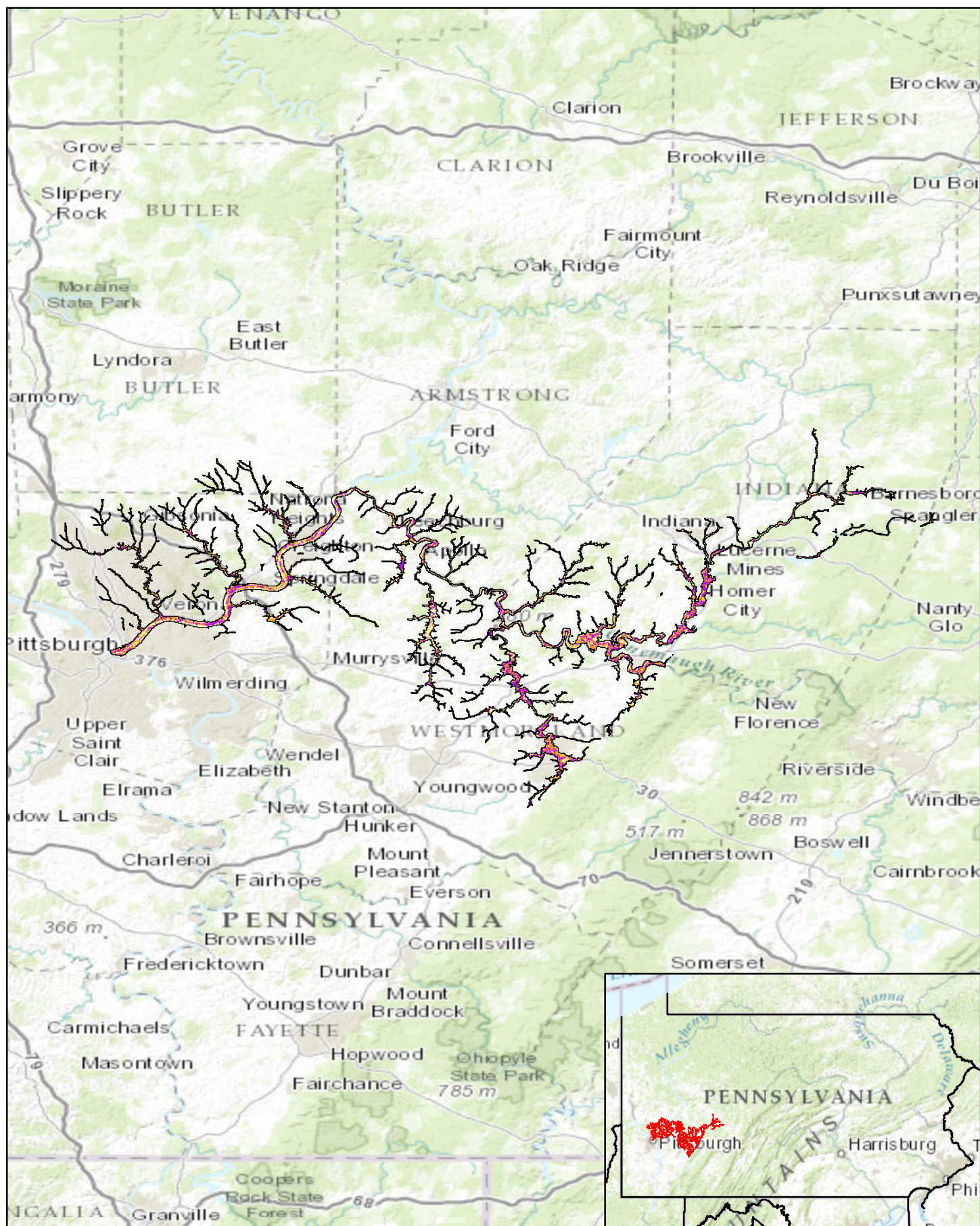


Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: riverine section 3

Sensitivity

- High
- Moderate
- Low

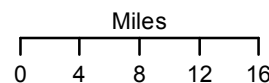


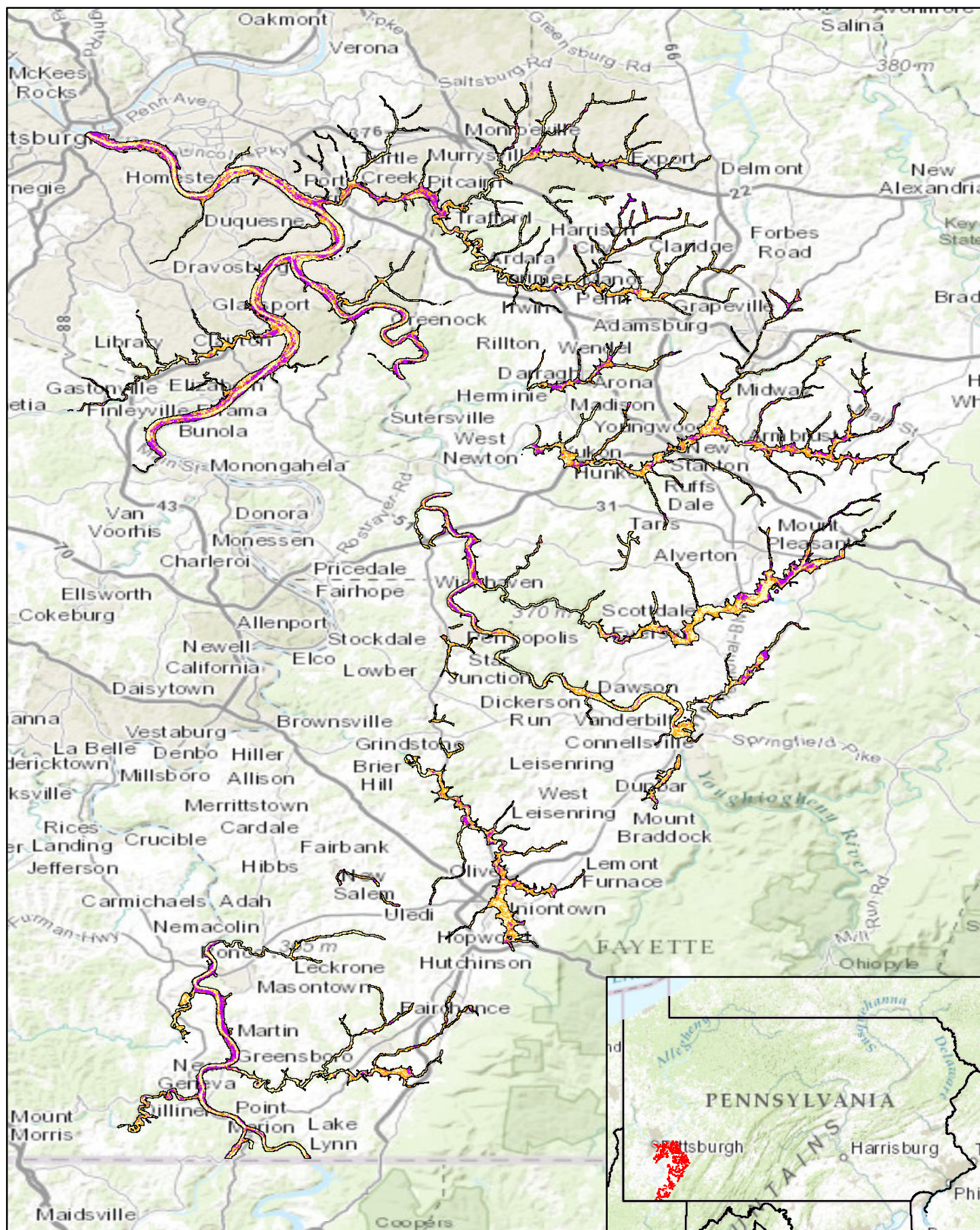


Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: riverine section 4

Sensitivity

- High
- Moderate
- Low



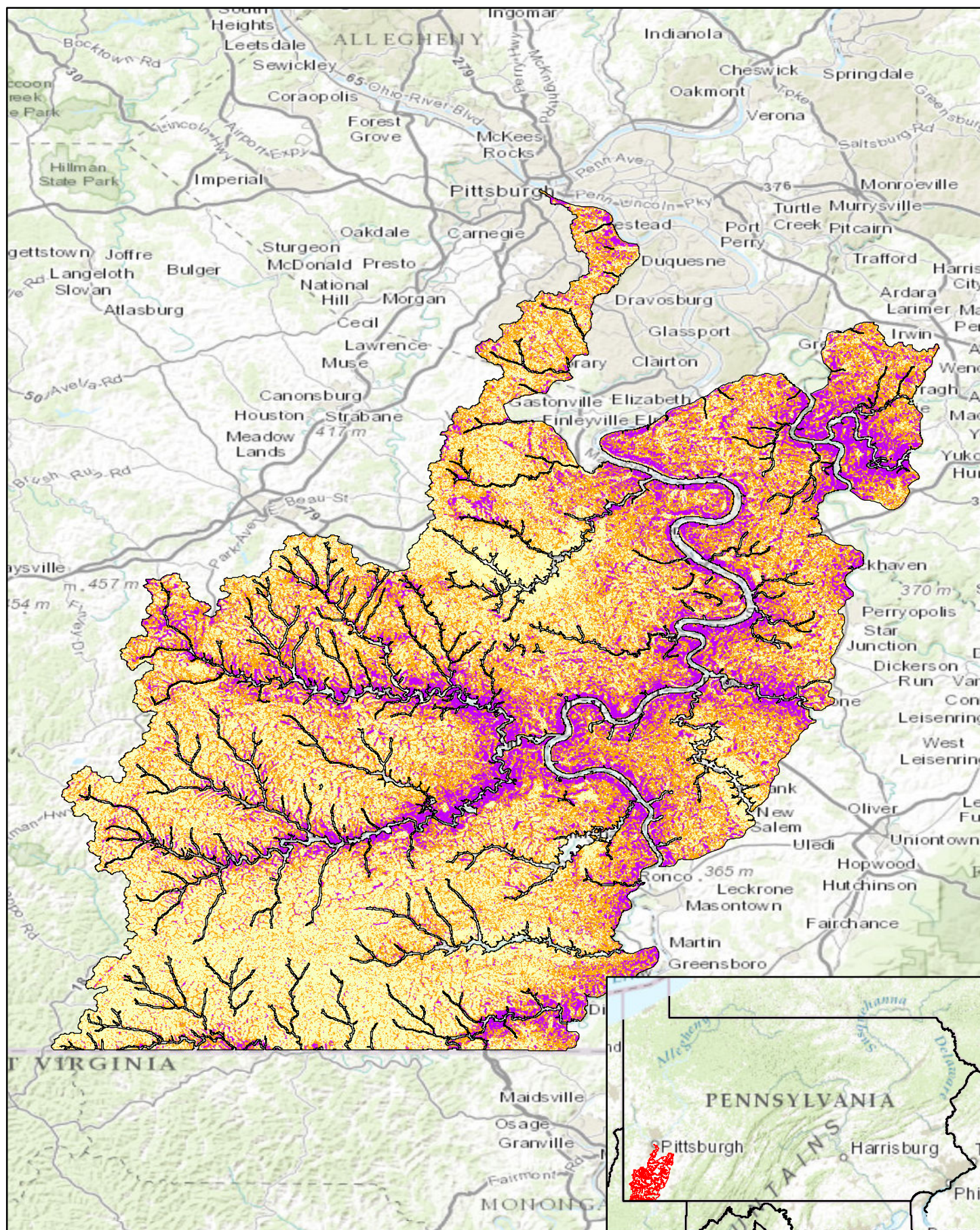


Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: riverine section 5

Sensitivity
 High
 Moderate
 Low

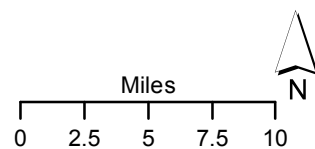
Miles
 0 2 4 6 8

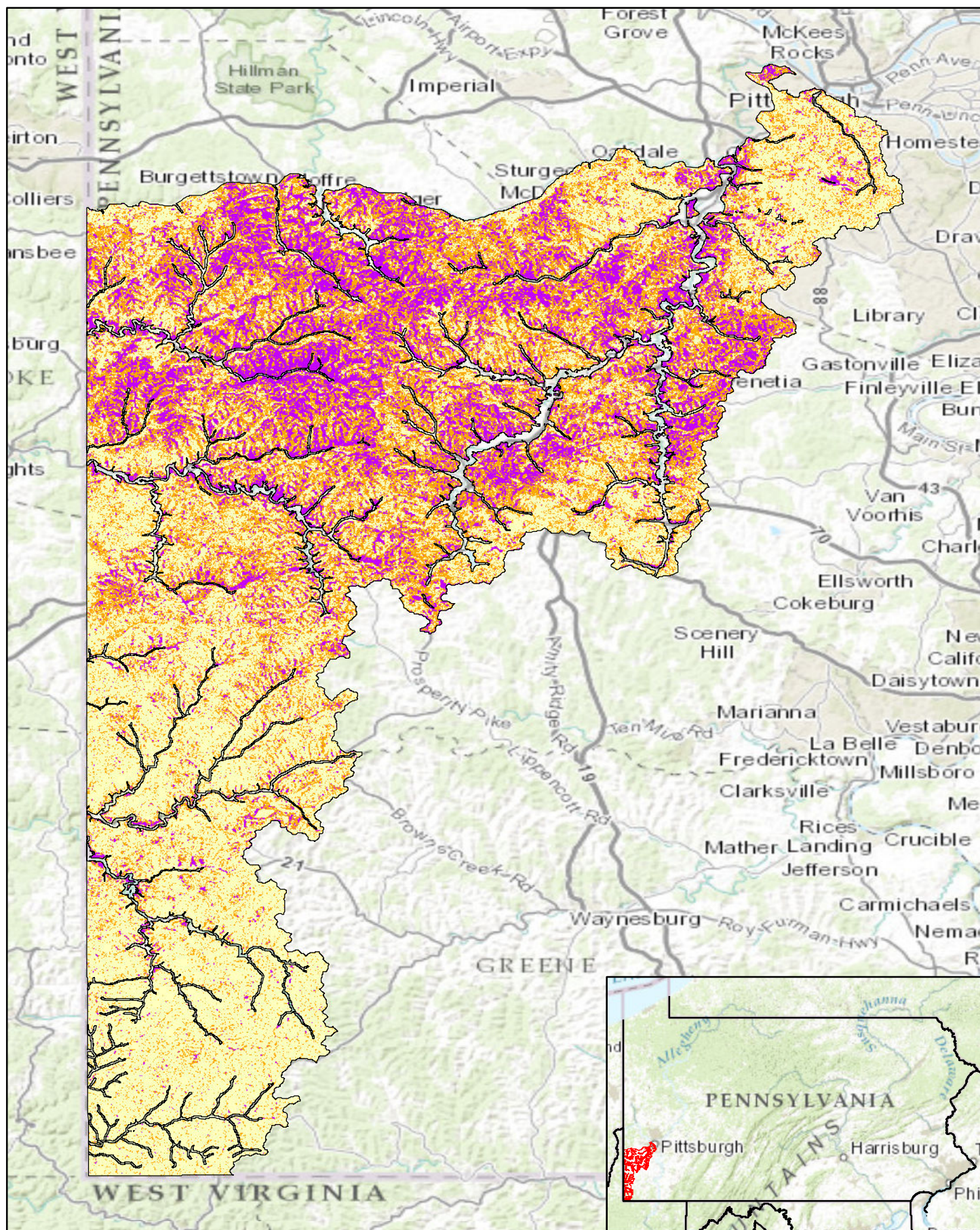




Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: upland section 1

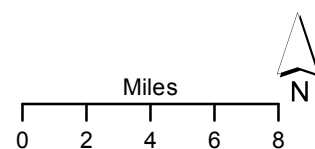
Sensitivity
 High
 Moderate
 Low

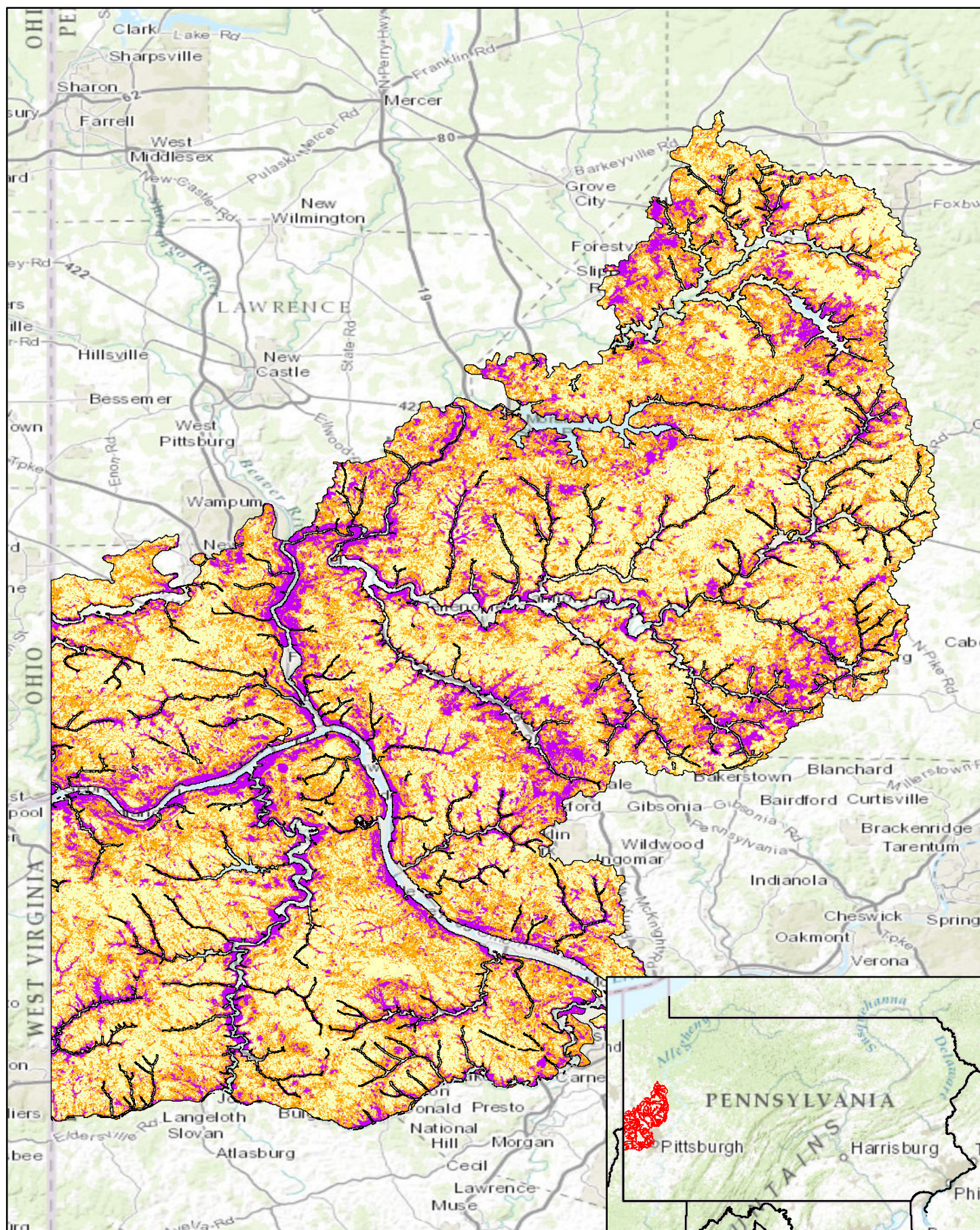




Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: upland section 2

Sensitivity
 High
 Moderate
 Low

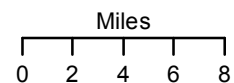


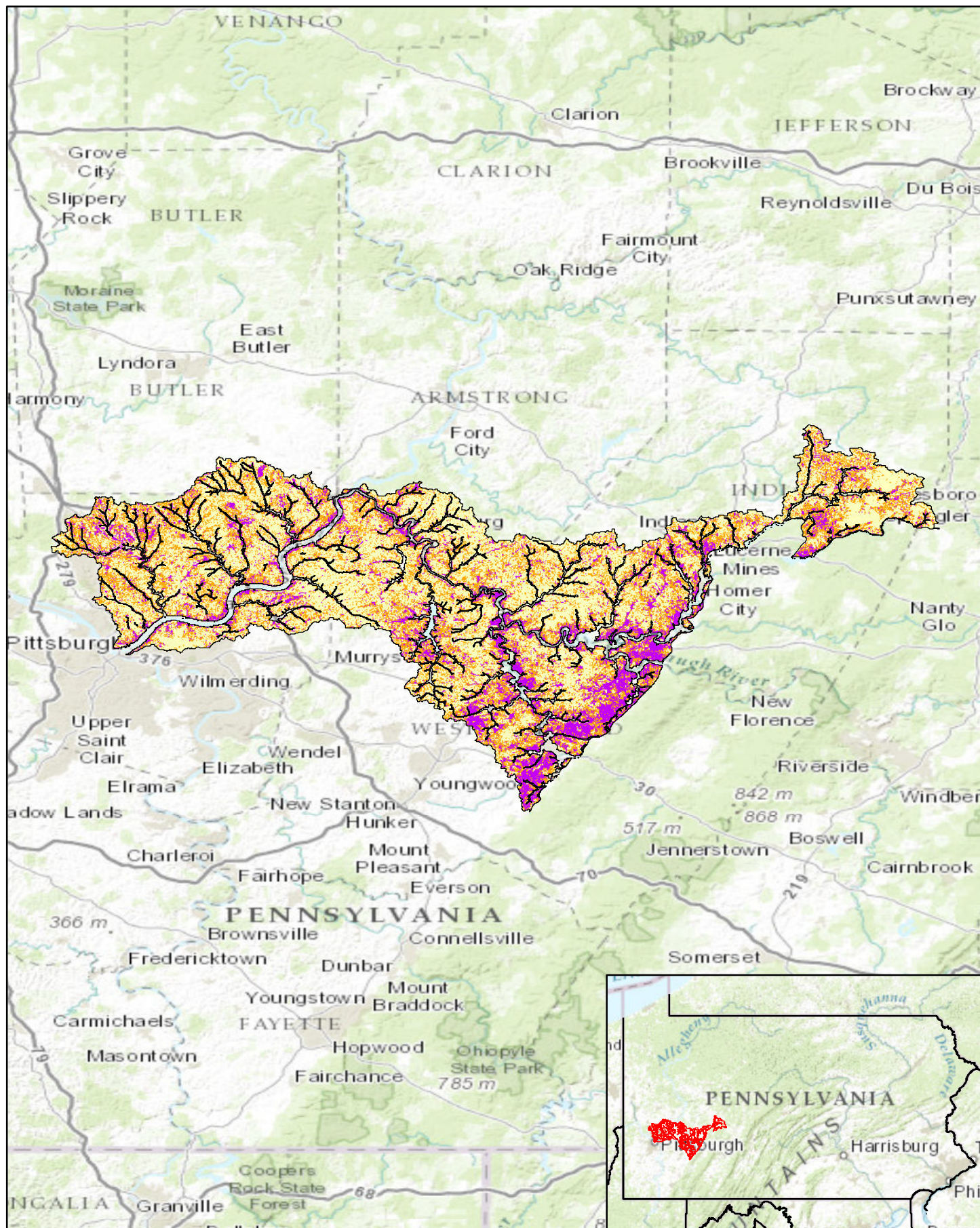


Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: upland section 3

Sensitivity

- High
- Moderate
- Low





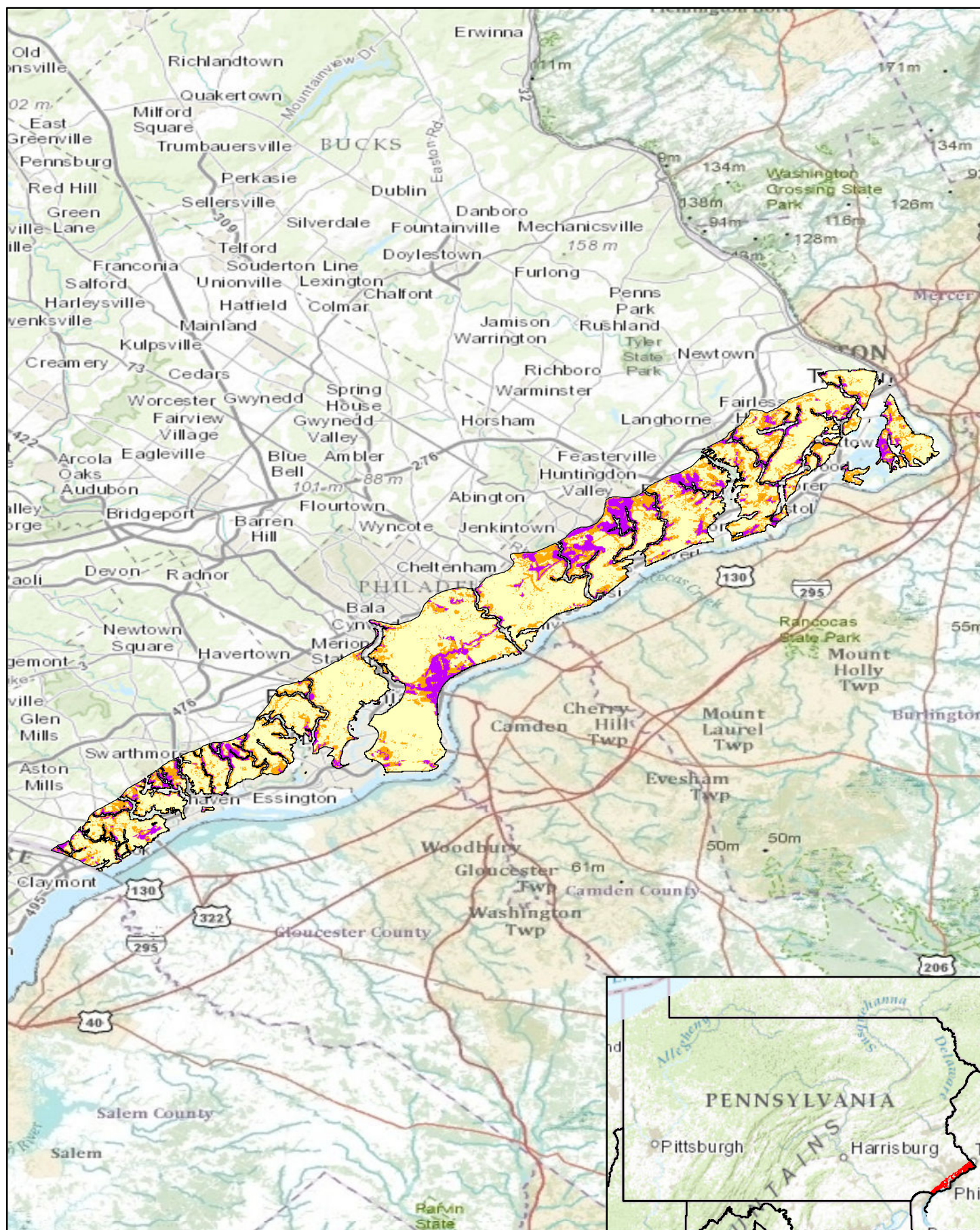
Pennsylvania Predictive Model Set
 Region: 1, Zone: west, Subarea: upland section 4

Sensitivity

- High
- Moderate
- Low

Miles
 0 4 8 12 16



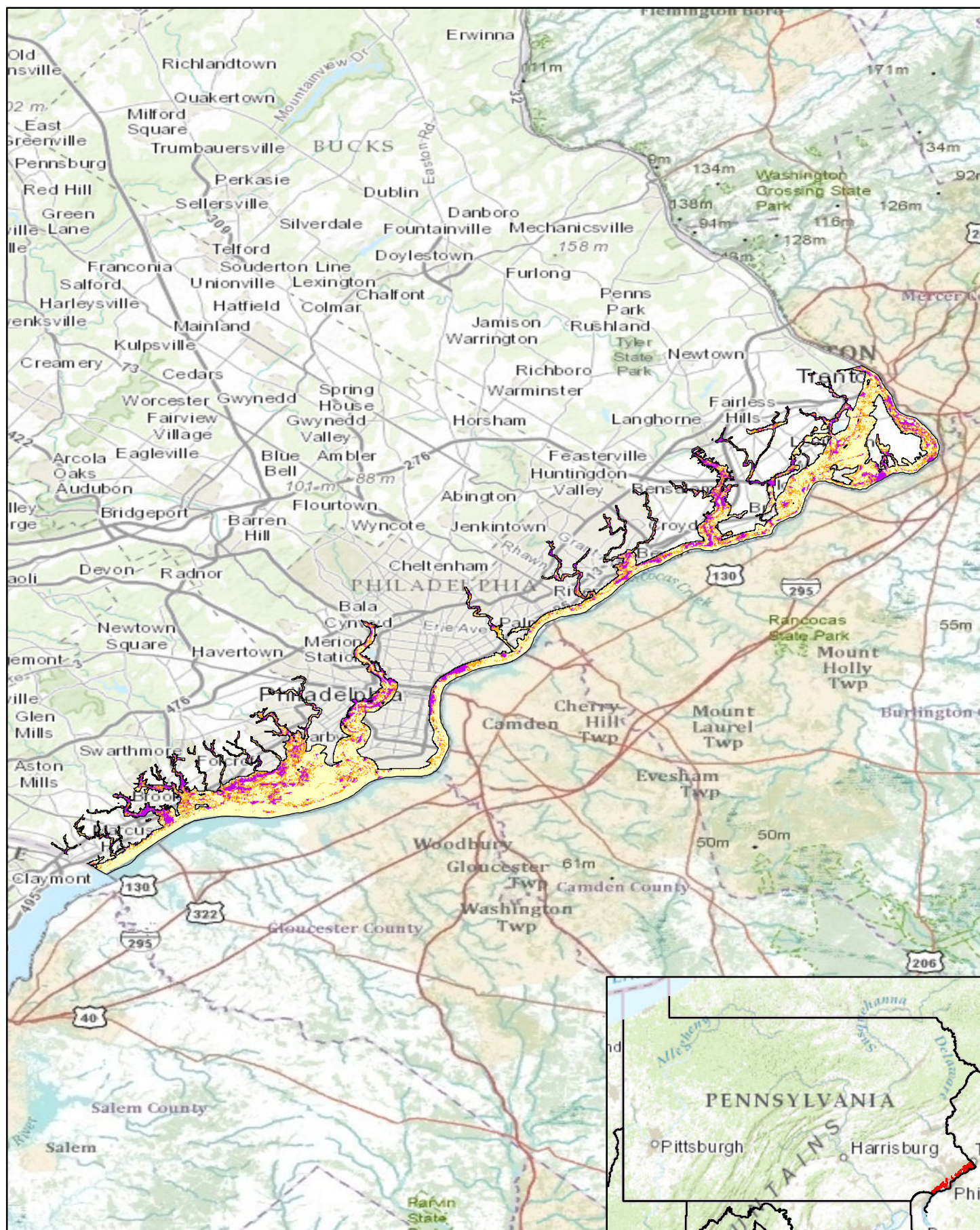


Pennsylvania Predictive Model Set
 Region: 10, Zone: all, Subarea: upland section 9

Sensitivity
 High
 Moderate
 Low

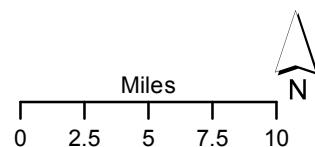
Miles
 0 2 4 6 8

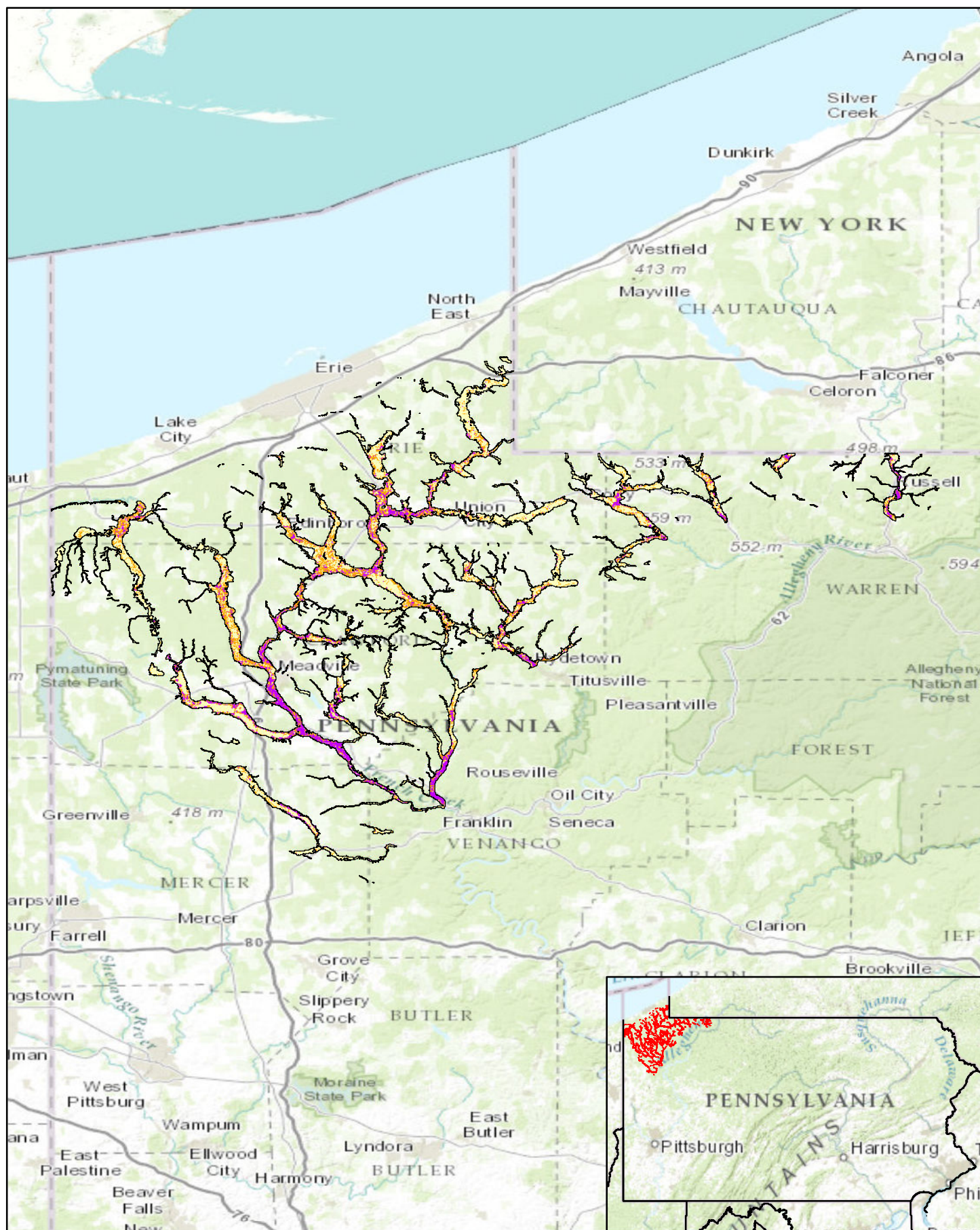




Pennsylvania Predictive Model Set
 Region: 10, Zone: all, Subarea: riverine section 9

Sensitivity
 High
 Moderate
 Low

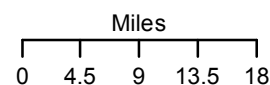


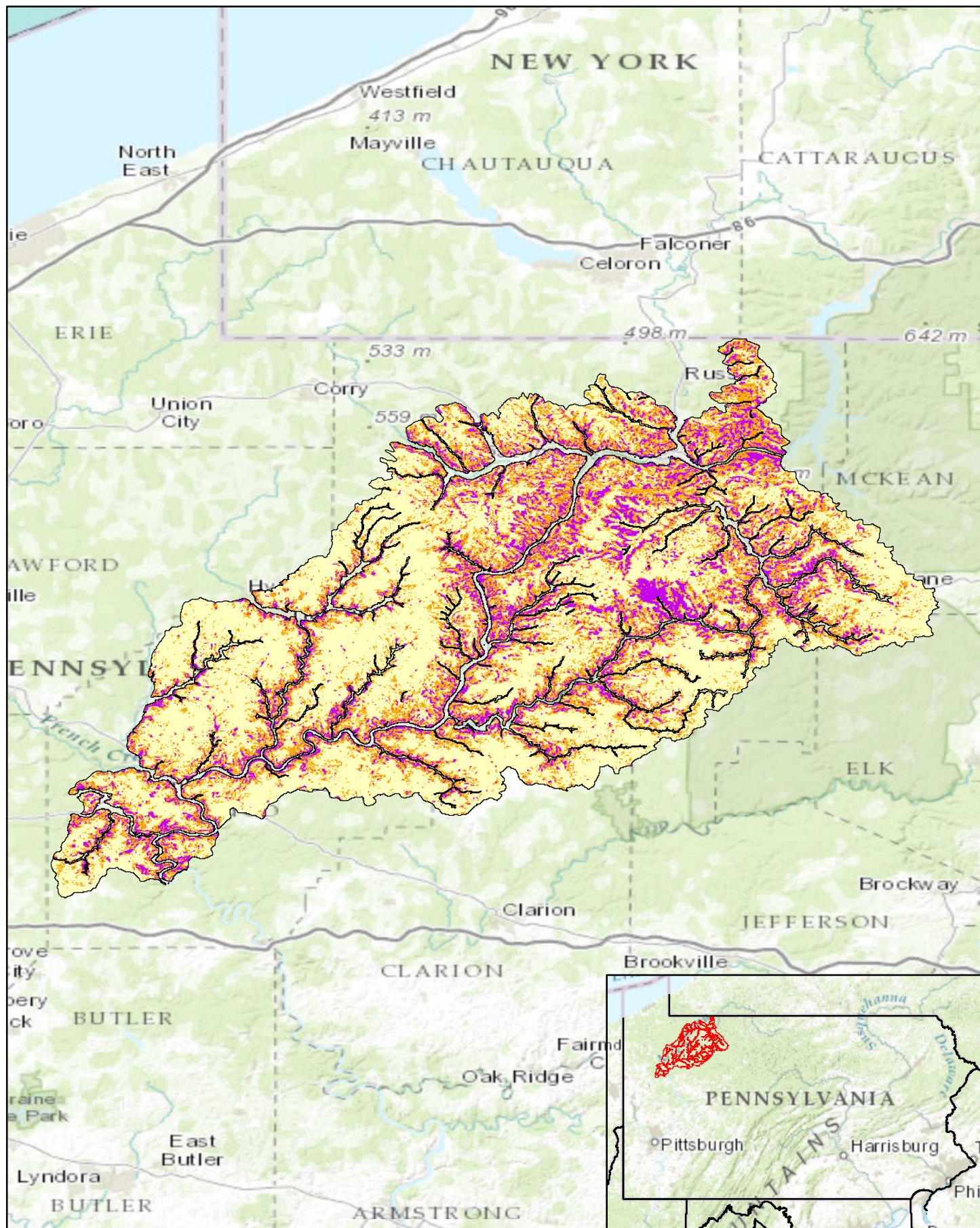


Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: riverine section 1

Sensitivity

- High
- Moderate
- Low

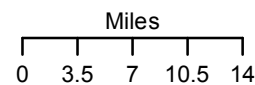


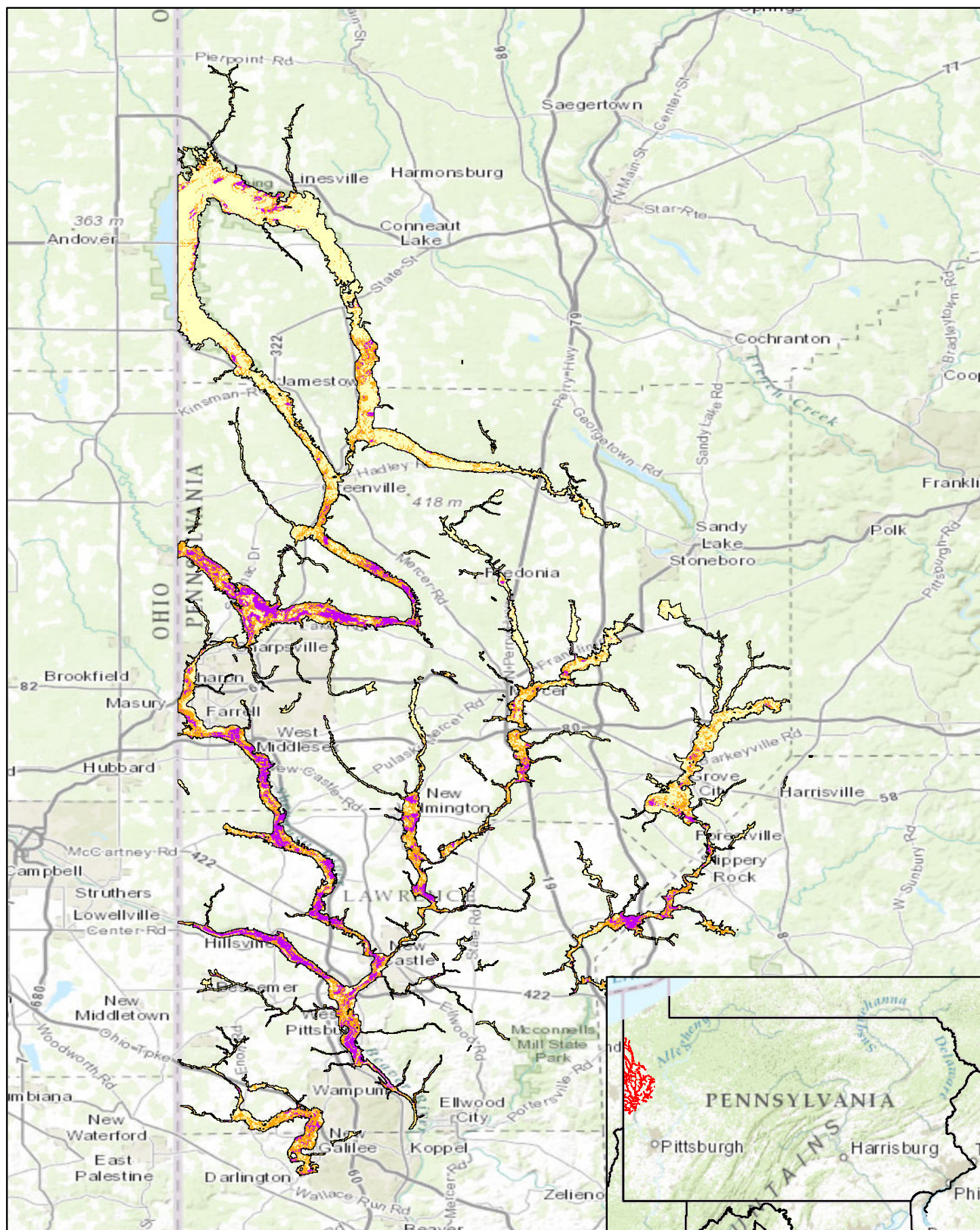


Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: upland section 5

Sensitivity

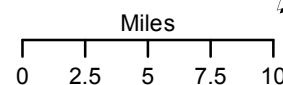
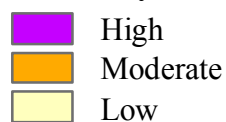
- High
- Moderate
- Low

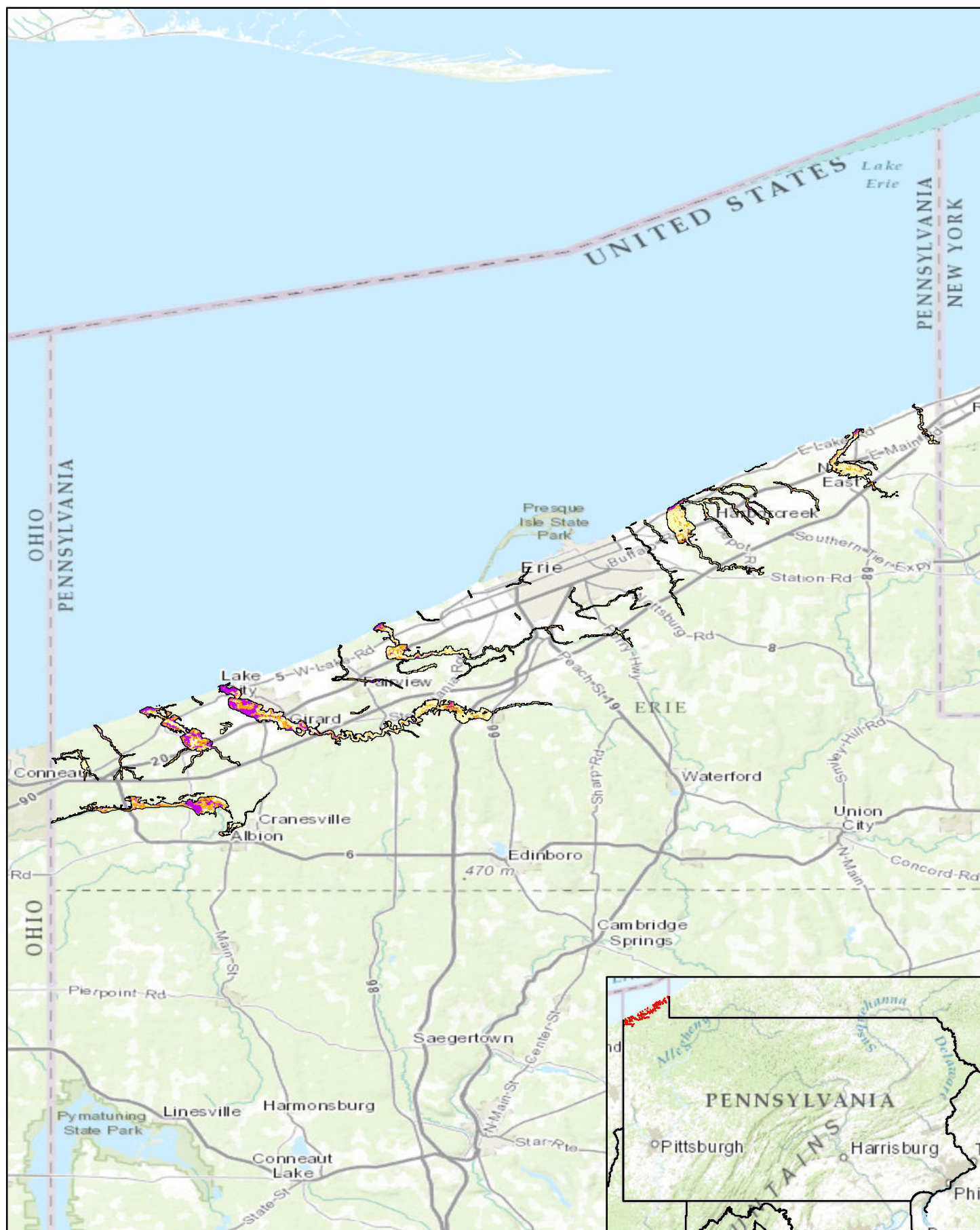




Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: riverine section 2

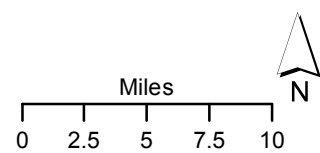
Sensitivity

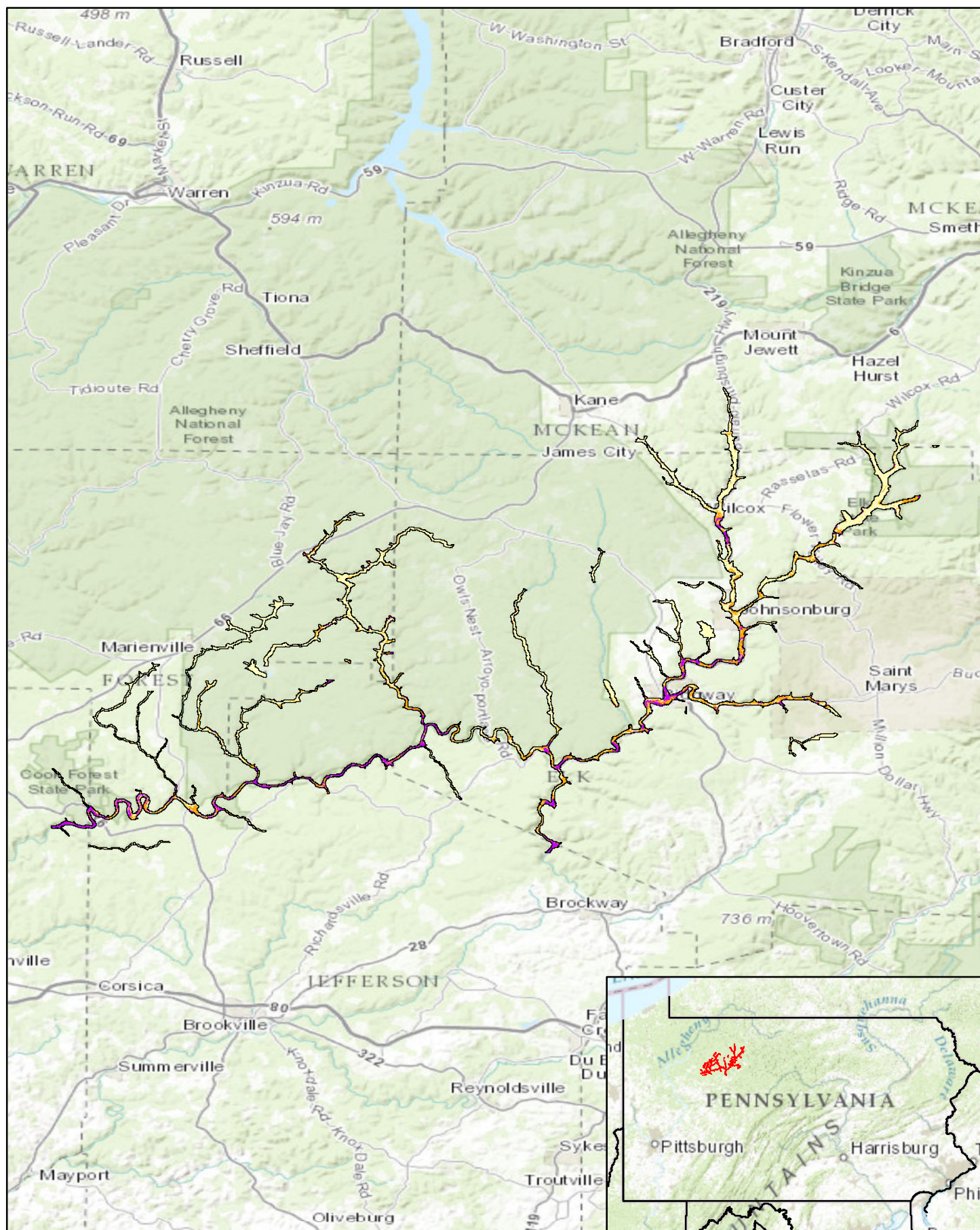




Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: riverine section 3

Sensitivity
 High
 Moderate
 Low

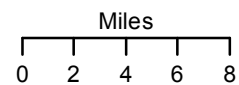


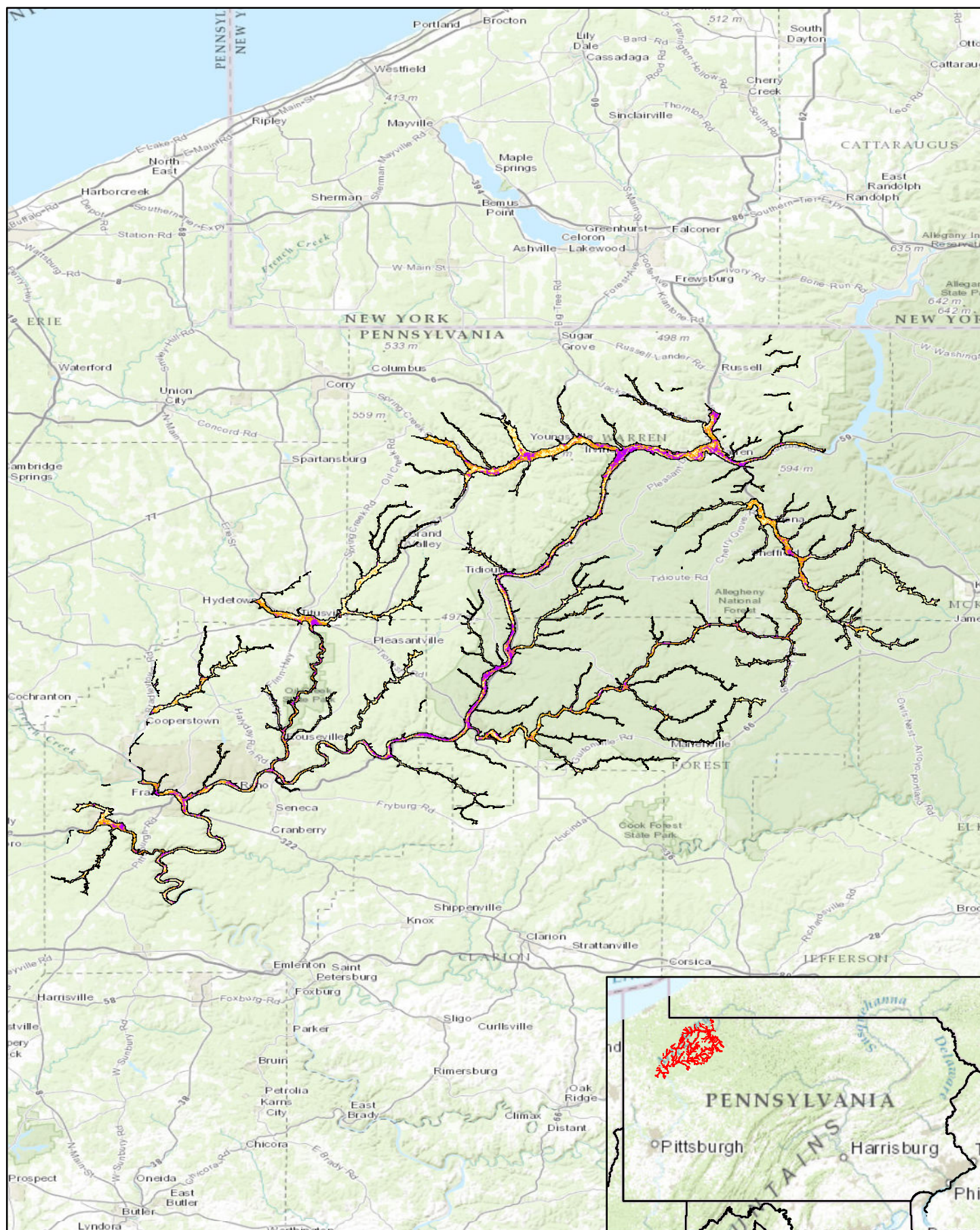


Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: riverine section 4

Sensitivity

- High
- Moderate
- Low

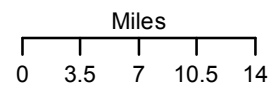


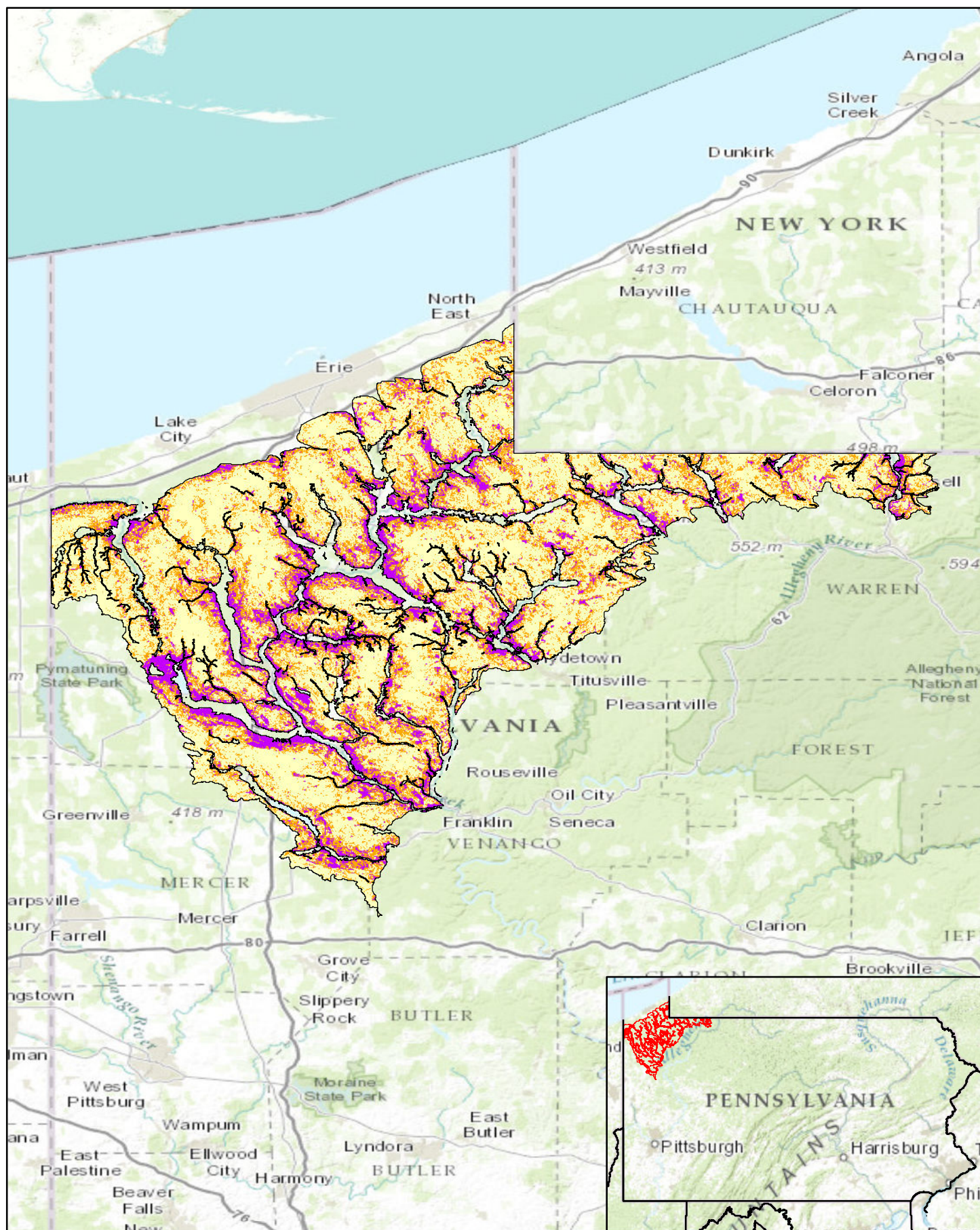


Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: riverine section 5

Sensitivity

- High
- Moderate
- Low

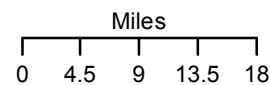


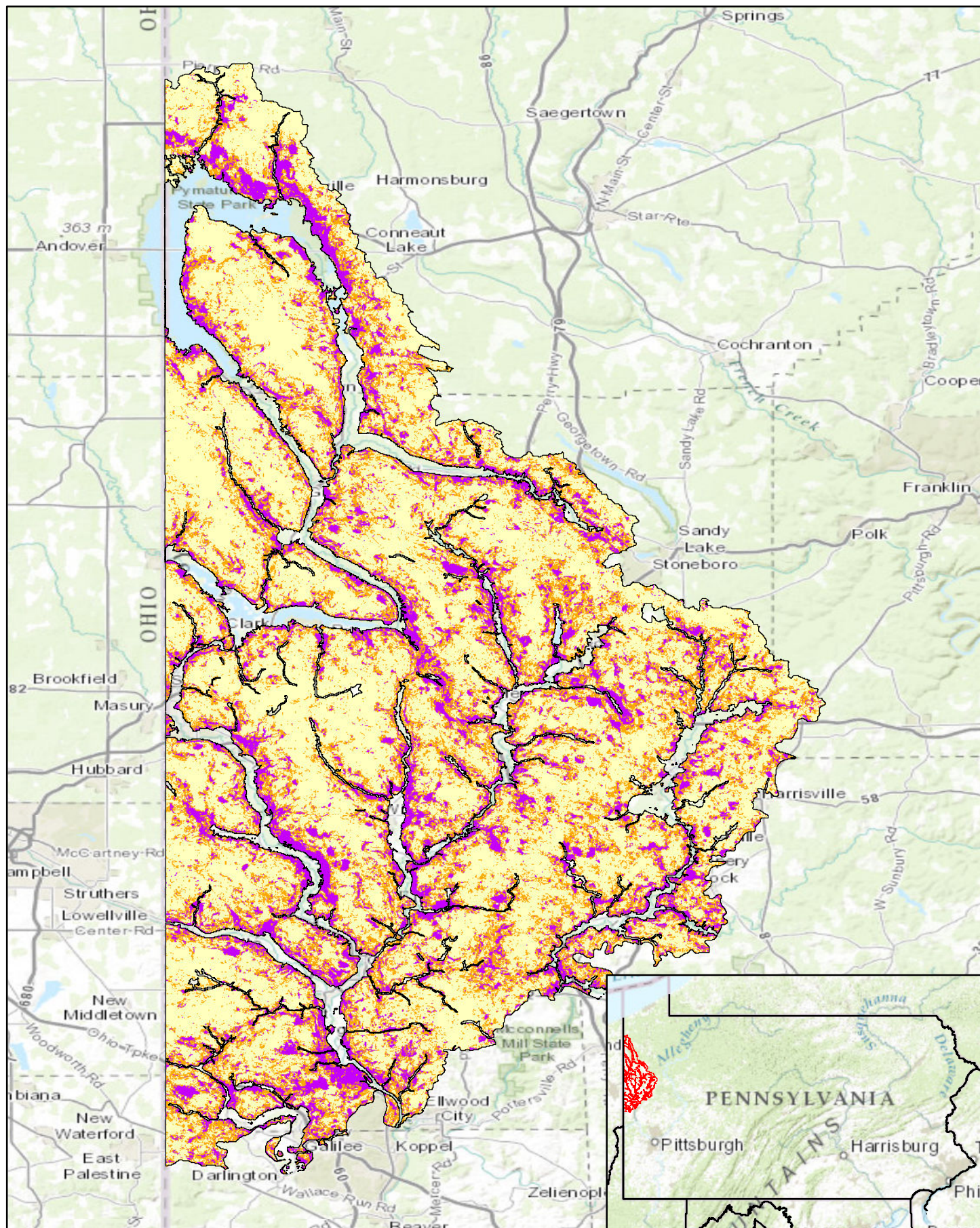


Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: upland section 1

Sensitivity

- High
- Moderate
- Low

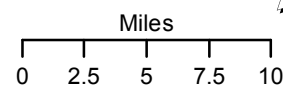


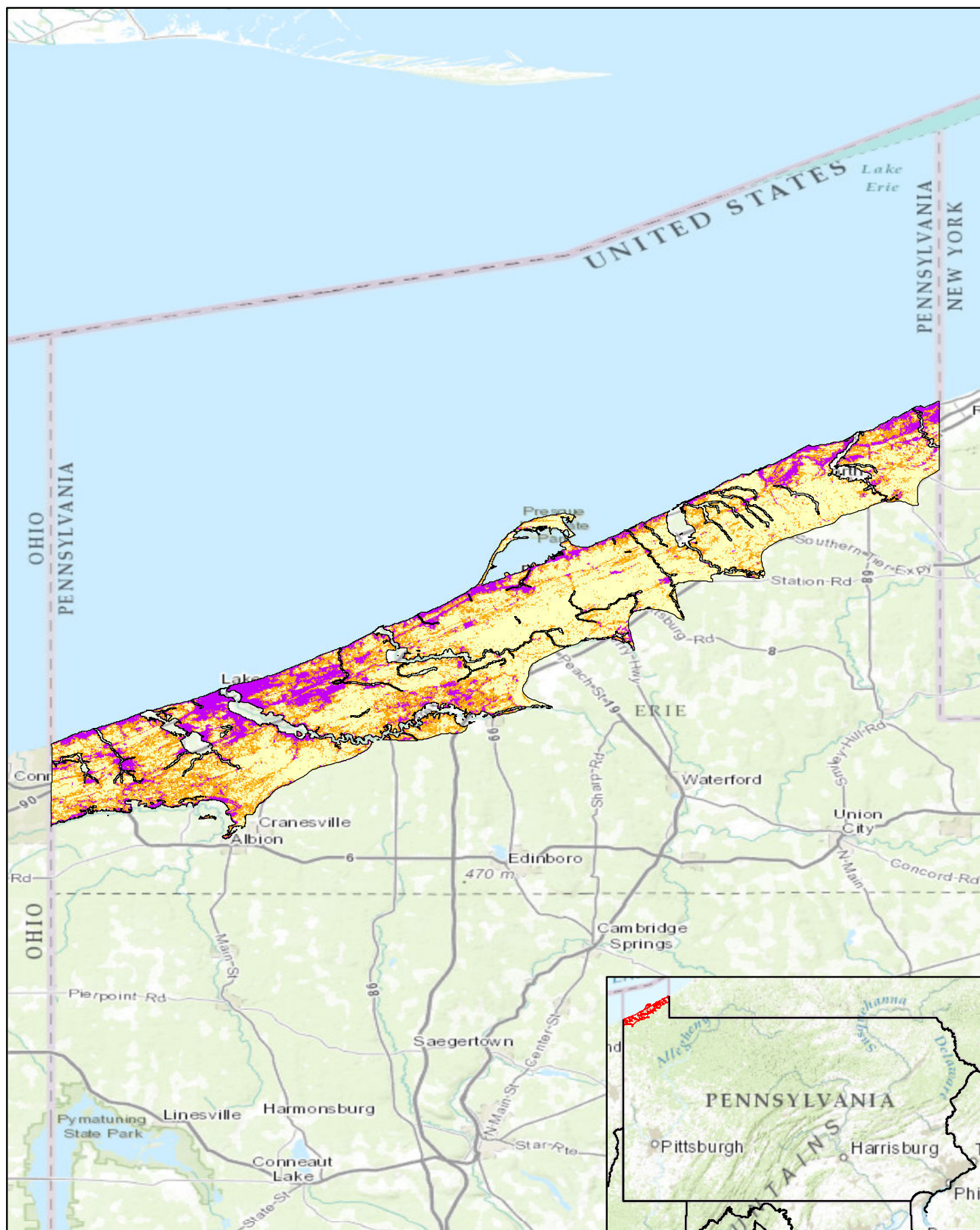


Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: upland section 2

Sensitivity

- High
- Moderate
- Low

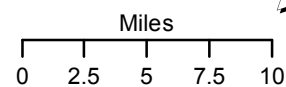


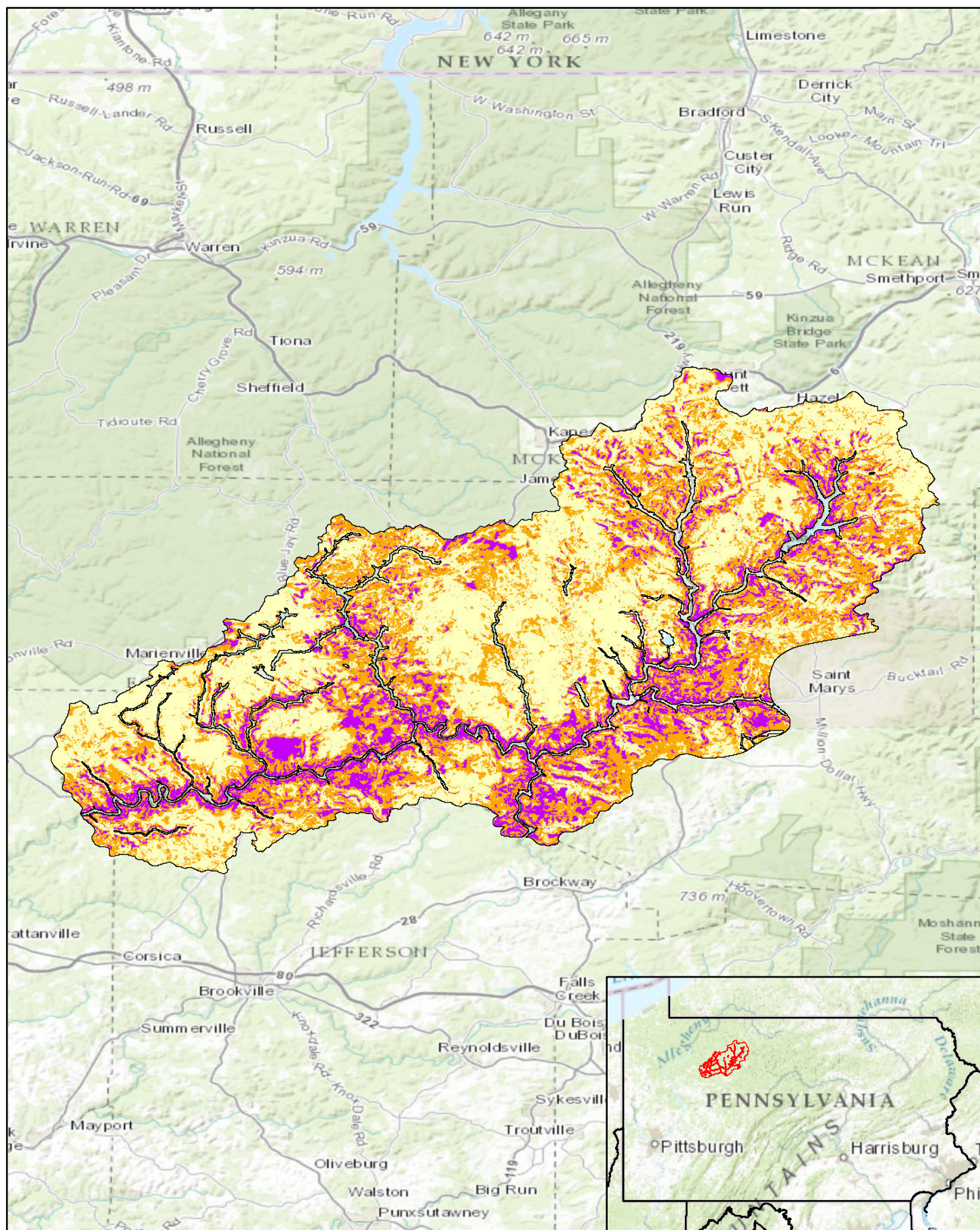


Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: upland section 3

Sensitivity

- High
- Moderate
- Low

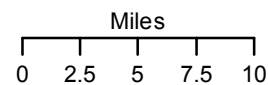


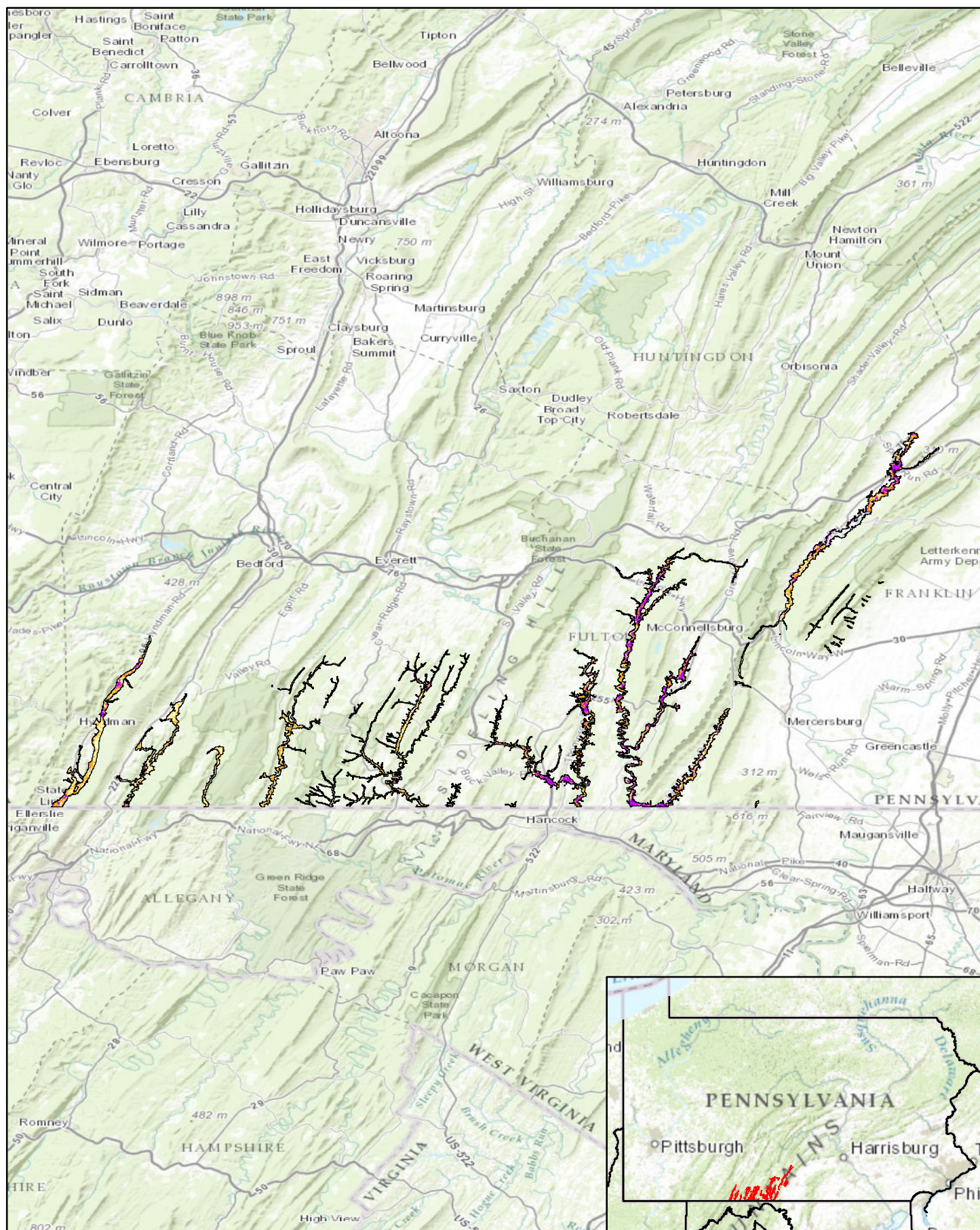


Pennsylvania Predictive Model Set
 Region: 23, Zone: all, Subarea: upland section 4

Sensitivity

- High
- Moderate
- Low

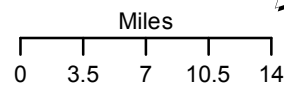


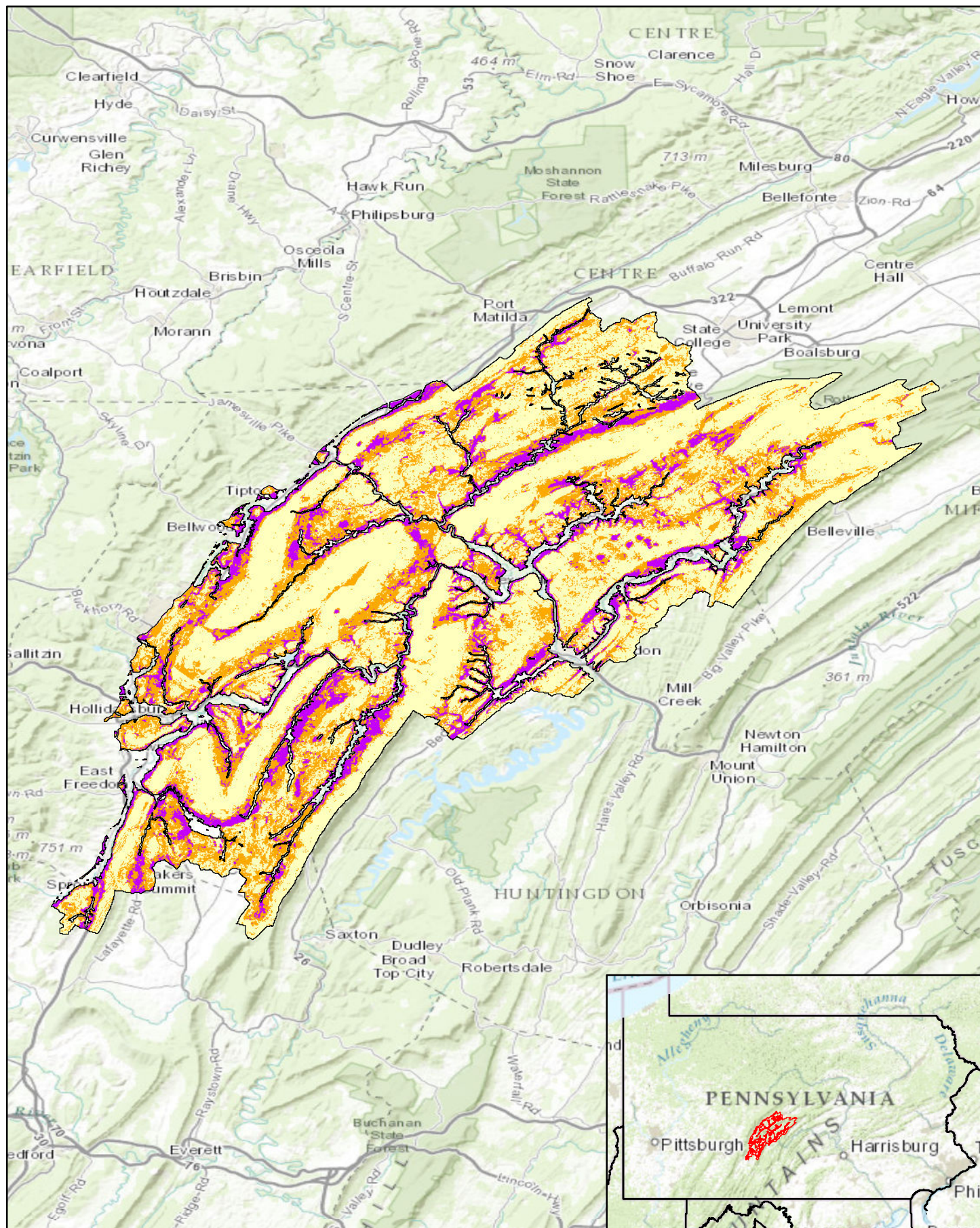


Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: riverine section 1

Sensitivity

- High
- Moderate
- Low



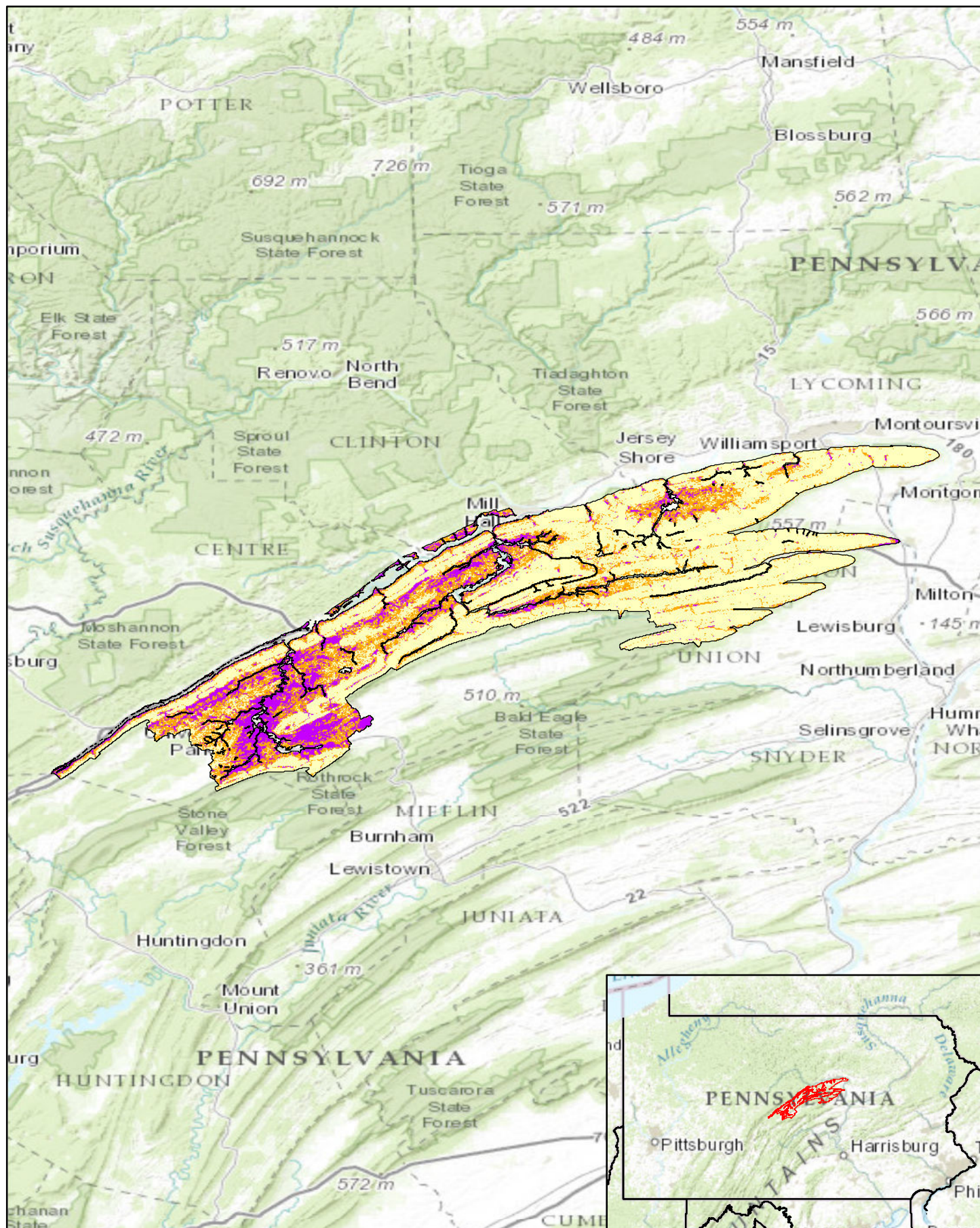


Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: upland section 4

Sensitivity
 High
 Moderate
 Low

Miles
 0 2.5 5 7.5 10

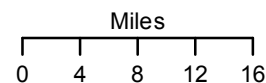


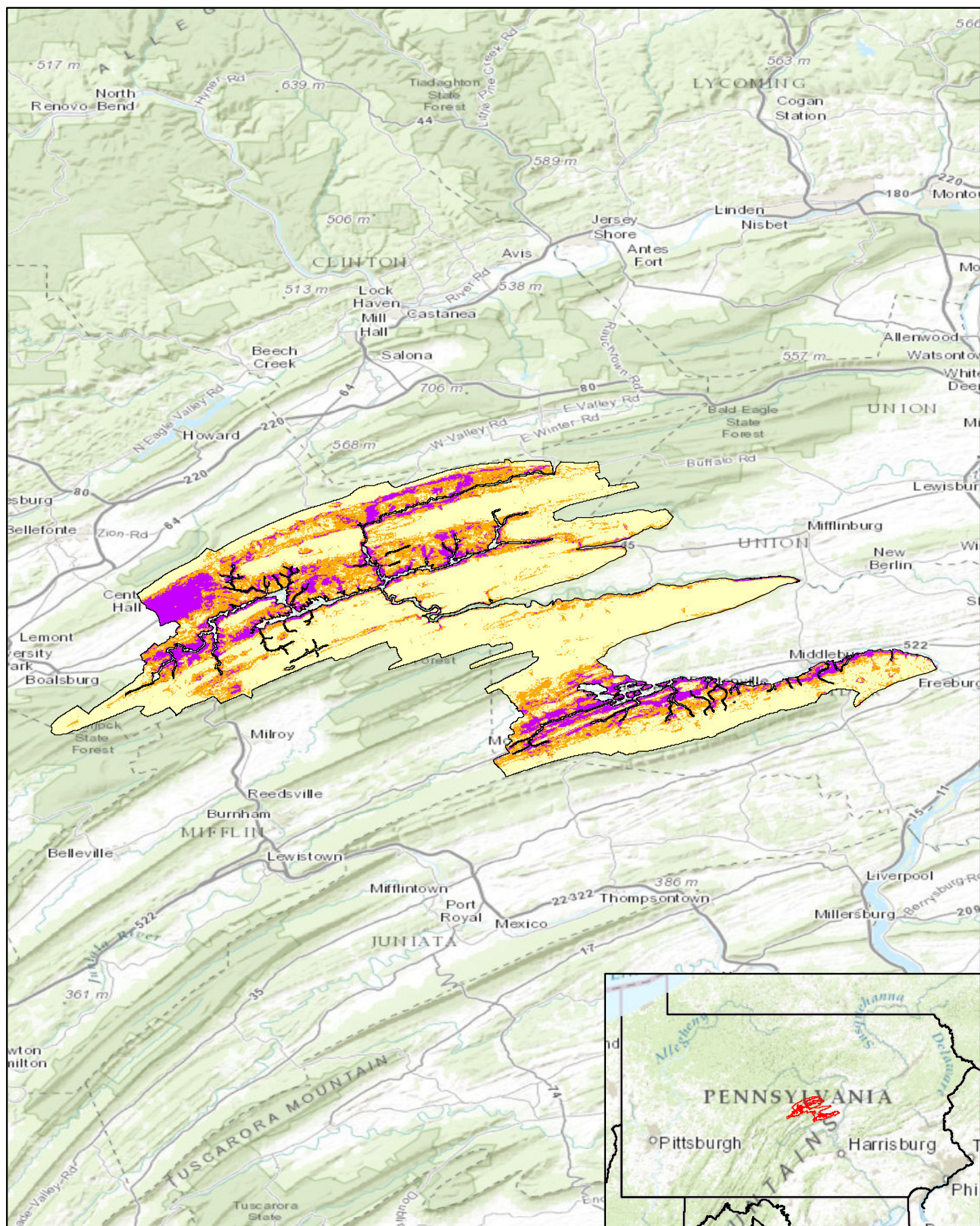


Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: upland section 5

Sensitivity

- High
- Moderate
- Low





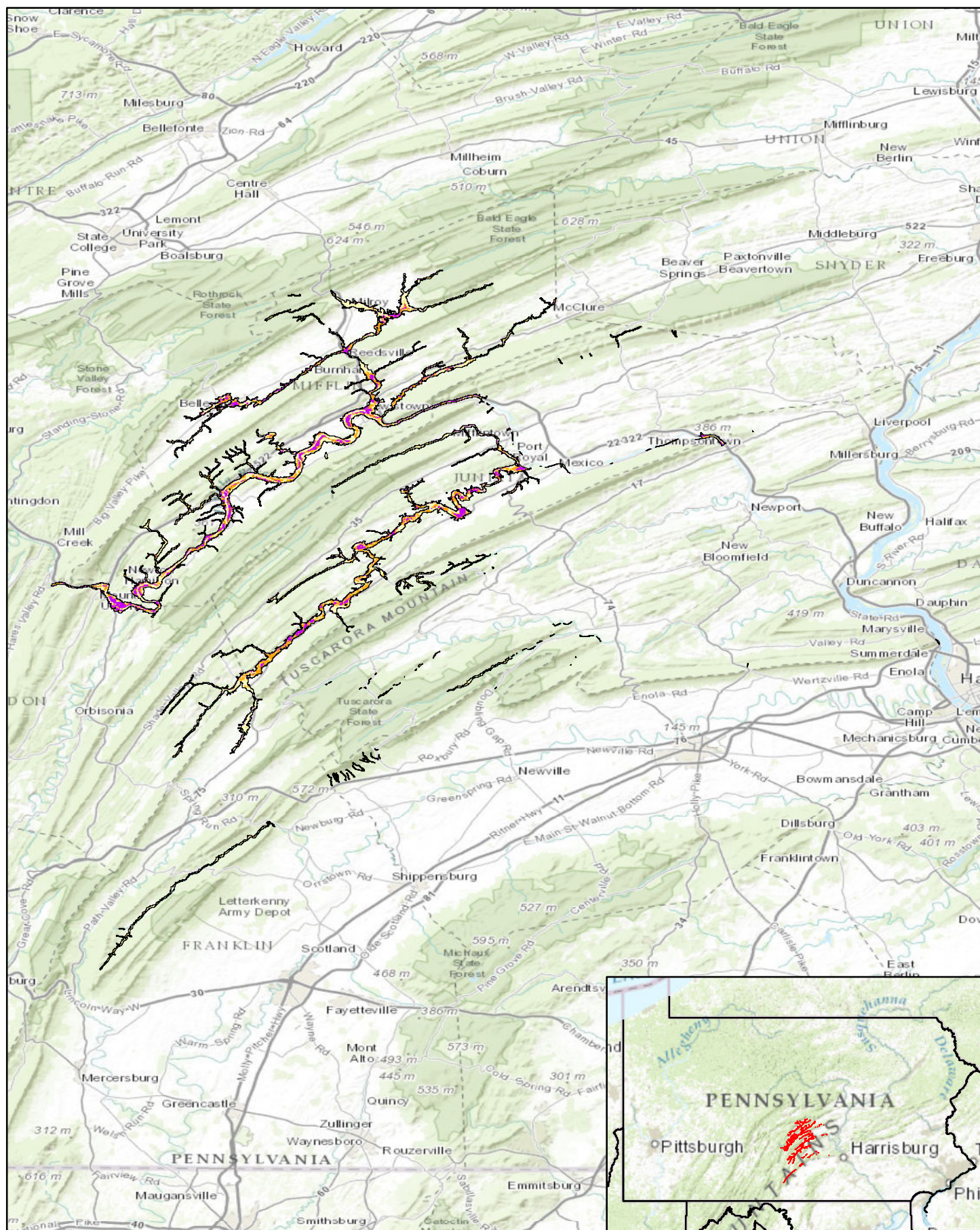
Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: upland section 6

Sensitivity

- High
- Moderate
- Low

Miles
 0 2.5 5 7.5 10

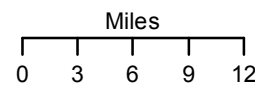


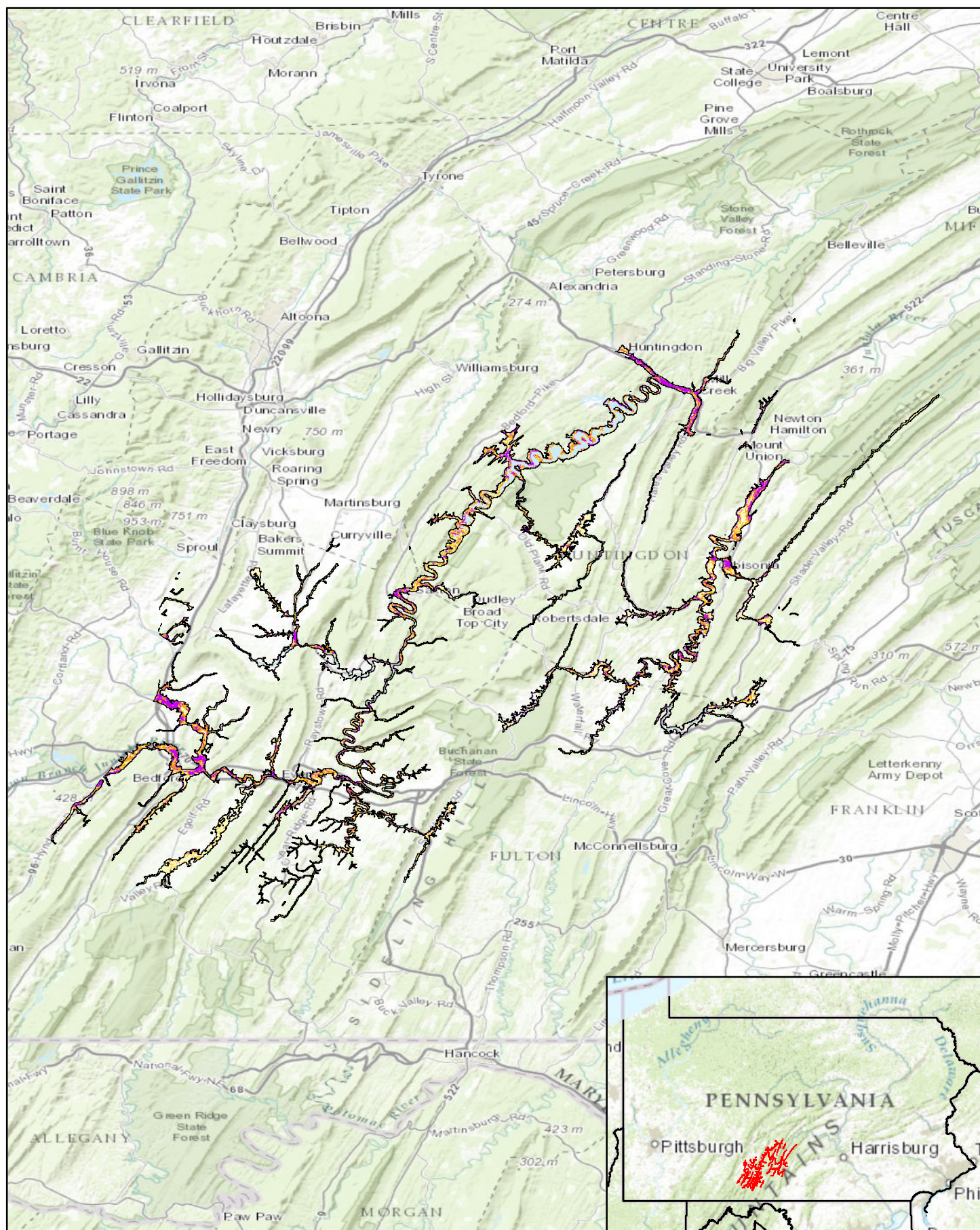


Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: riverine section 2

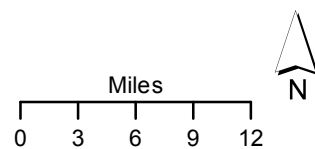
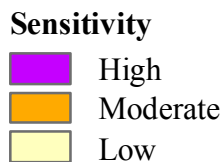
Sensitivity

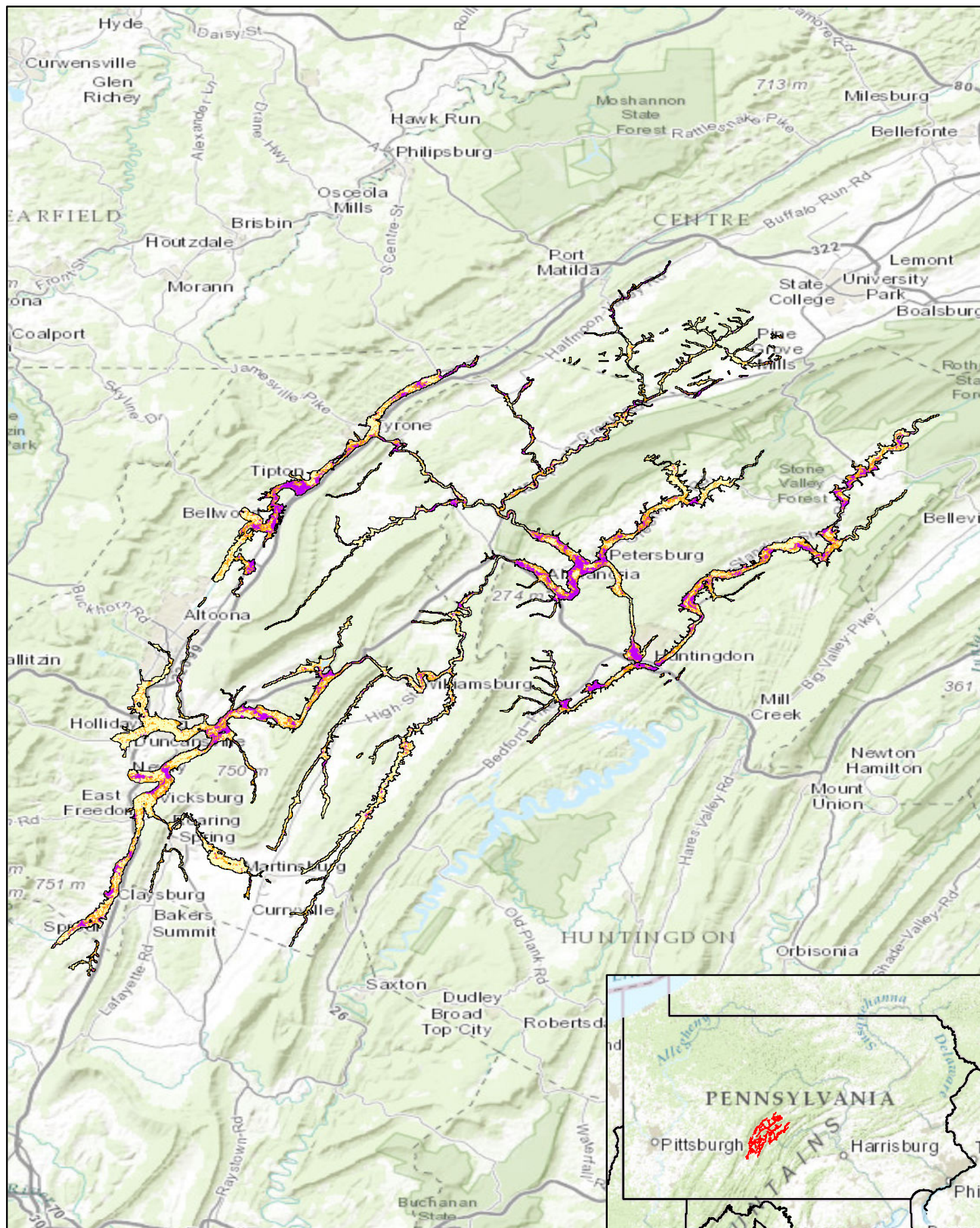
- High
- Moderate
- Low





Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: riverine section 3

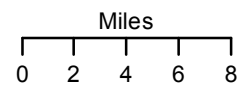


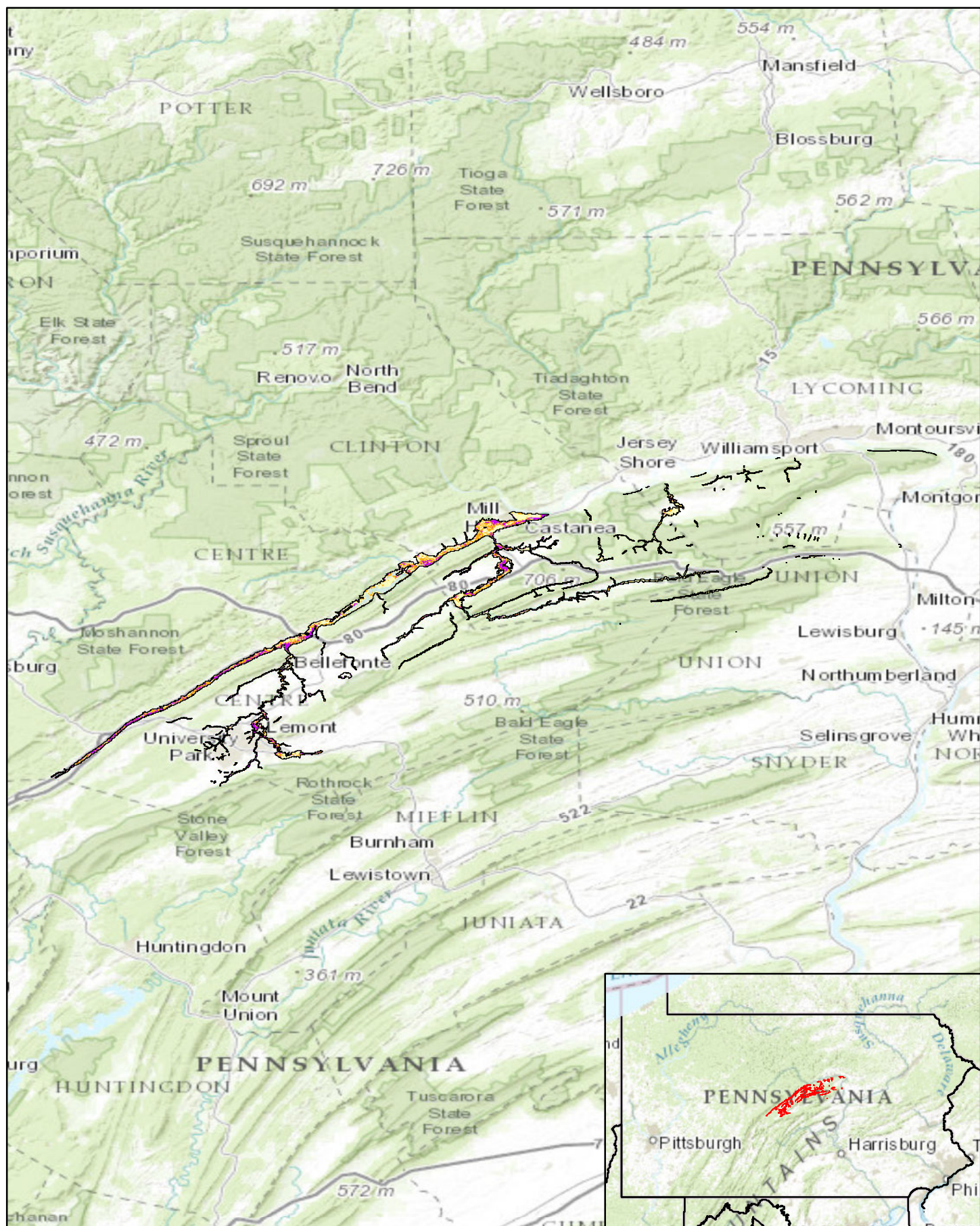


Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: riverine section 4

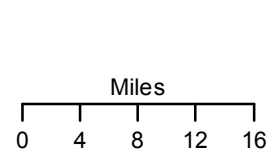
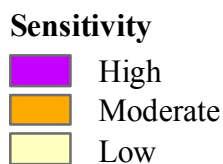
Sensitivity

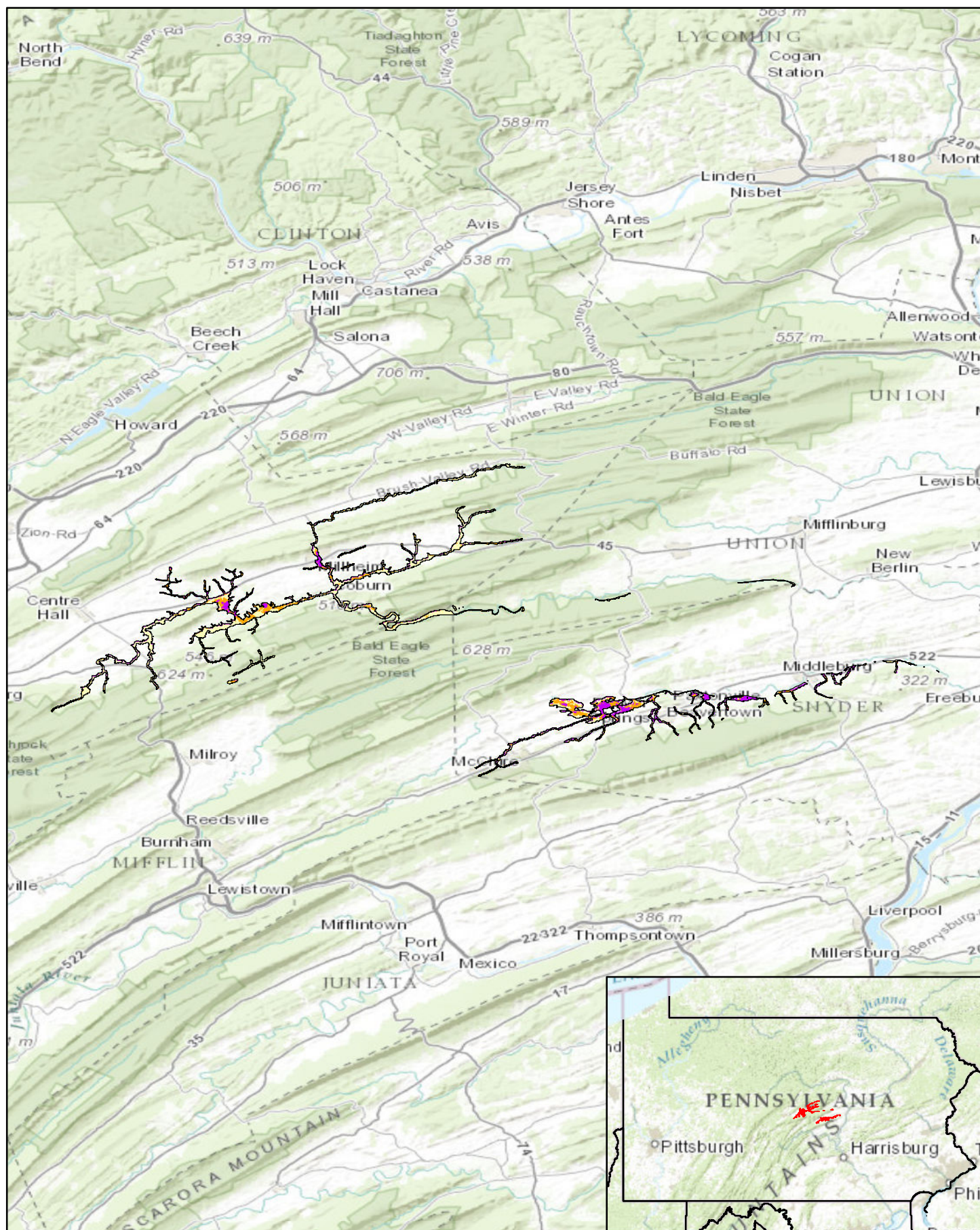
- High
- Moderate
- Low





Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: riverine section 5

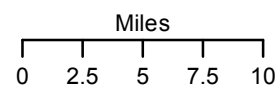


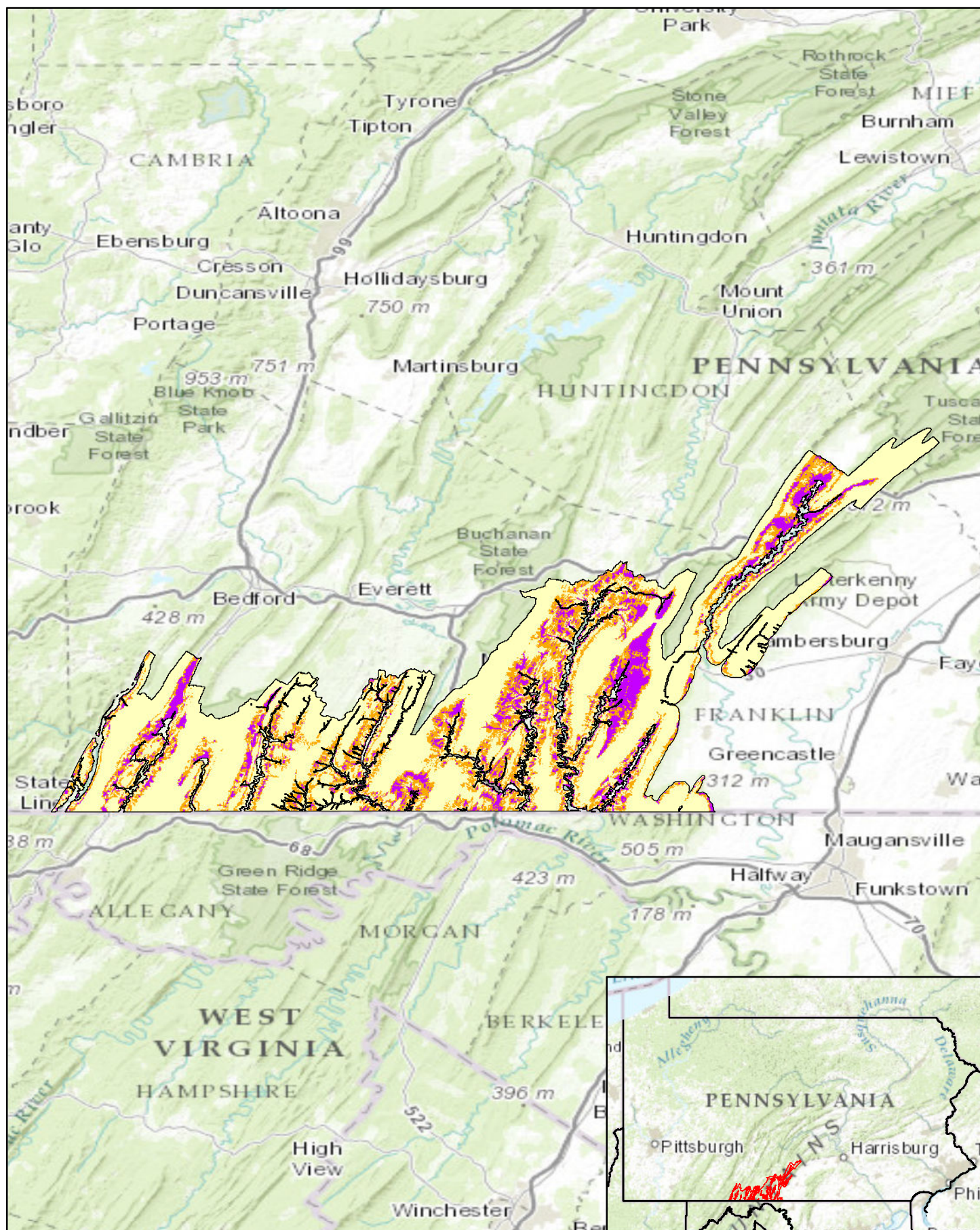


Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: riverine section 6

Sensitivity

- High
- Moderate
- Low

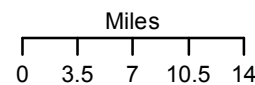


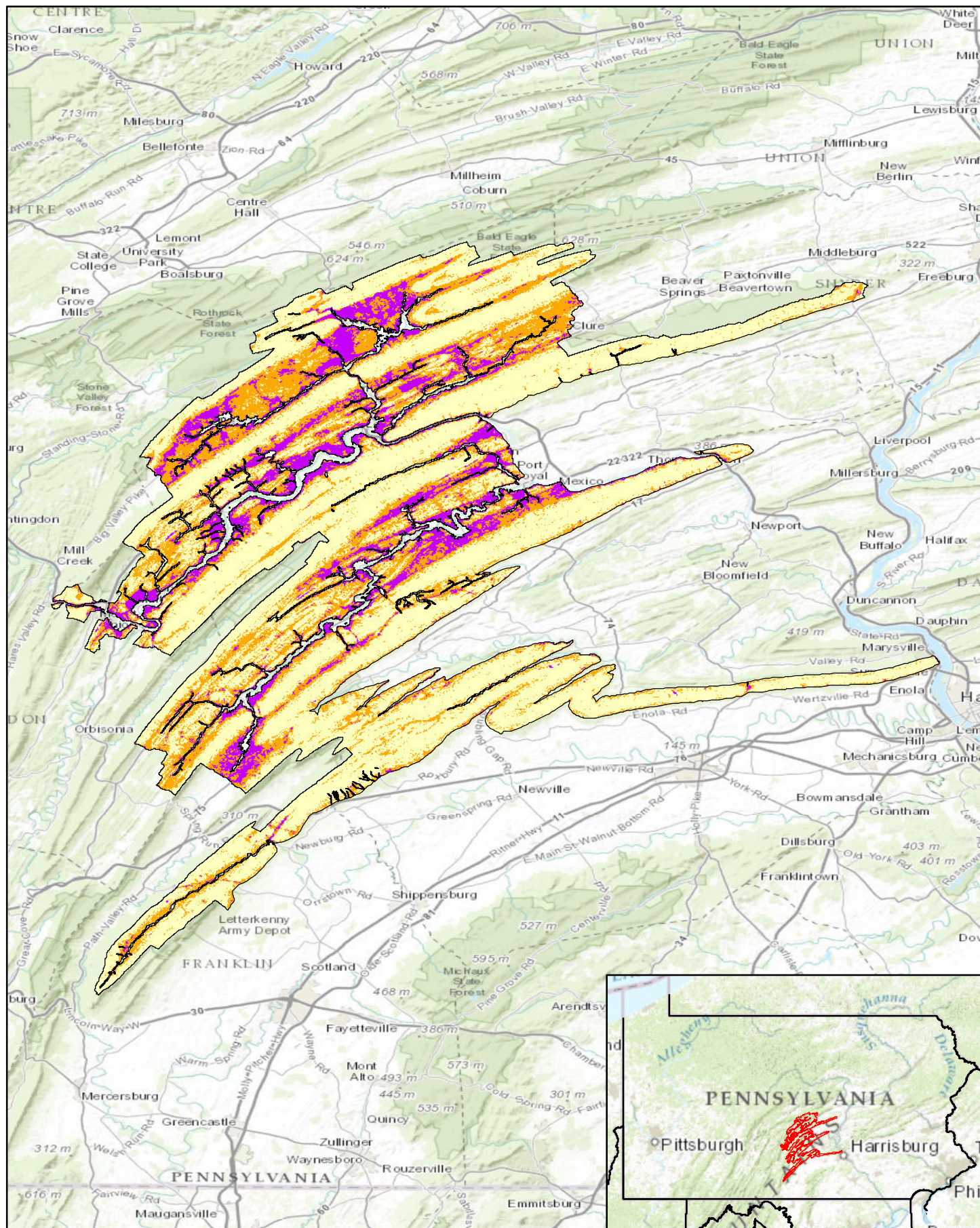


Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: upland section 1

Sensitivity

- High
- Moderate
- Low

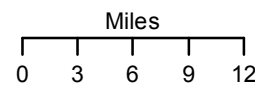


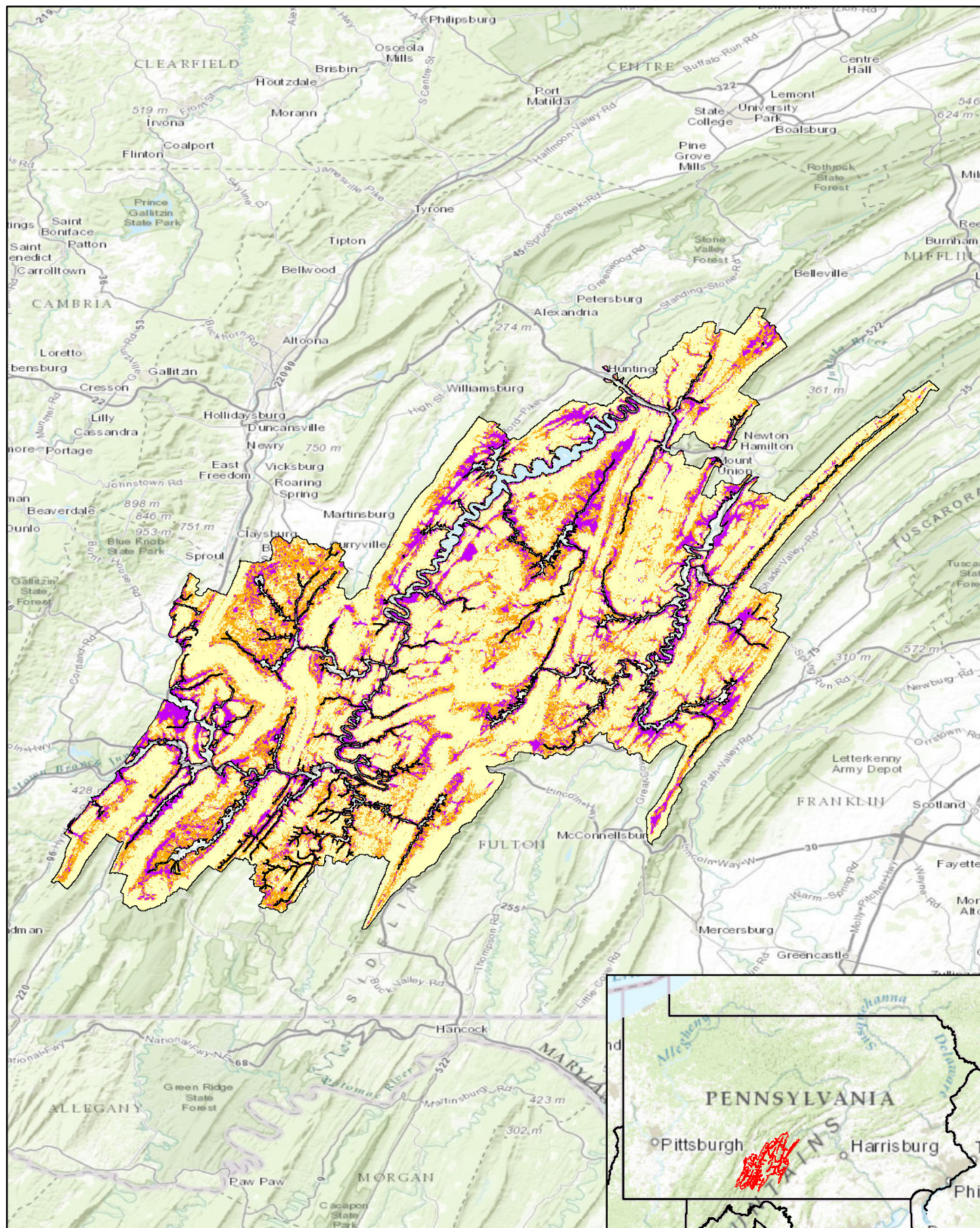


Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: upland section 2

Sensitivity

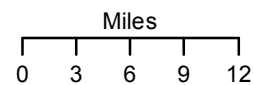
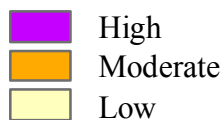
- High
- Moderate
- Low

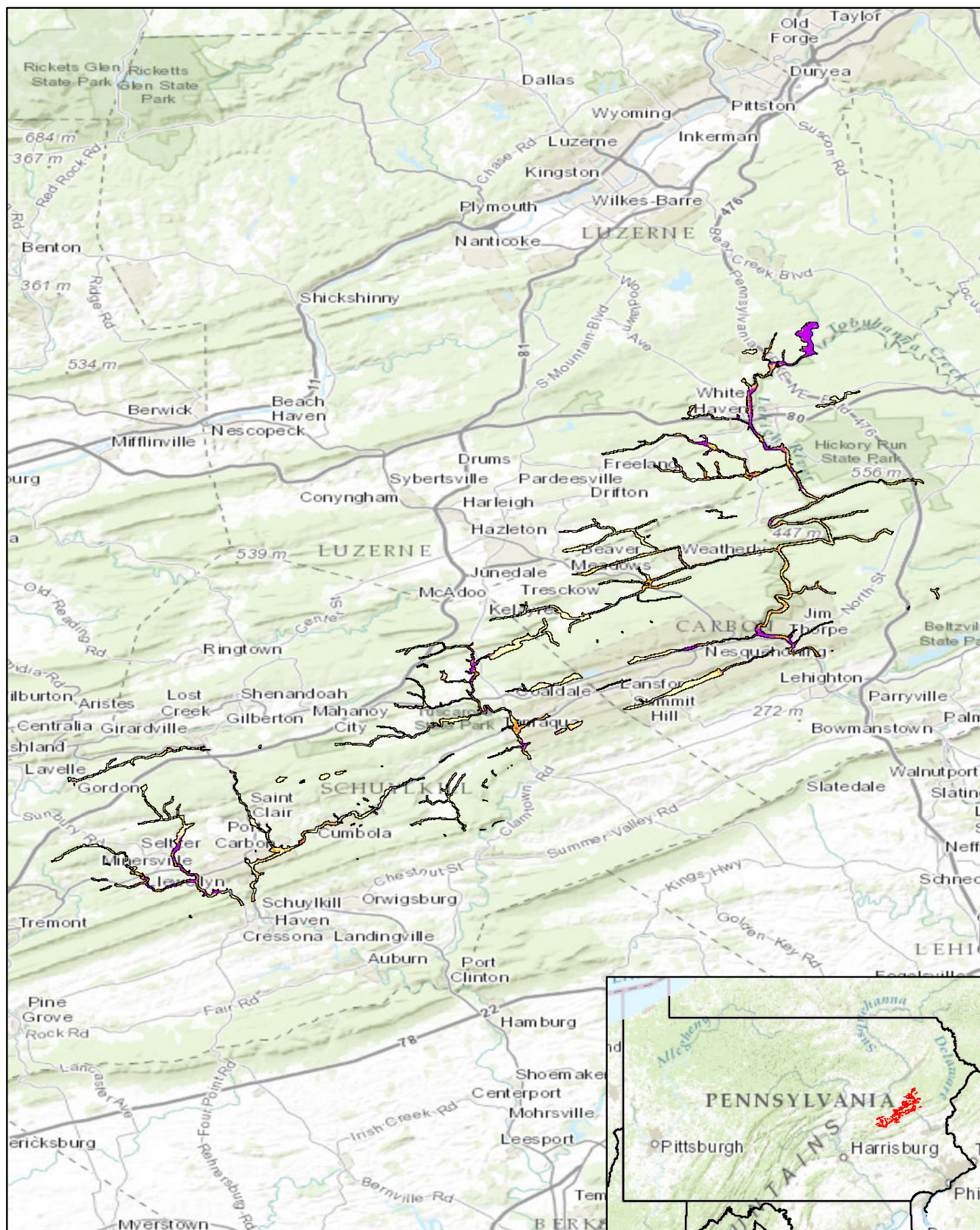




Pennsylvania Predictive Model Set
 Region: 4, Zone: west, Subarea: upland section 3

Sensitivity





Pennsylvania Predictive Model Set

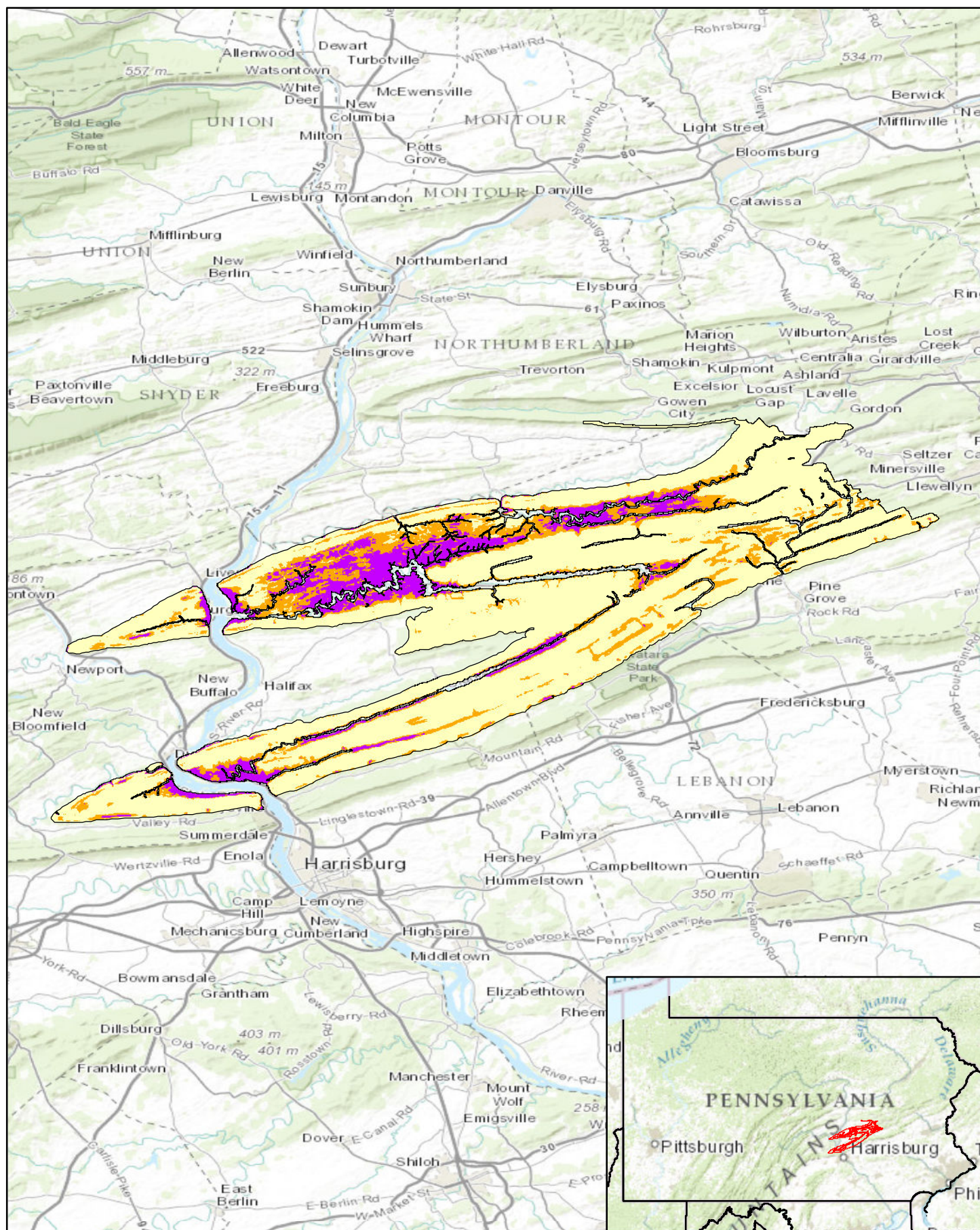
Region: 5, Zone: east, Subarea: riverine section 1

Sensitivity

- High
- Moderate
- Low

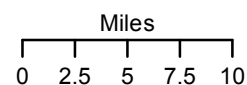
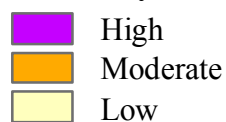
Miles
0 2 4 6 8

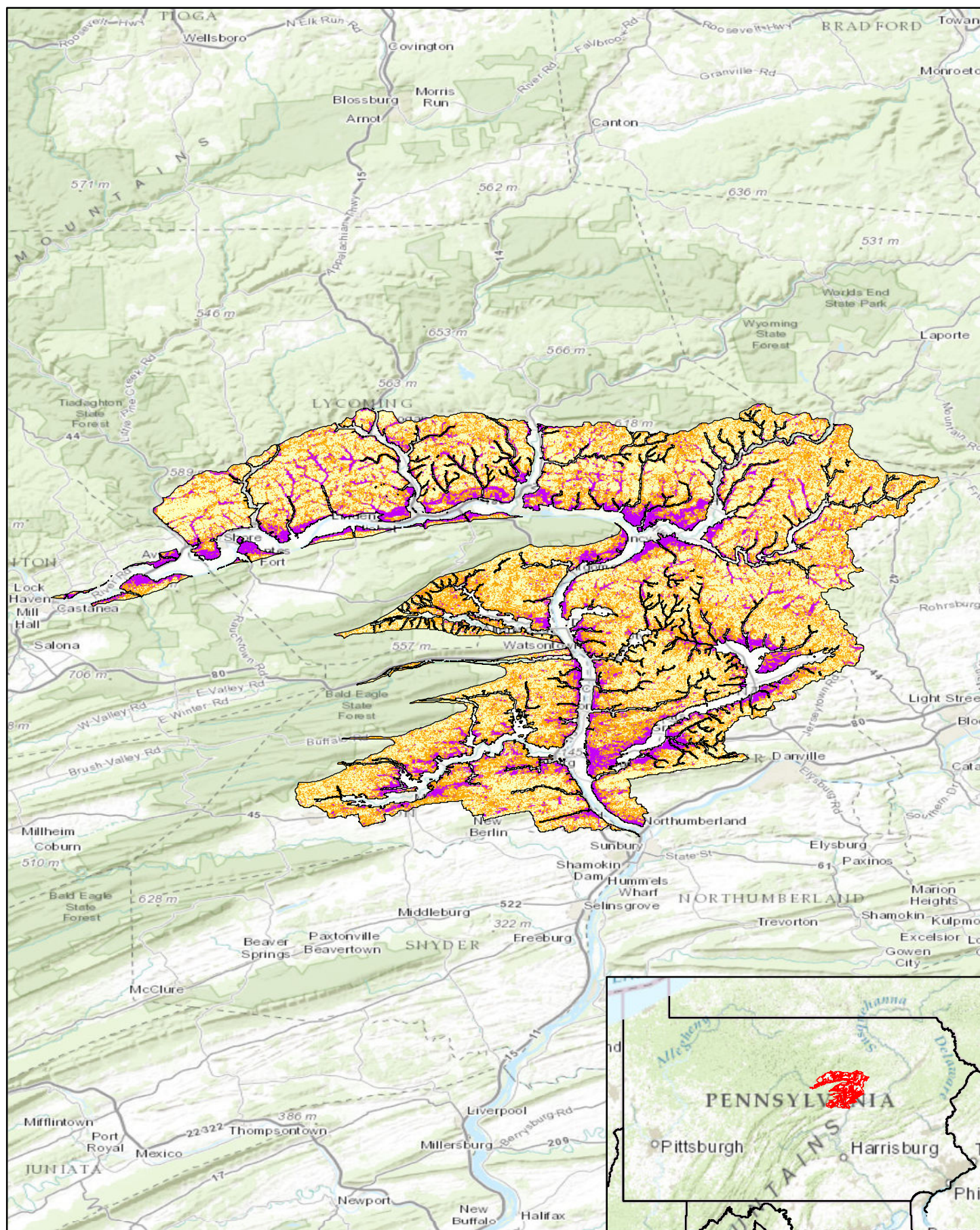




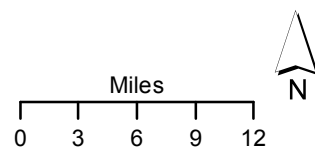
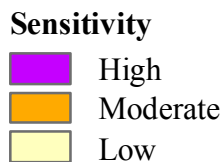
Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: upland section 3

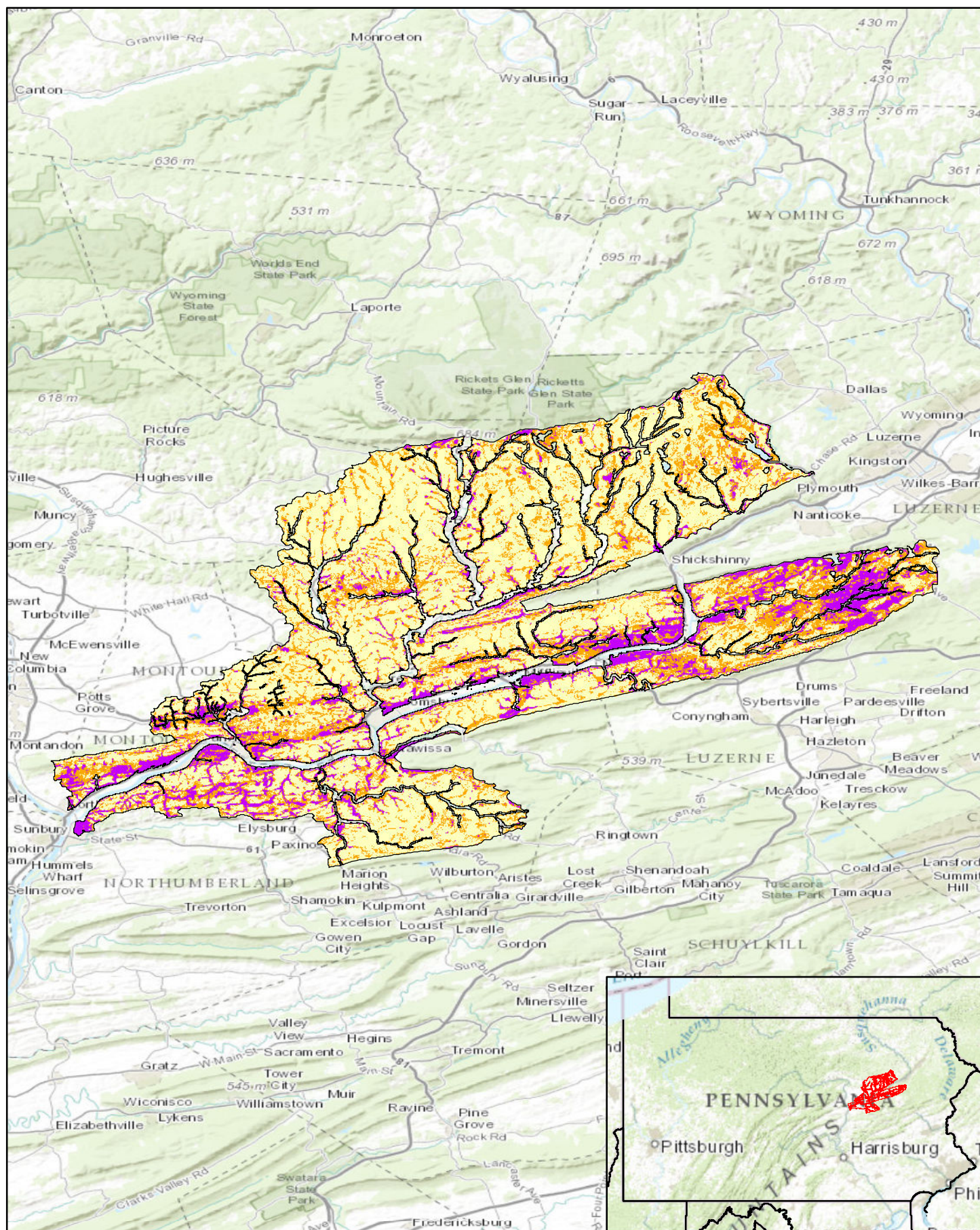
Sensitivity





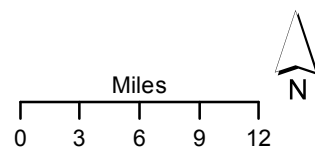
Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: upland section 4

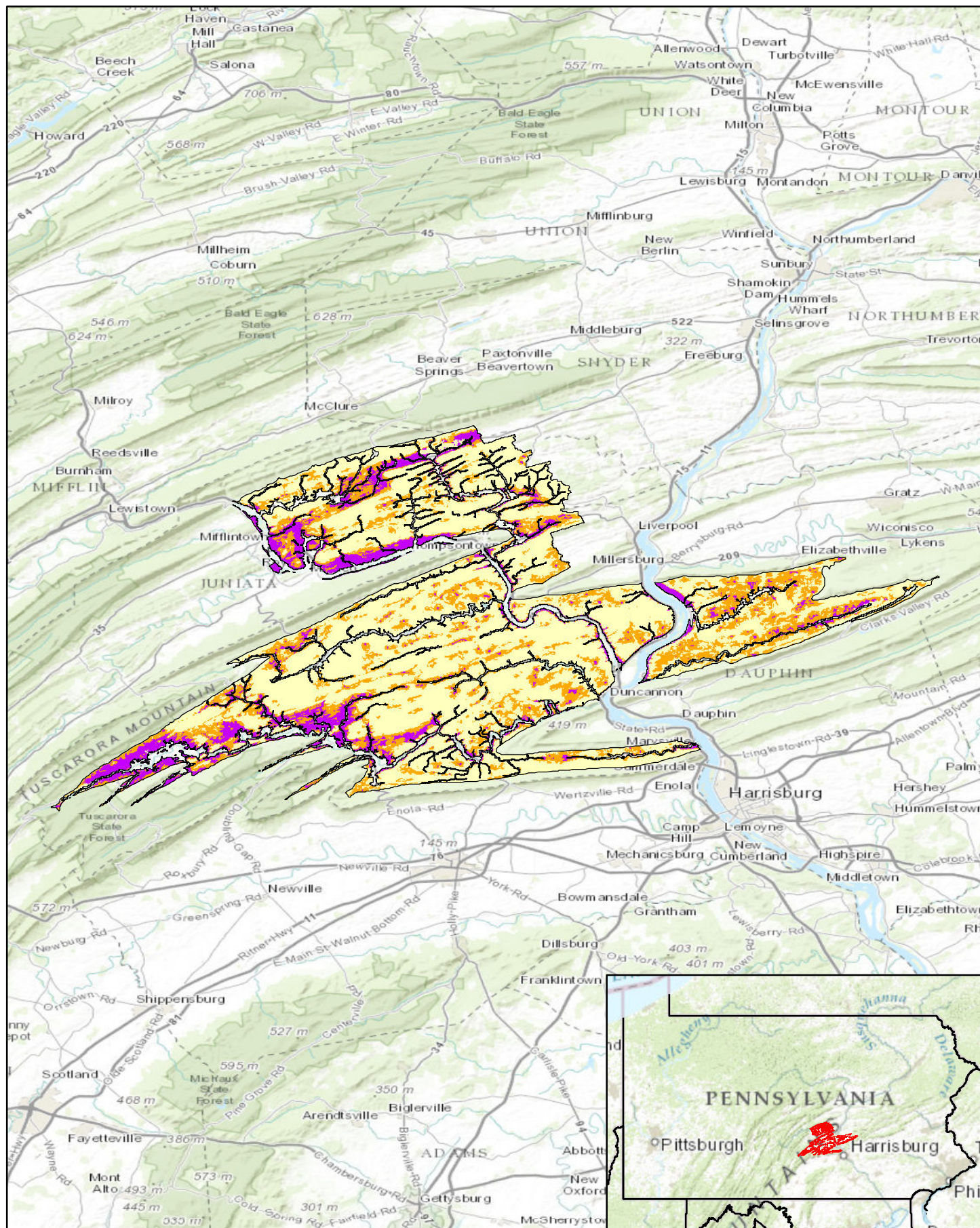




Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: upland section 5

Sensitivity
 High
 Moderate
 Low

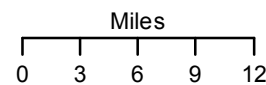


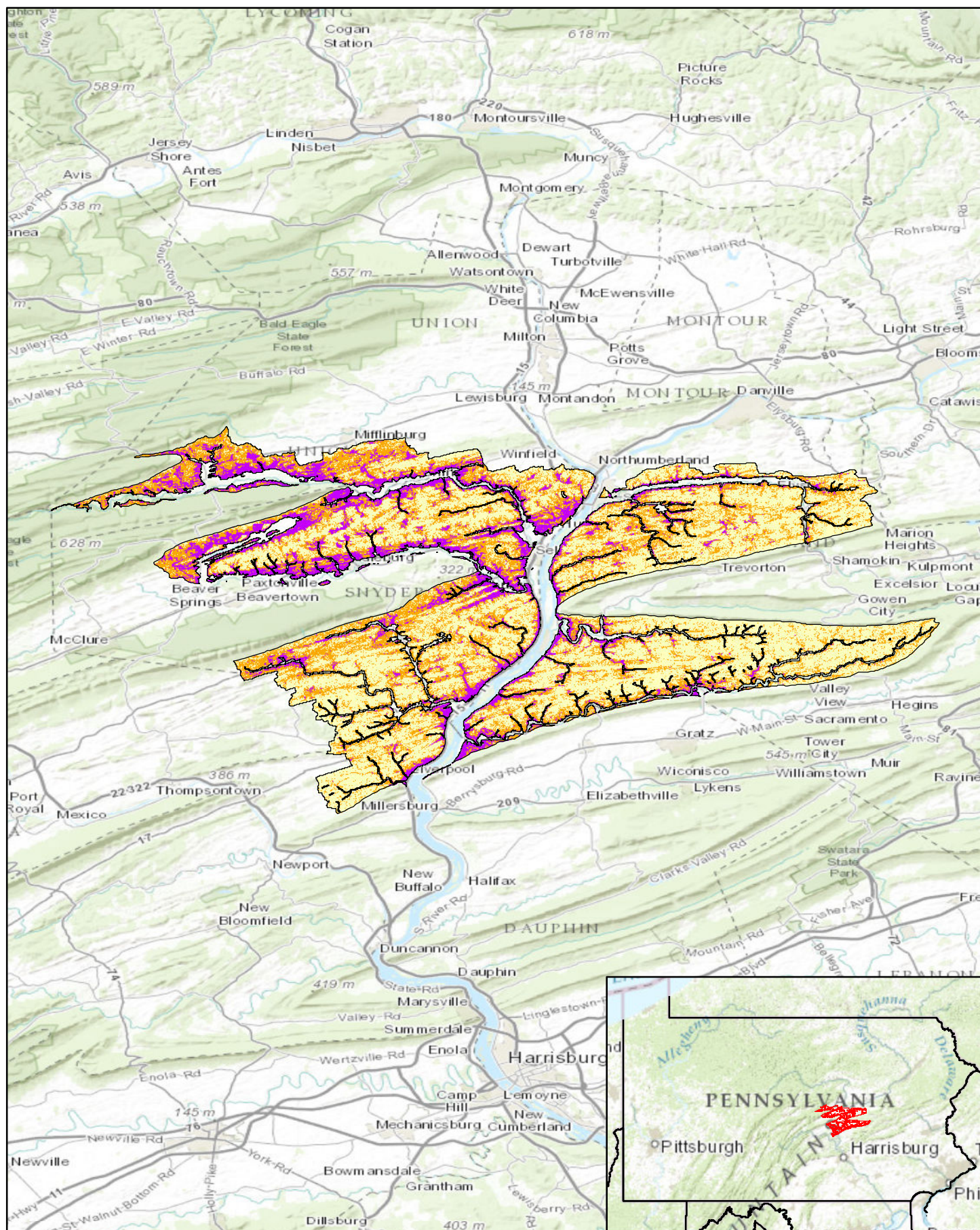


Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: upland section 6

Sensitivity

- High
- Moderate
- Low

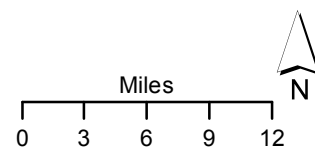


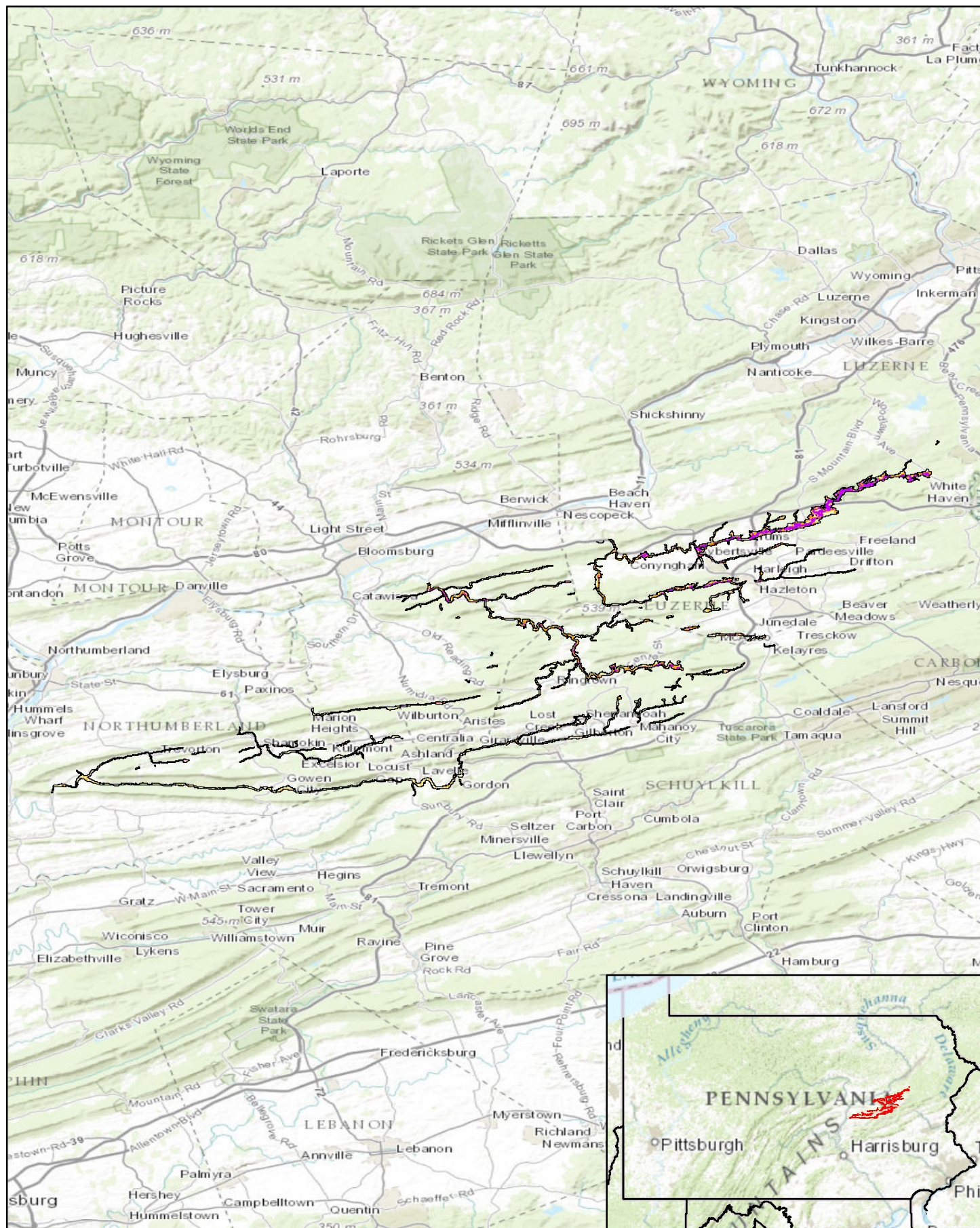


Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: upland section 7

Sensitivity

- High
- Moderate
- Low

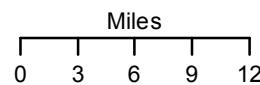


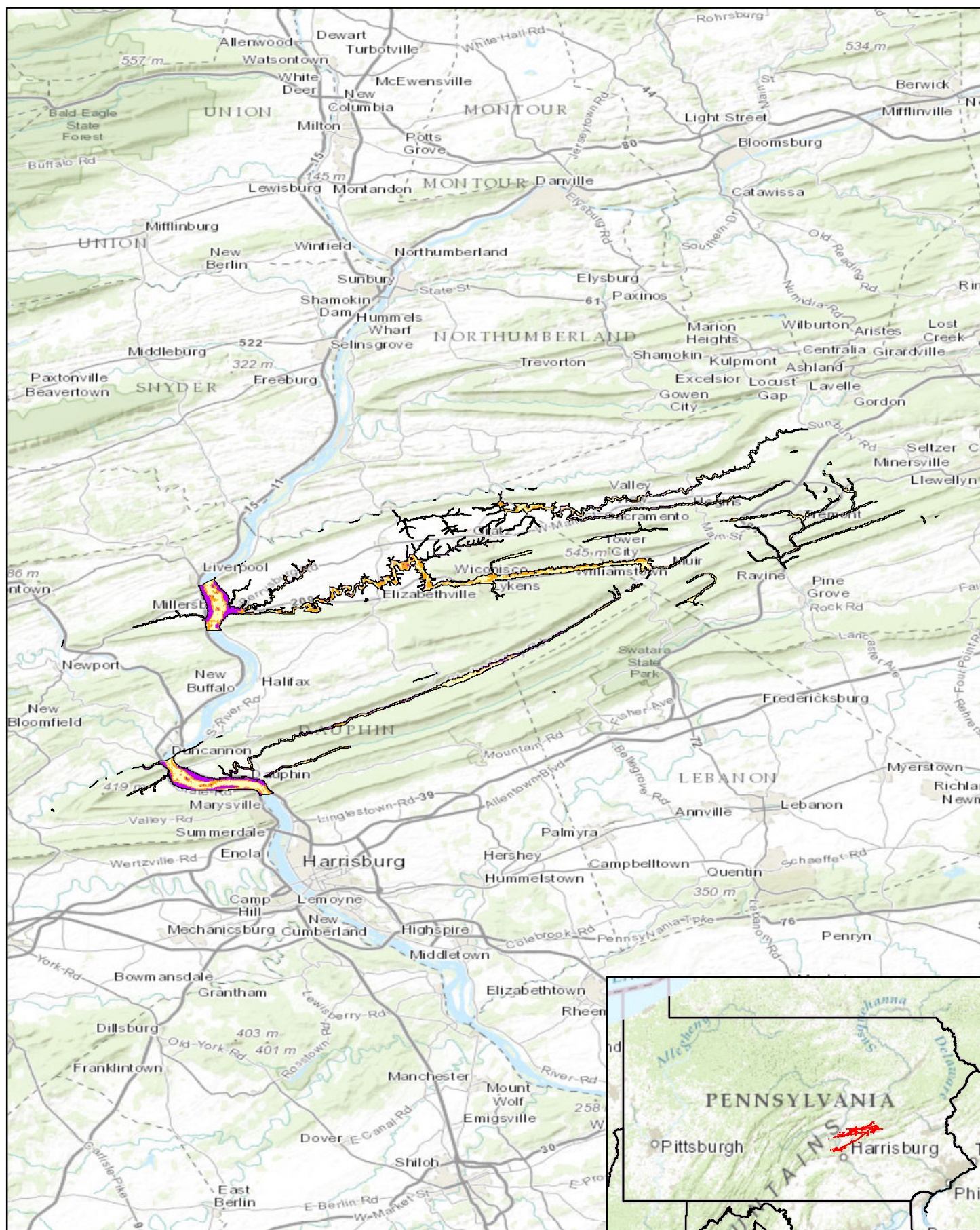


Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: riverine section 2

Sensitivity

- High
- Moderate
- Low





Pennsylvania Predictive Model Set

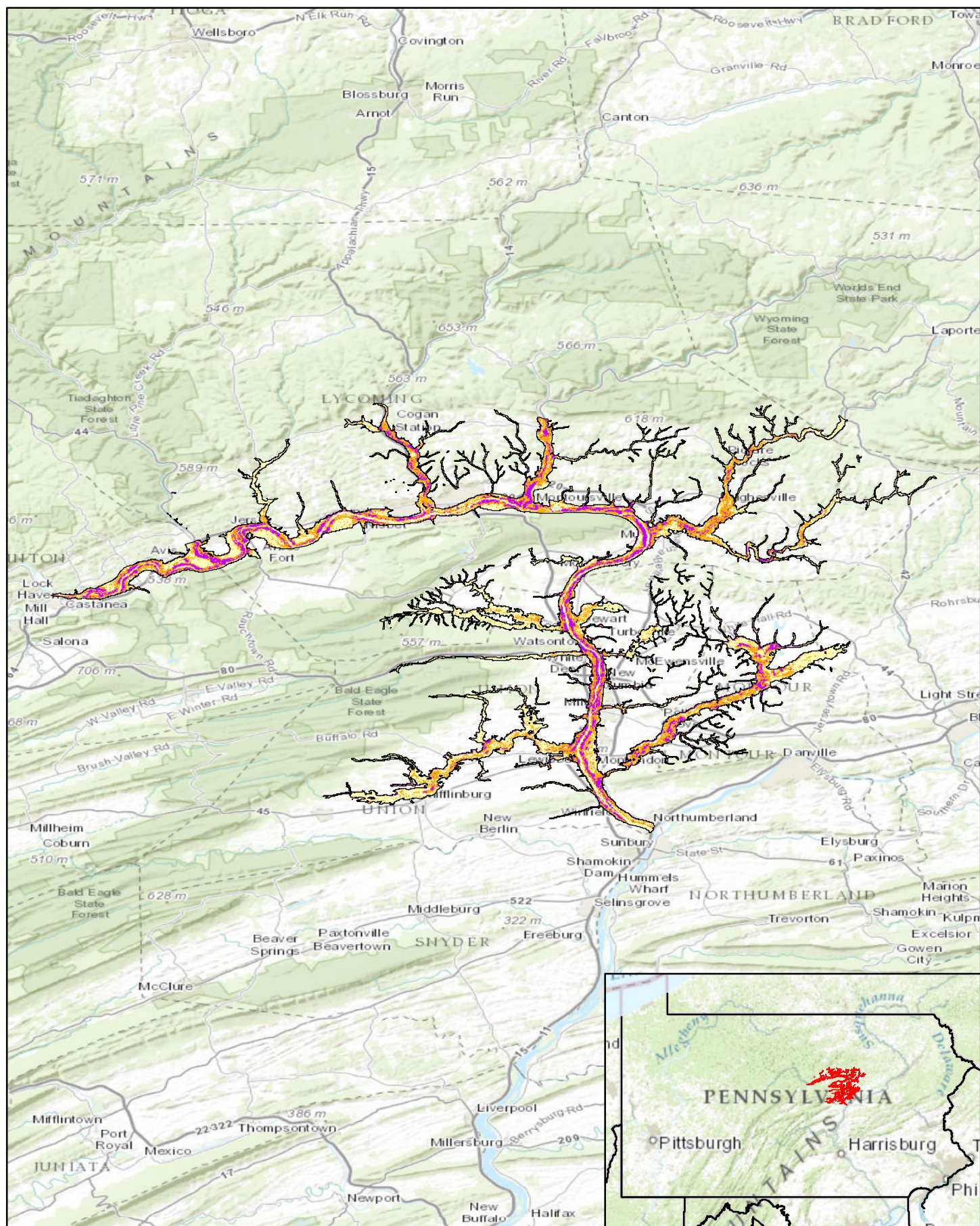
Region: 5, Zone: east, Subarea: riverine section 3

Sensitivity

- High
- Moderate
- Low

Miles
0 2.5 5 7.5 10





Pennsylvania Predictive Model Set

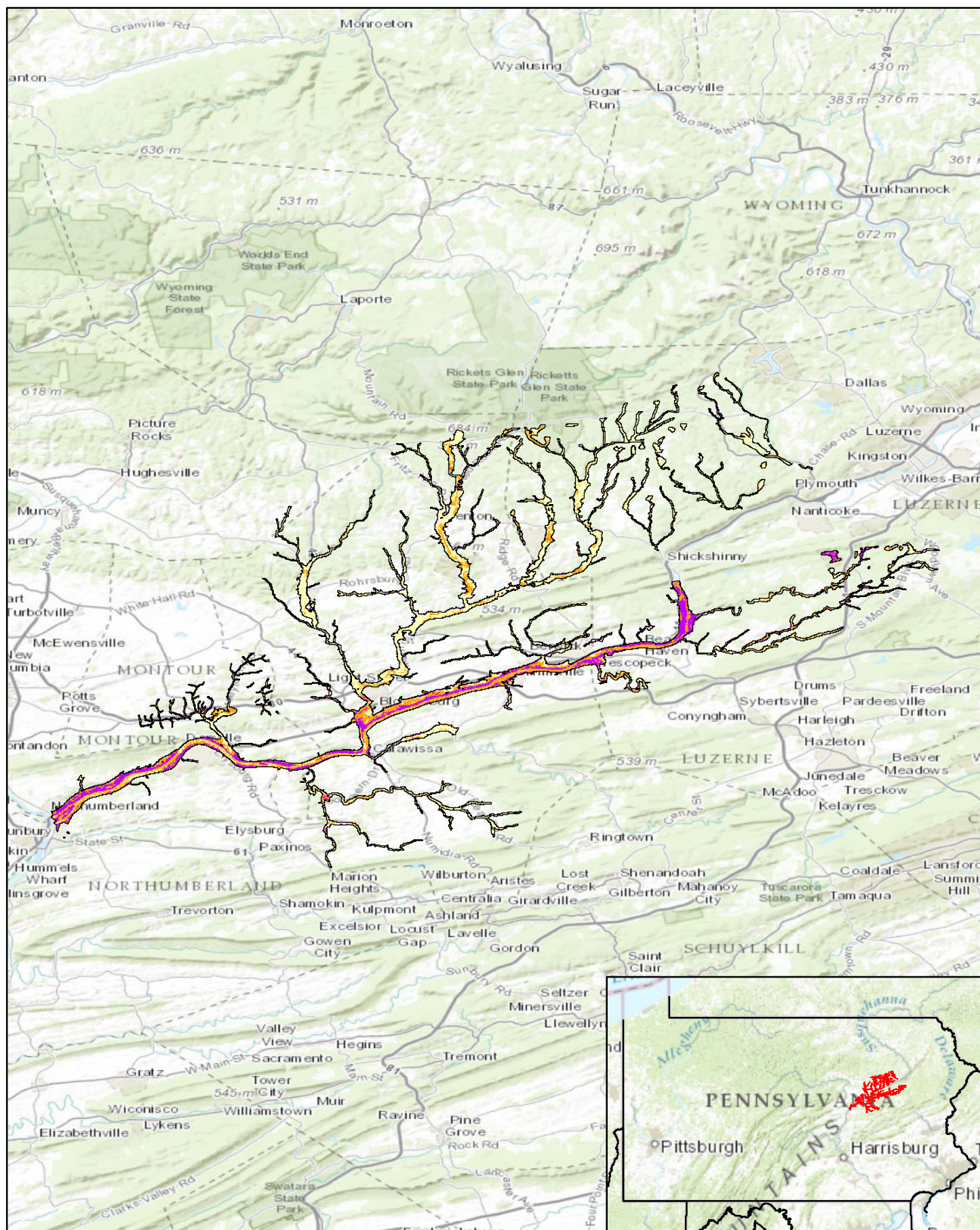
Region: 5, Zone: east, Subarea: riverine section 4

Sensitivity

- High
- Moderate
- Low

Miles
0 3 6 9 12



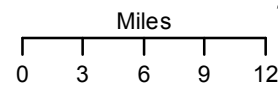


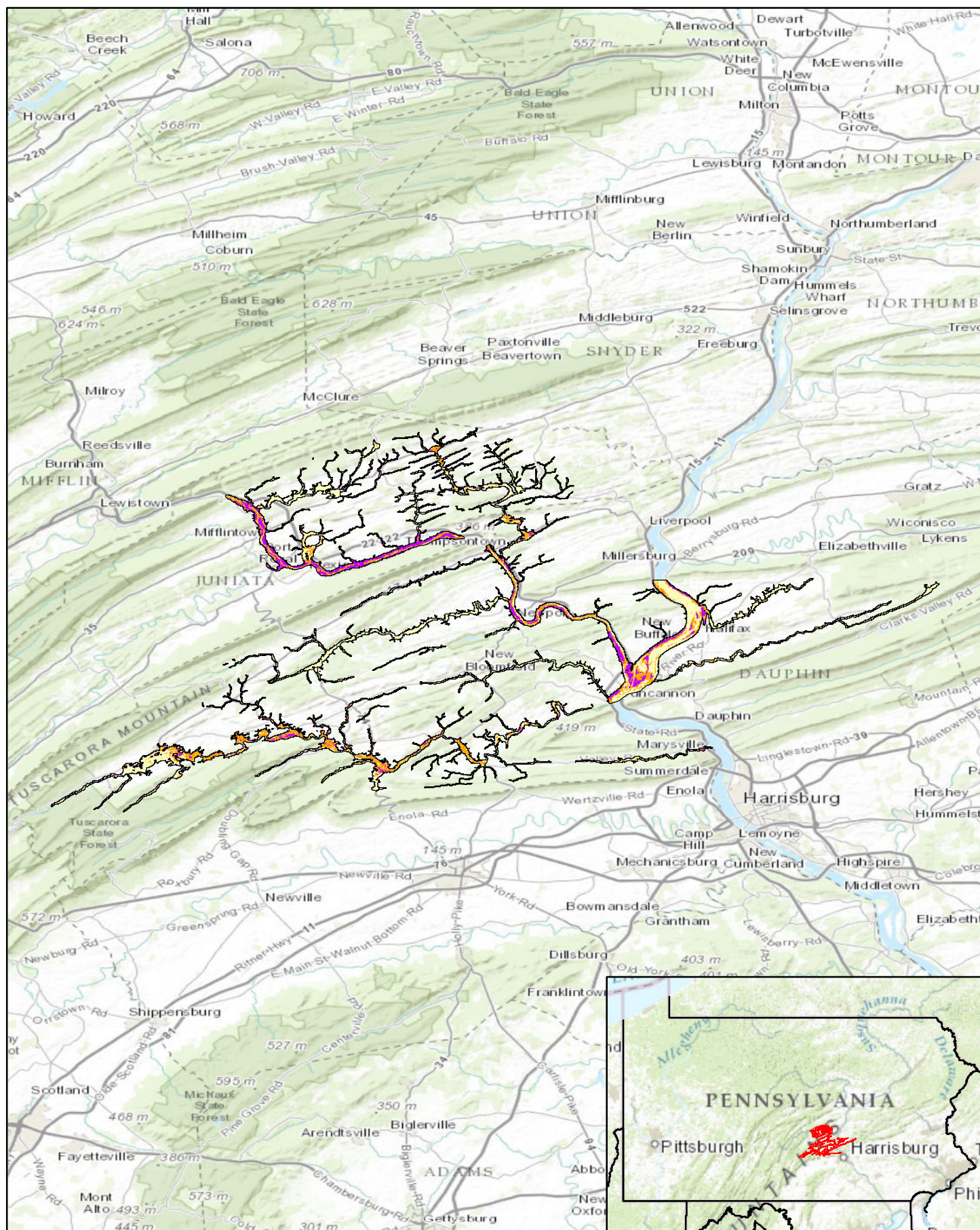
Pennsylvania Predictive Model Set

Region: 5, Zone: east, Subarea: riverine section 5

Sensitivity

- High
- Moderate
- Low

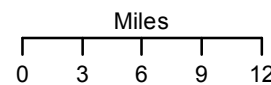


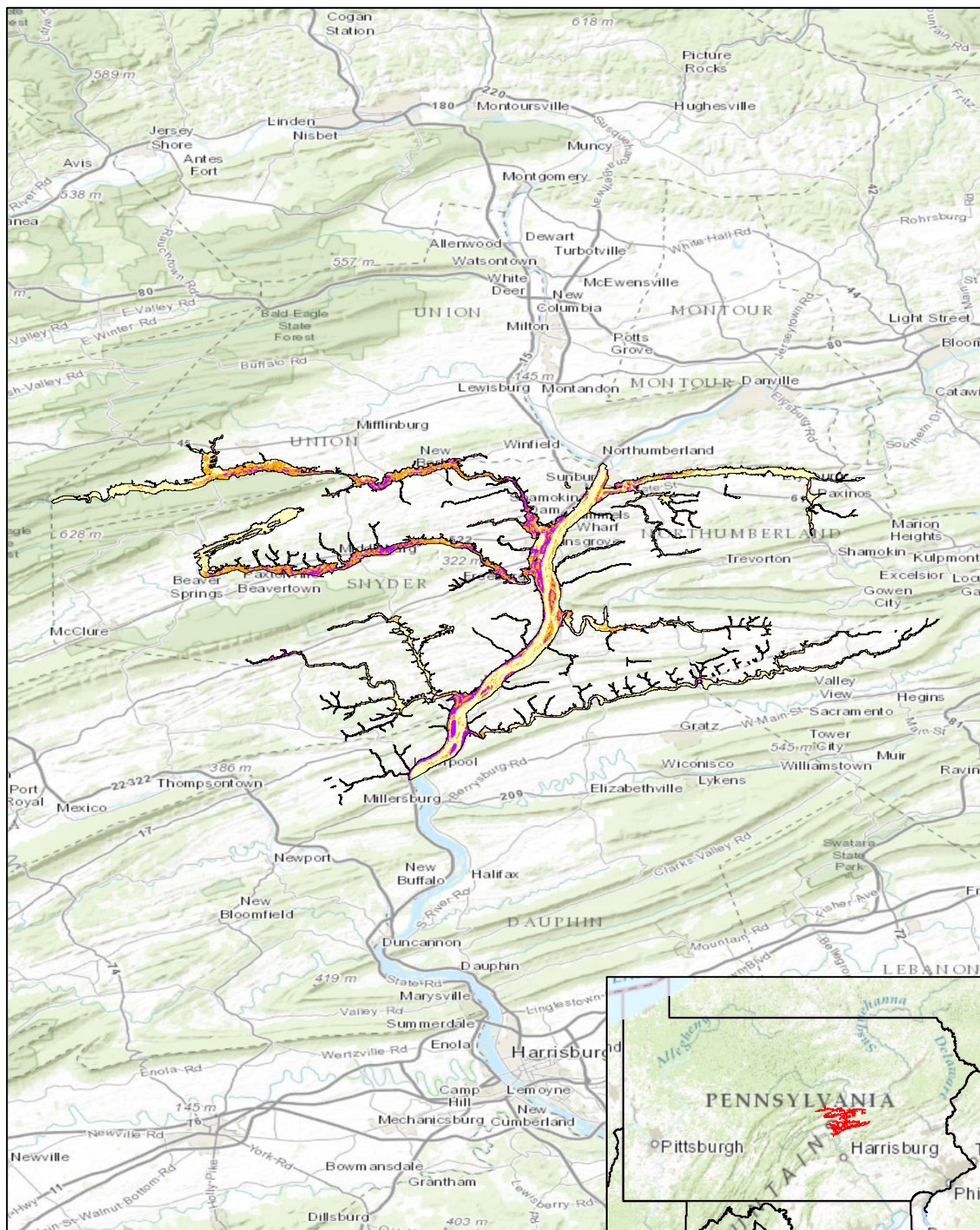


Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: riverine section 6

Sensitivity

- High
- Moderate
- Low

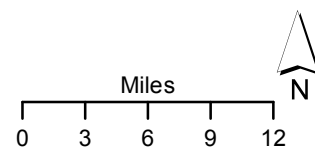


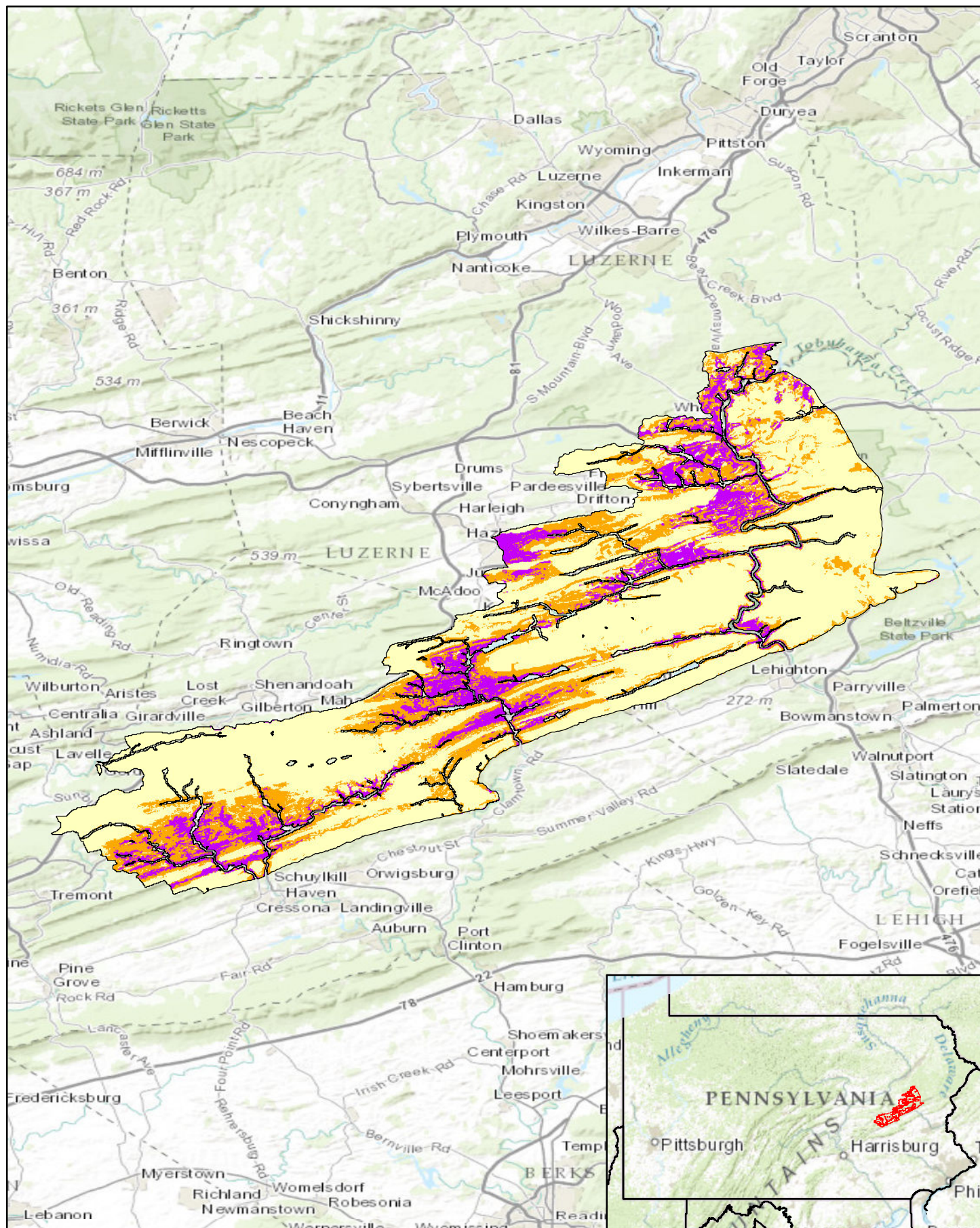


Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: riverine section 7

Sensitivity

- High
- Moderate
- Low



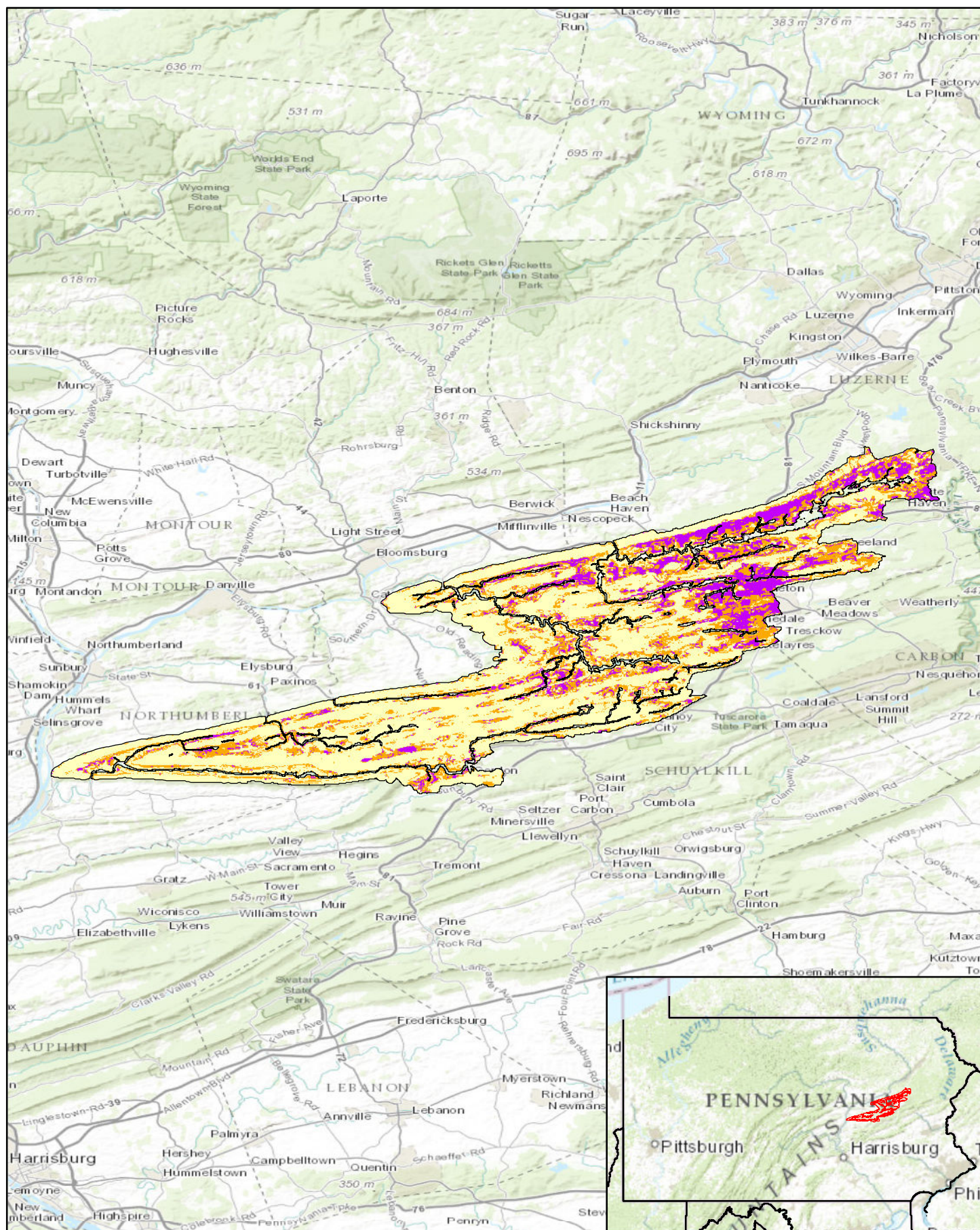


Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: upland section 1

Sensitivity
 High
 Moderate
 Low

Miles
 0 2.5 5 7.5 10

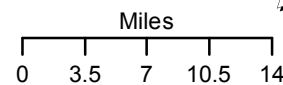


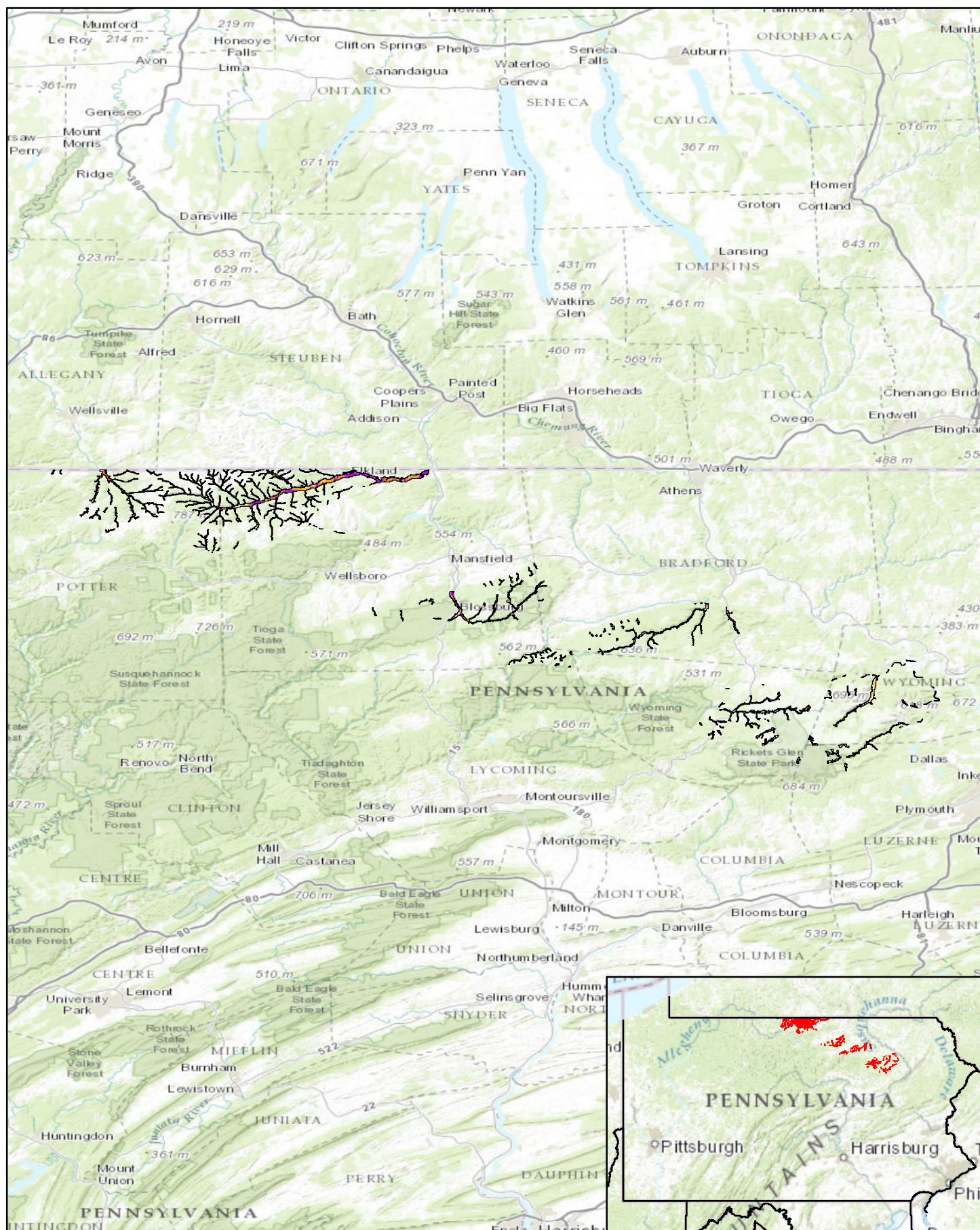


Pennsylvania Predictive Model Set
 Region: 5, Zone: east, Subarea: upland section 2

Sensitivity

- High
- Moderate
- Low

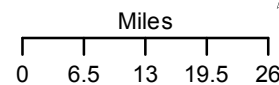


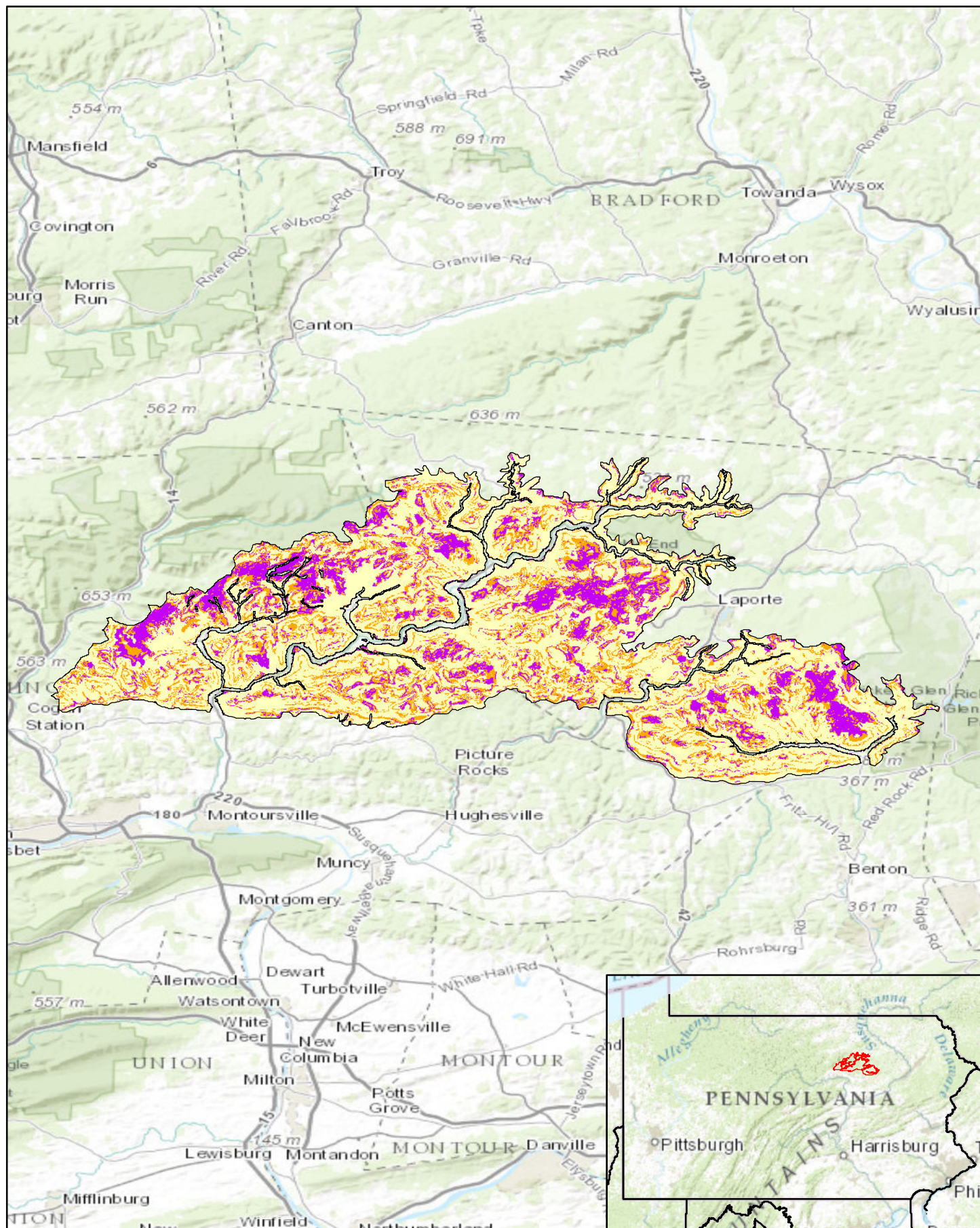


Pennsylvania Predictive Model Set
 Region: 6, Zone: all, Subarea: riverine section 1

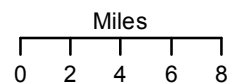
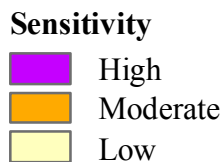
Sensitivity

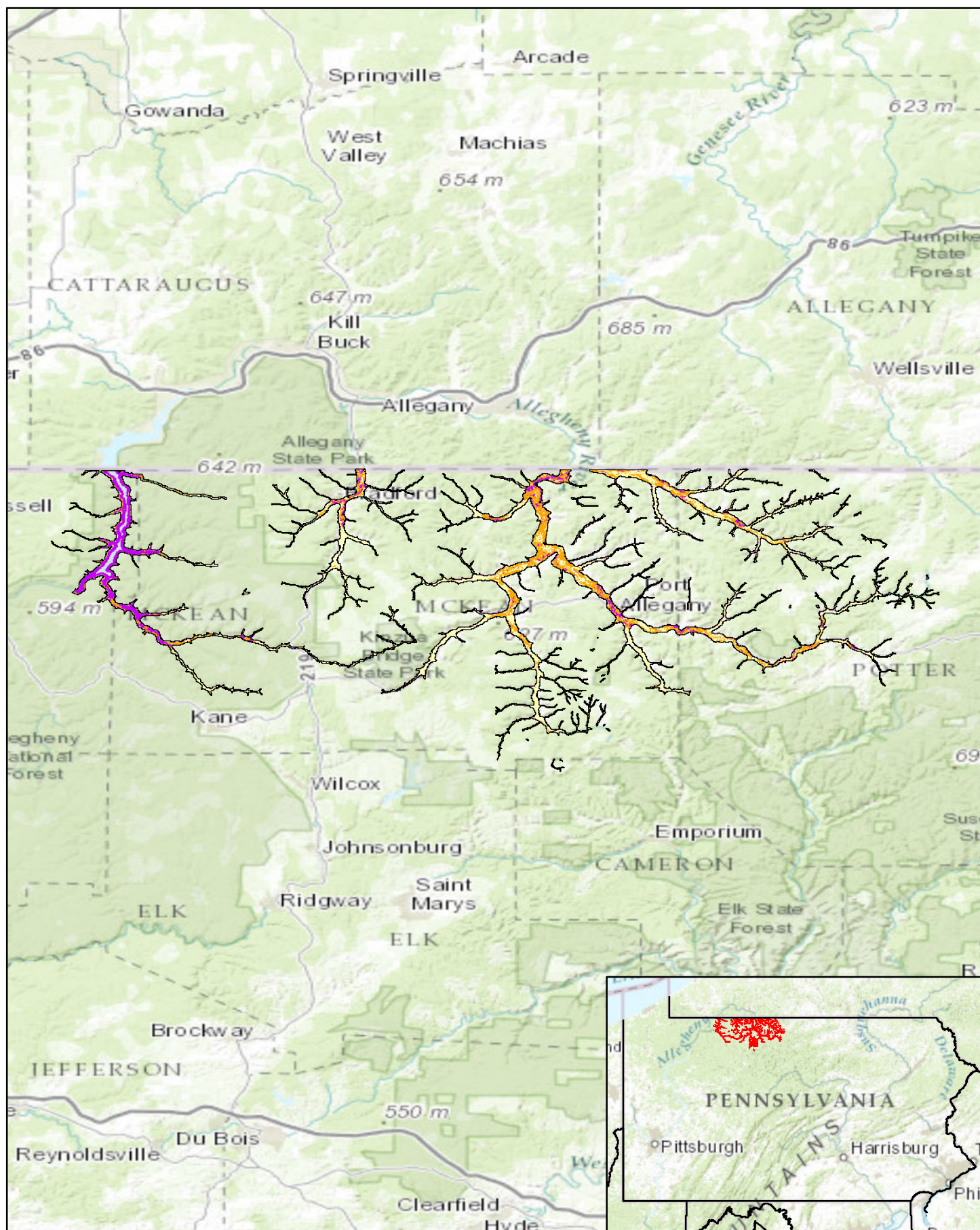
- High
- Moderate
- Low





Pennsylvania Predictive Model Set
 Region: 6, Zone: all, Subarea: upland section 5

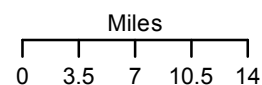


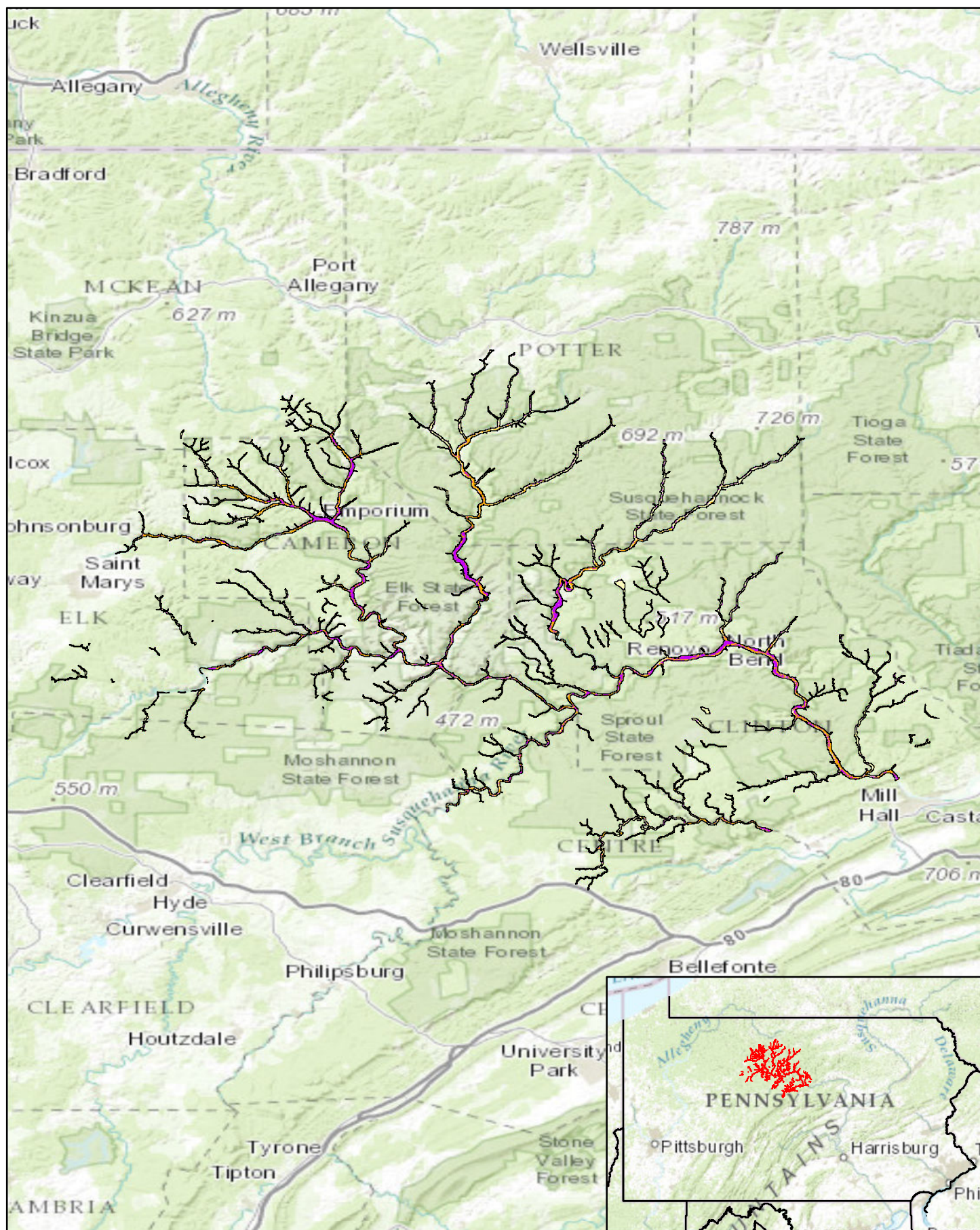


Pennsylvania Predictive Model Set
 Region: 6, Zone: all, Subarea: riverine section 2

Sensitivity

- High
- Moderate
- Low

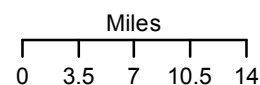


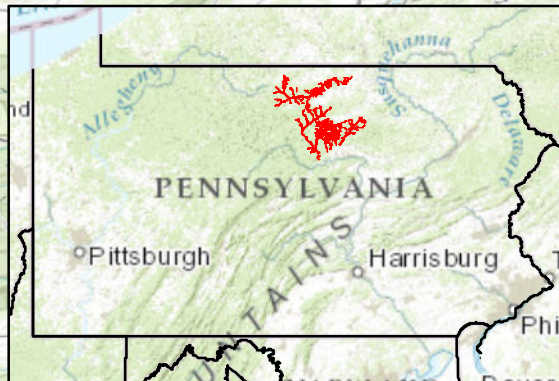


Pennsylvania Predictive Model Set
 Region: 6, Zone: all, Subarea: riverine section 3




Sensitivity

- High
- Moderate
- Low





Region: 6, Zone: all, Subarea: riverine section 4

	High
	Moderate
	Low

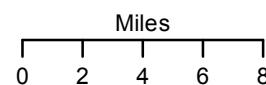


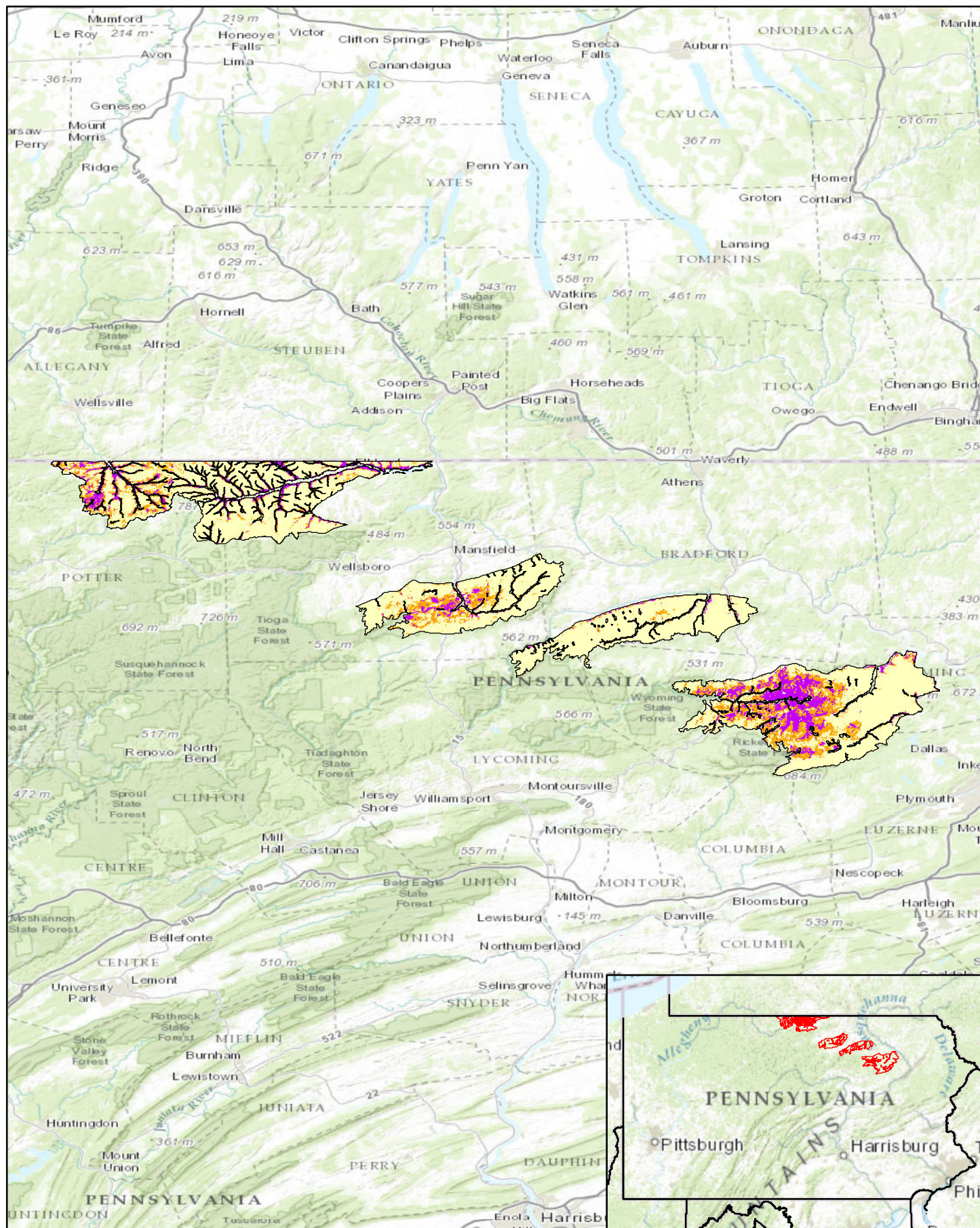


Pennsylvania Predictive Model Set
 Region: 6, Zone: all, Subarea: riverine section 5

Sensitivity

- High
- Moderate
- Low

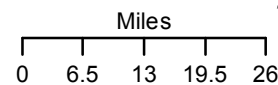


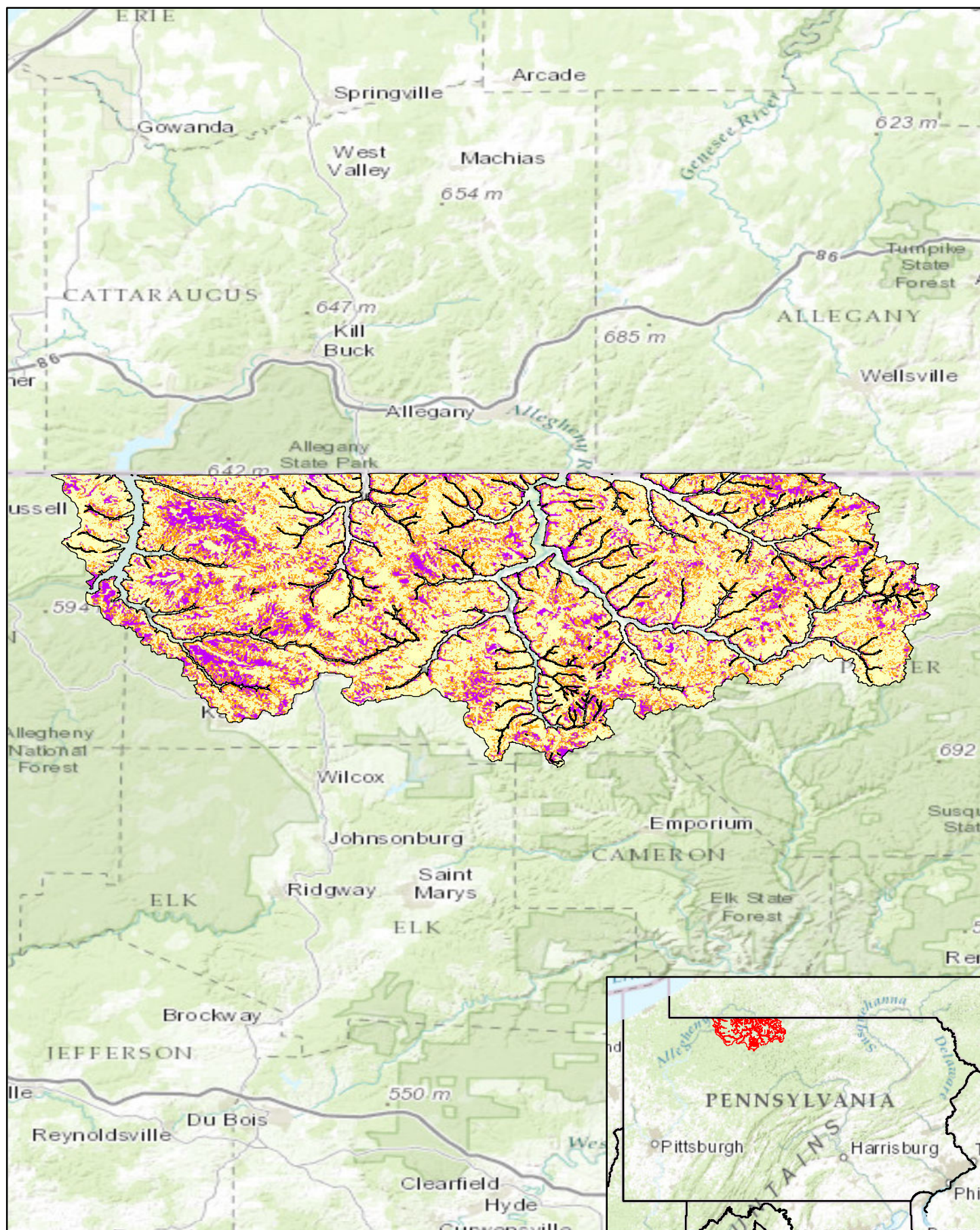


Pennsylvania Predictive Model Set
 Region: 6, Zone: all, Subarea: upland section 1

Sensitivity

- High
- Moderate
- Low

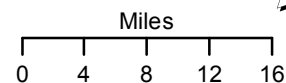


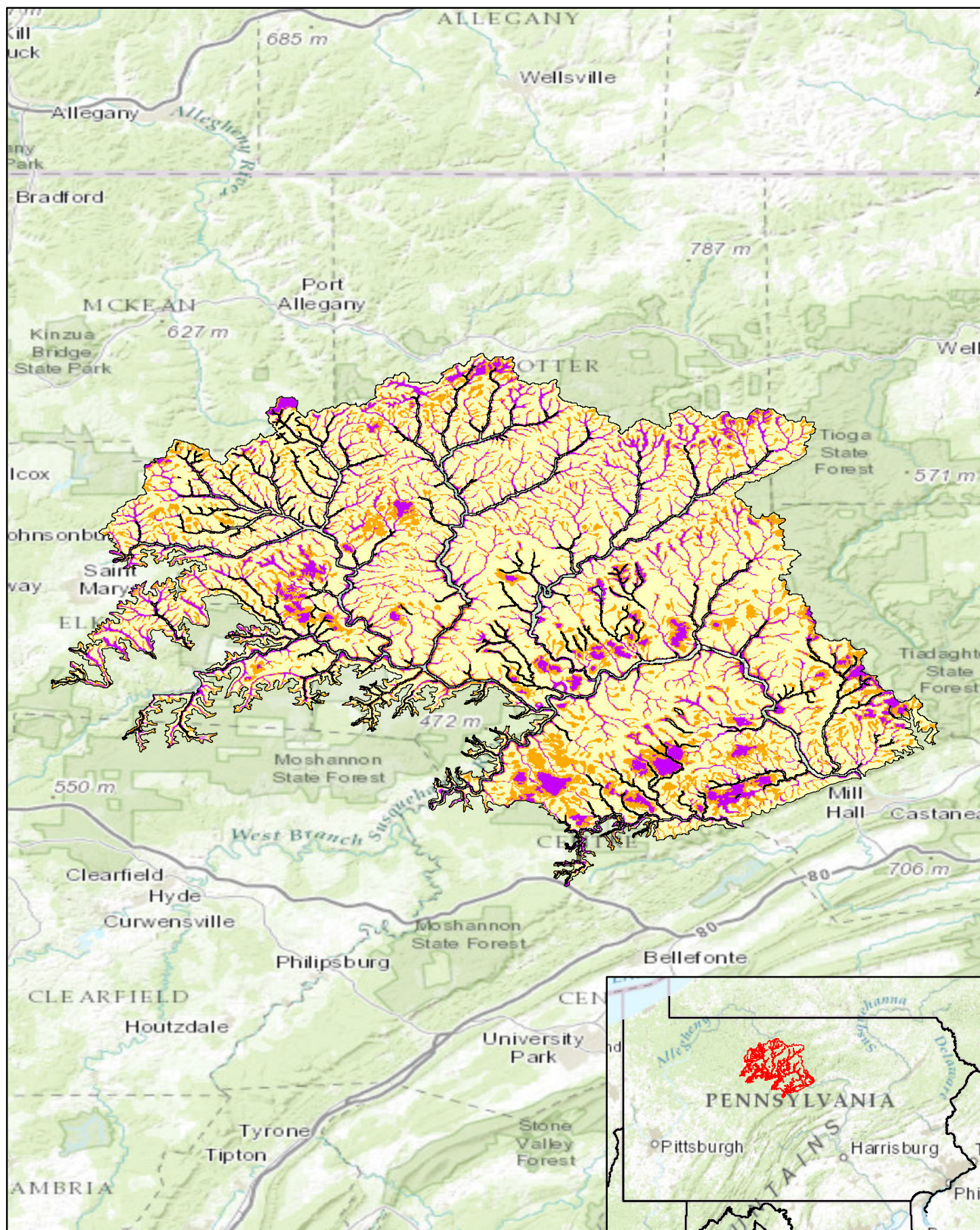


Pennsylvania Predictive Model Set
 Region: 6, Zone: all, Subarea: upland section 2

Sensitivity

- High
- Moderate
- Low

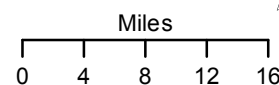


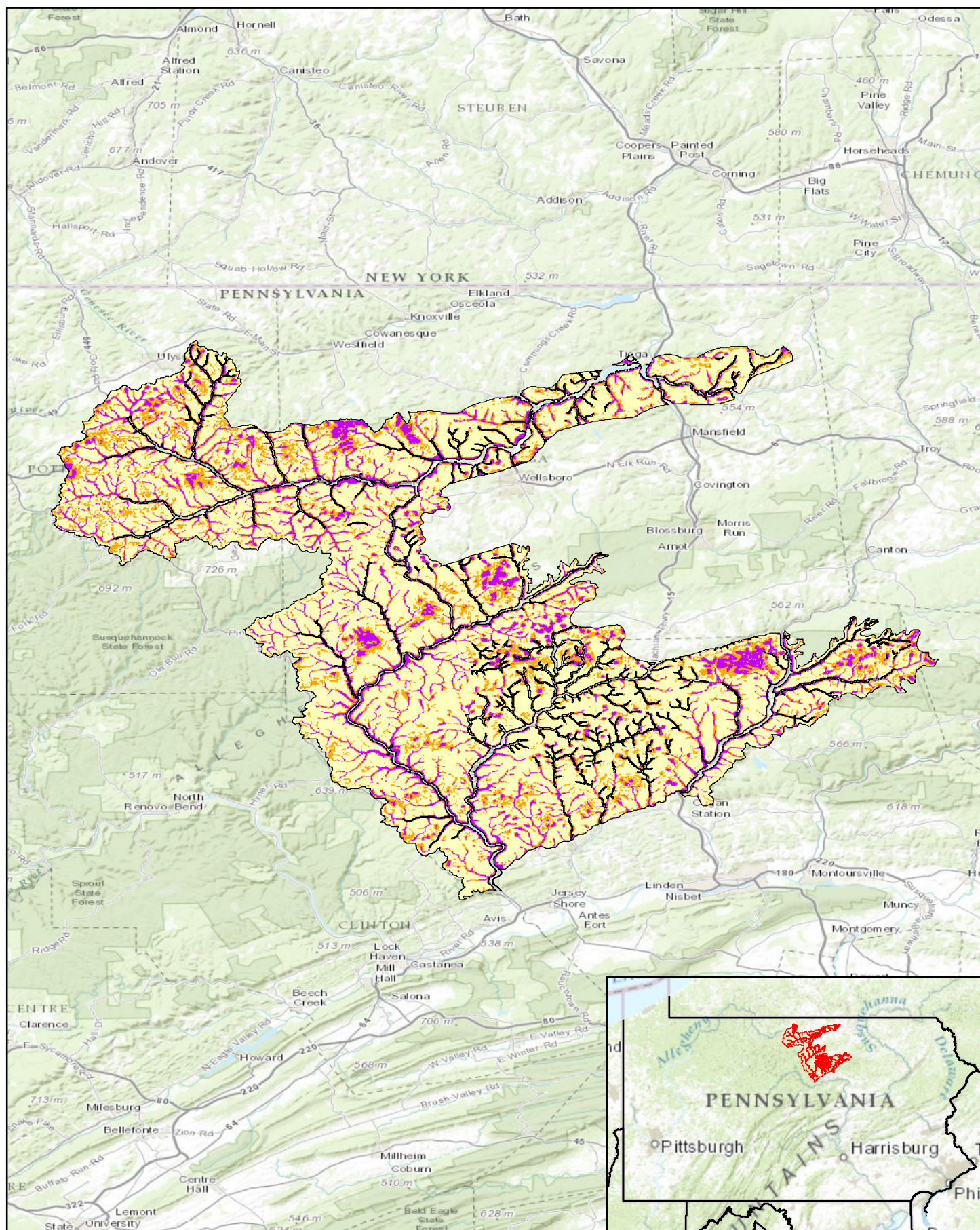


Pennsylvania Predictive Model Set
 Region: 6, Zone: all, Subarea: upland section 3

Sensitivity

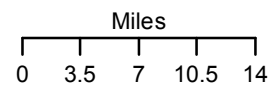
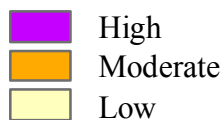
- High
- Moderate
- Low

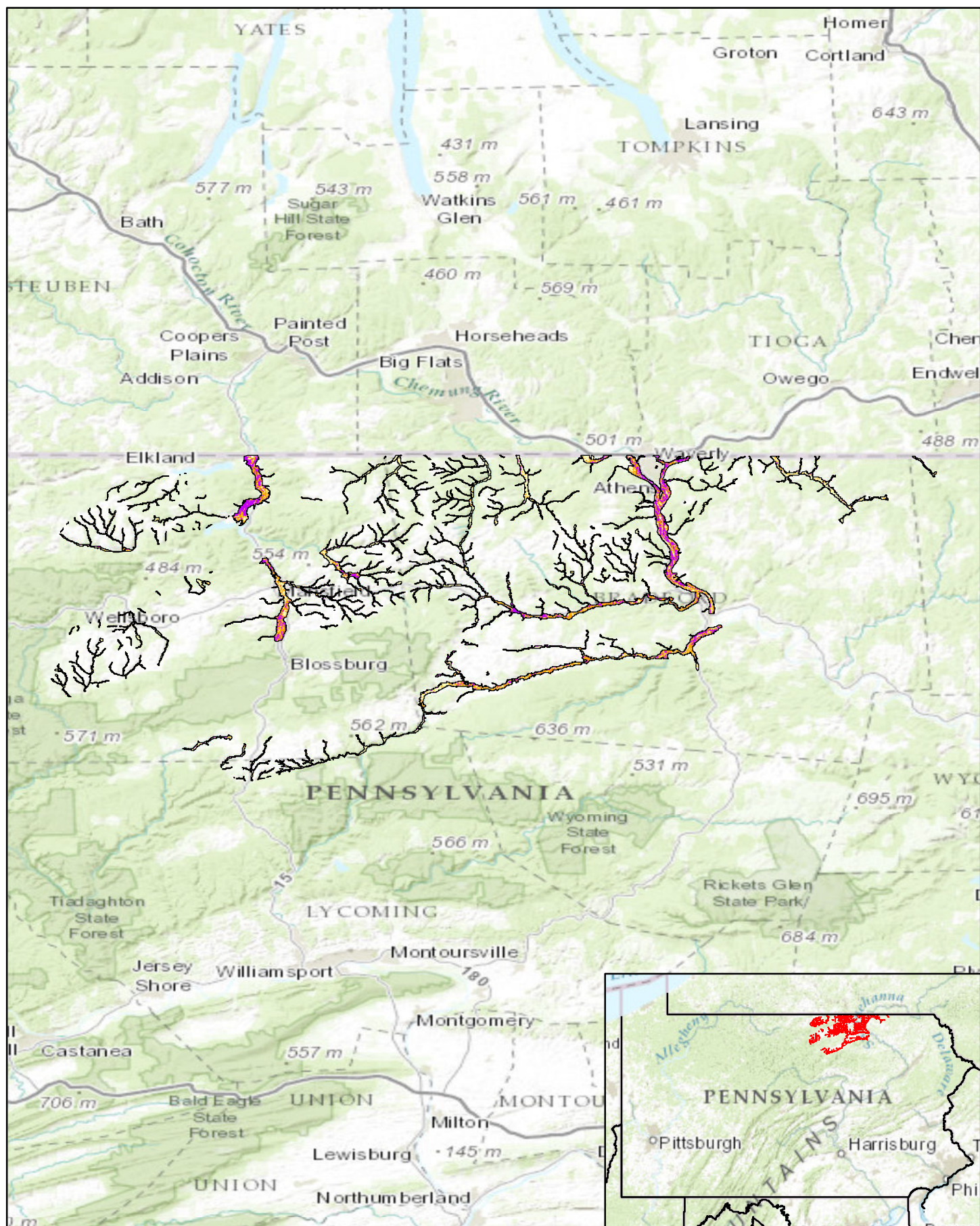




Pennsylvania Predictive Model Set
 Region: 6, Zone: all, Subarea: upland section 4

Sensitivity

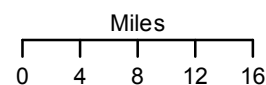


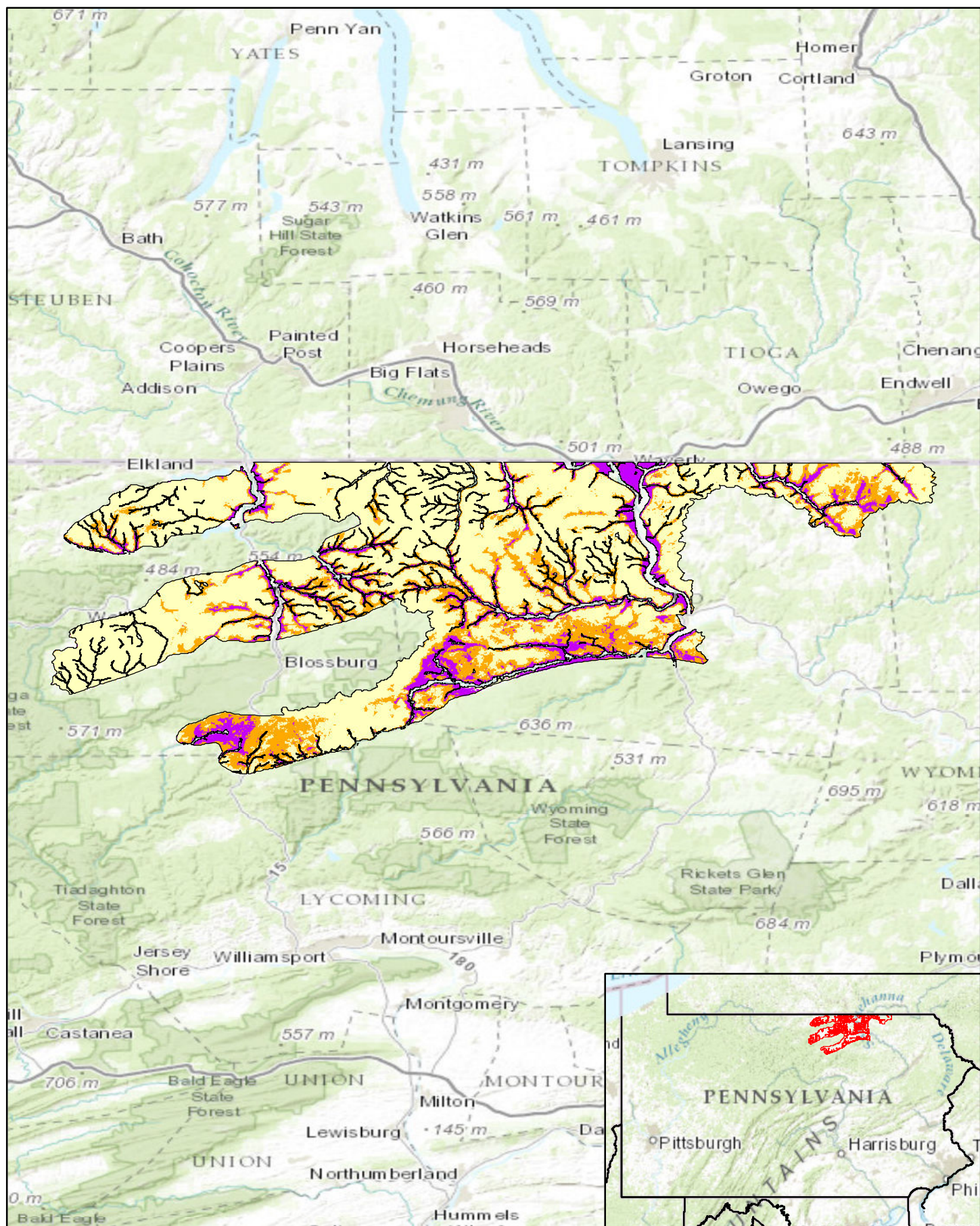


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: riverine section 1

Sensitivity

- High
- Moderate
- Low





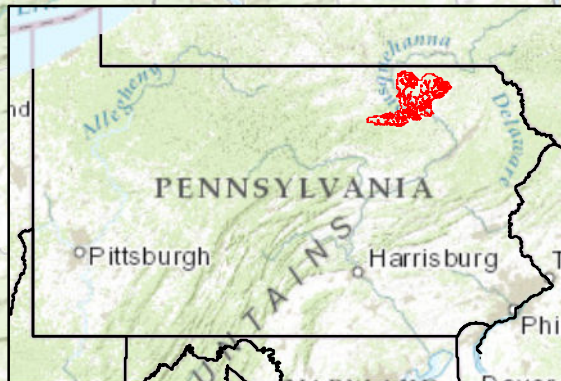
Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: upland section 1

Sensitivity




- High
- Moderate
- Low

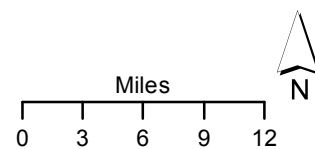
Miles
 0 4.5 9 13.5 18

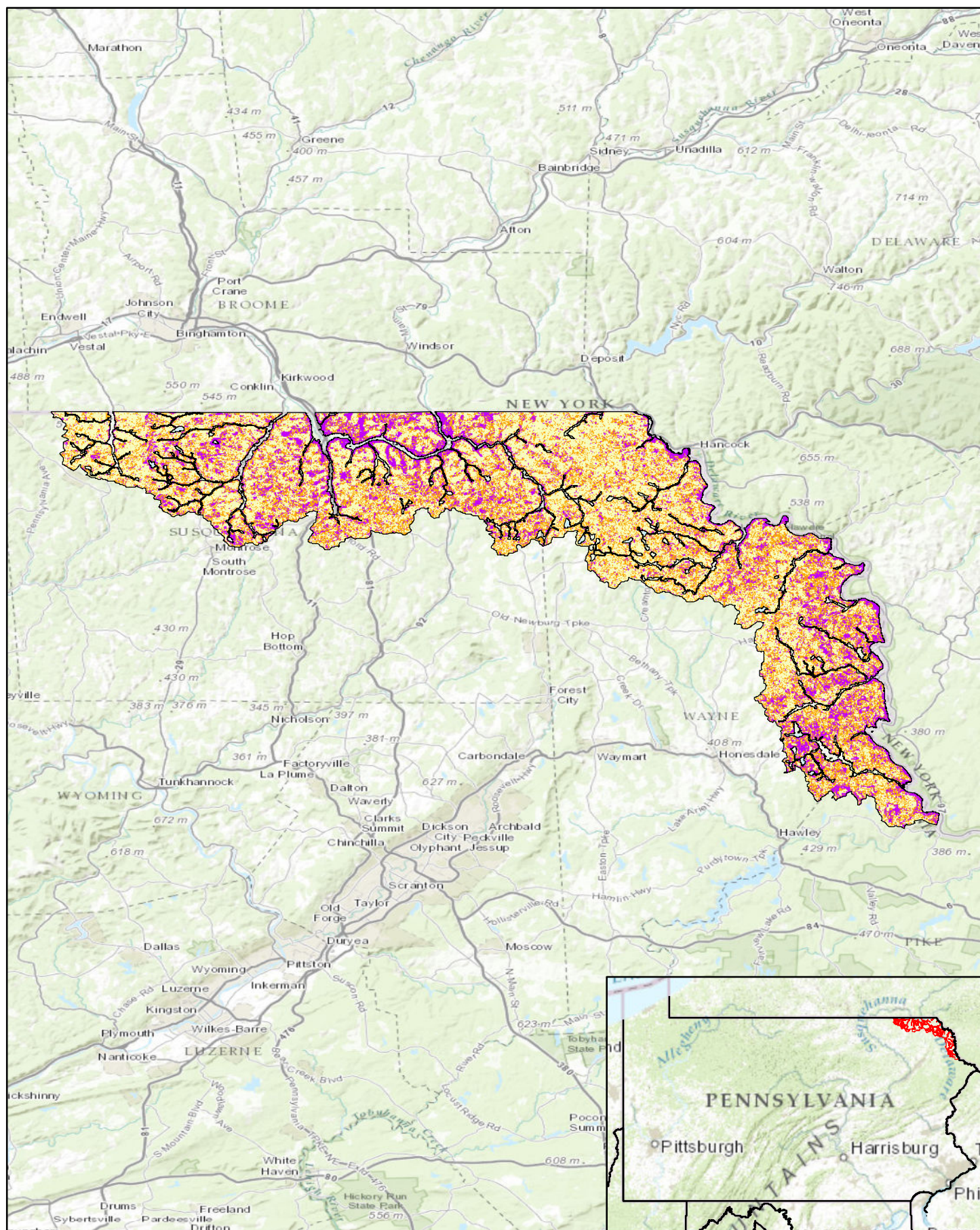




Region: 7, Zone: all, Subarea: upland section 2

	High
	Moderate
	Low

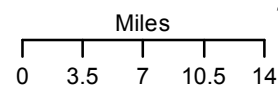


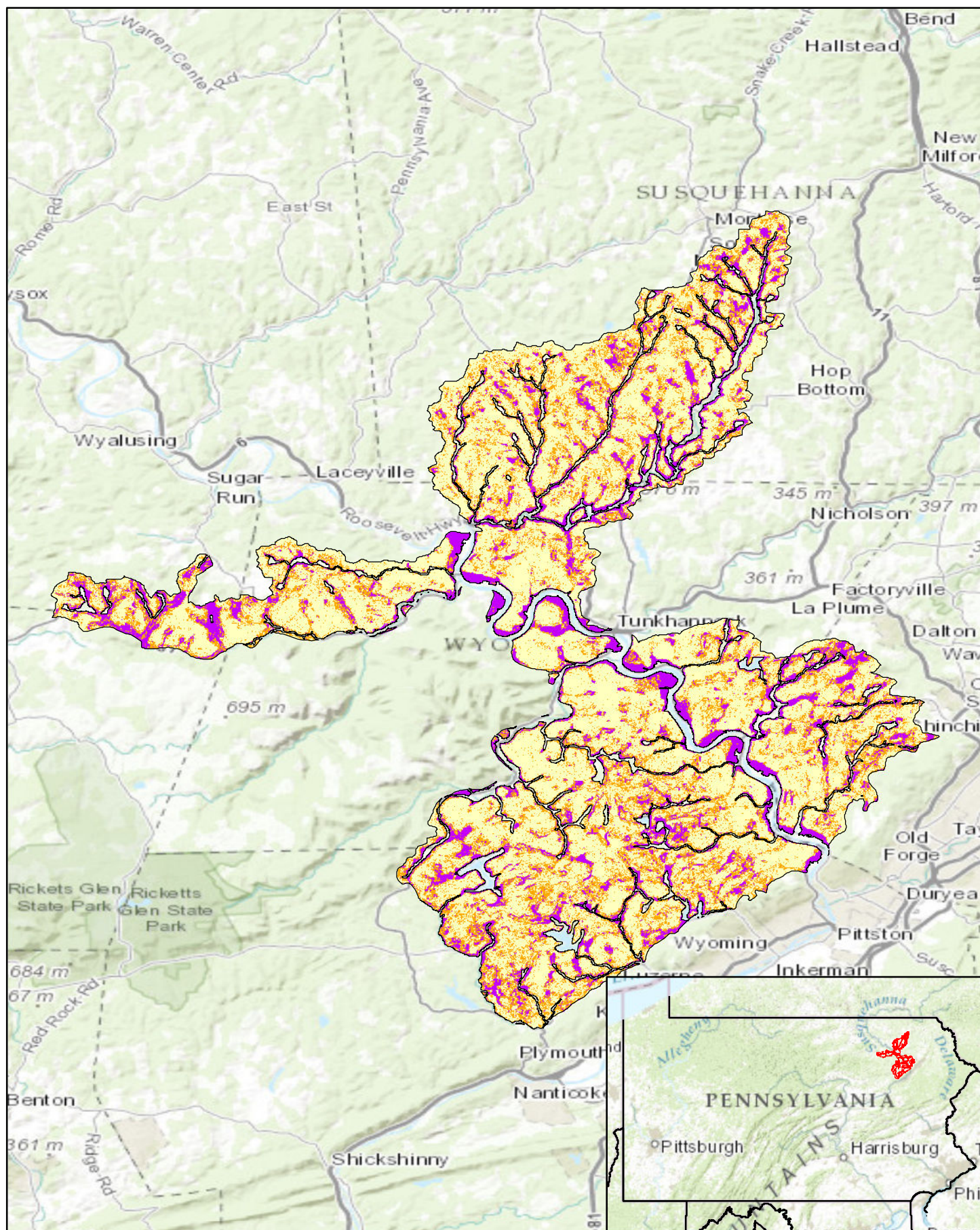


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: upland section 3

Sensitivity

- High
- Moderate
- Low

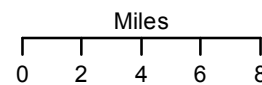


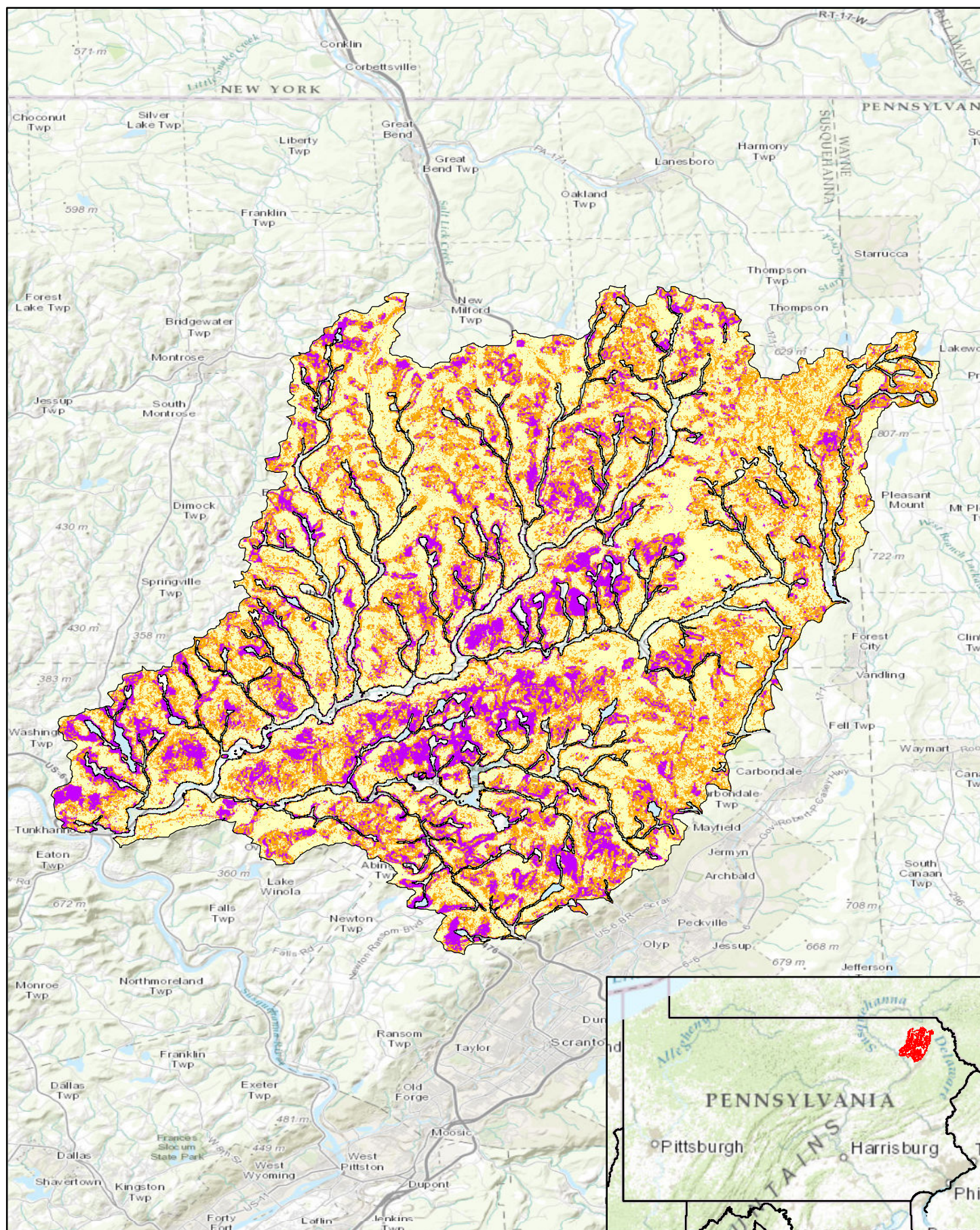


Pennsylvania Predictive Model Set
Region: 7, Zone: all, Subarea: upland section 4

Sensitivity

- High
- Moderate
- Low





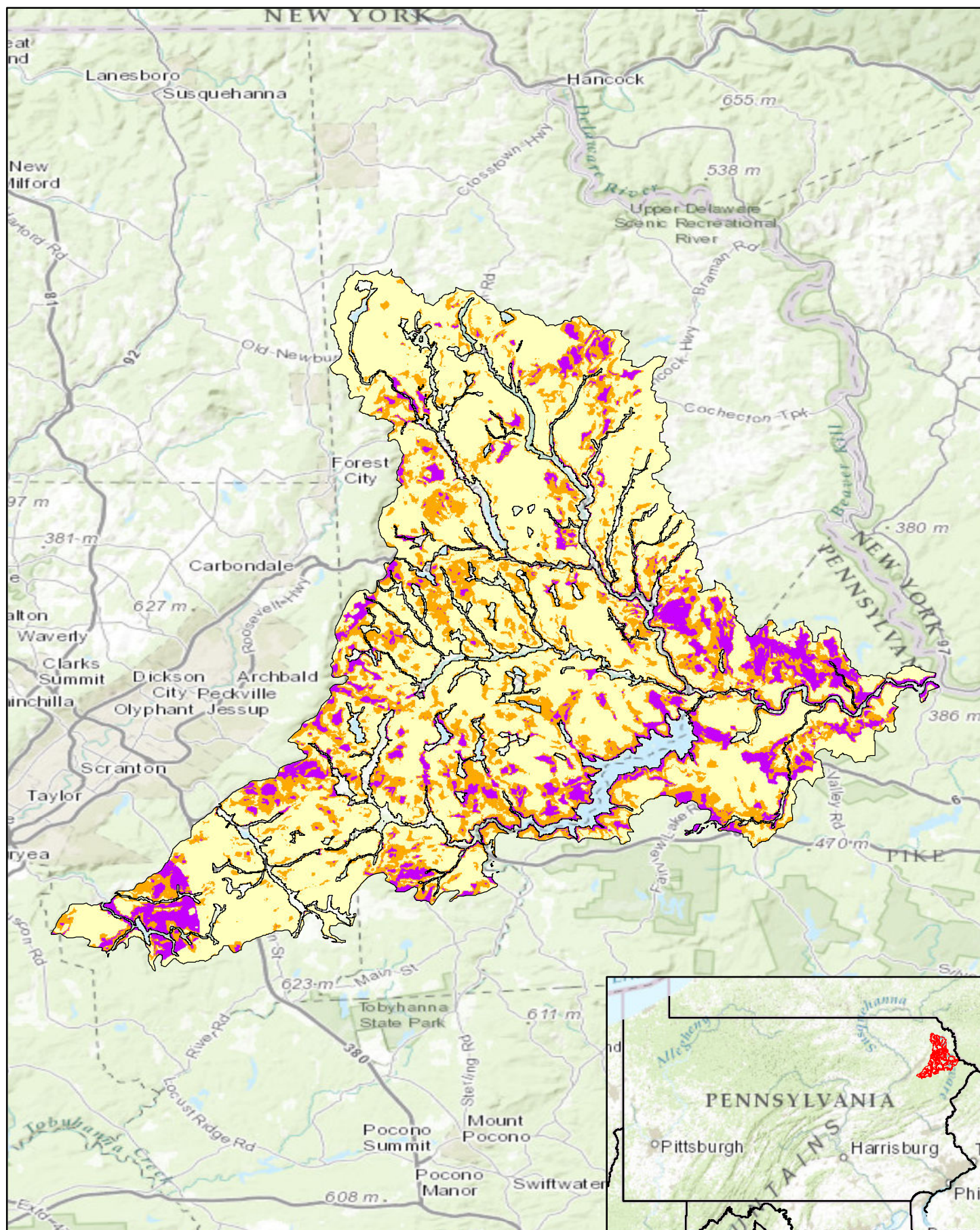
Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: upland section 5

Sensitivity

- High
- Moderate
- Low

Miles
 0 1.5 3 4.5 6

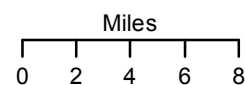


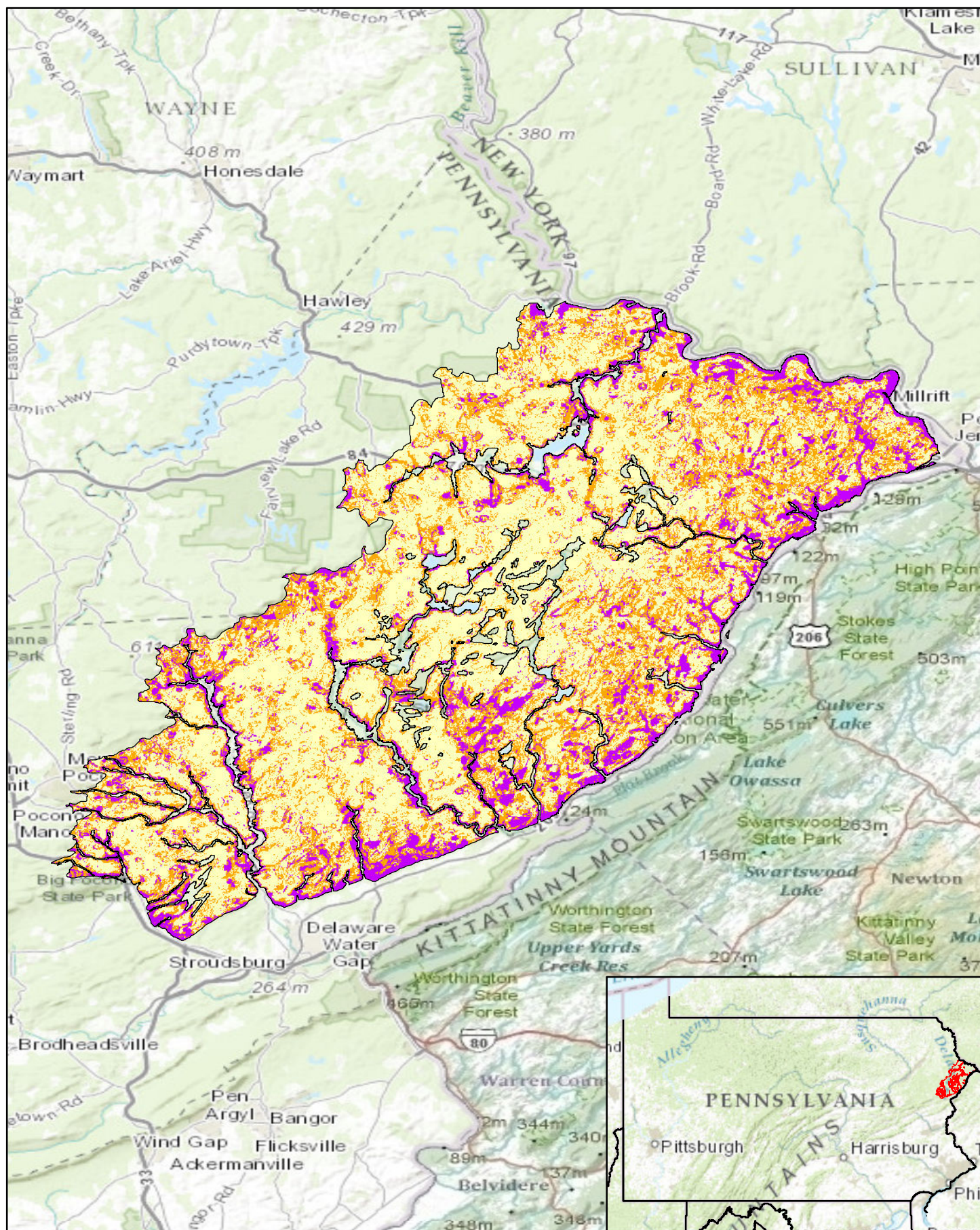


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: upland section 6

Sensitivity

- High
- Moderate
- Low



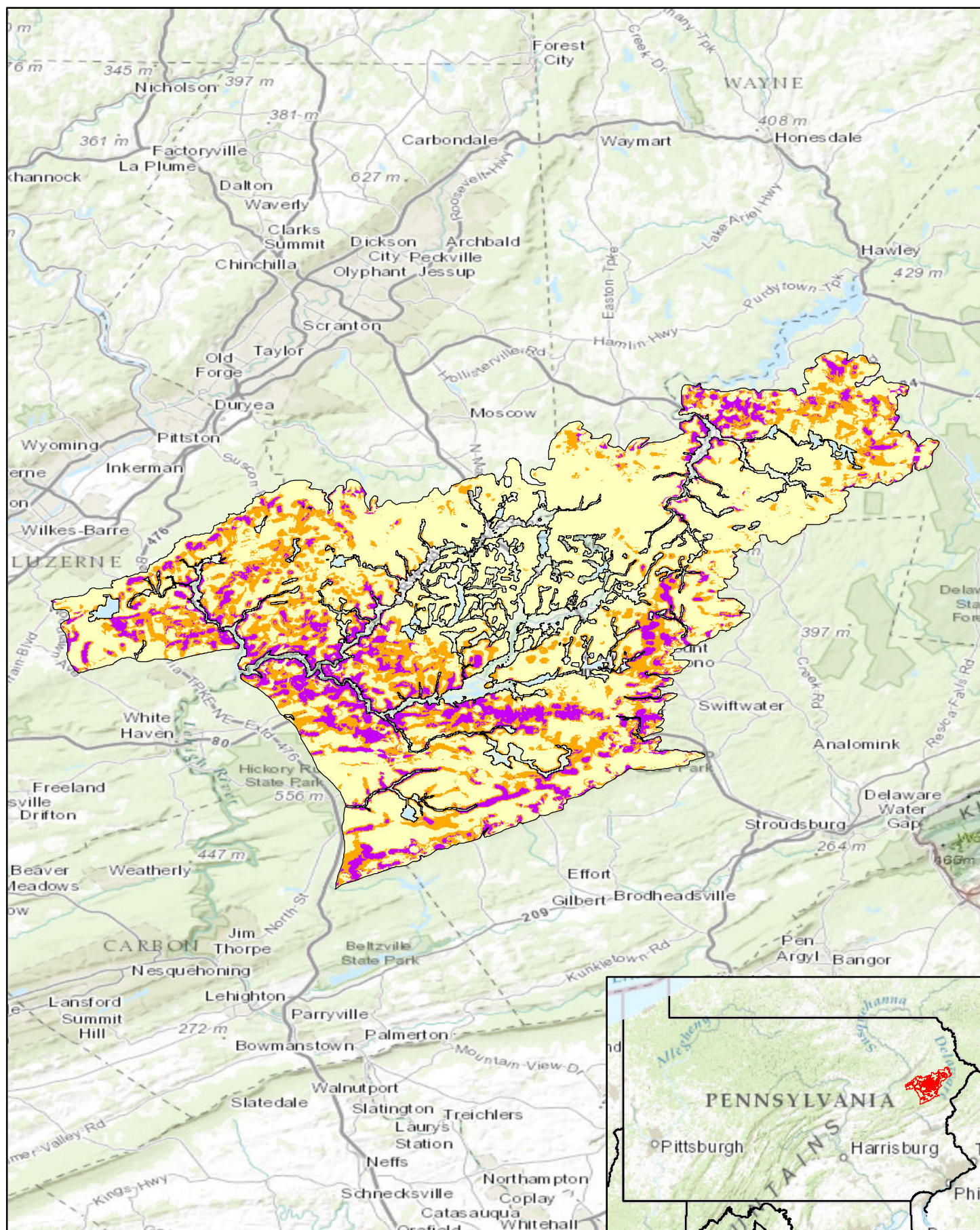


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: upland section 7

Sensitivity
 High
 Moderate
 Low

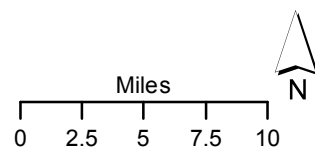
Miles
 0 2 4 6 8

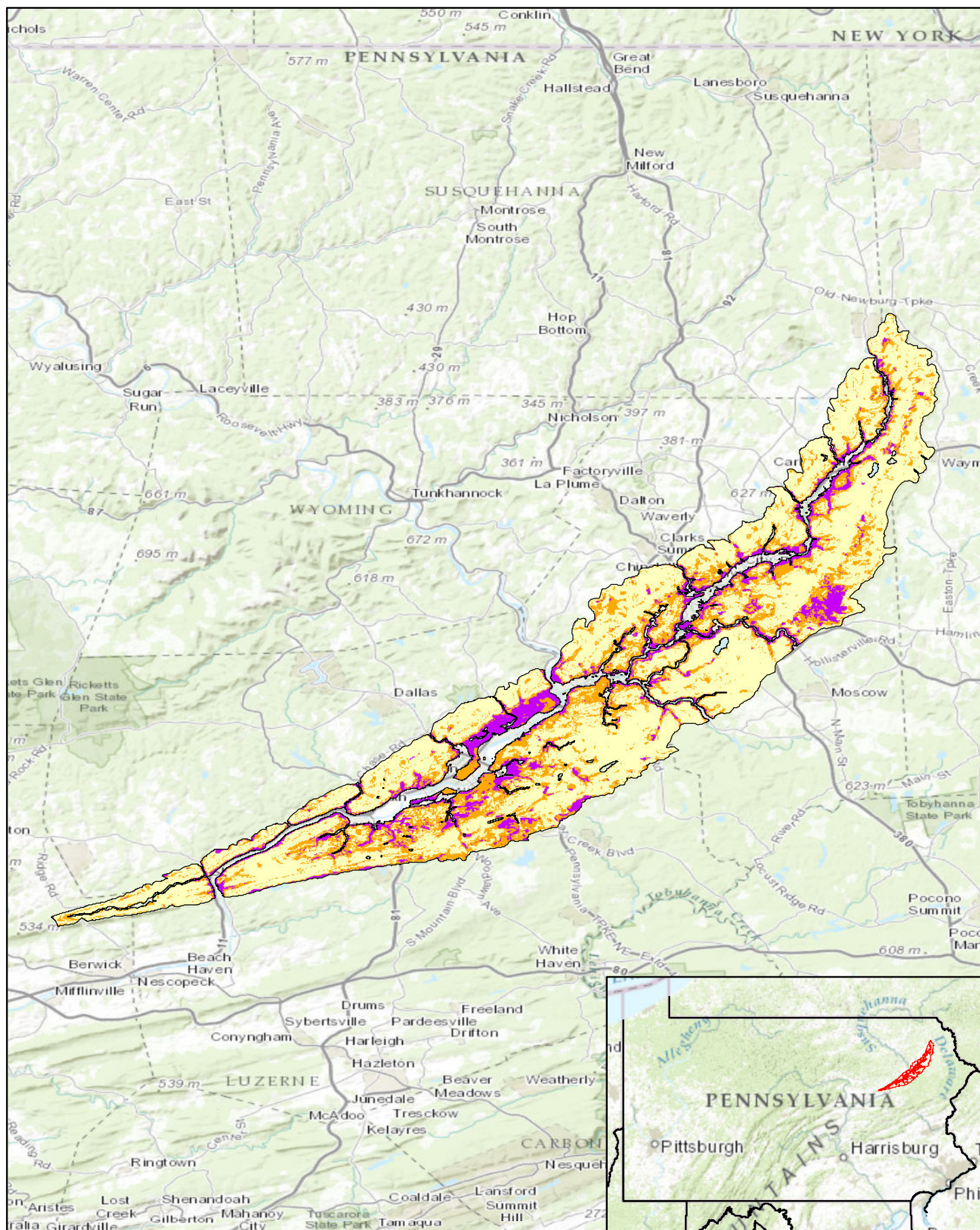




Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: upland section 8

Sensitivity
 High
 Moderate
 Low

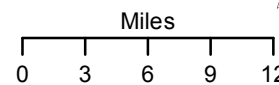


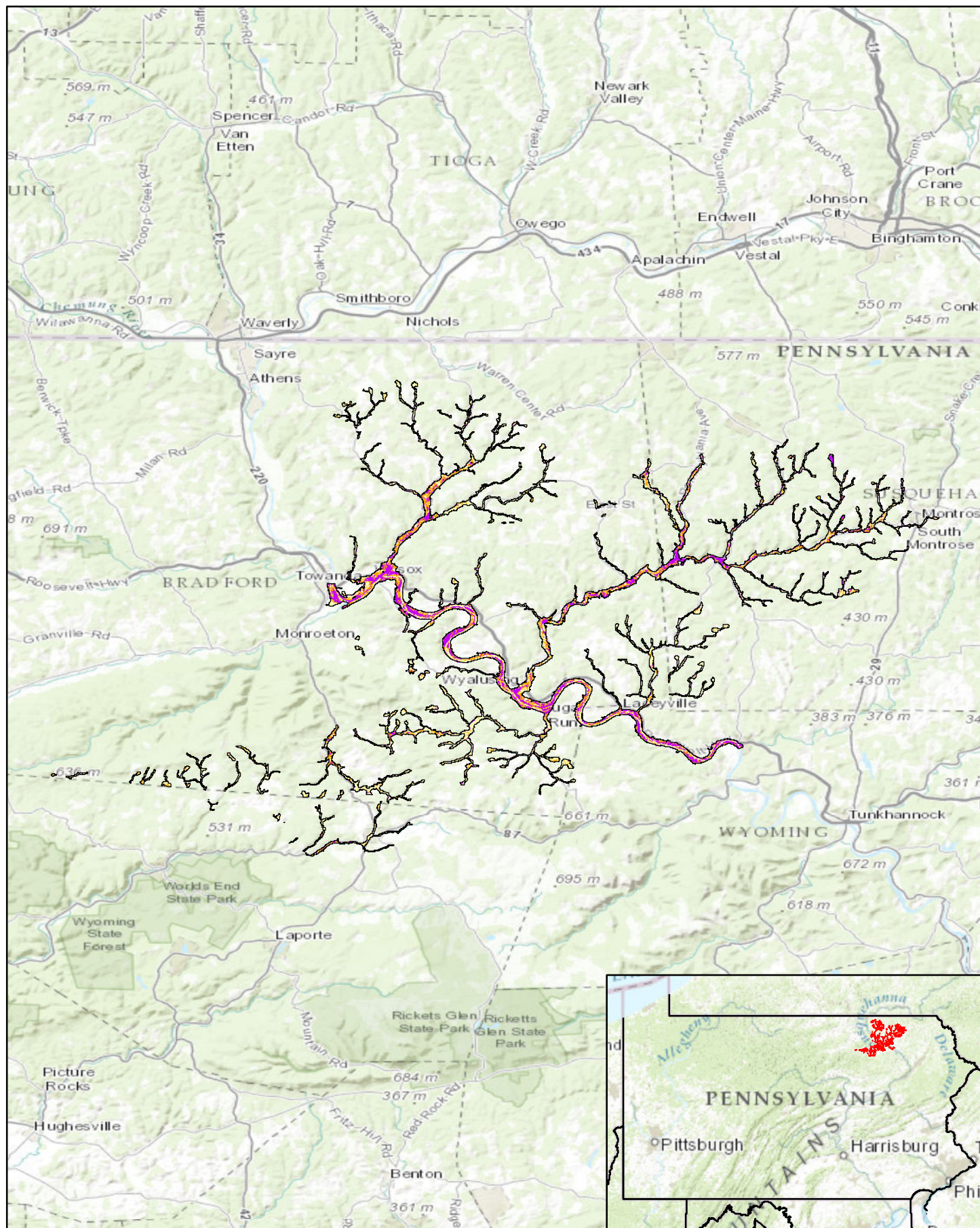


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: upland section 9

Sensitivity

- High
- Moderate
- Low

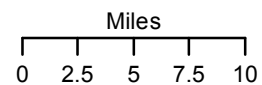


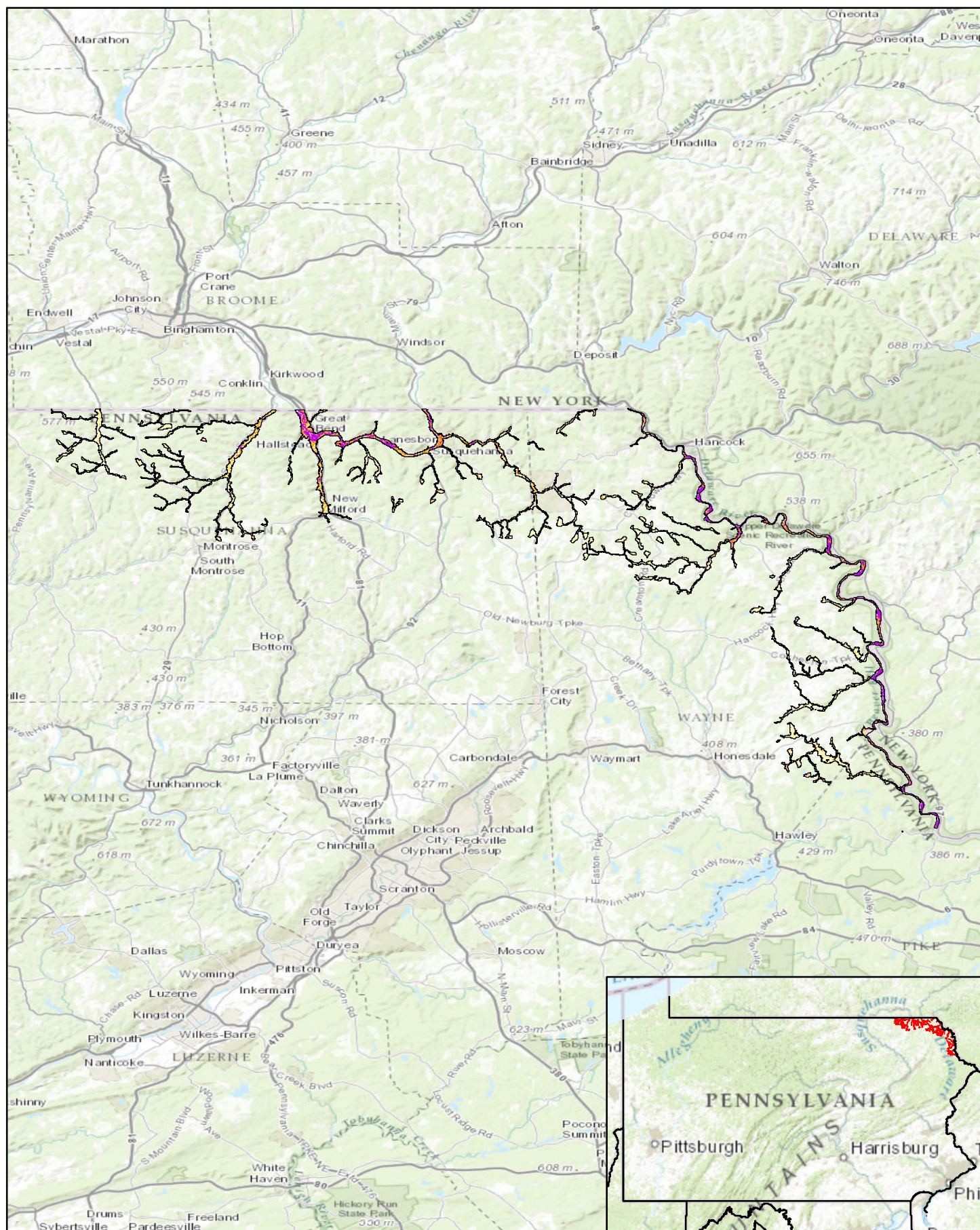


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: riverine section 2

Sensitivity

- High
- Moderate
- Low

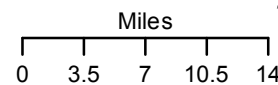


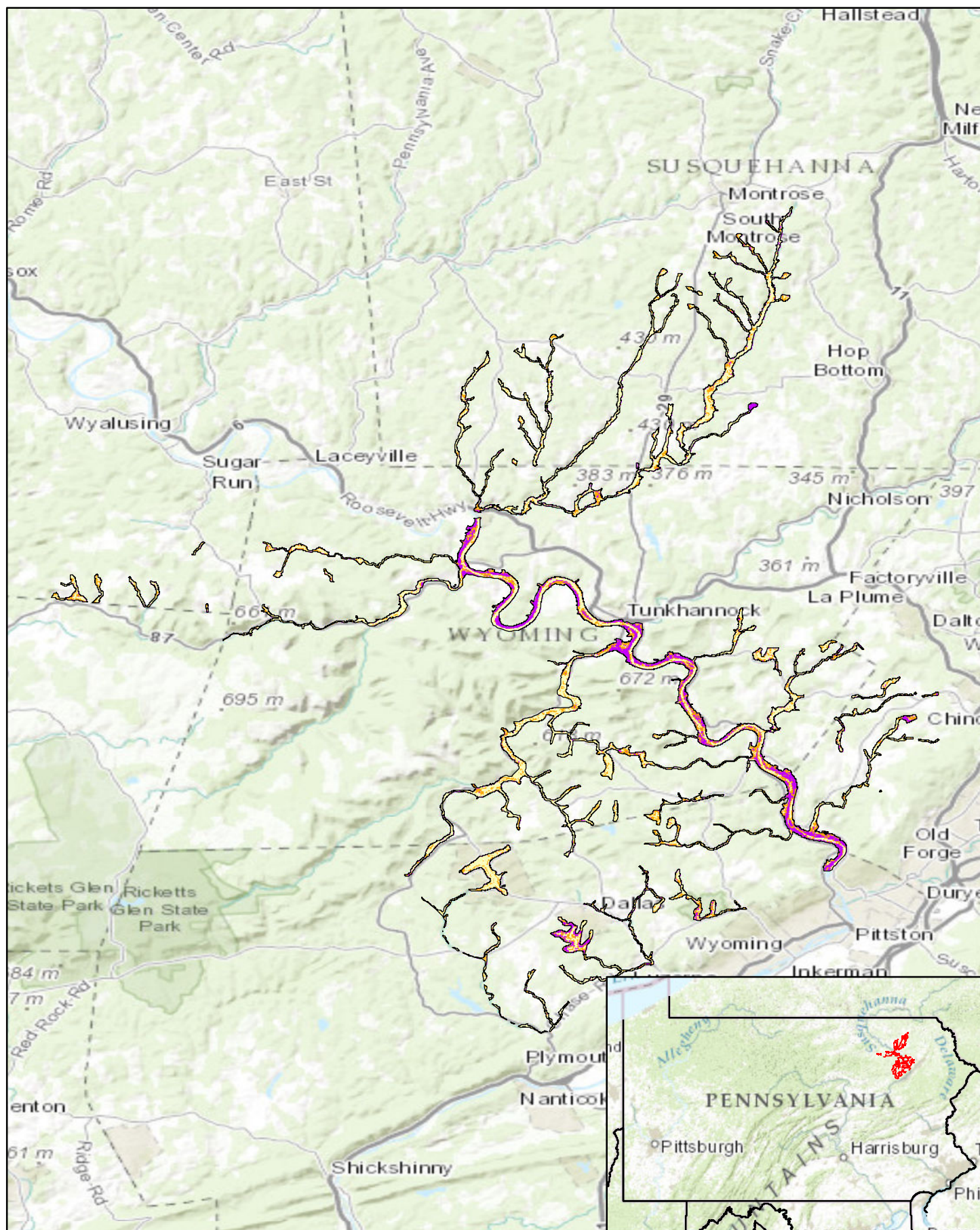


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: riverine section 3

Sensitivity

- High
- Moderate
- Low

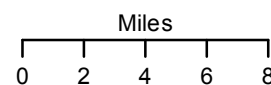


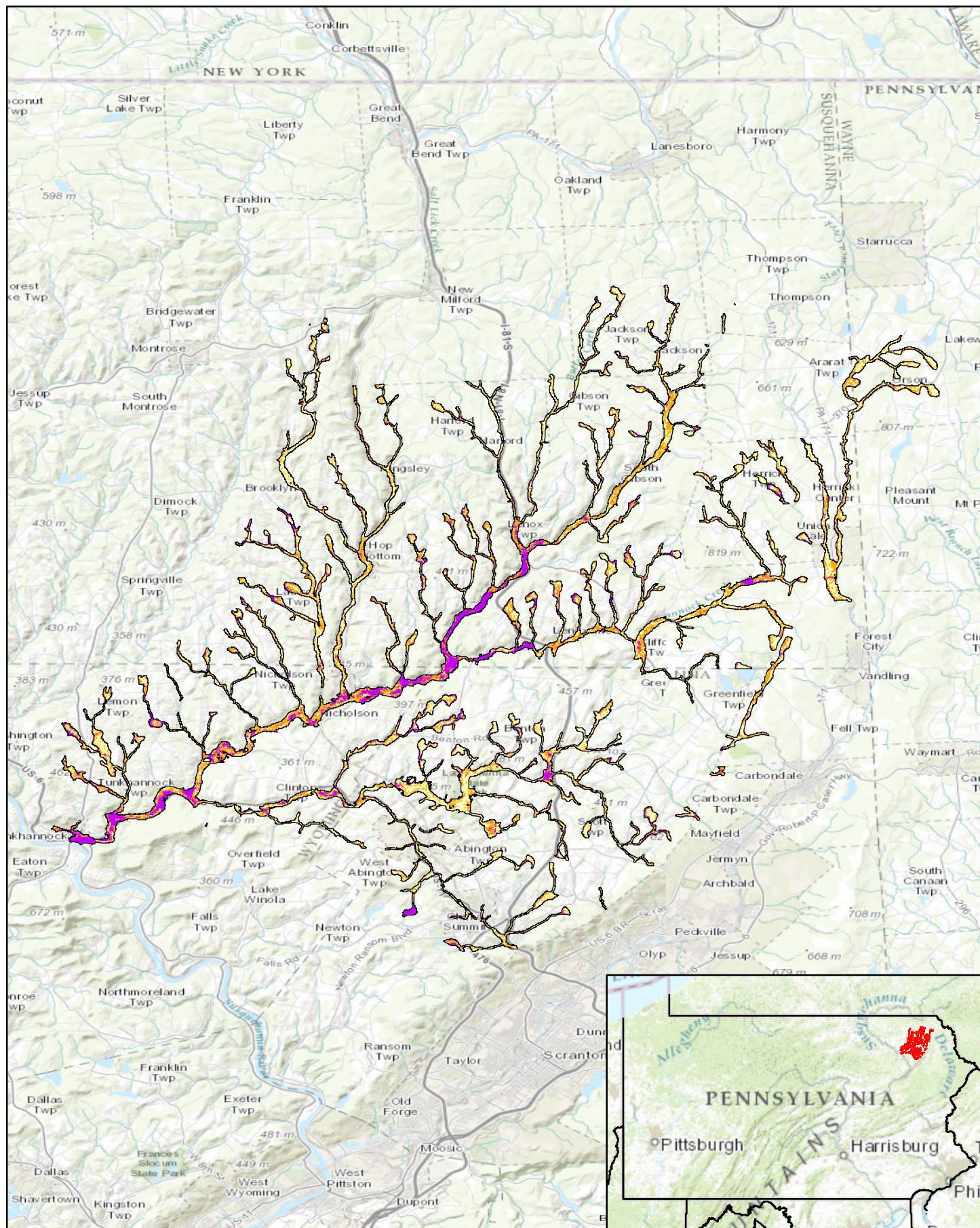


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: riverine section 4

Sensitivity

- High
- Moderate
- Low

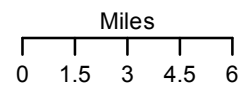


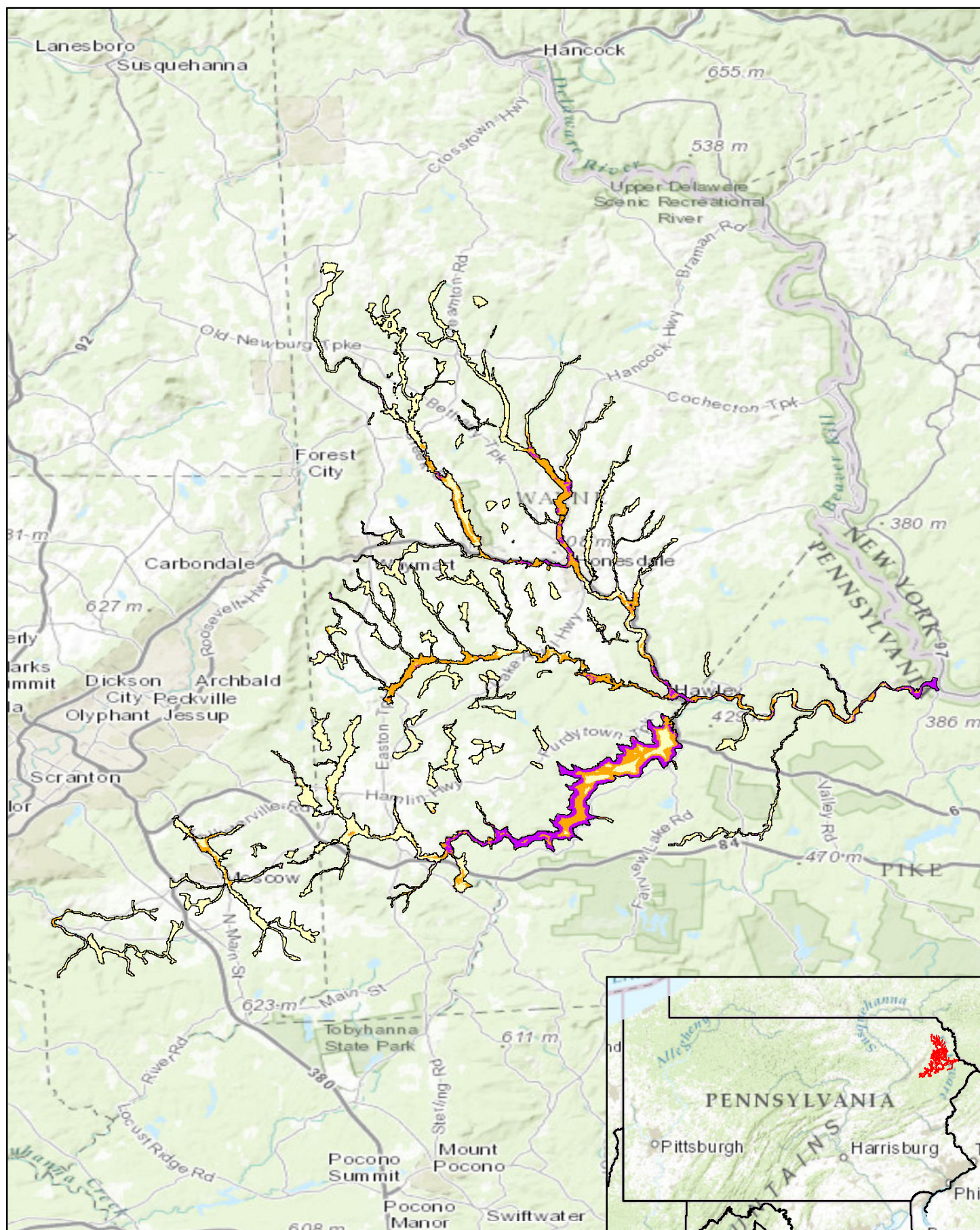


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: riverine section 5

Sensitivity

- High
- Moderate
- Low

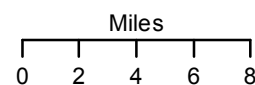


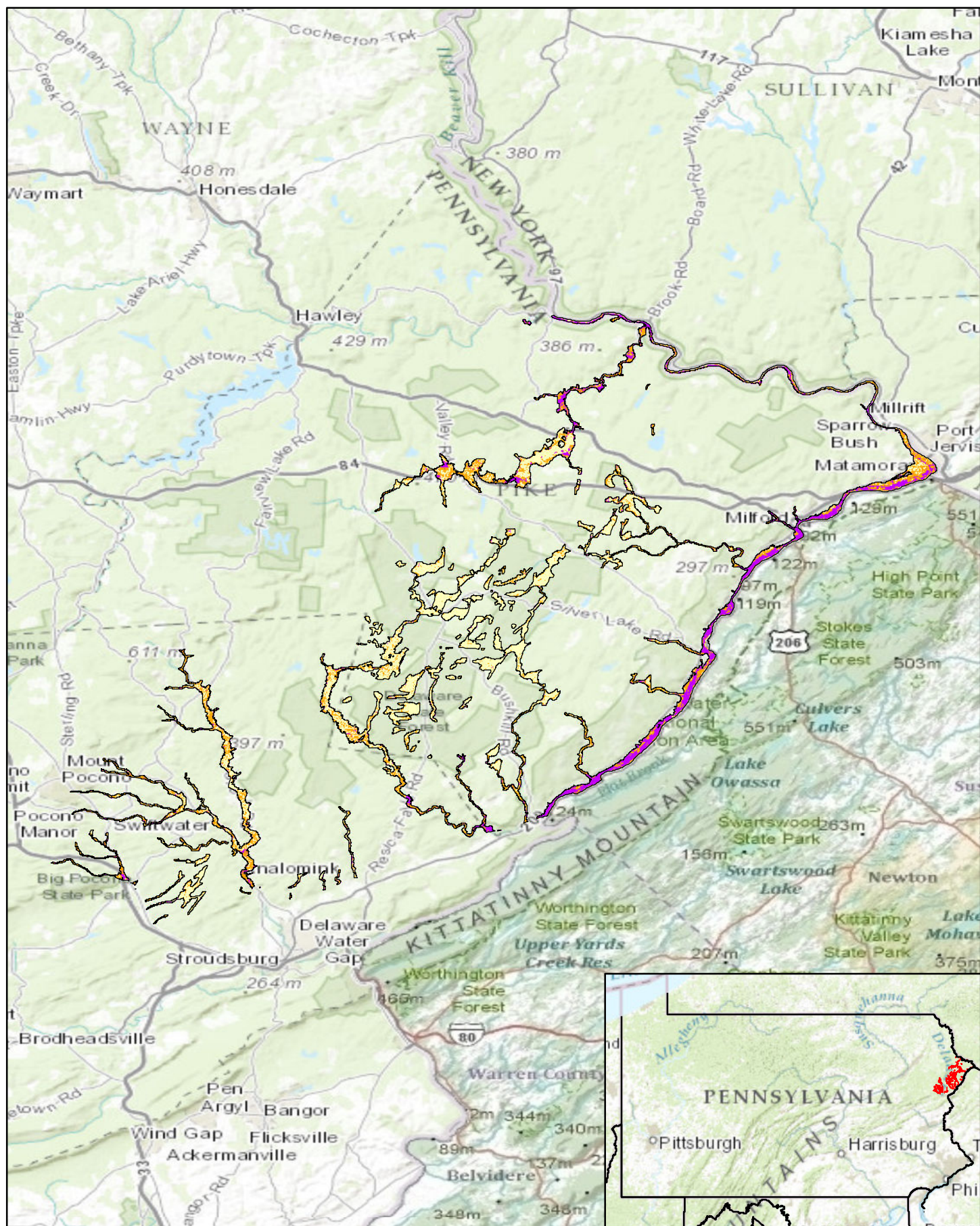


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: riverine section 6

Sensitivity

- High
- Moderate
- Low





Pennsylvania Predictive Model Set

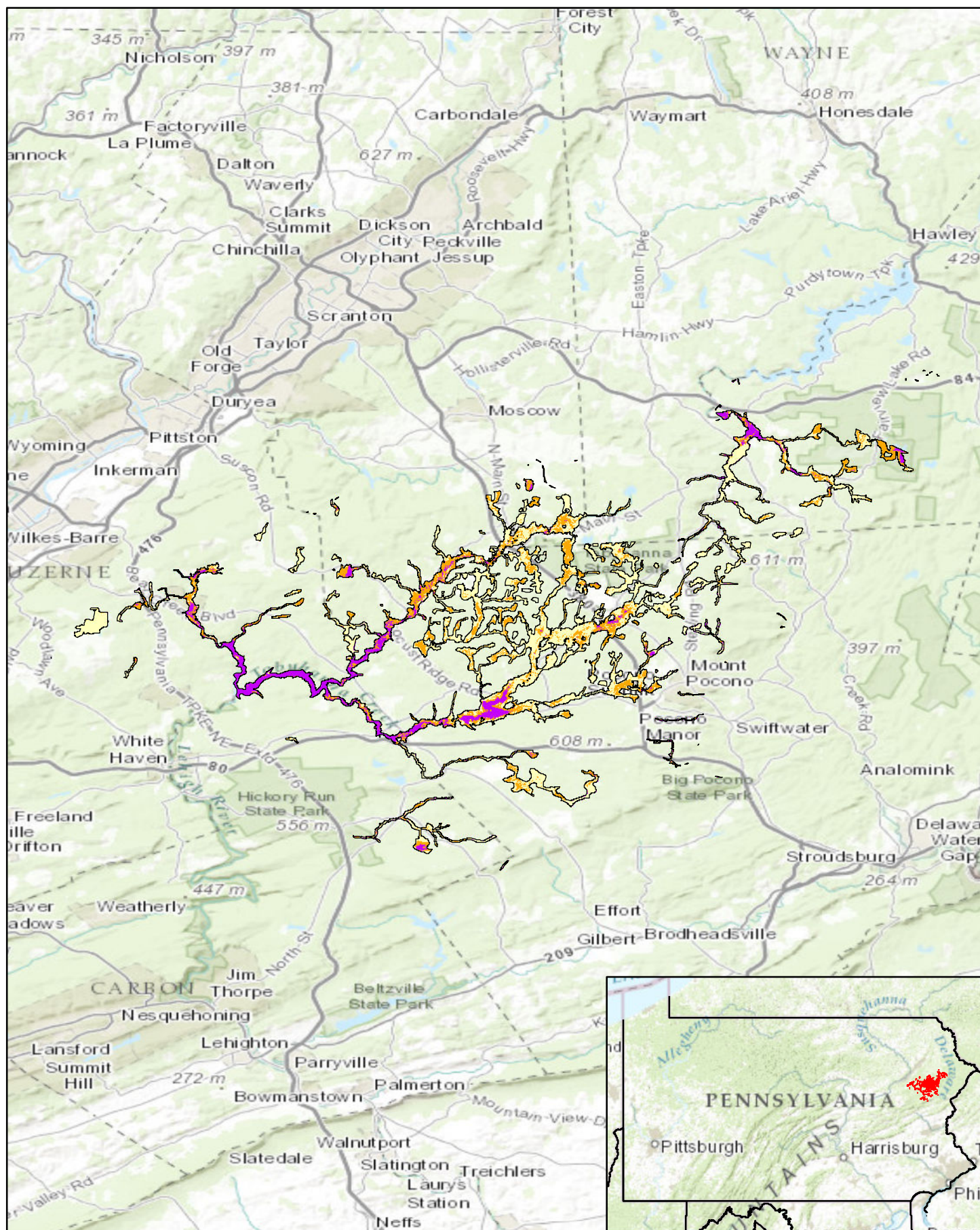
Region: 7, Zone: all, Subarea: riverine section 7

Sensitivity

- High
- Moderate
- Low

Miles
0 2 4 6 8



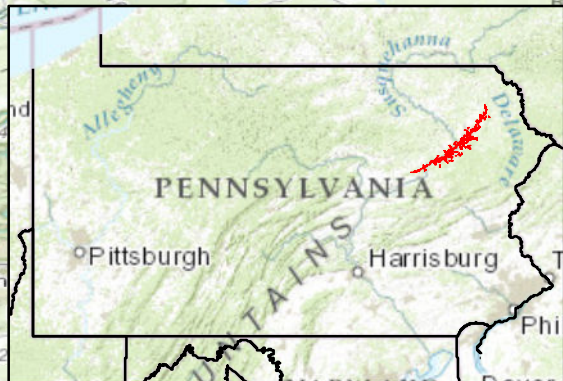


Pennsylvania Predictive Model Set
 Region: 7, Zone: all, Subarea: riverine section 8




Sensitivity
 High
 Moderate
 Low

Miles
 0 2 4 6 8





Region: 7, Zone: all, Subarea: riverine section 9

	High
	Moderate
	Low

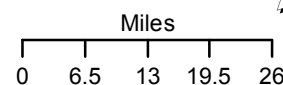


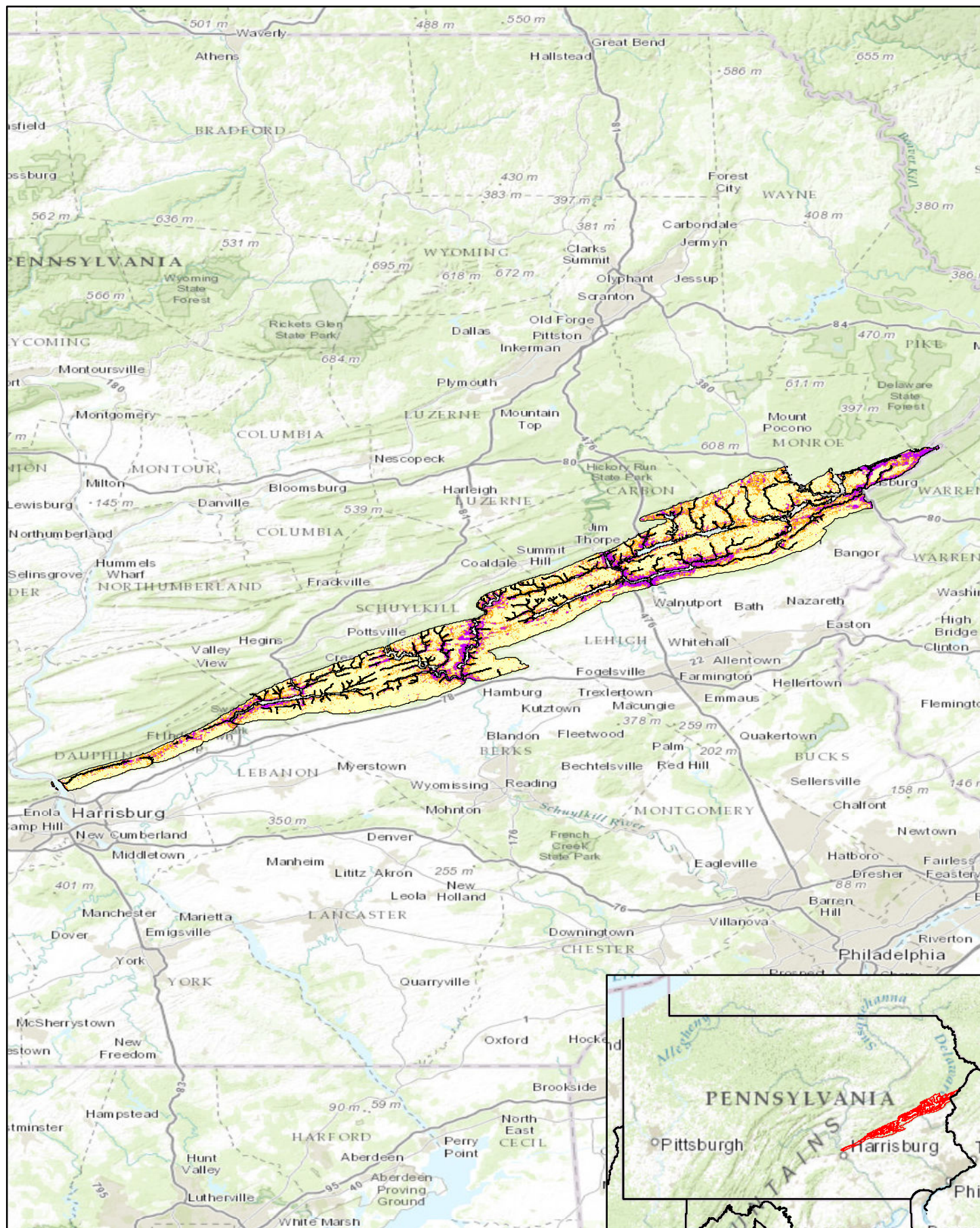


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: riverine section 1

Sensitivity

- High
- Moderate
- Low

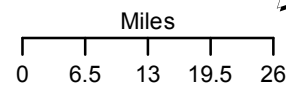


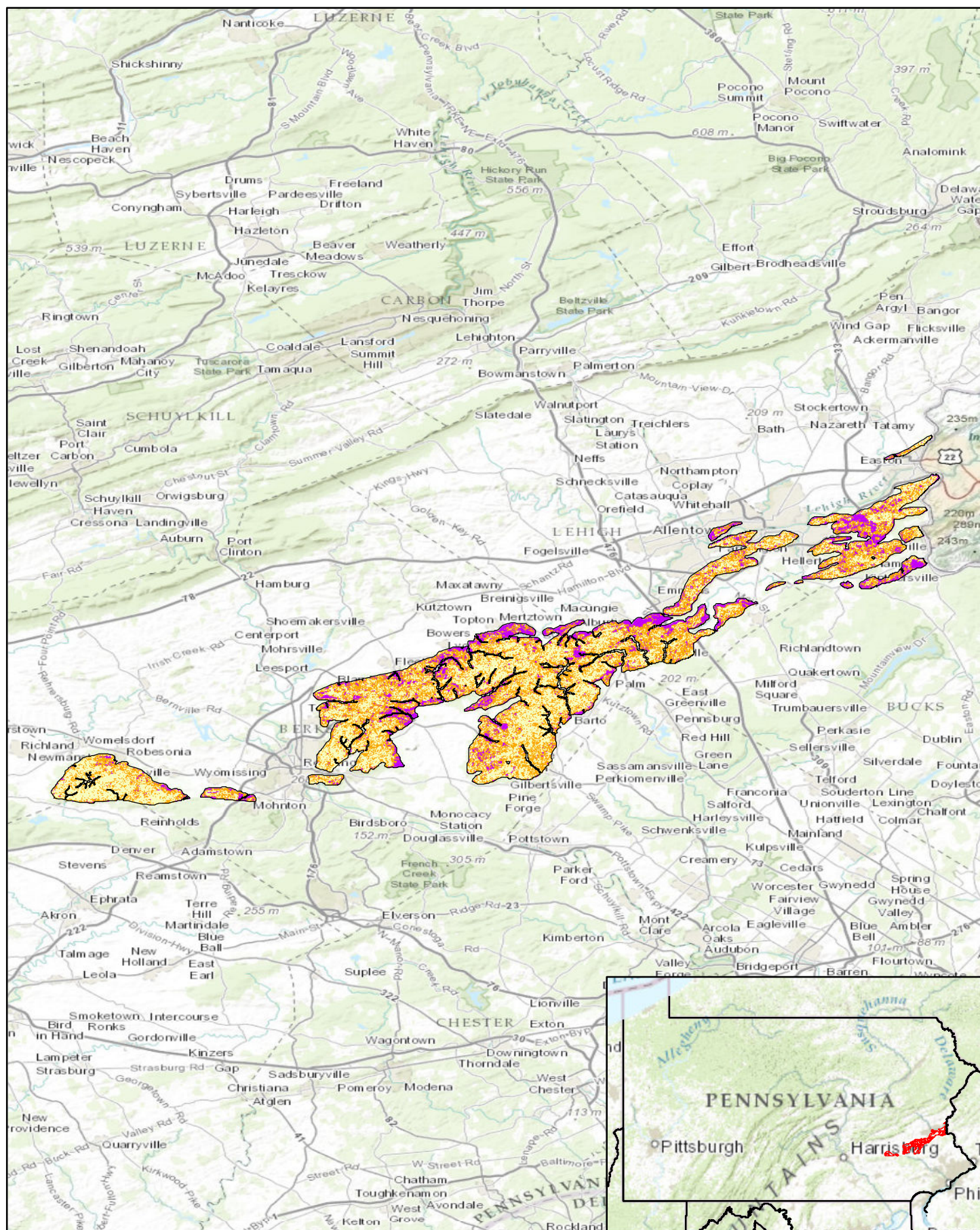


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: upland section 1

Sensitivity

- High
- Moderate
- Low

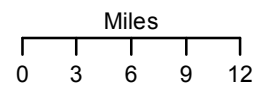


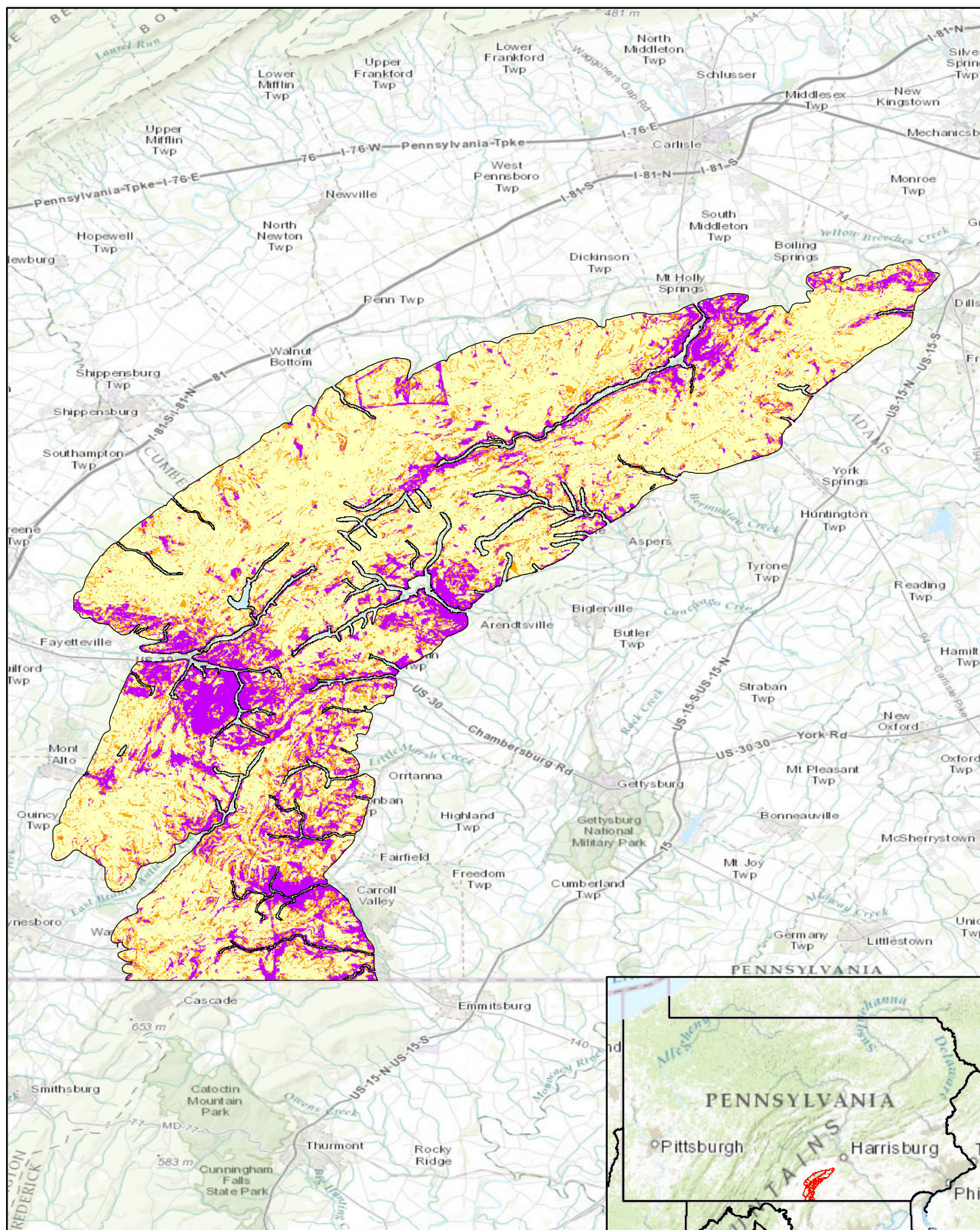


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: upland section 2

Sensitivity

- High
- Moderate
- Low



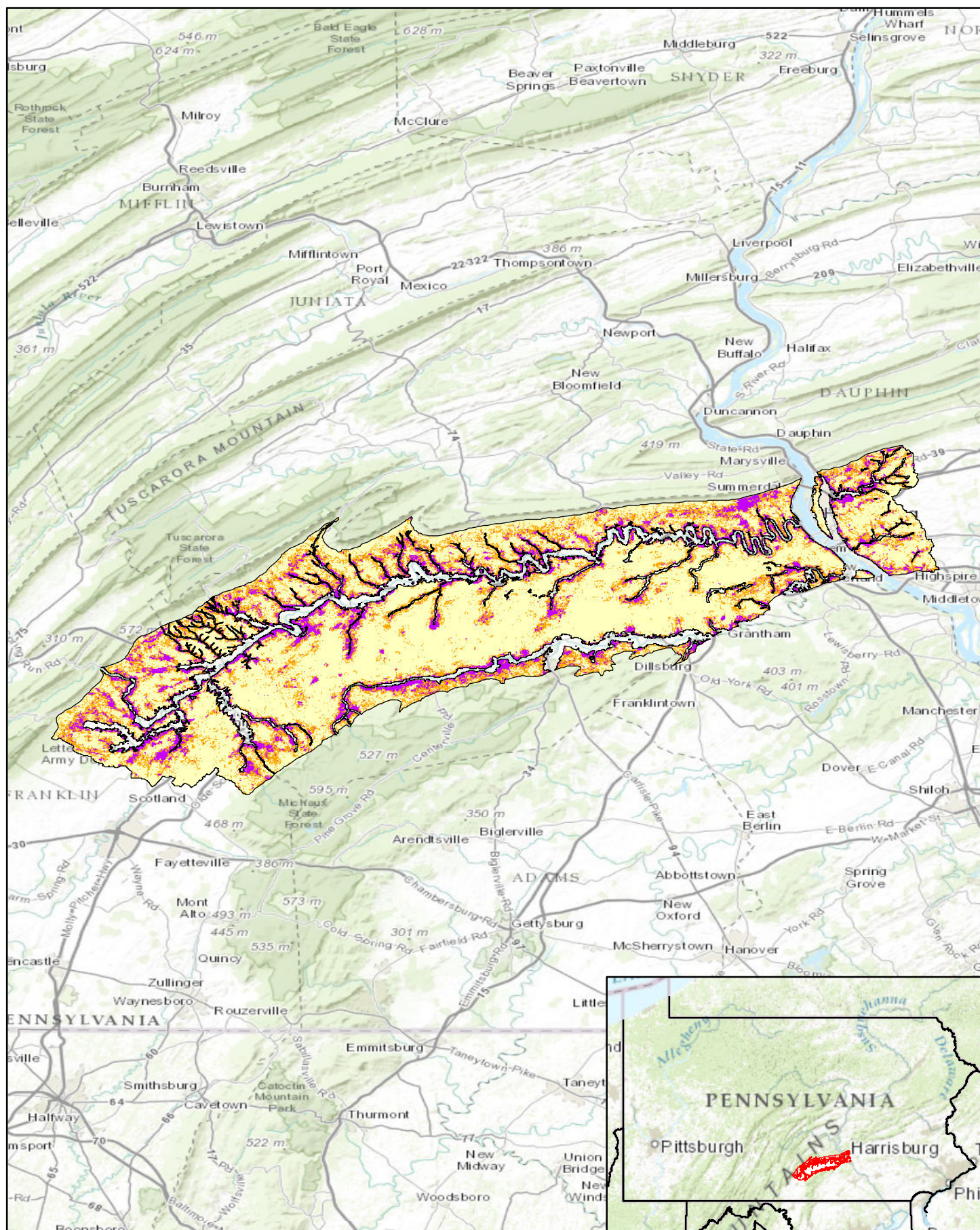


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: upland section 3

Sensitivity
 High
 Moderate
 Low

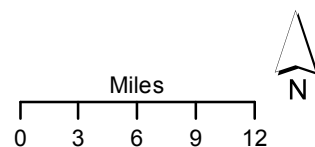
Miles
 0 1.5 3 4.5 6

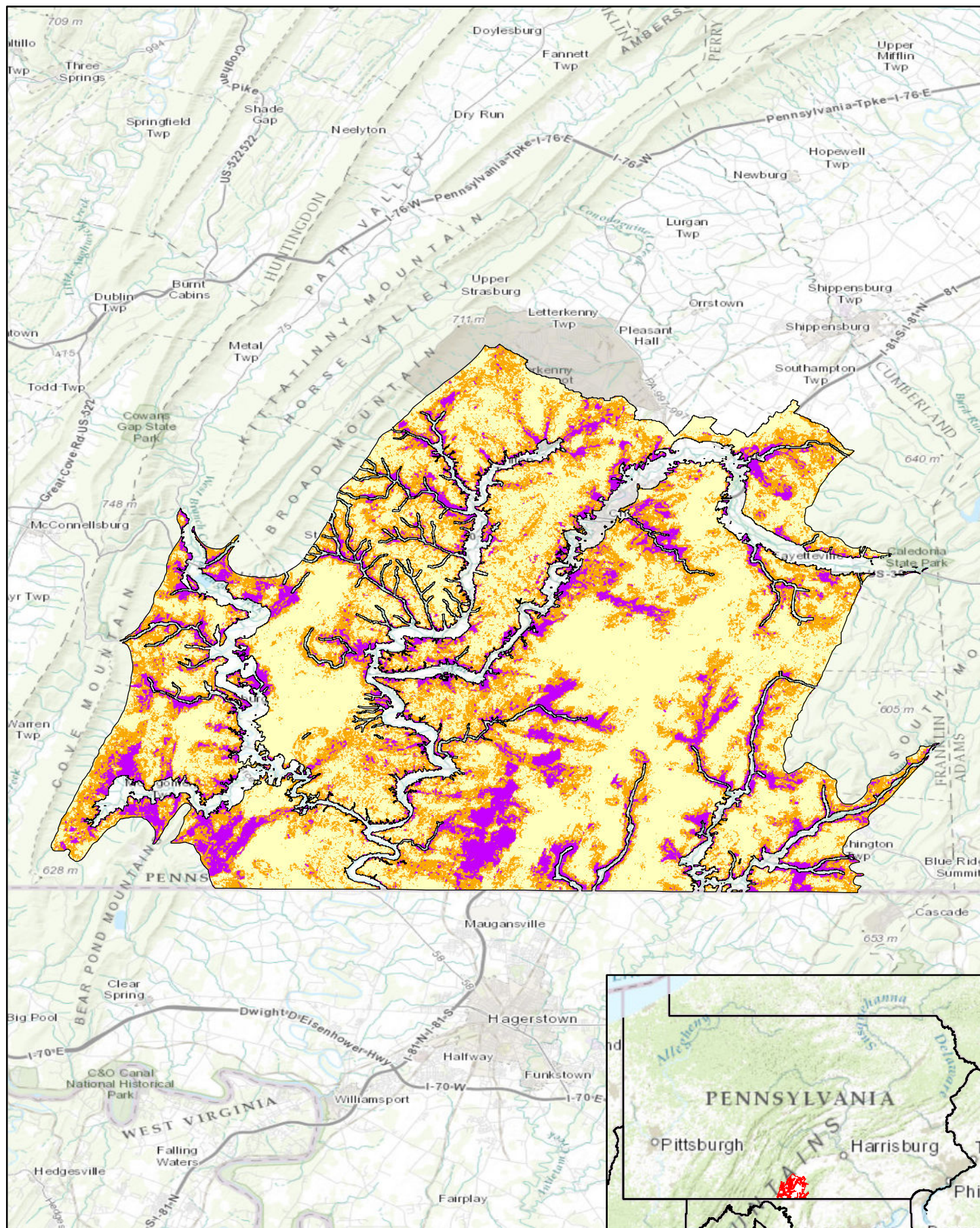




Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: upland section 4

Sensitivity
 High
 Moderate
 Low



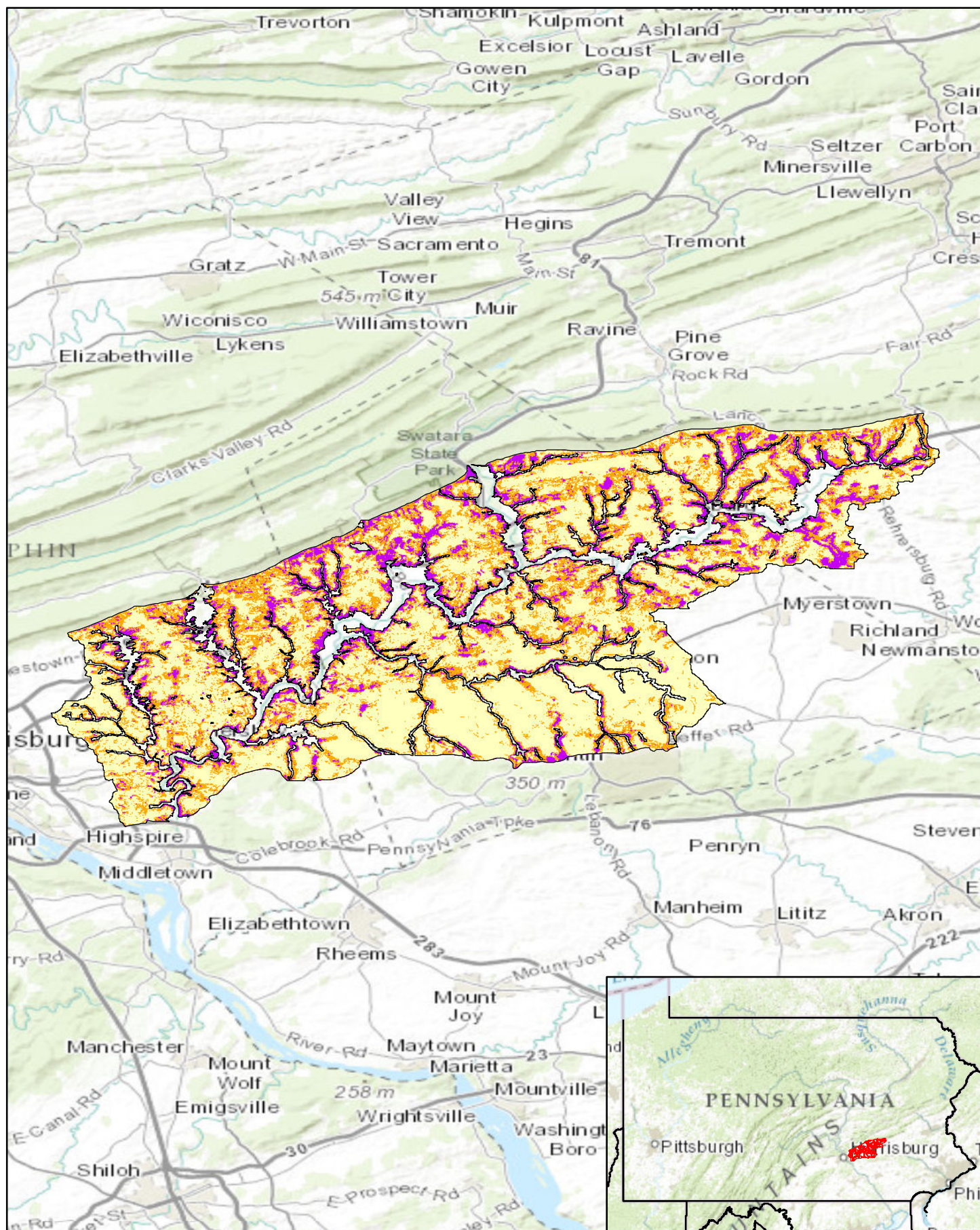


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: upland section 5

Sensitivity
 High
 Moderate
 Low

Miles
 0 1.5 3 4.5 6





Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: upland section 6




Sensitivity
 High
 Moderate
 Low

Miles
 0 2 4 6 8

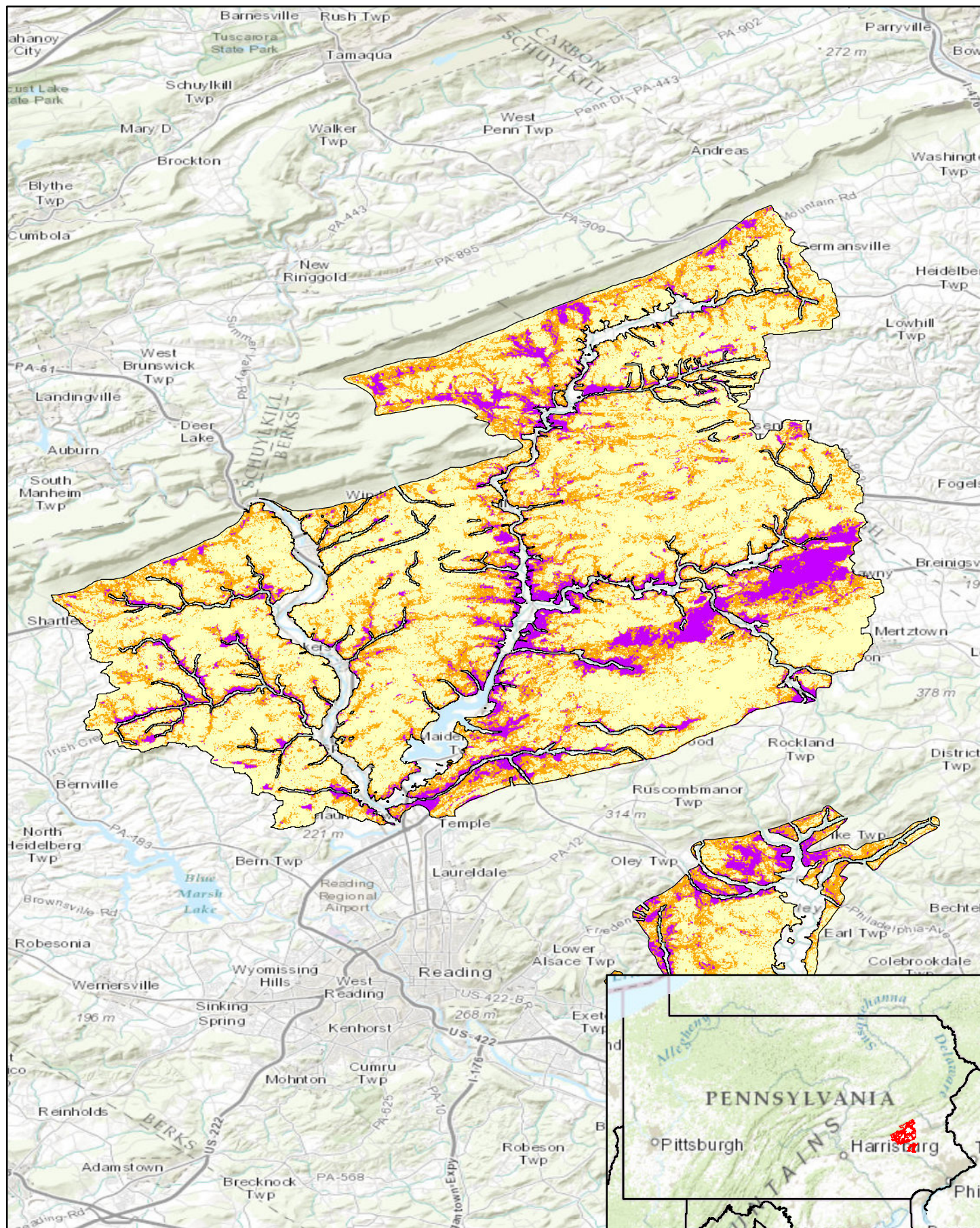




Region: 8, Zone: all, Subarea: upland section 7

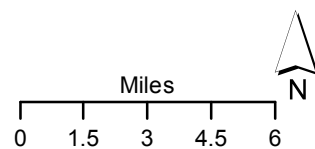
	High
	Moderate
	Low

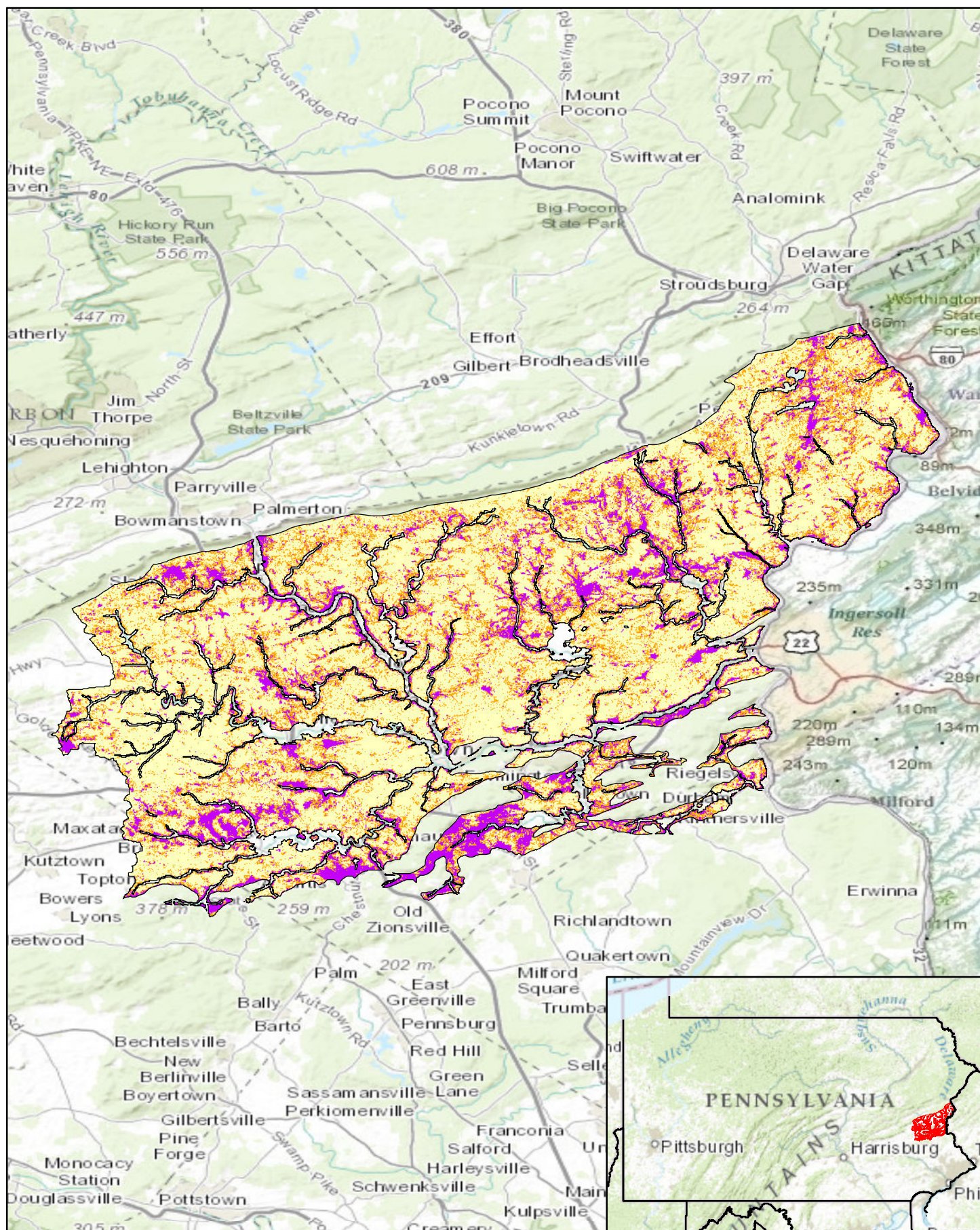




Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: upland section 8

Sensitivity
 High
 Moderate
 Low



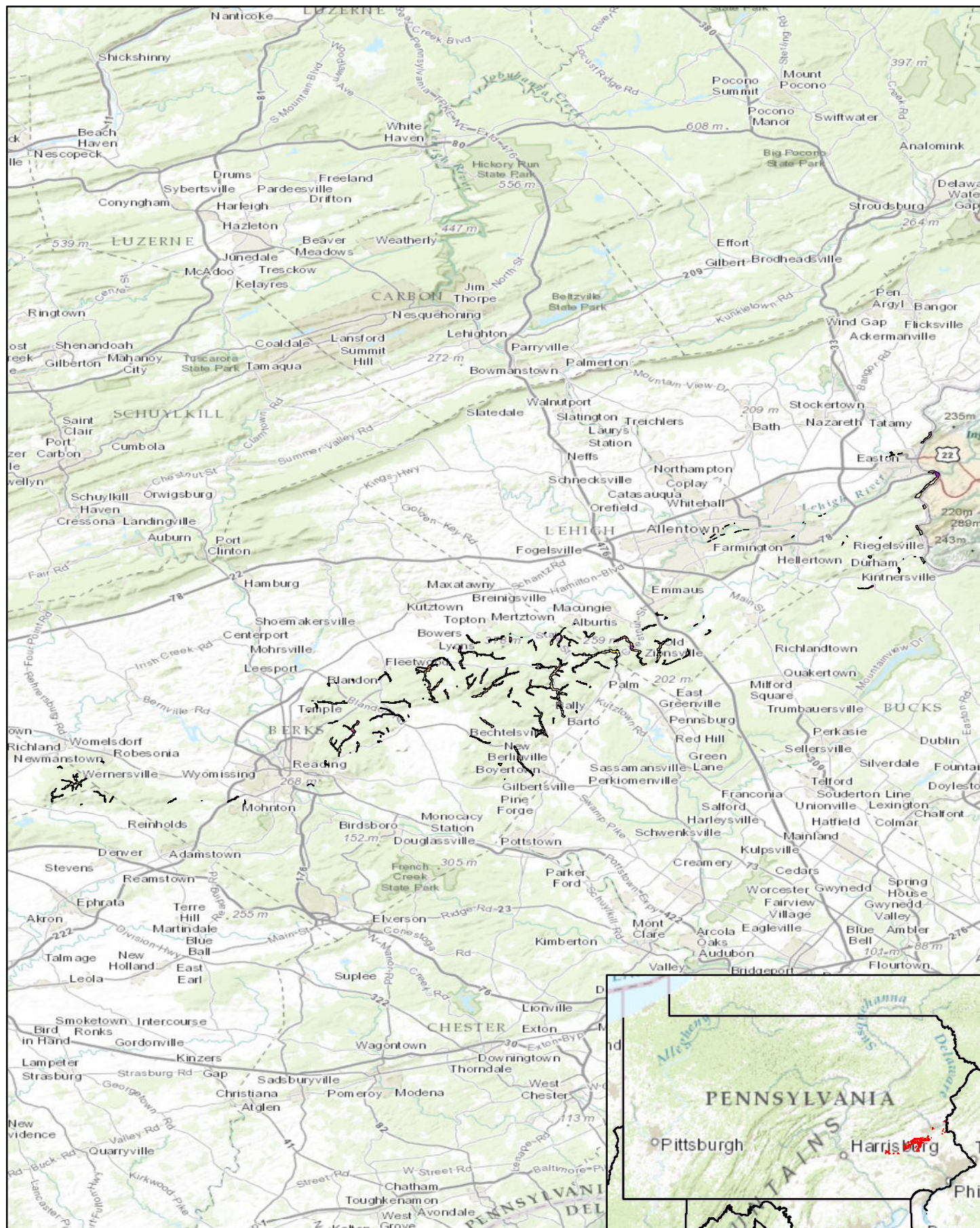


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: upland section 9

Sensitivity
 High
 Moderate
 Low

Miles
 0 2 4 6 8

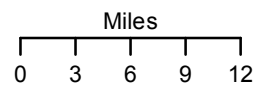


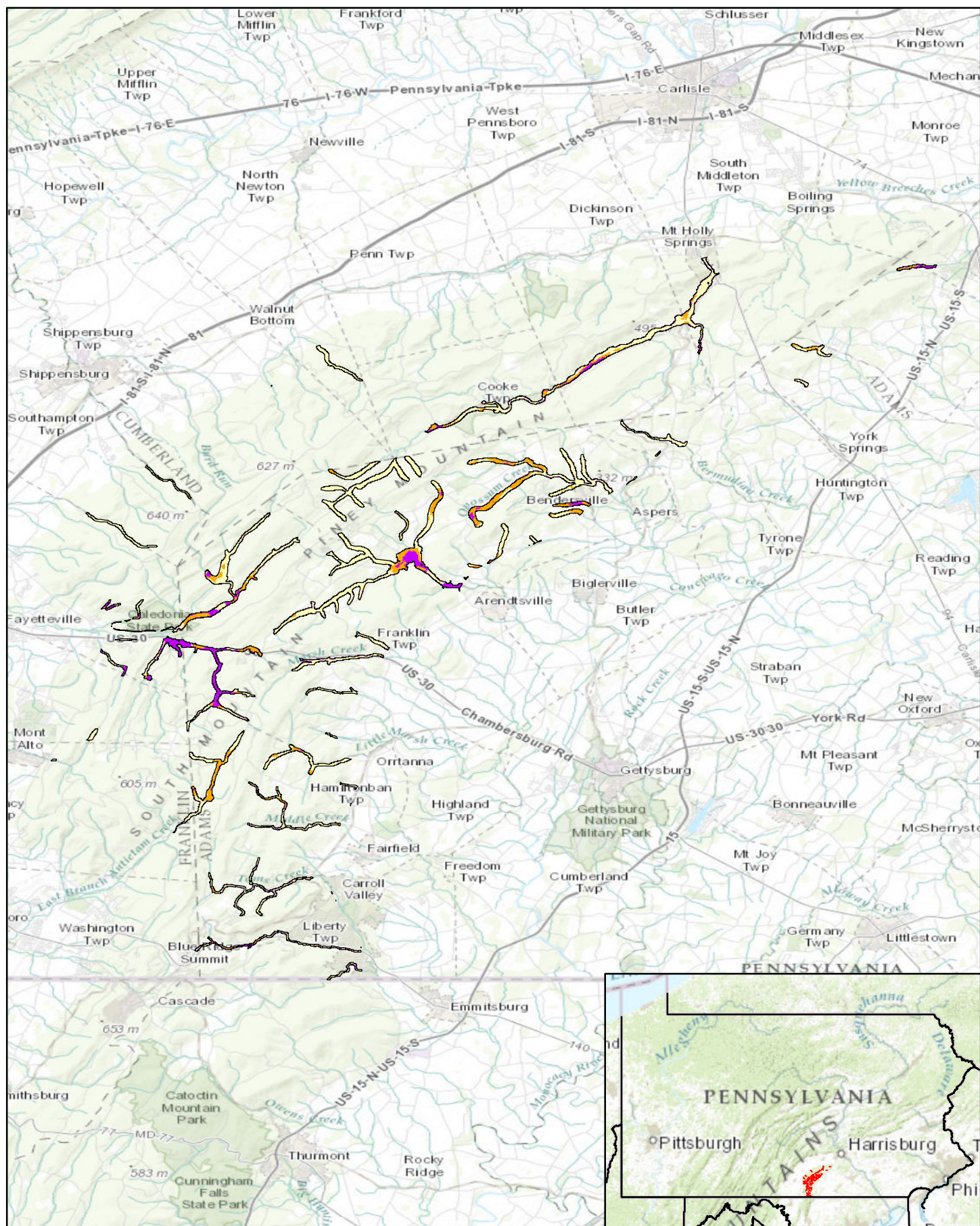


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: riverine section 2

Sensitivity

- High
- Moderate
- Low





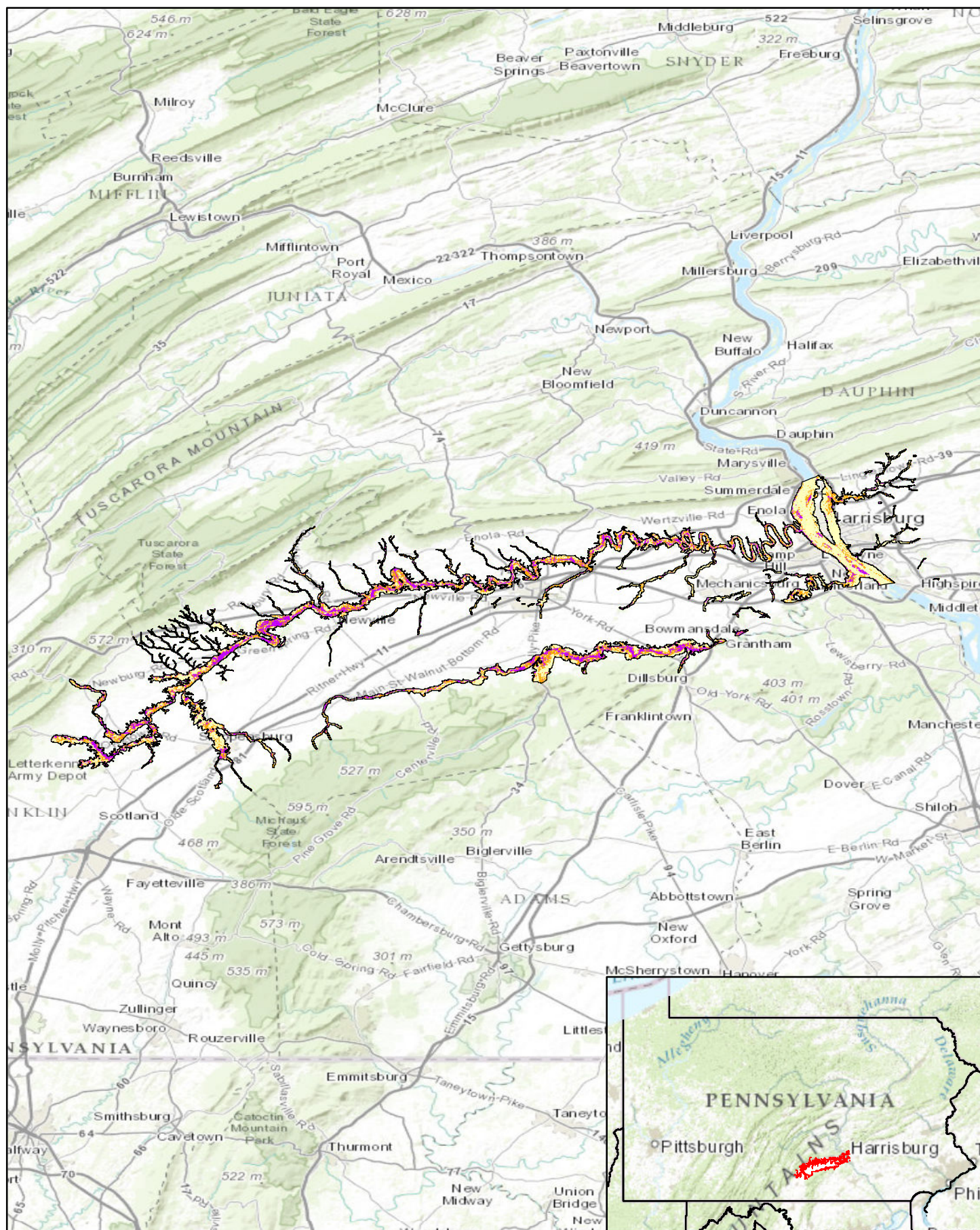
Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: riverine section 3

Sensitivity

- High
- Moderate
- Low

Miles
 0 1.5 3 4.5 6

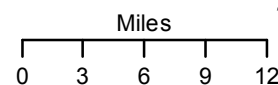


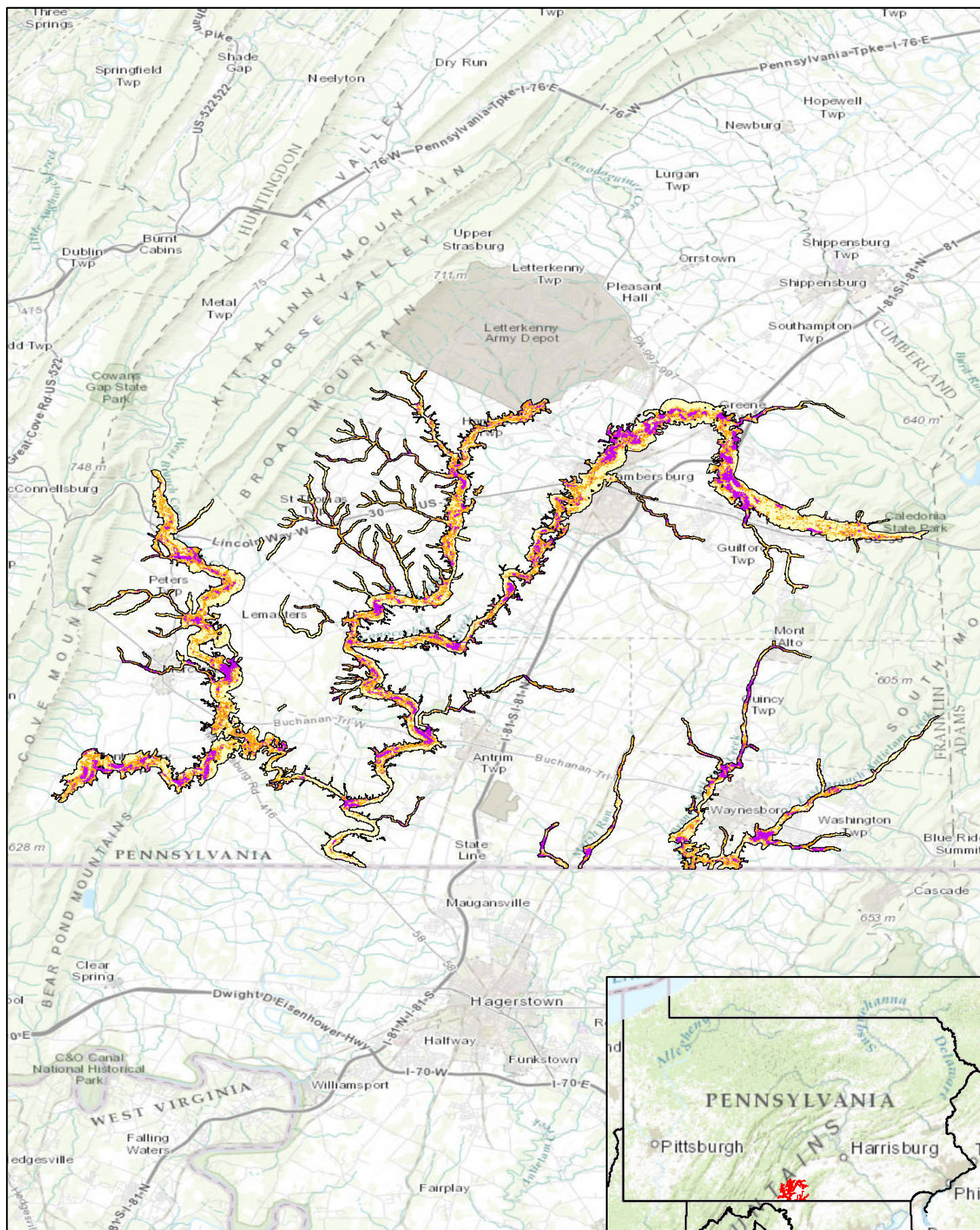


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: riverine section 4

Sensitivity

- High
- Moderate
- Low

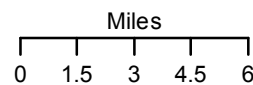


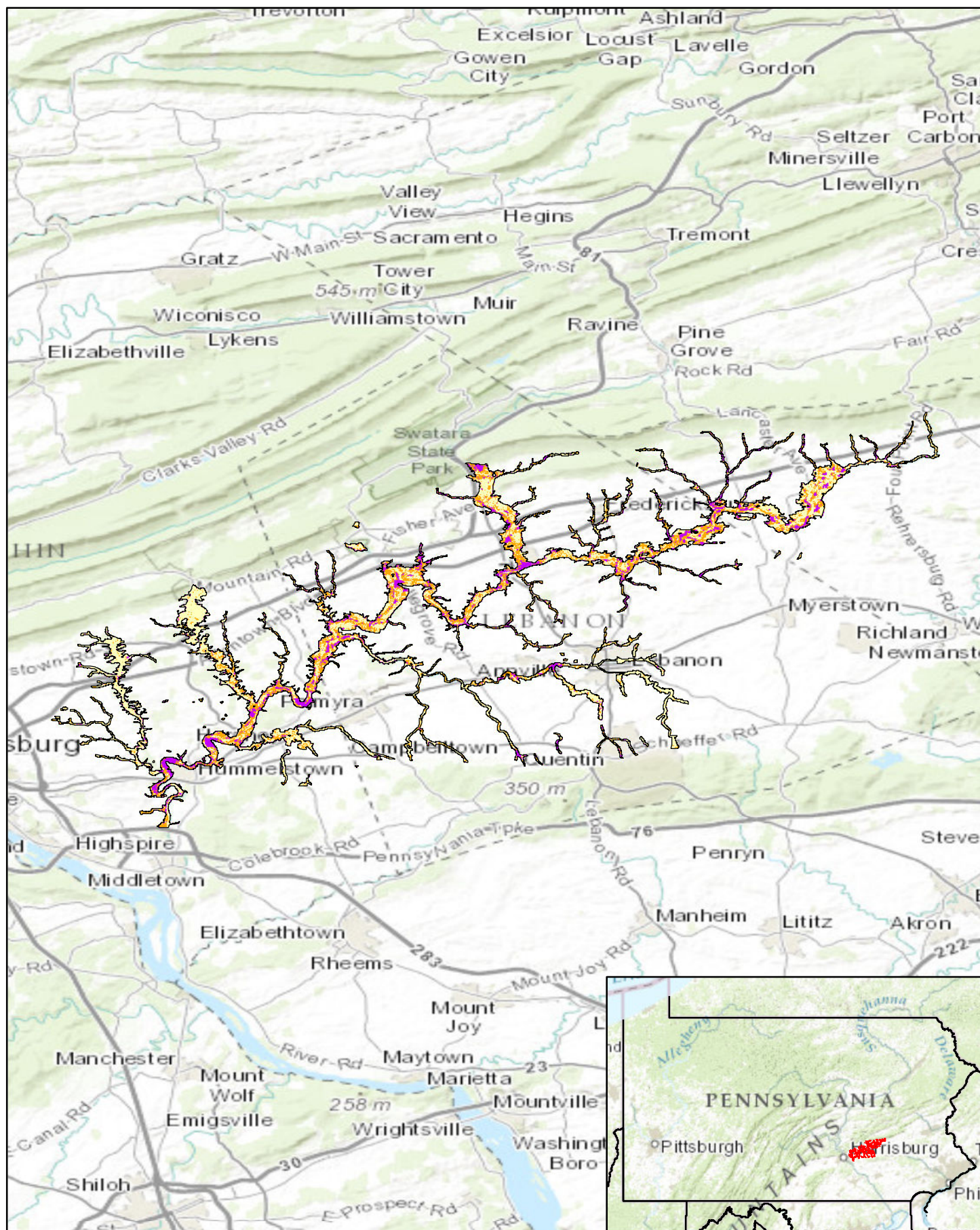


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: riverine section 5

Sensitivity

- High
- Moderate
- Low

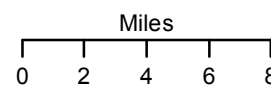


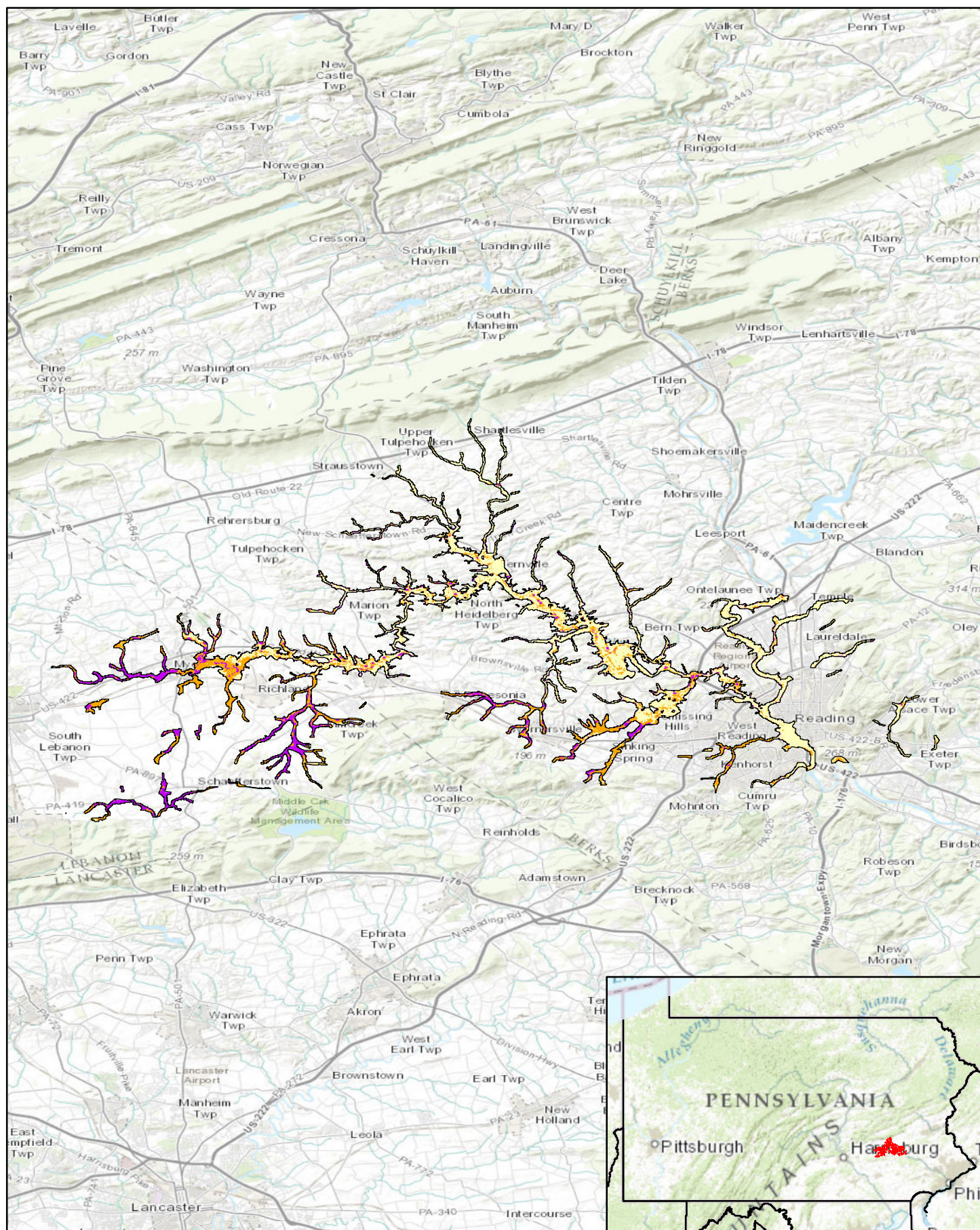


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: riverine section 6

Sensitivity

- High
- Moderate
- Low

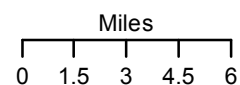


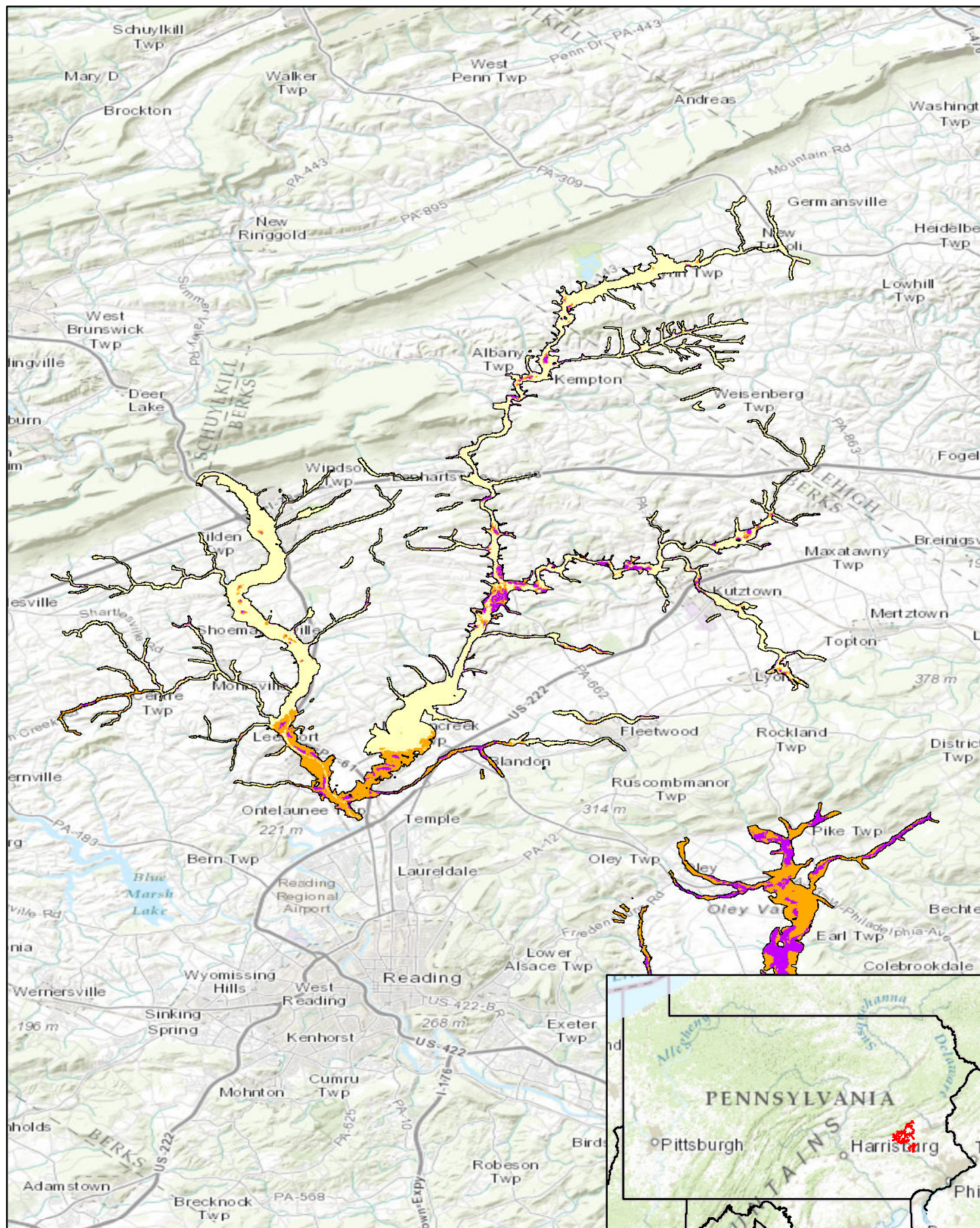


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: riverine section 7

Sensitivity

- High
- Moderate
- Low

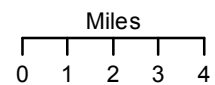


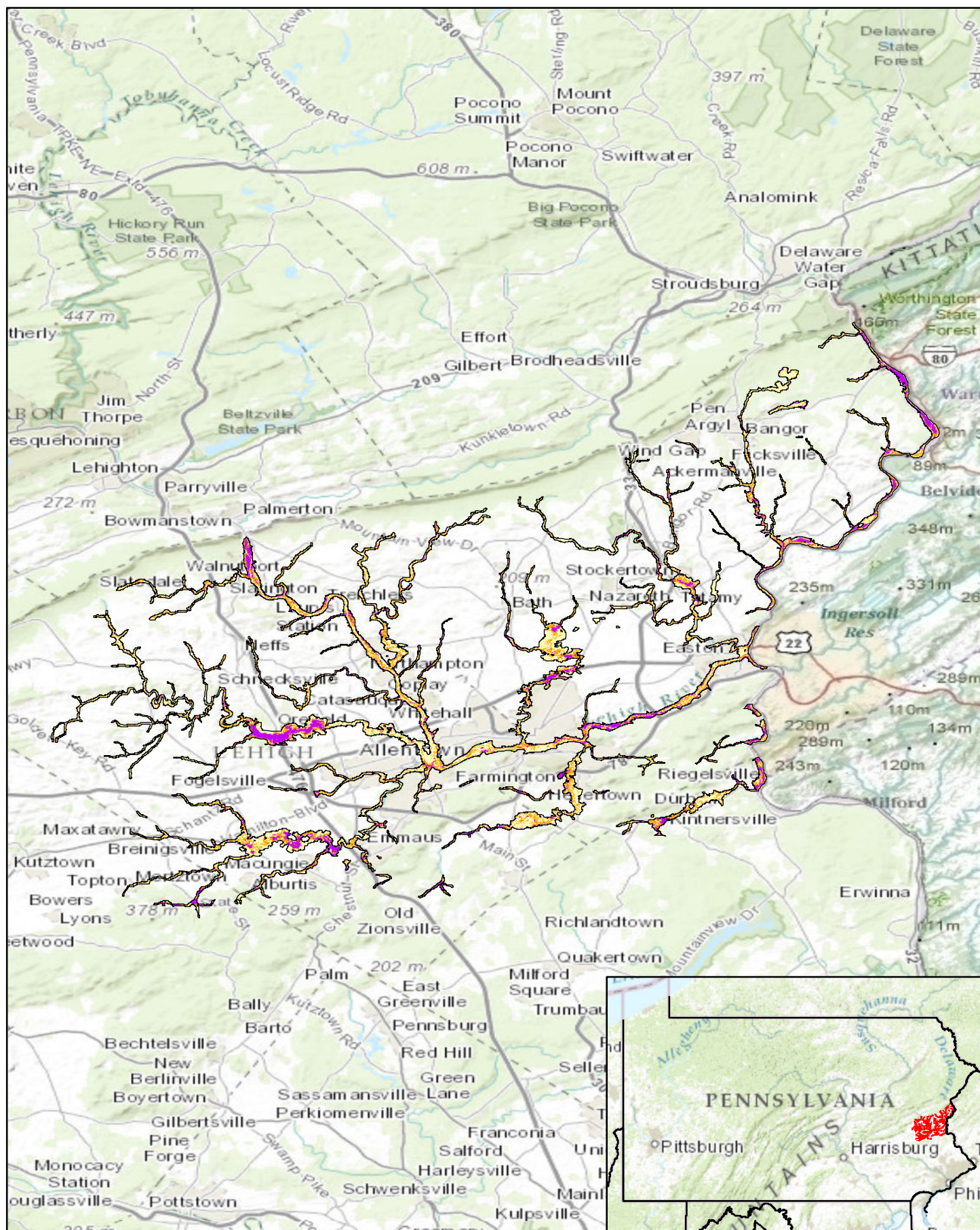


Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: riverine section 8

Sensitivity

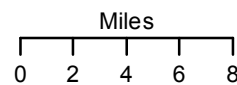
- High
- Moderate
- Low

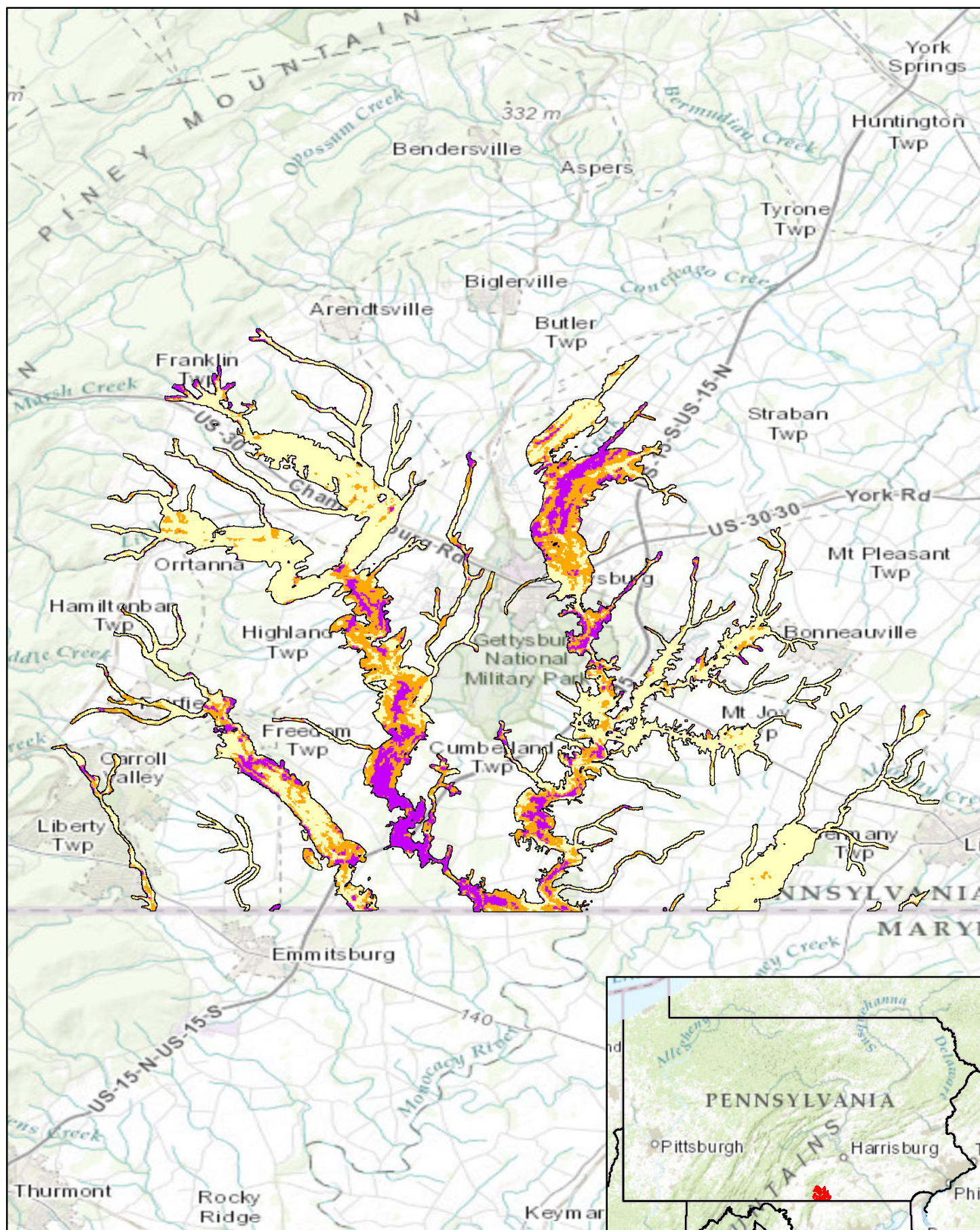




Pennsylvania Predictive Model Set
 Region: 8, Zone: all, Subarea: riverine section 9

Sensitivity
 High
 Moderate
 Low

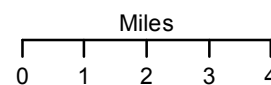


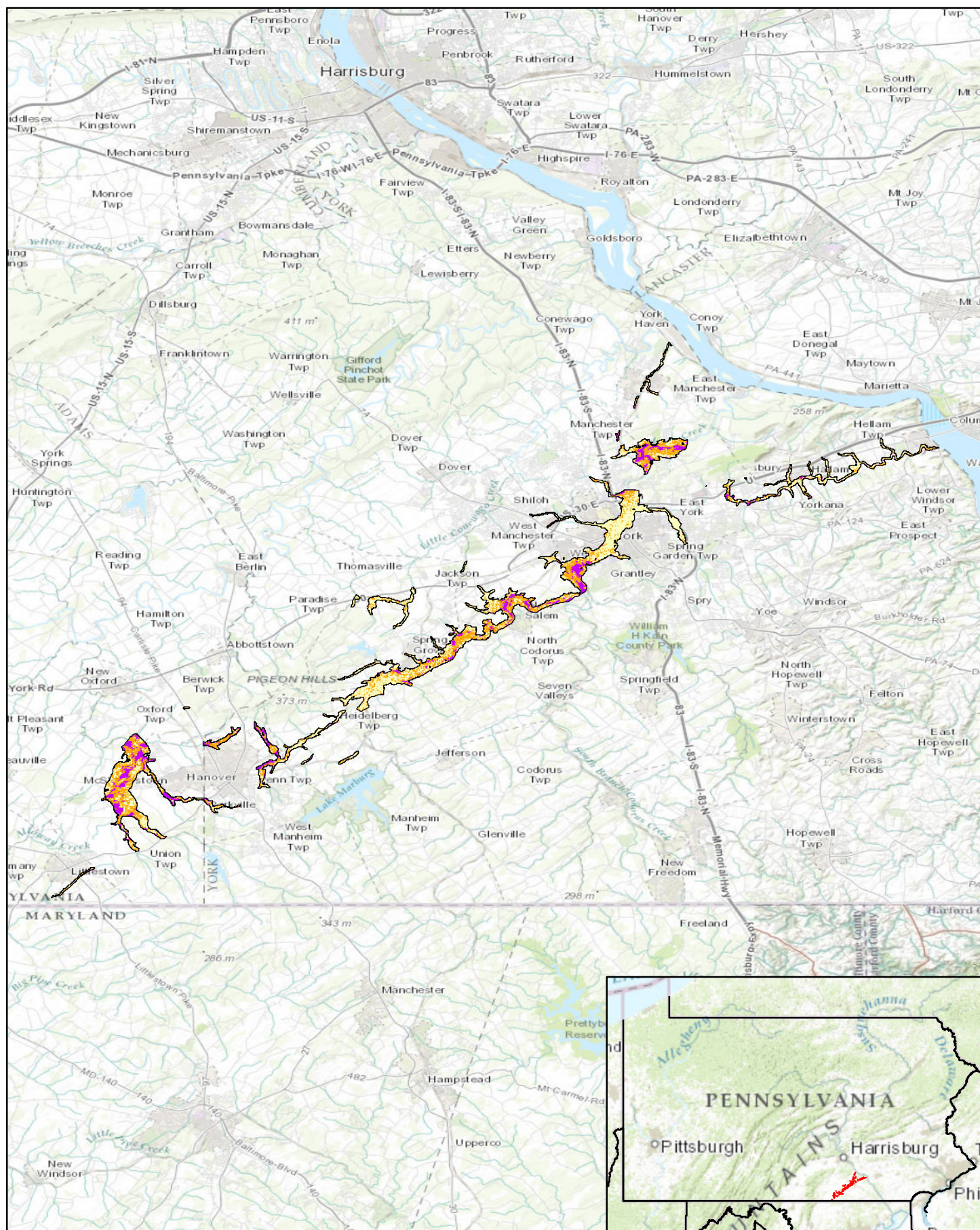


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 1

Sensitivity

- High
- Moderate
- Low



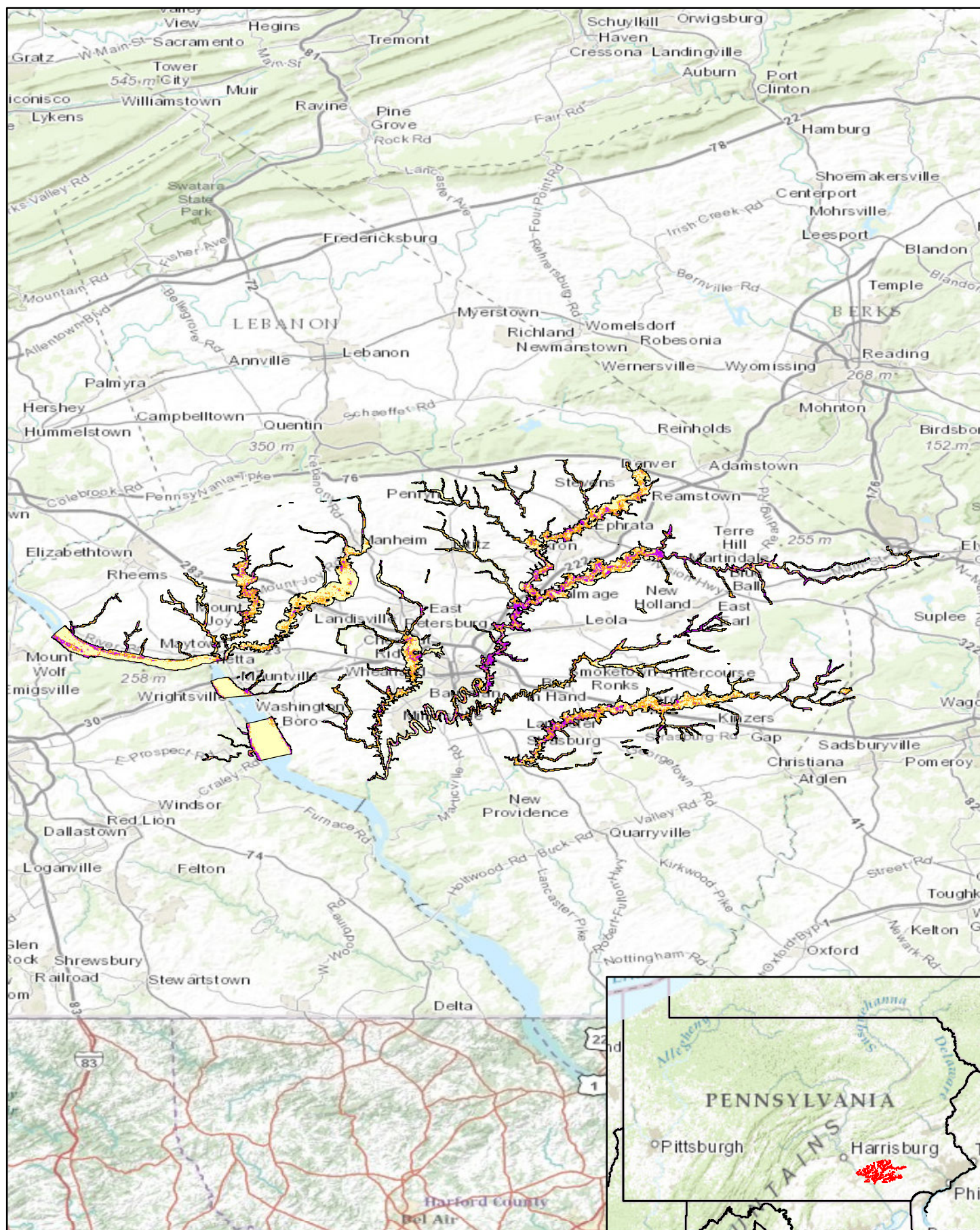


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 10

Sensitivity
 High
 Moderate
 Low

Miles
 0 1.5 3 4.5 6



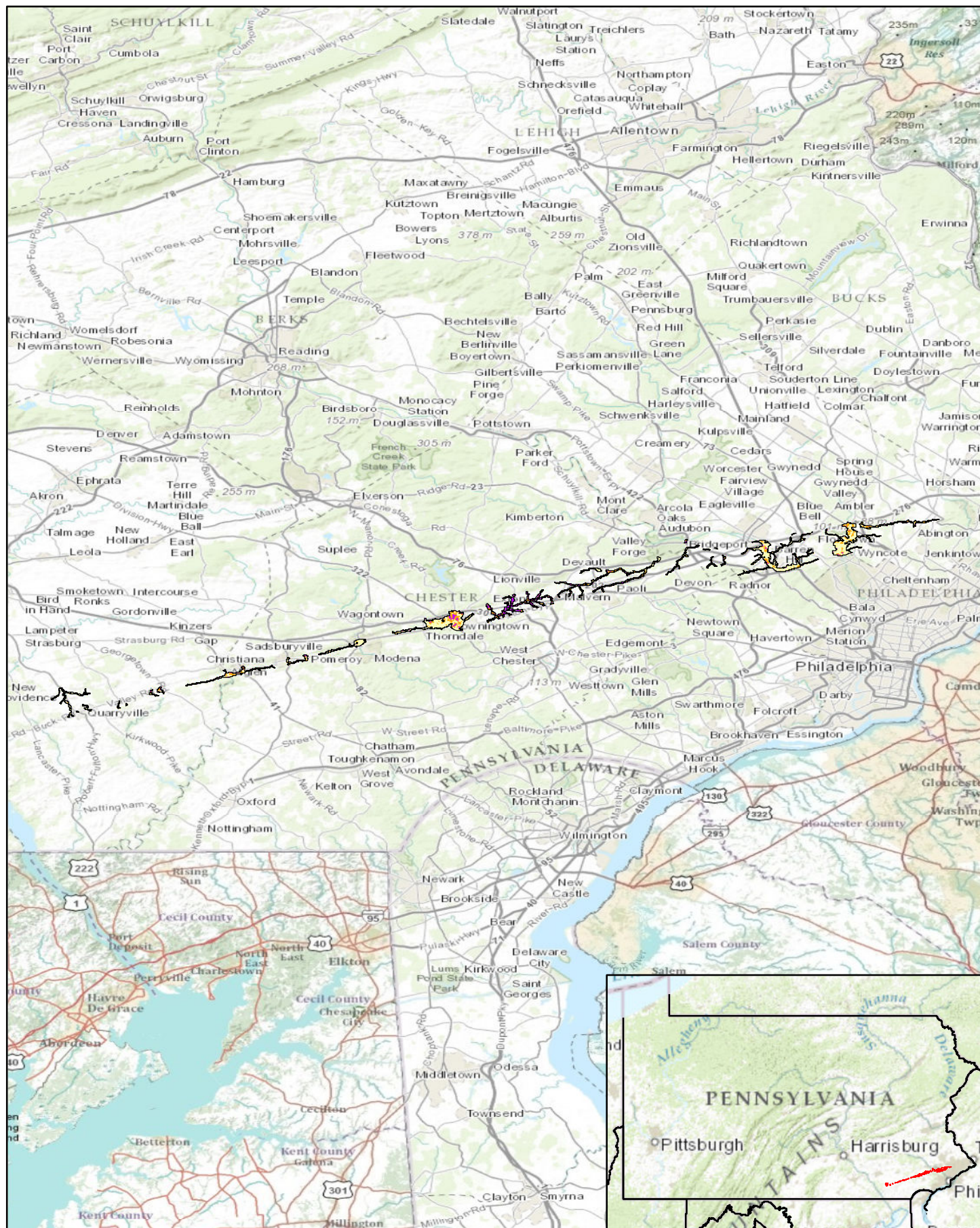


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 11

Sensitivity
 High
 Moderate
 Low

Miles
 0 2.5 5 7.5 10





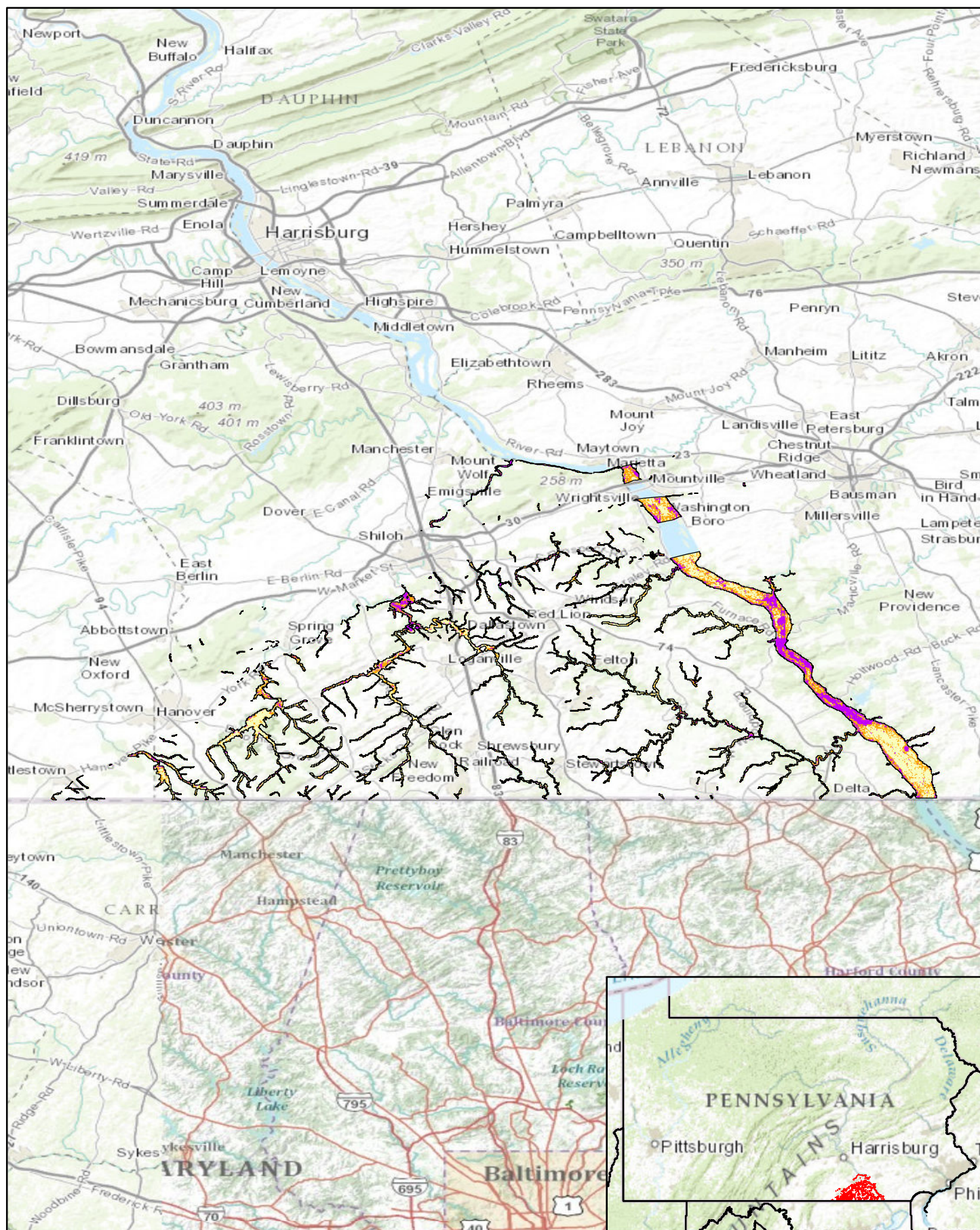
Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 12

Sensitivity

- High
- Moderate
- Low

Miles
 0 3.5 7 10.5 14





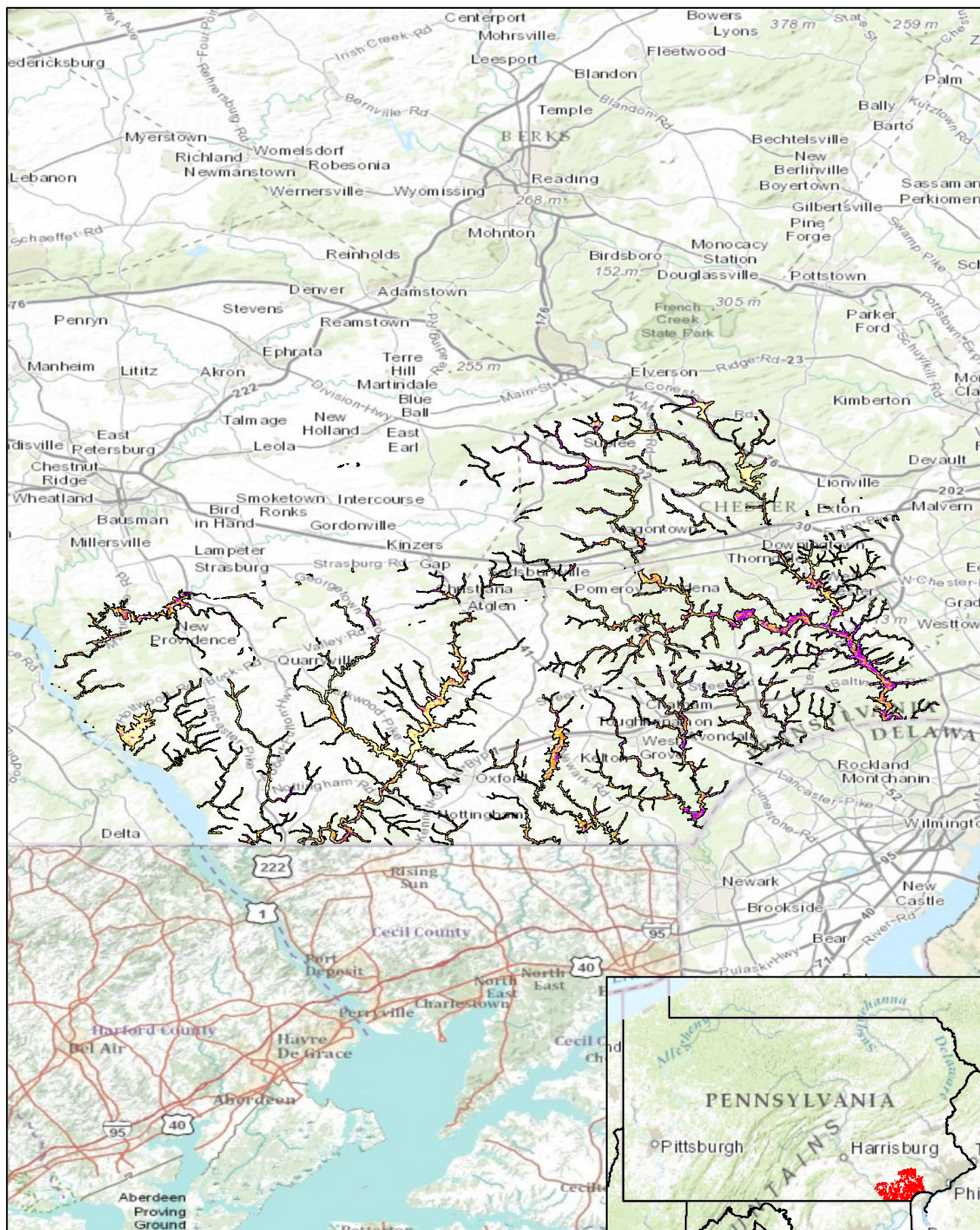
Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 13

Sensitivity

- High
- Moderate
- Low

Miles
 0 2.5 5 7.5 10





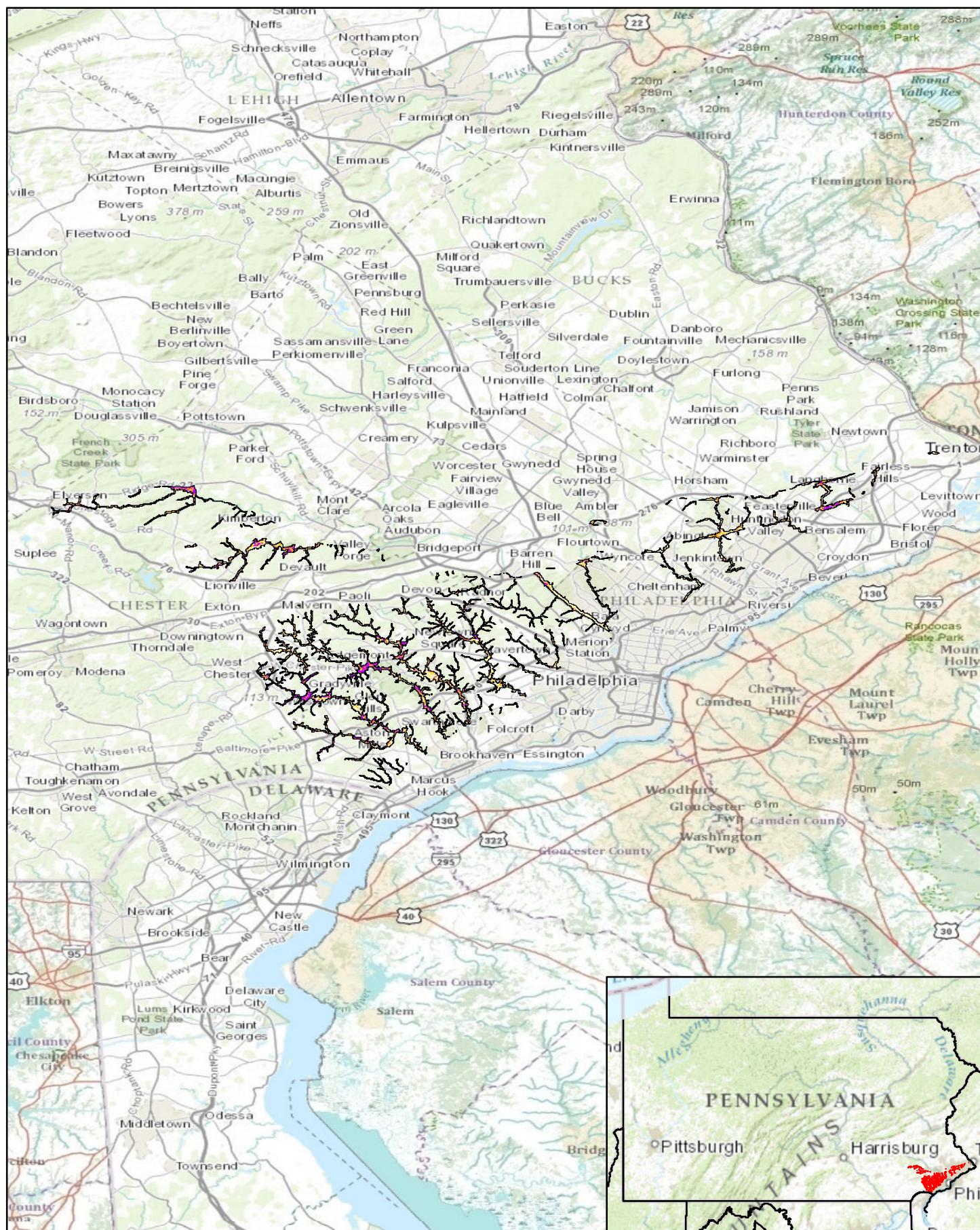
Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 14

Sensitivity

- High
- Moderate
- Low

Miles
 0 2.5 5 7.5 10



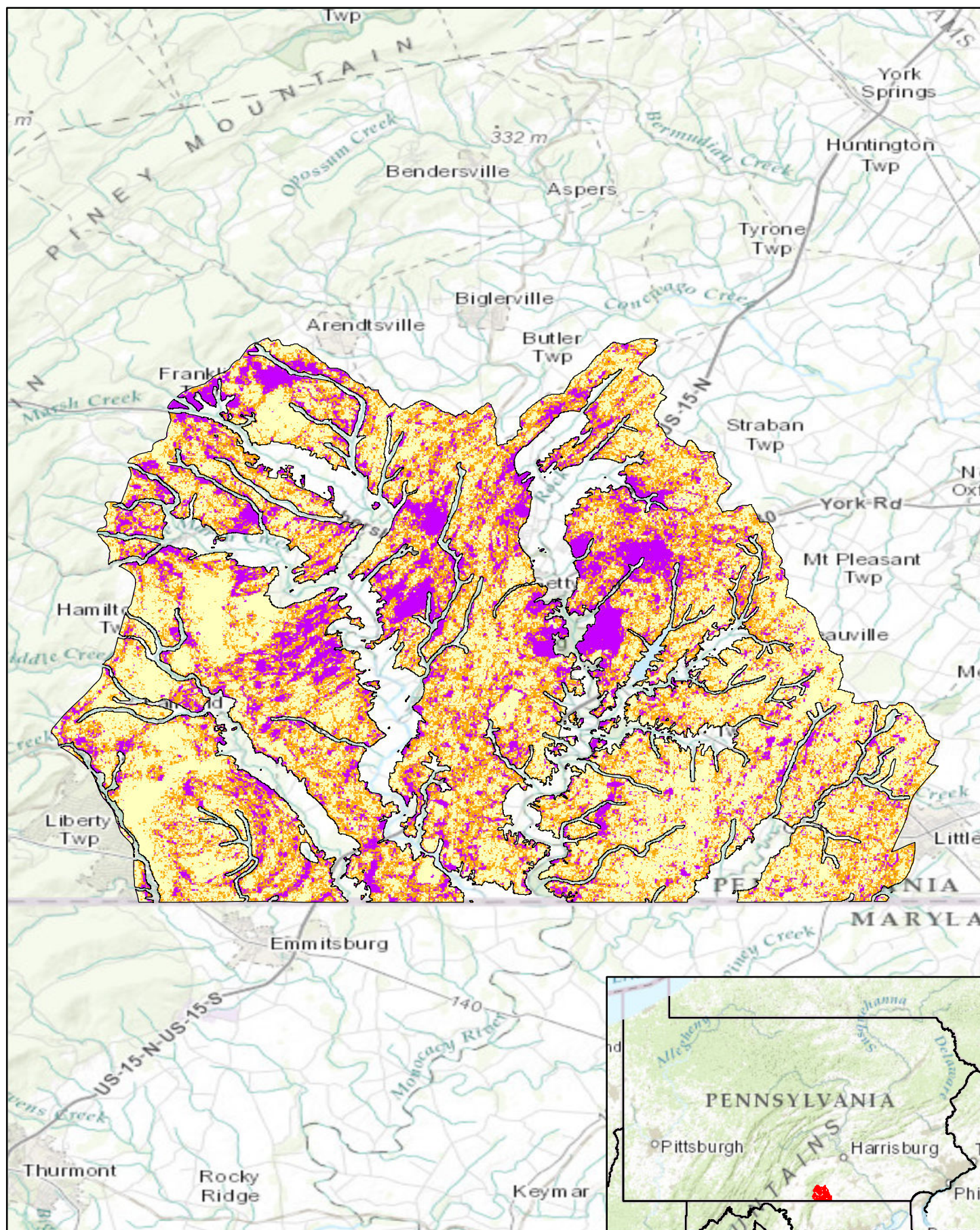


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 15

Sensitivity
 High
 Moderate
 Low

Miles
 0 3 6 9 12

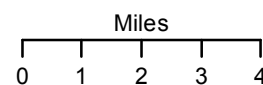


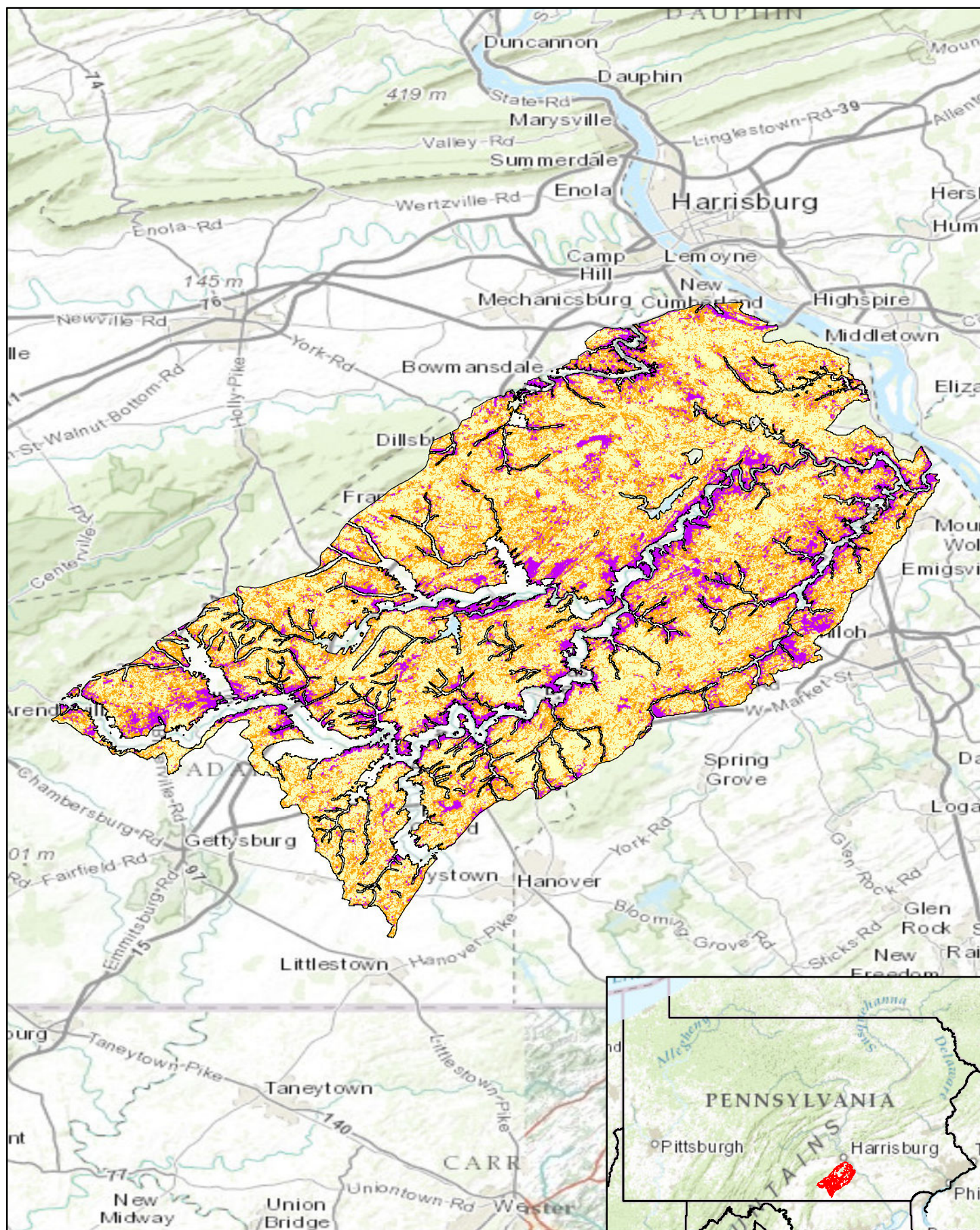


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 1

Sensitivity

- High
- Moderate
- Low

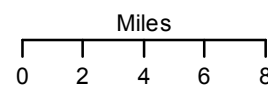


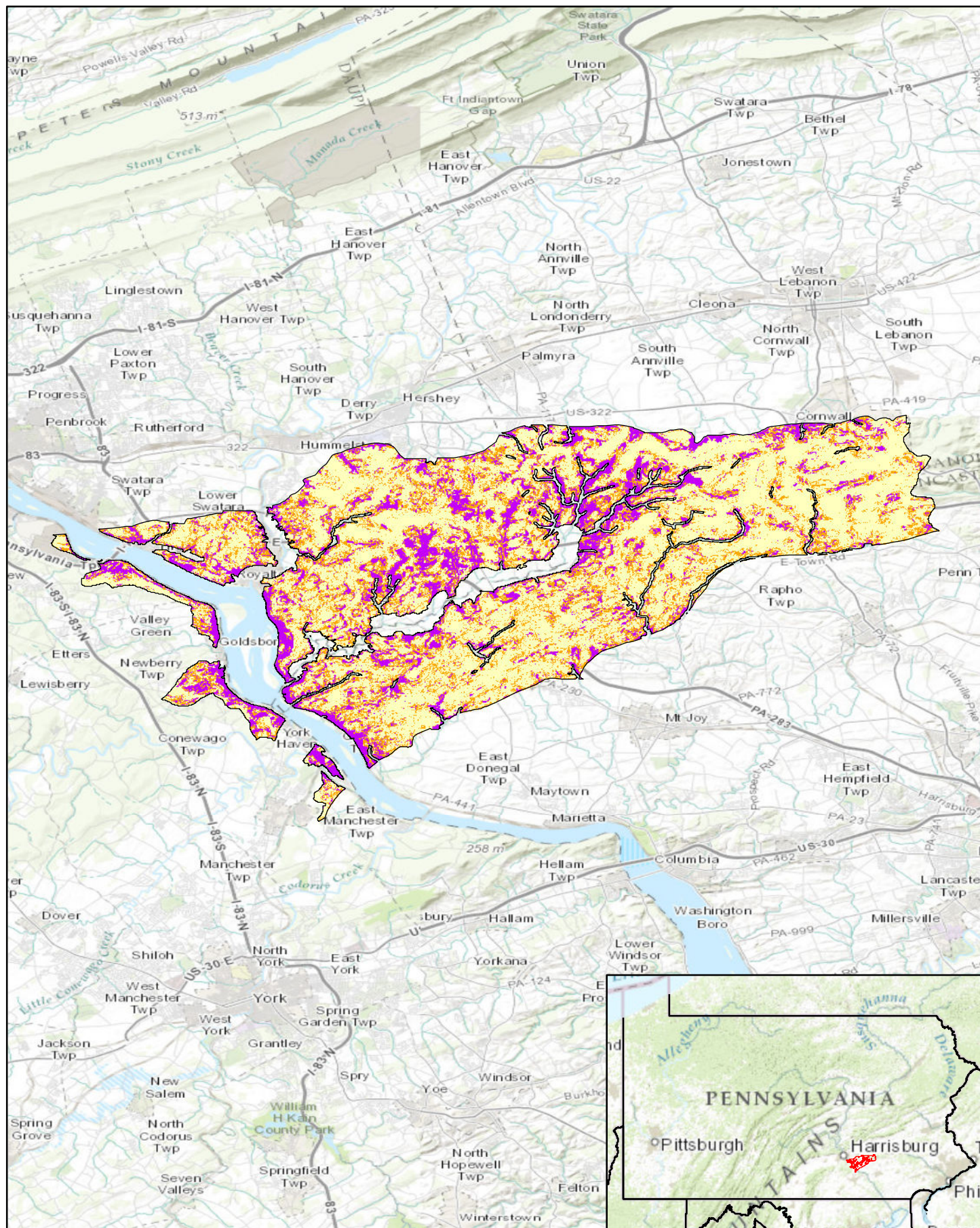


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 2

Sensitivity

- High
- Moderate
- Low



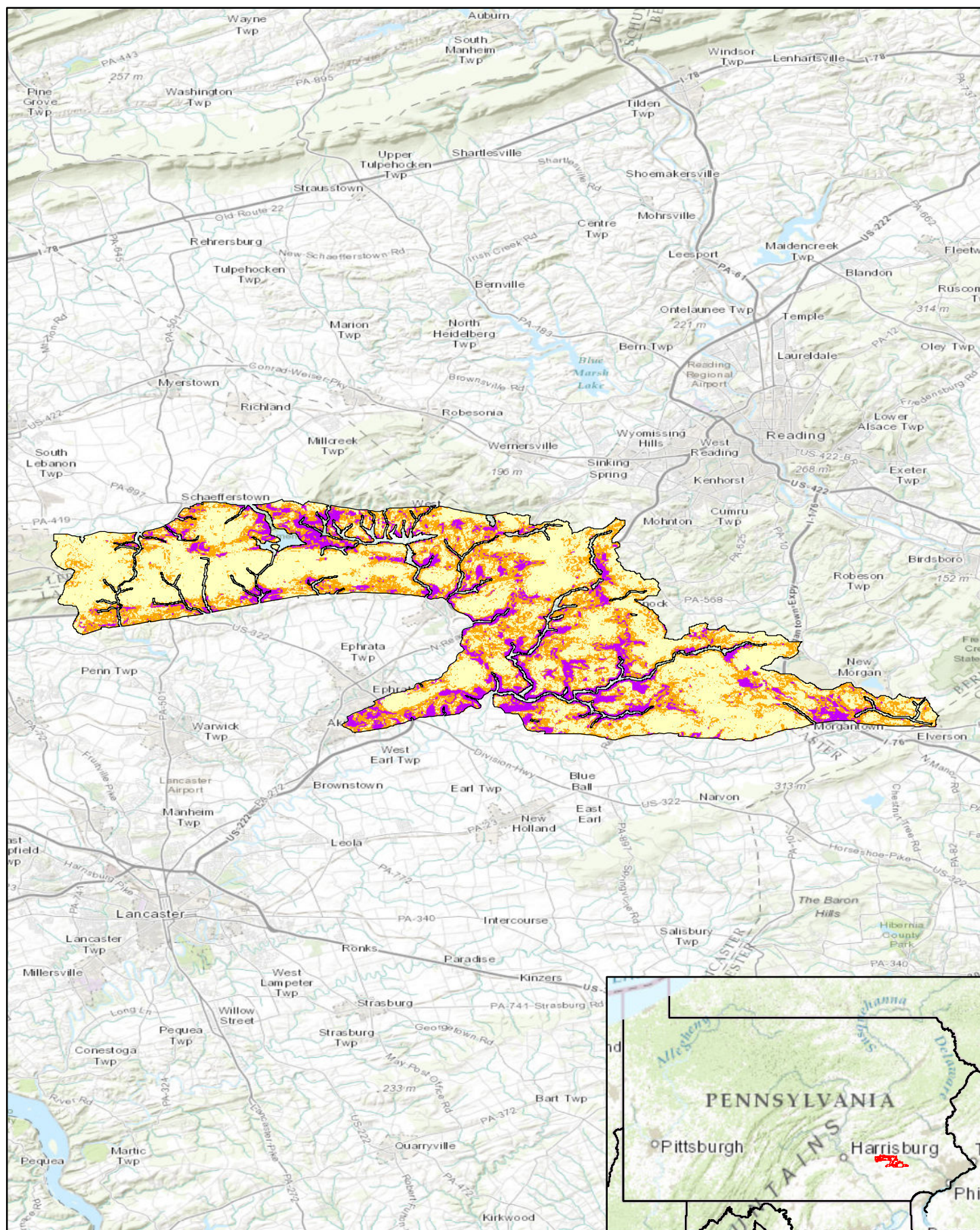


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 3

Sensitivity
 High
 Moderate
 Low

Miles
 0 1.5 3 4.5 6





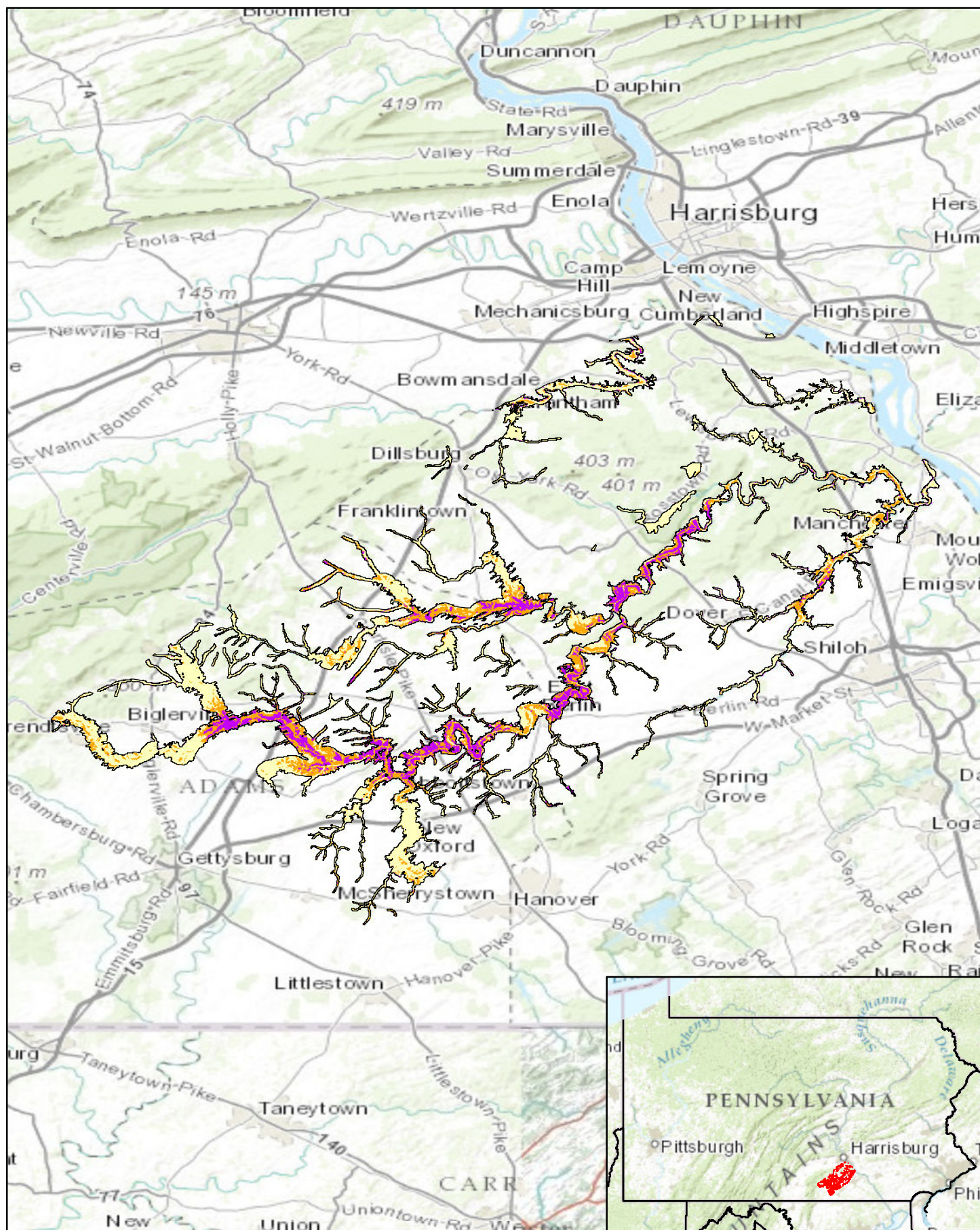
Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 4

Sensitivity

- High
- Moderate
- Low

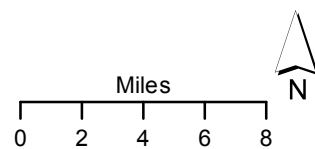
Miles
 0 1.5 3 4.5 6

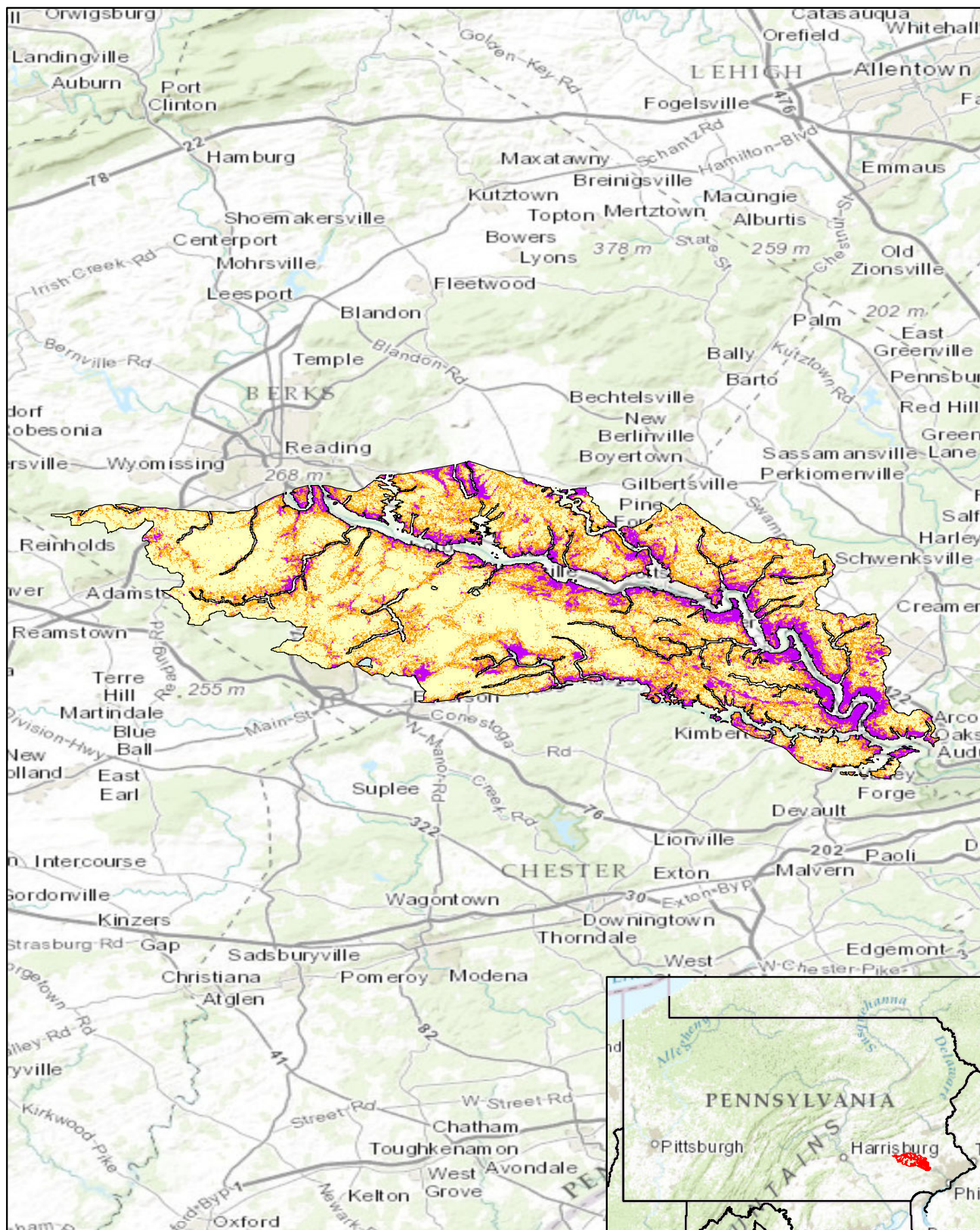




Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 2

Sensitivity
 High
 Moderate
 Low

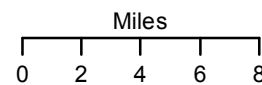


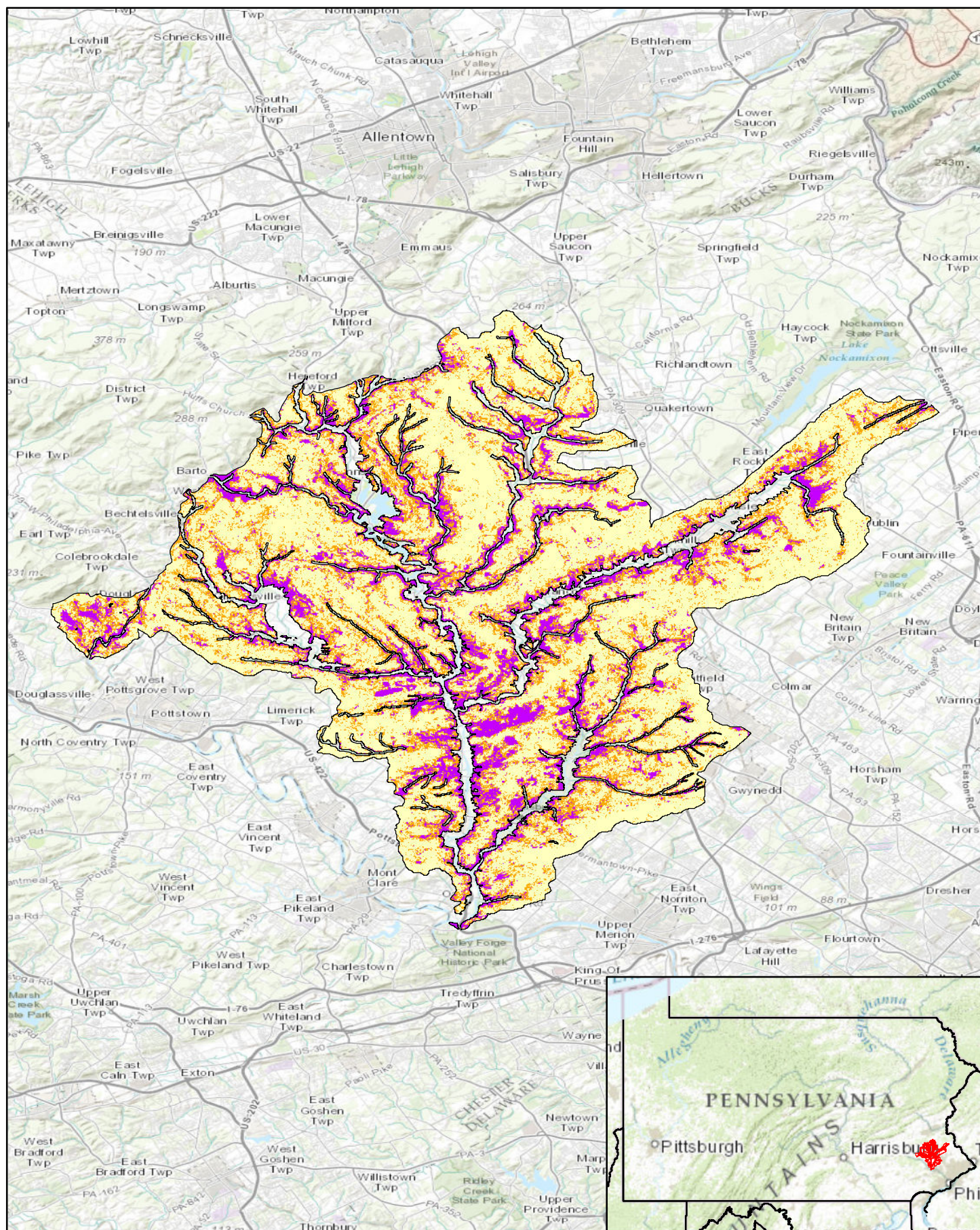


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 5

Sensitivity

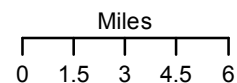
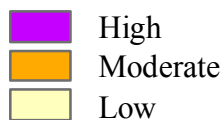
- High
- Moderate
- Low

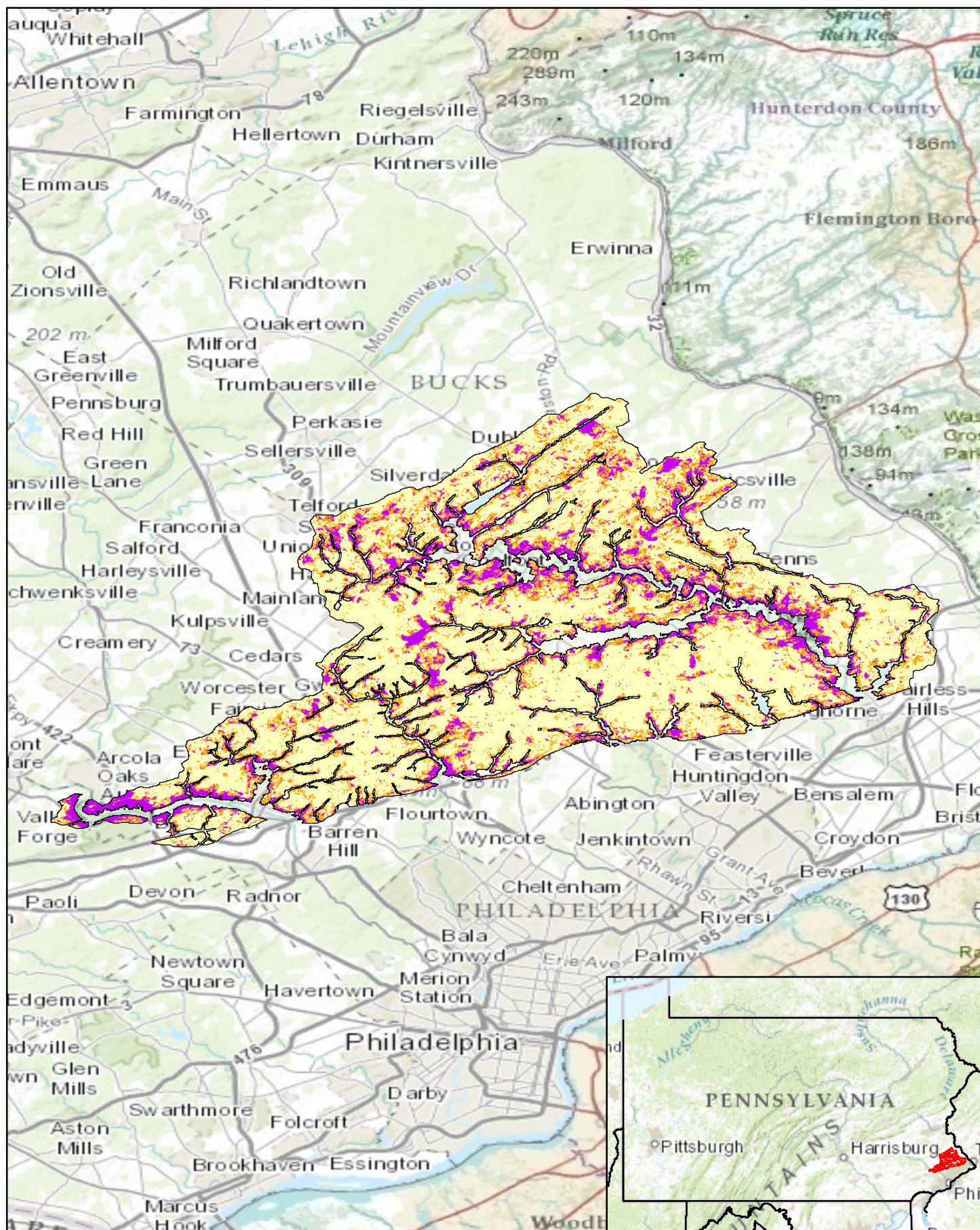




Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 6

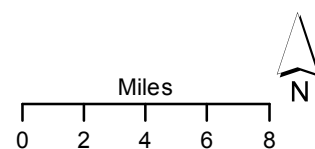
Sensitivity

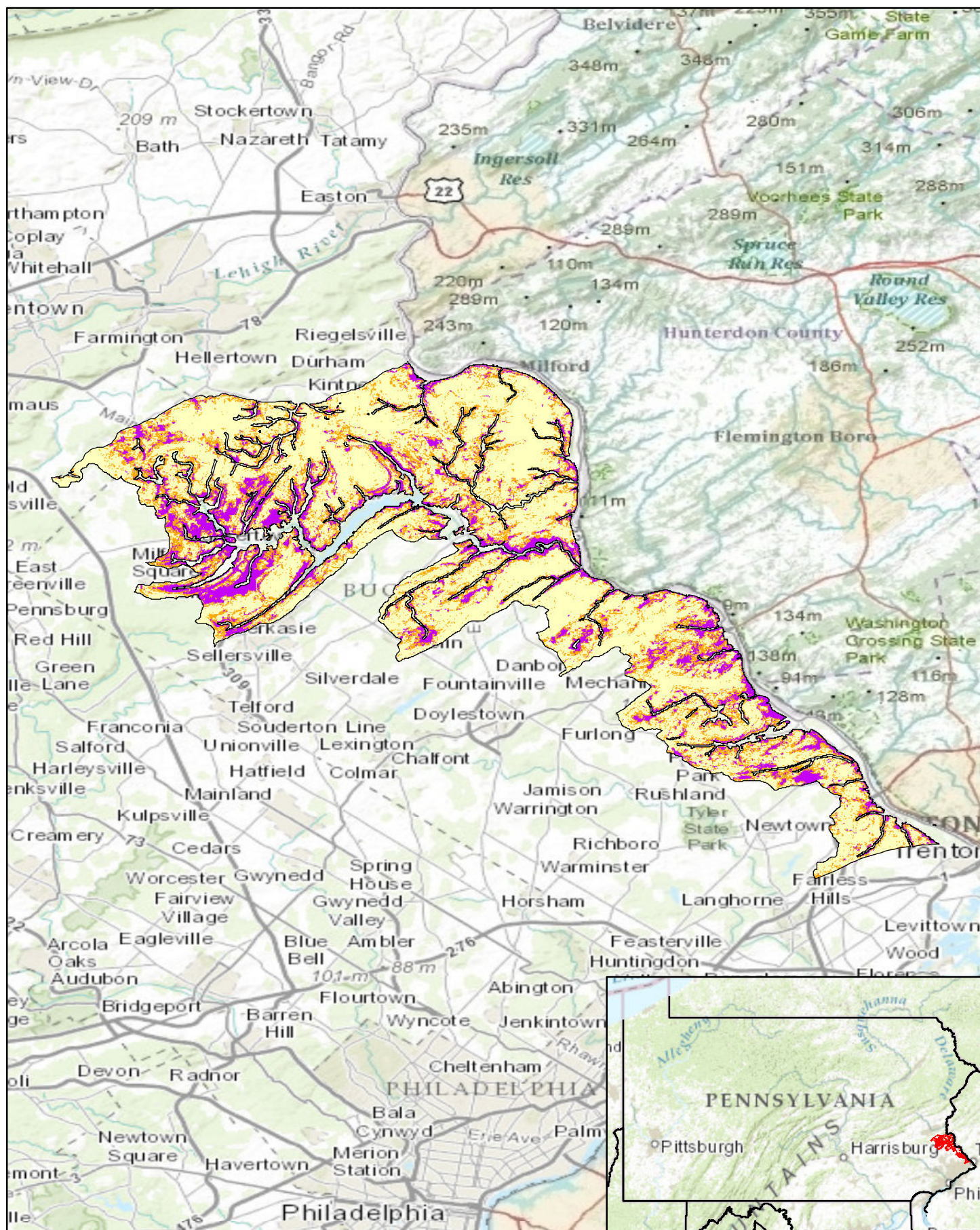




Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 7

Sensitivity
 High
 Moderate
 Low

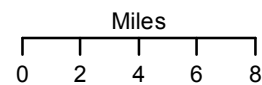


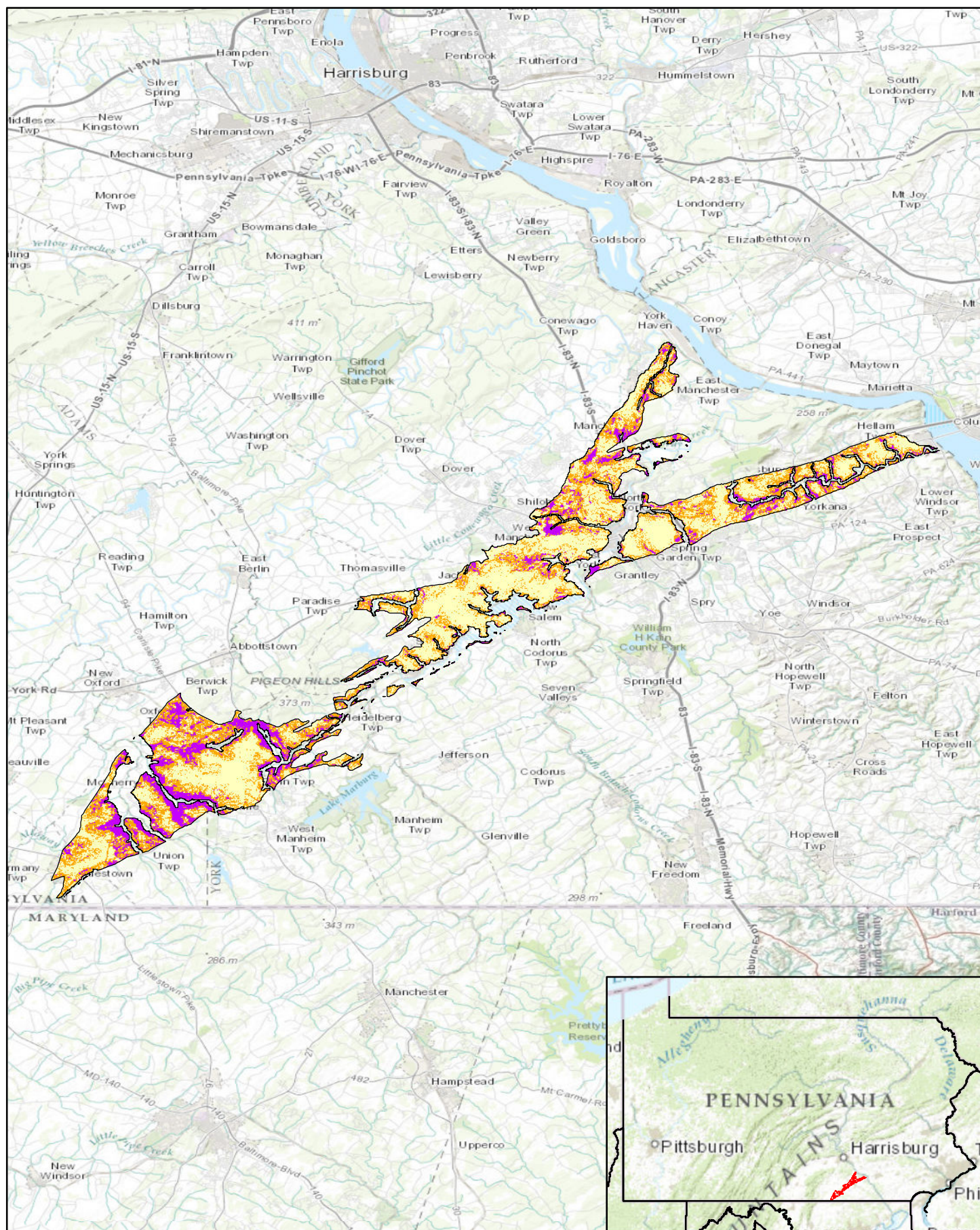


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 8

Sensitivity

- High
- Moderate
- Low

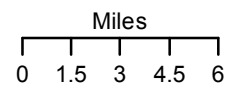


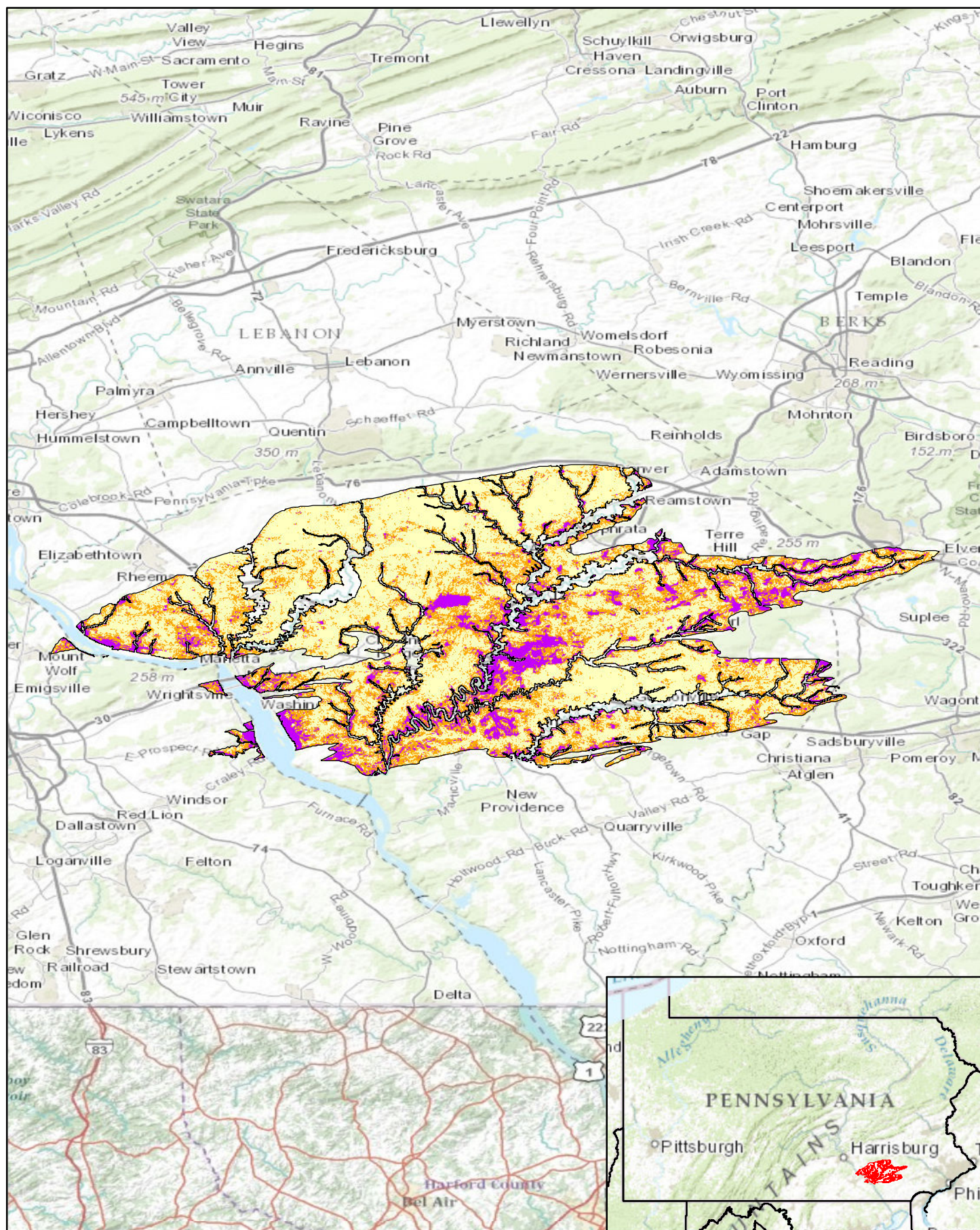


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 10

Sensitivity

- High
- Moderate
- Low



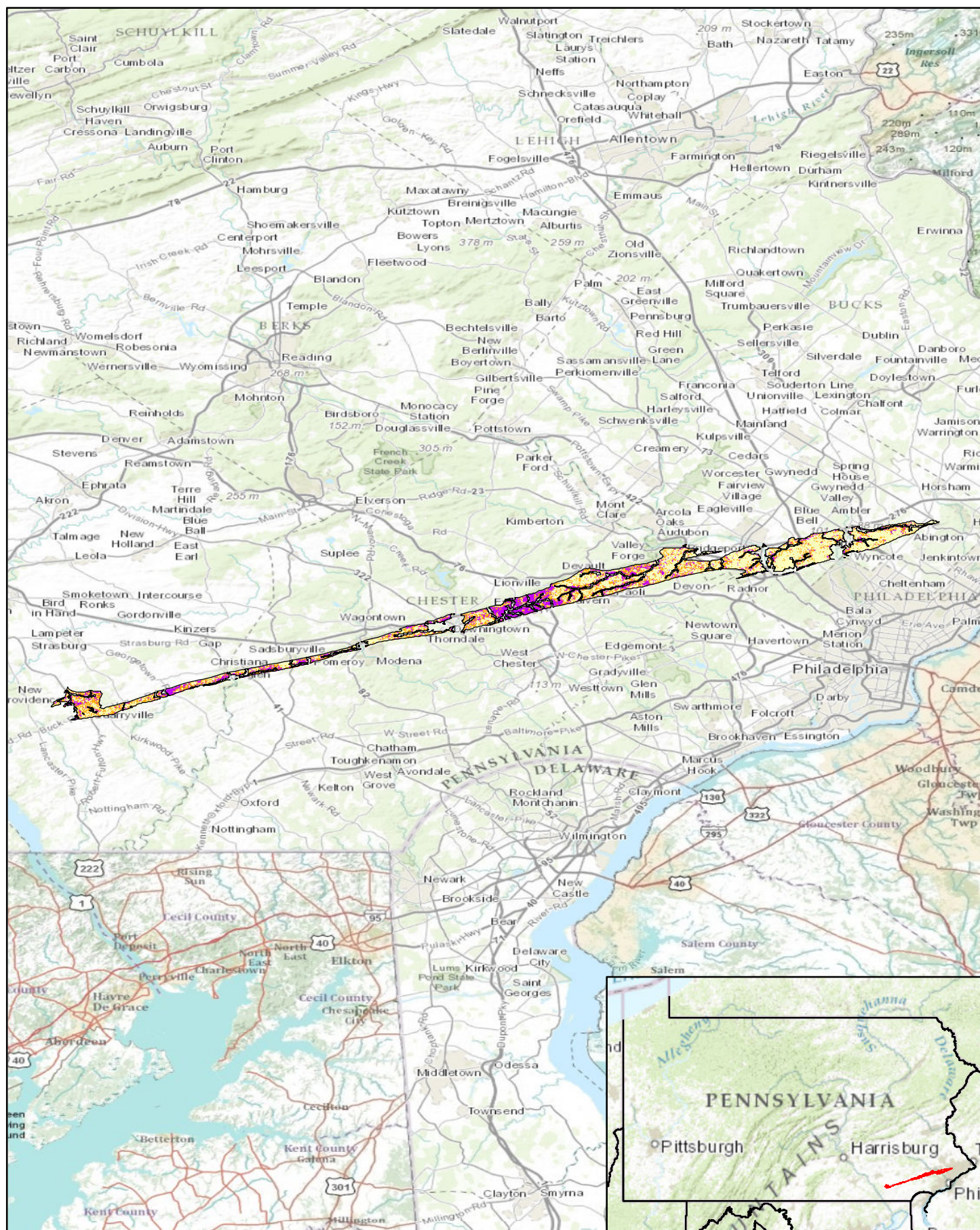


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 11

Sensitivity
 High
 Moderate
 Low

Miles
 0 2.5 5 7.5 10





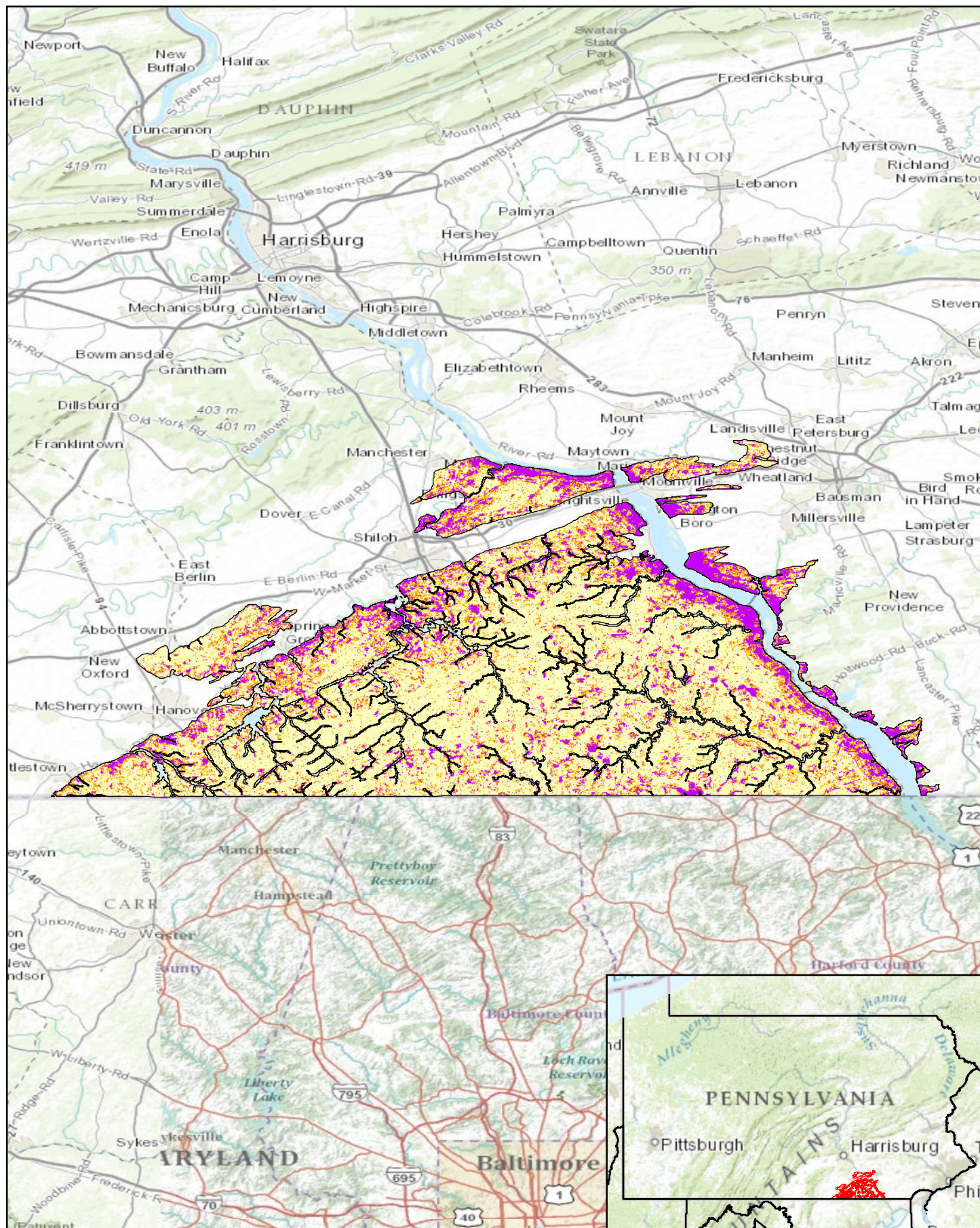
Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 12

Sensitivity

- High
- Moderate
- Low

Miles
 0 3.5 7 10.5 14



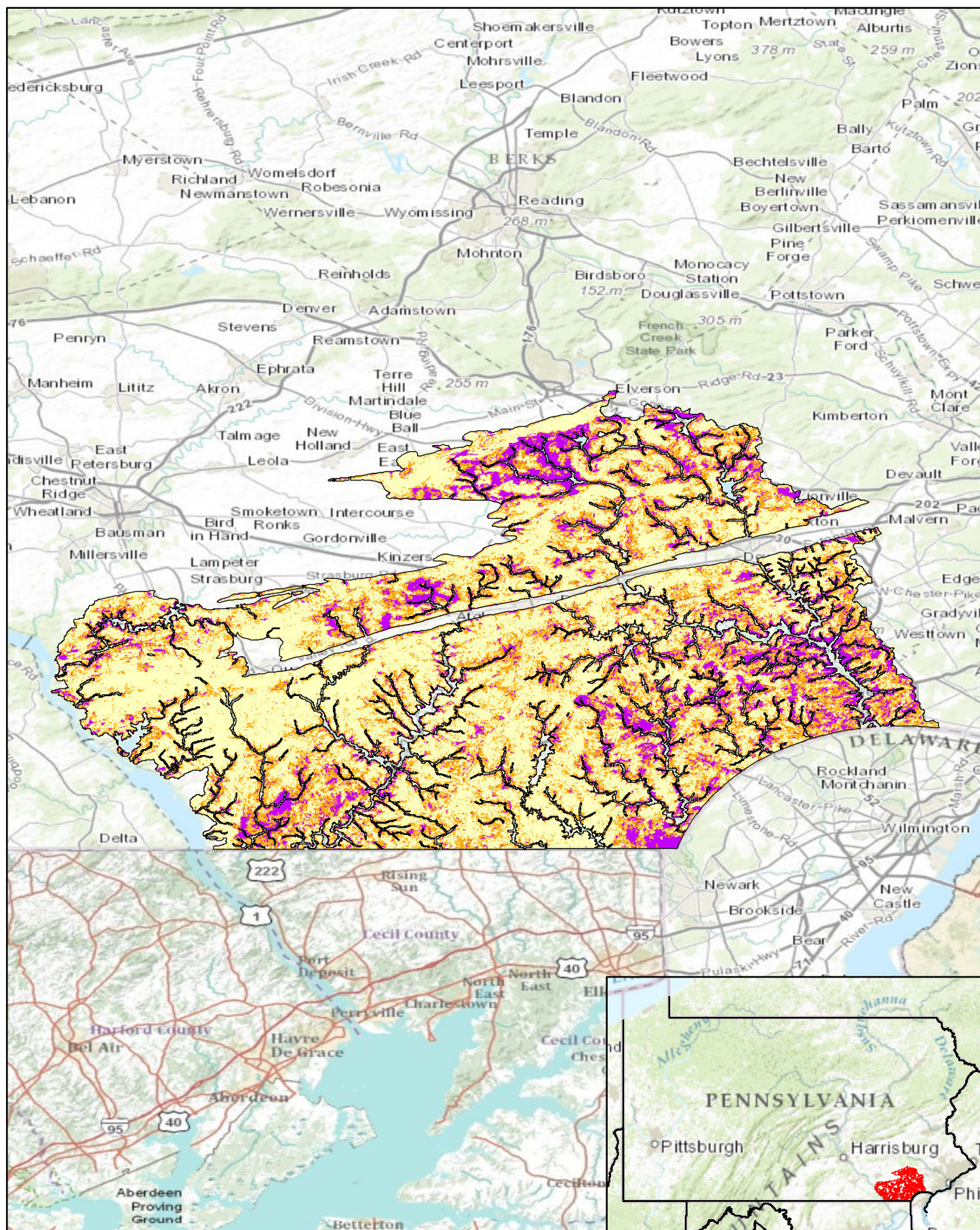


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 13

Sensitivity
 High
 Moderate
 Low

Miles
 0 2.5 5 7.5 10





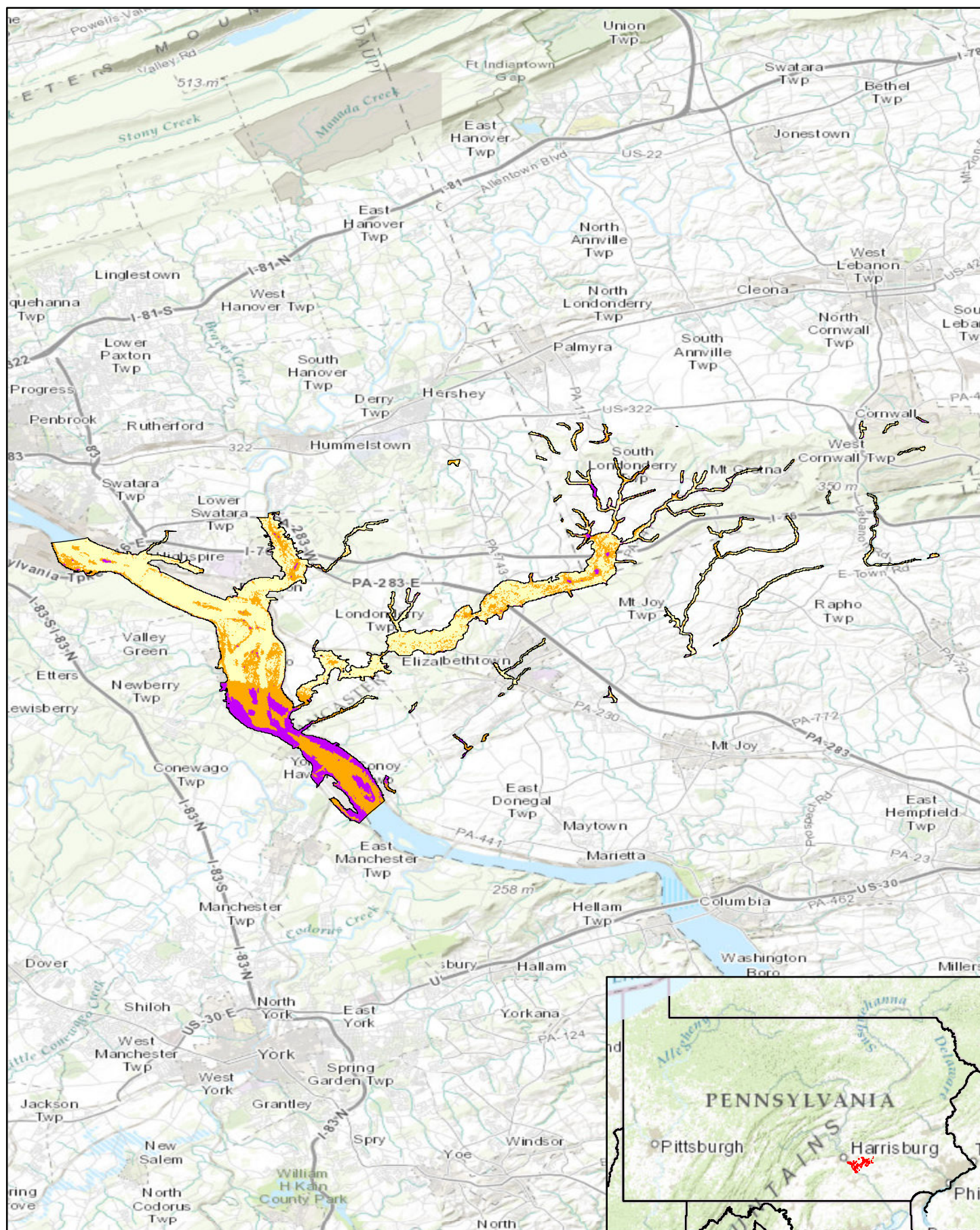
Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 14

Sensitivity

- High
- Moderate
- Low

Miles
 0 2.5 5 7.5 10





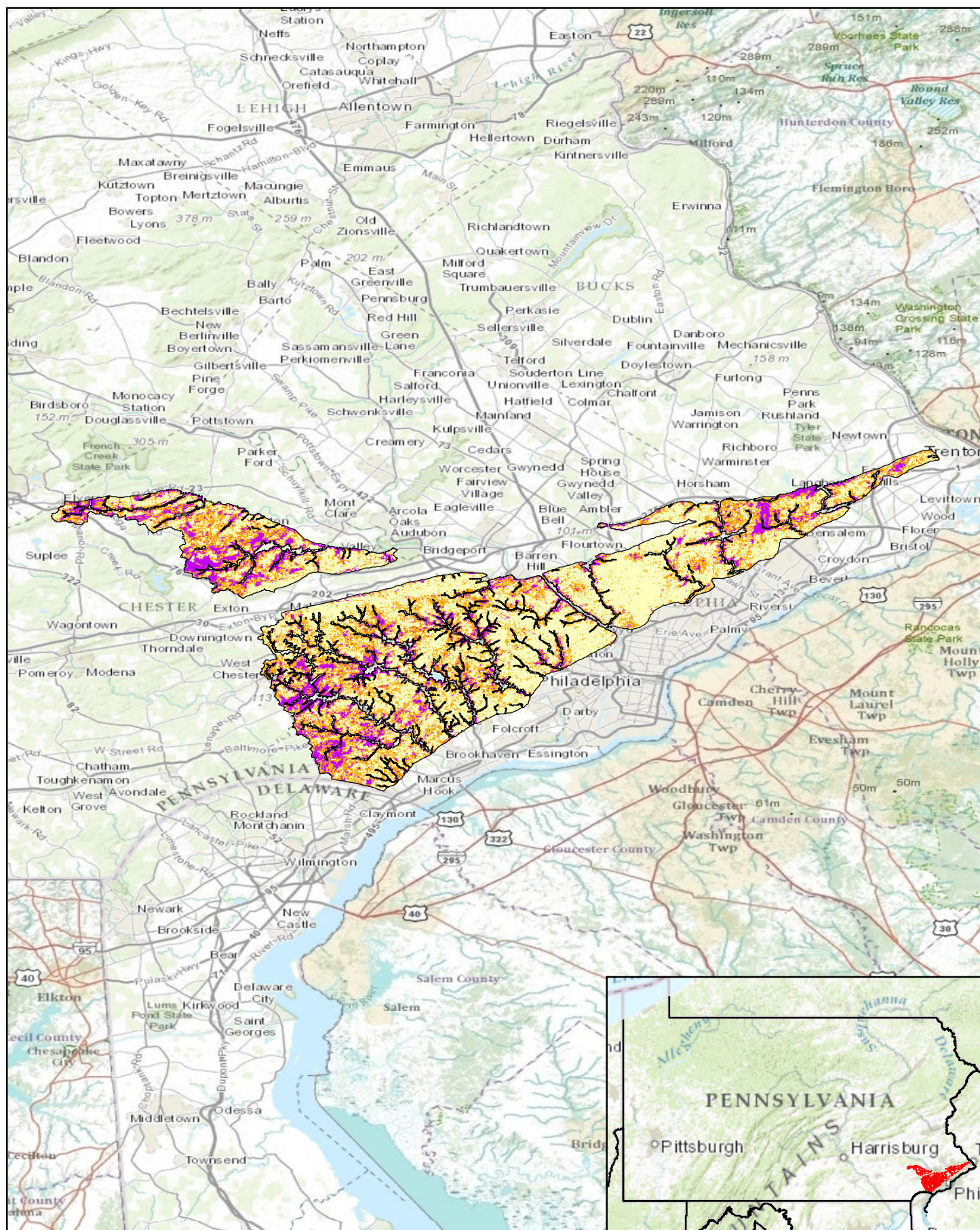
Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 3

Sensitivity

- High
- Moderate
- Low

Miles
 0 1 2 3 4

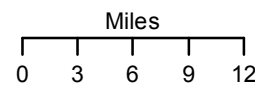


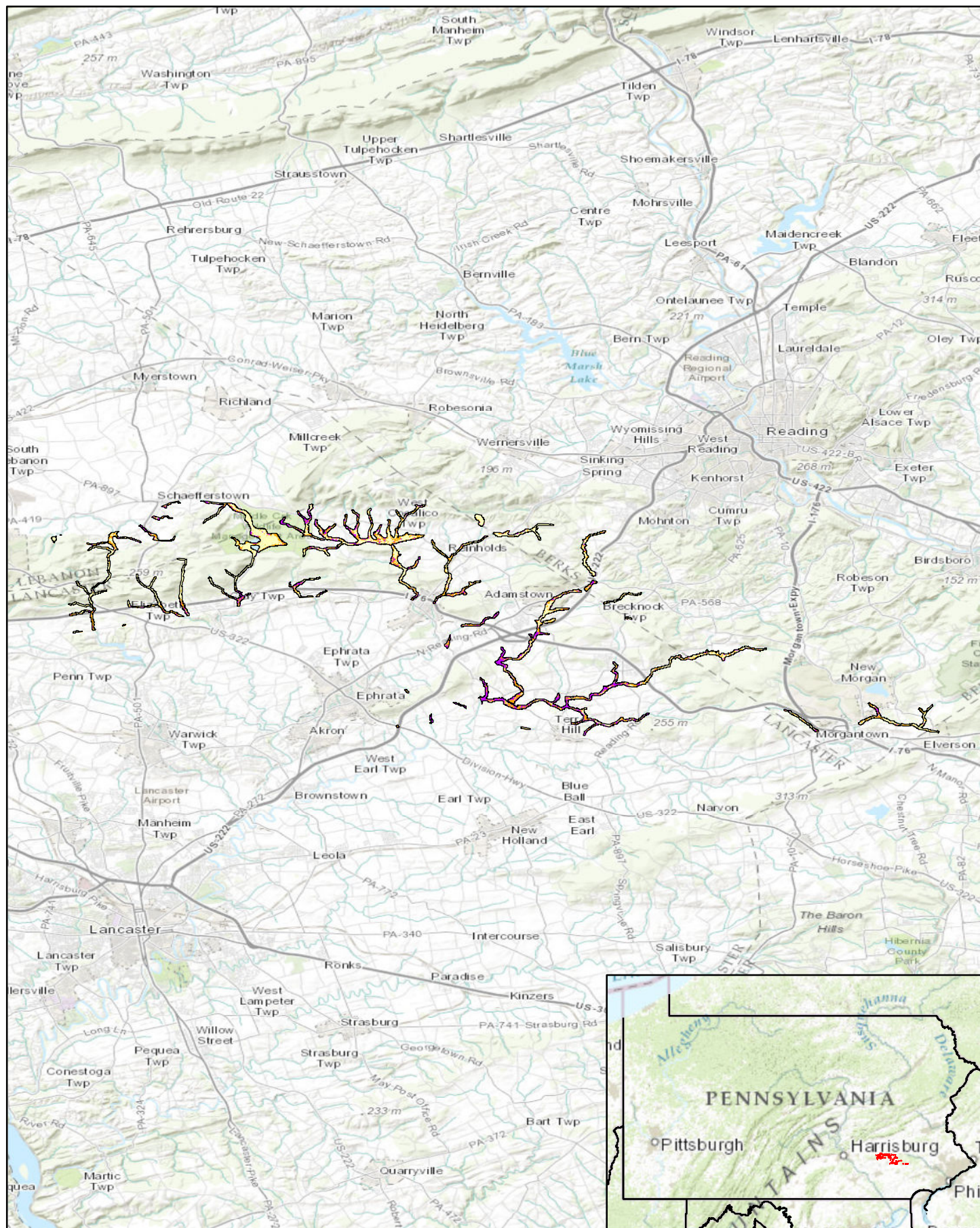


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: upland section 15

Sensitivity

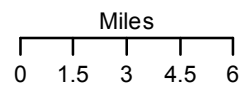
- High
- Moderate
- Low

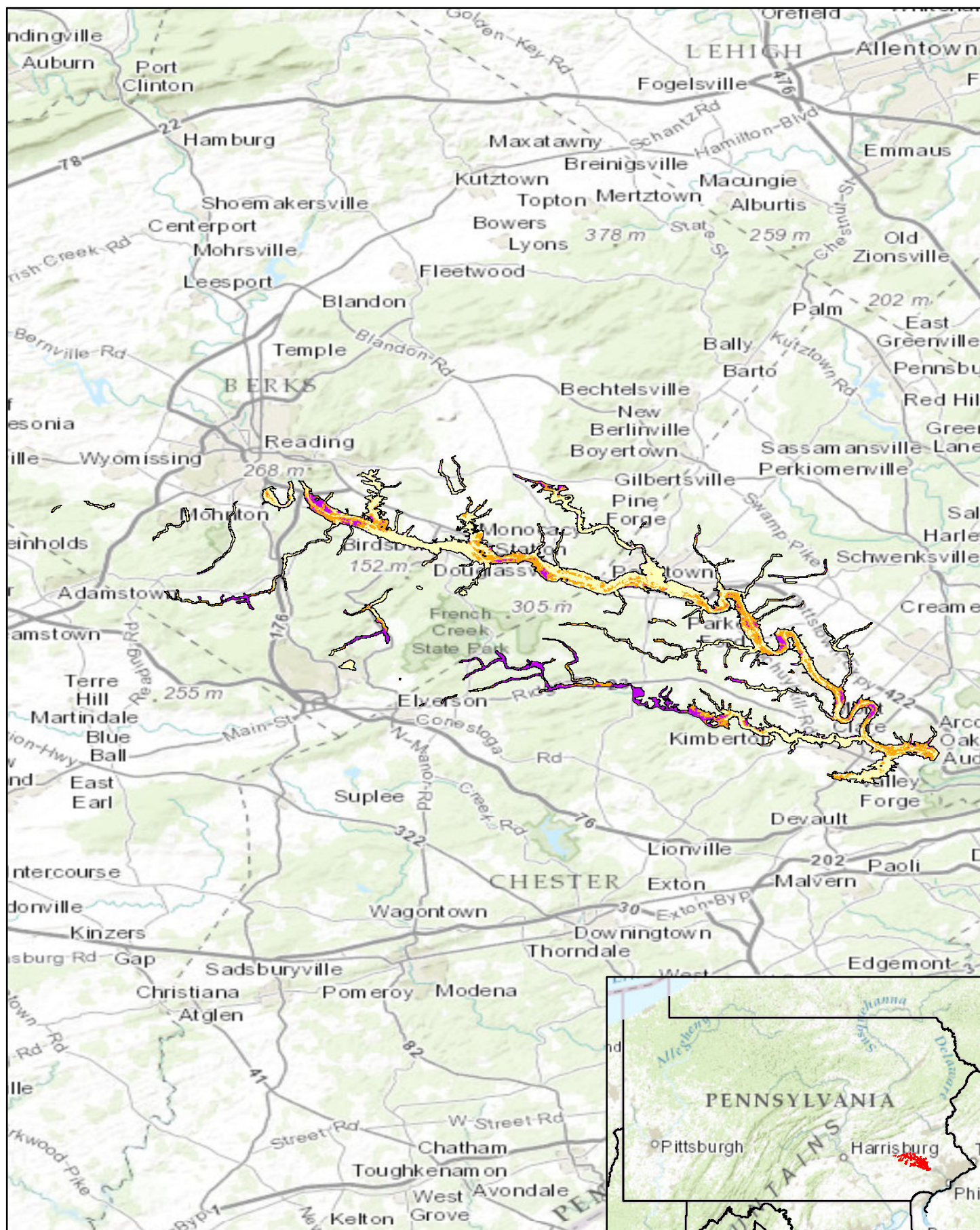




Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 4

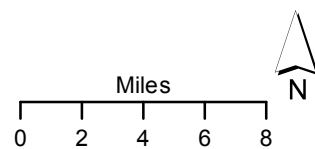
Sensitivity
 High
 Moderate
 Low

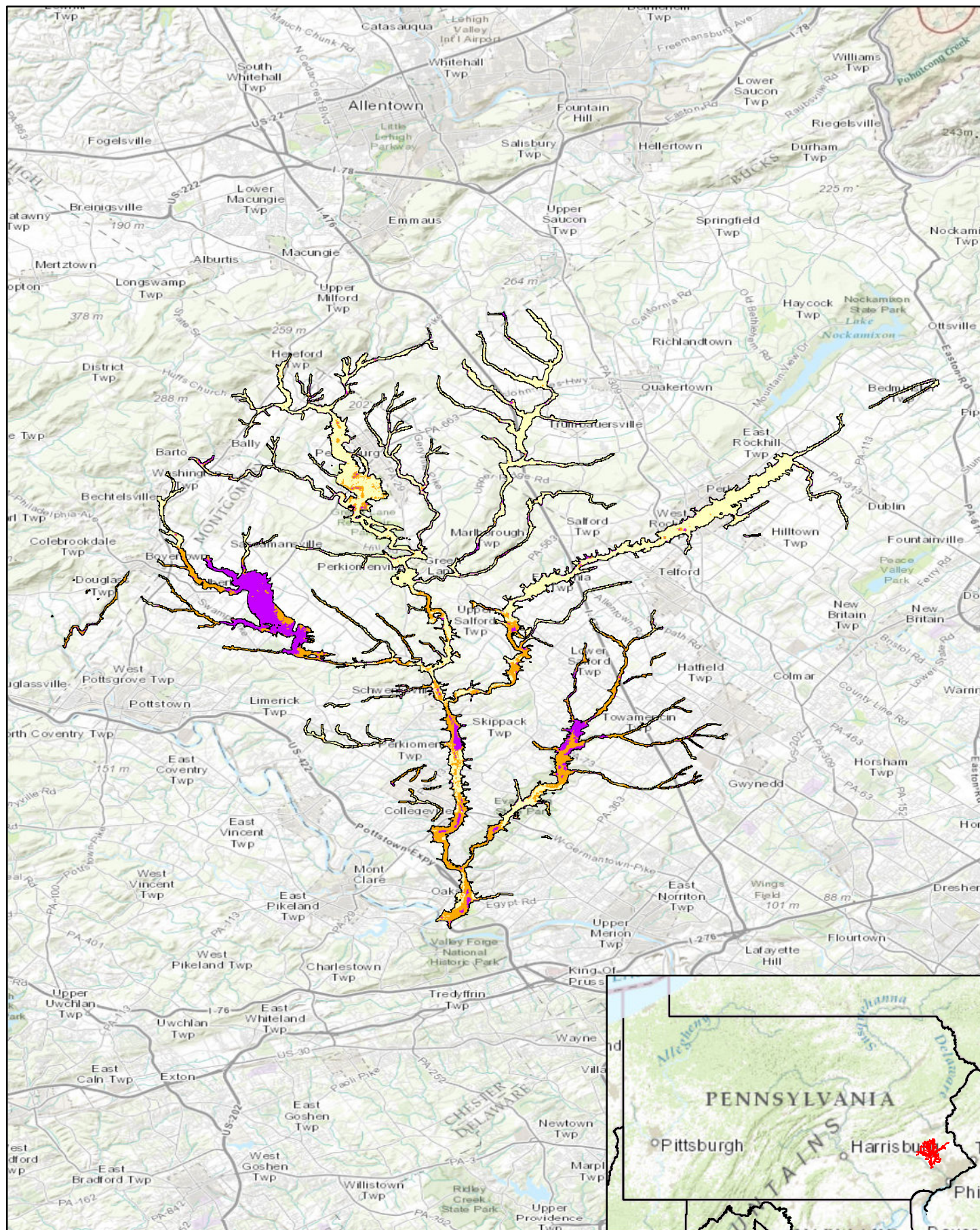




Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 5

Sensitivity
 High
 Moderate
 Low

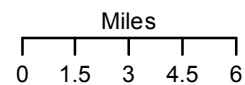


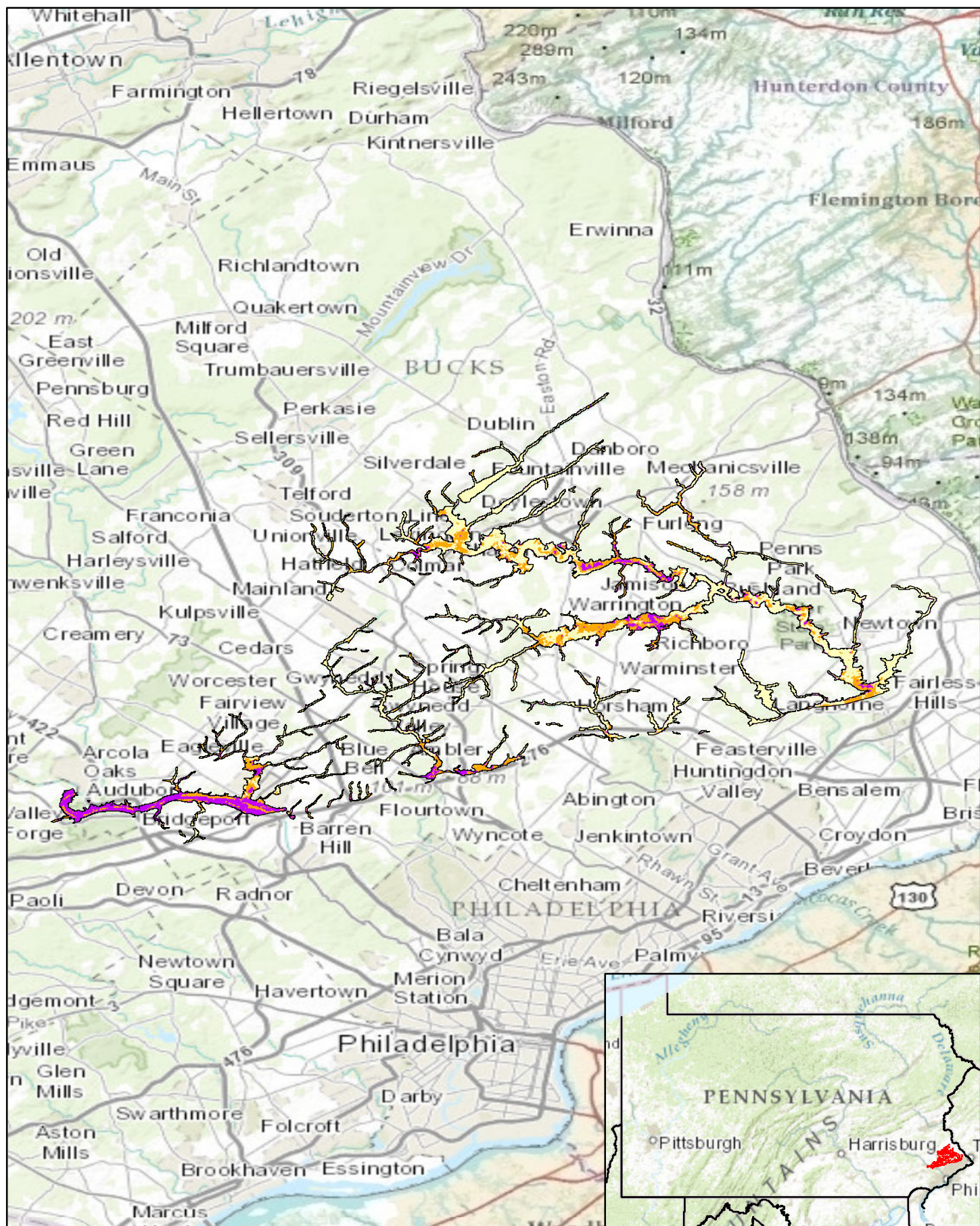


Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 6

Sensitivity

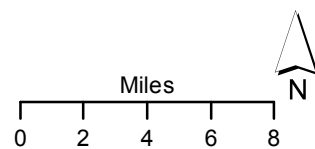
- High
- Moderate
- Low

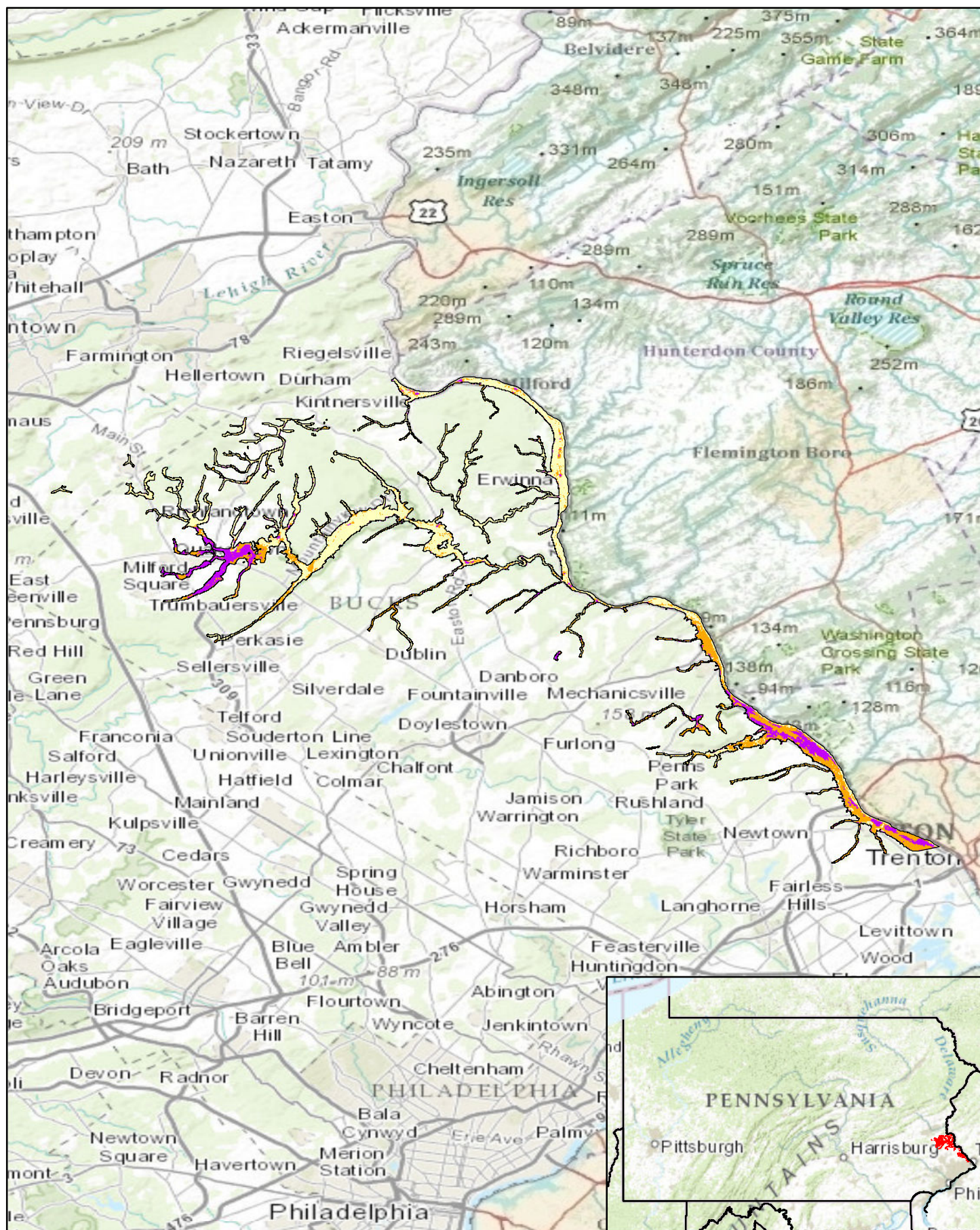




Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 7

Sensitivity
 High
 Moderate
 Low





Pennsylvania Predictive Model Set
 Region: 9, Zone: all, Subarea: riverine section 8

Sensitivity
 High
 Moderate
 Low

