To: Appalachian Basin Geothermal Play Fairway Analysis Group

From: Jared Smith

Date: 16 October, 2015

Subject: Exploratory Data Analysis and Interpolation Methodology for Thermal Field Estimation.

Applicability: This memo describes the methods used to interpolate the geotherm data at each well to create the thermal risk factor and uncertainty maps for the project. Included in this memo is an exploratory data analysis on wells after processing in the thermal model.

Introduction

The Appalachian Basin Geothermal Play Fairway Analysis team needs to have a method for predicting the thermal field throughout the basin using the geotherms calculated for each well as control points (see Thermal Model Methods and Well Database Organization in GPFA-AB). The method described in this memo was used to compute the geothermal resource and risk factor maps, and quantify associated uncertainties of the spatial prediction.

A valuable precursor to any regression is an exploratory data analysis (EDA) that scans the data for potentially rogue or anomalous observations, and retains only those points that are deemed to be of sufficient quality. The EDA used in this project involved assessment of the data quality according to the depth of the bottomhole temperature (BHT) measurement, followed by the identification and removal of spatial outliers. The details of the outlier analysis algorithm developed for this project are provided in the memo entitled Thermal Outlier Assessment in GPFA-AB. This memo presents general results from the outlier analysis. The EDA also included an evaluation of the spatial autocorrelation of the variable to be predicted. Results of the spatial correlation analysis are provided for the Depth to 80 °C. All other thermal resource variables were subject to the same EDA methodology.

Many interpolation algorithms may be suitable for prediction of the thermal field. The results of the EDA were used to inform the selection of the interpolation algorithm. One major investigation in the EDA that provided insight to the choice of the interpolation algorithm was the spatial correlation of the thermal variables based on potential “heat flow provinces” (Roy, Blackwell, and Birch, 1968; (see the Thermal Model Methods and Well Database Organization in GPFA-AB memo for further details). Statistically, in order to retain the assumption that these heat flow provinces represent different data generating regions, the spatial regression algorithm had to preserve the province boundaries. This memo discusses the creation of the interpolation boundaries based on the gravity and magnetic potential field edges that were calculated as part of this project (see Identifying Potentially Activatable Faults in GPFA-AB memo for details). Based on this effort, the interpolation algorithm selected was a spatially stratified ordinary kriging interpolation implemented in the gstat package of R (Pebesma, 1998). The details of the interpolation are discussed below.

Finally, evaluation of the performance of the kriging algorithm was tested using two cross validation techniques. The two techniques used were a Leave-One-Out-Cross-Validation (LOOCV), and an evaluation of the predicted mean temperature at 1.5 km depth compared to the set of 47 equilibrium temperature measurements near 1.5 km depth throughout the basin. Results and details of the LOOCV are provided in the memo entitled Selection of Four Counties in Each State with the Best Thermal Resources. Results are presented for counties rather than for the full region because spatial clustering of wells clutters the regional map so much that utility of the map is lost.

Well Data Processing and Exploratory Data Analysis

After the wells were run through the thermal model, a few processing steps were required prior to doing an EDA on the calculated geotherms.

Wells in the Same Spatial Location

Post processing in the thermal model, the well database was checked for exact duplicate coordinates that would interfere with spatial exploratory data analyses. For example, in a spatial outlier analysis a location should only be counted once, assuming that the information about the thermal field in that provided by the multiple measurements in the same location is the same. Exact duplicate locations in the database may result from a well having more than one measurement at different depths (same API number), from multiple wells starting at the same offshoot but branching off from each other (different API numbers), from wells being located on the same drill pad and assigned the average coordinates of the pad (different API numbers), or from a duplicate of a record in the database. In all of these cases, only the deepest measurement for a location is retained for analysis. The deepest measurement is used as a method of rapid quality control. Other methods the include all of the temperature measurements at different depths may be more accurate to the true geotherm.

In a few locations the deepest depth had two or more different BHTs. The measurement date and time were not available for these wells, so it was not possible to tell if these temperature differences were spread out in time, and thereby indicative of temporal variations in the thermal field. For these data, only the wells with BHTs at the same depth that were within 2 °C of each other were retained – the remainder were dropped from the dataset for quality concerns. The retained wells were checked for potential errors in recording the depth of measurement that could explain the difference in BHTs at same depth. Each well in the database has up to three depths: a driller depth, a true vertical depth, and a depth of temperature measurement. The depth of temperature measurement is used for wells in this project; however the database could contain errors in this depth. Therefore, the quality control was that the smaller of the multiple BHTs should correspond to the lesser of the depth of measurement and the true vertical depth or the driller depth. For wells that had sufficient information to make this assignment, the BHT corresponding to the depth of measurement was retained. For all other wells, the average of the multiple BHTs at the same depth were taken, and the geotherm was recomputed to reflect the

average BHT. Only 5 wells (10 records) needed to be rerun, so an adjustment of uncertainty in the BHT measurement was not made to reflect that 2 measurements were taken at the same depth in these 5 wells, rather than 1.

In summary, after taking only the deepest of the 20,750 well records for each location, the number of data points for analysis in the EDA was reduced to 19,975 unique locations.

Negative Thermal Gradients

Of the remaining wells, 39 had negative values of the geothermal gradient computed between the surface and the depth of measurement. Calculated values of the geothermal gradient are negative if the assumed annual average surface temperature is greater than the BHT measurement in the well. A negative geothermal gradient is not reasonable, thus for these wells a calculated negative gradient may indicate that 1) the annual average surface temperature is too high, 2) the BHT is reduced as a result of advection of heat via groundwater, or 3) the BHT or depth of measurement was not properly recorded. Case 1) would affect shallower wells more than deeper wells, whereas case 2) and 3) could affect any well. Lacking information to test these cases, all 39 records with negative gradients were removed from the dataset.

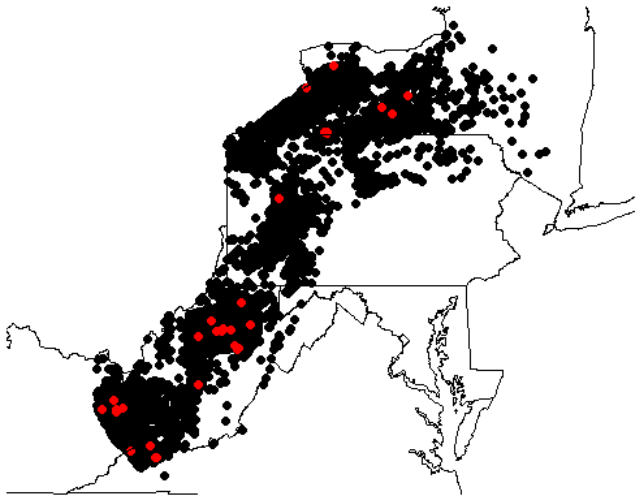

Figure 1: Locations of all 19,750 wells (black) and the 39 wells with negative gradients (red). Some clustering of wells with negative gradients exists, which may indicate advection of heat, but could also be a result of improper recording of data from the drilling company. Given that many more wells around the wells with negative gradients are positive, and further information is not available to quality control these data, removing them from the dataset is justified.

Depth Cutoff for Quality Purposes

Each of the thermal variables were plotted against the depth of measurement to determine if there were any biases based on depth. The plot for surface heat flow is provided in Figure 2, and all thermal variables of interest exhibit the same trend as surface heat flow. Clearly, values of heat flow greater than 100 mW/m^2 are alarmingly high for the Appalachian Basin, which is considered to be stable continent. Based on Figure 2, it was determined that many wells with BHT measurements shallower than 1,000 m were likely unreliable, and they were removed from the database for quality control. Aside from a few high outliers for West Virginia, the heat flow values for wells with a depth of BHT measurement greater than 1000 m appear to be in agreement.

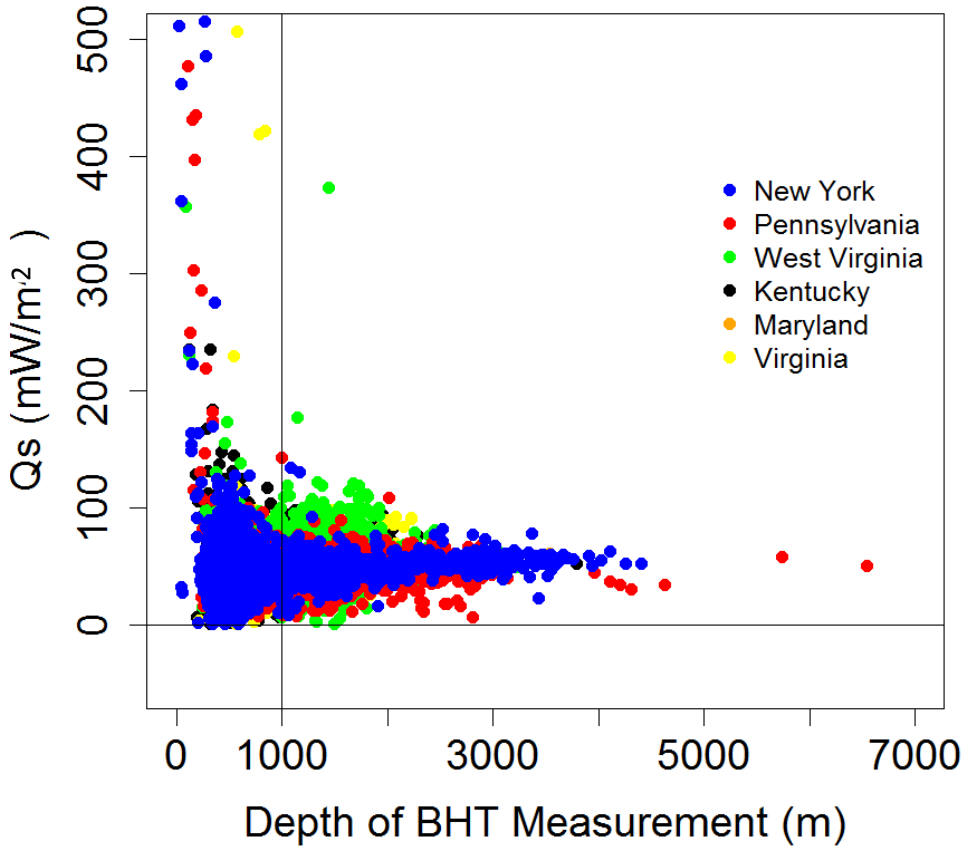

Figure 2: Surface heat flow versus the depth of BHT measurement for the 19,750 remaining wells. Only wells deeper than the 1000 m cutoff (vertical black line) were retained for further analysis.

For the large majority of the region these processing steps did not cause major data gaps to appear on the map (Figure 3). An exception is in northwestern Pennsylvania near Allegheny

National Forest, which did not have sufficient deep well data. After these processing steps, 13,312 BHT measurements remained for further analysis.

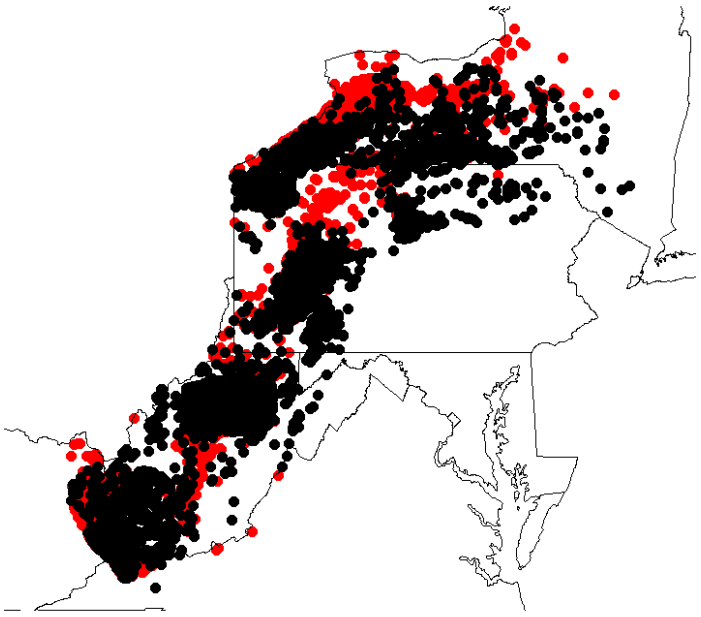

Figure 3: Wells remaining for outlier analysis (black) and wells removed based on the 1000 m cutoff depth (red). Generally, northern New York wells are removed because of a shallow depth to basement, and the oil and gas resource of interest was shallower than 1000 m. For most locations other than northern New York and a data gap near Allegheny National Forest in Pennsylvania, there is good data coverage even with the depth cutoff.

Outlier Analysis

Each thermal variable was subject to a spatial outlier analysis prior to being used in the interpolation. The memo entitled Thermal Outlier Assessment in GPFA-AB contains the details of the outlier algorithms developed for this project. The selected algorithm used the nearest 25 points within a 32 km radius for outlier analysis. For the roughly 200 points that did not have 25 points within 32 km, the outlier analysis was not conducted in the algorithm. Instead, manual inspection of these points is performed by looking at the values of the available nearest neighbor points. Only two wells were removed based on manual inspection. Table 1 summarizes the number of outliers removed for each thermal variable. It is interesting that wells that are outliers at one depth may not be outliers at another depth.

Table 1: Outliers identified for each thermal variable of interest for this project.

Variable	Number of Outliers	Percentage of Data Removed
Surface Heat Flow	1010	7.6%
Depth to 80 °C	1181	8.9%
Depth to 100 °C	1117	8.4%
Temperature at 1.5 km	979	7.3%
Temperature at 2.5 km	944	7.1%
Temperature at 3.5 km	970	7.3%

Spatial Autocorrelation

Previous studies of the Appalachian Basin thermal field have shown that there is significant spatial autocorrelation that ought to be captured in prediction of the thermal field (Aguirre, 2014). A study by Smith, Whealton and Stedinger (2014) showed that local small-scale variation in the structure of thermal field spatial autocorrelation in New York and Pennsylvania makes a region-wide model of spatial autocorrelation inappropriate. Therefore, the hypothesis for this project was that the spatial autocorrelation would exhibit similar local structure in West Virginia. In order to capture local structure in spatial autocorrelation, a different model of autocorrelation is needed for sub-regions of the Appalachian Basin. Sub-regions were originally defined by physiographic sections in New York, Pennsylvania, West Virginia, and surrounding states, but with the development of gravity and magnetic potential field edges in this project, they have since been redefined. We believe that the use of potential fields to define these sub-regions better reflects physical differences in rock properties at depth than the physiographic provinces. The following section describes how sub-regions were created from the multiscale potential field edges.

Creating Interpolation Boundaries

The concept of a heat flow province was first discussed in Roy, Blackwell, and Birch (1968) and referred to the apparent transition zone in heat flow originating from the mantle across a major continental structural divide on the order of about 100 km width. This concept of a heat flow province is applied on a smaller scale to the interpolation boundaries defined in this project. The physical and statistical justification for using sub-regions is that rocks with different properties may contribute different amounts of heat to the surface, thereby acting as different data generating processes that must be modeled separately in prediction of the thermal field.

The interpolation boundaries used in this project are provided in Figure 4. The primary source for determining the division of interpolation boundaries is from the gravity edges. The magnetic edges were used to refine the interpolation boundaries when large anomalies appeared within any section, but were not the primary source of the delineation.

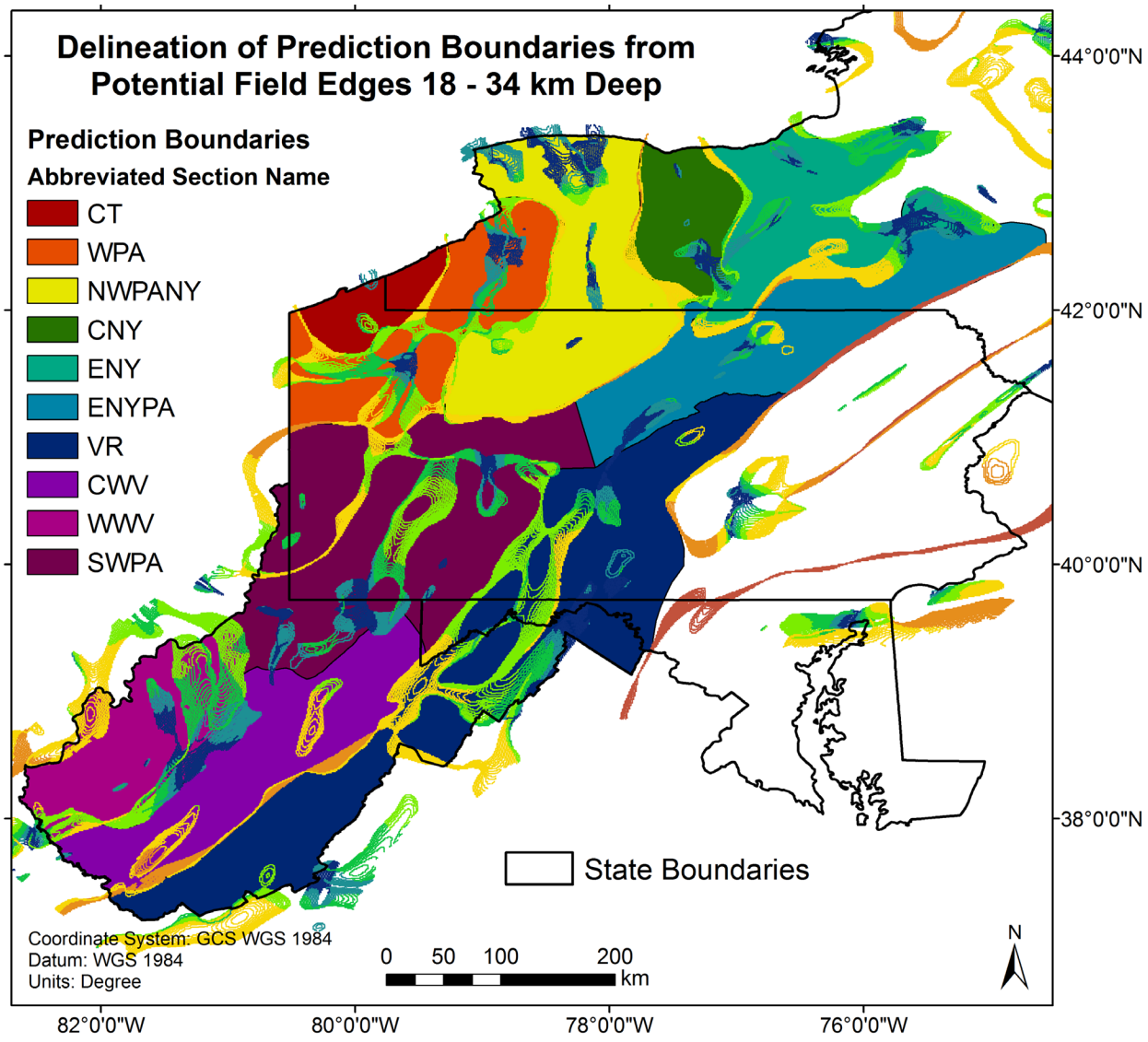

Figure 4: Interpolation boundaries used in this project. Boundaries are defined by gravity and magnetic potential field edges at depths from 18 km – 32 km. Only gravity edges are shown because interpolation boundaries were primarily defined by the gravity edges, with the exception being SWPA. Gravity edges are colored red/yellow/green/blue, with red having higher contrast in rock properties on either side of the boundary, and thus higher influence in the creation of the boundaries. The interpolation sub-regions are colored in rainbow colors starting in the top left, proceeding clockwise. CT: Chautauqua, NY; WPA: Western Pennsylvania; NWPANY: Northwest Pennsylvania and New York; CNY: Central New York; ENY: Eastern New York; ENYPA: Eastern New York and Pennsylvania; VR: Valley and Ridge; CWV: Central West Virginia; WWV: Western West Virginia; and SWPA: Southwest Pennsylvania.

Stratified Ordinary Kriging Interpolation

The sub-regions defined above act as laterally defined strata in an ordinary Kriging interpolation for the region. Ordinary Kriging assumes that the mean is unknown within the estimation window, taken as 50 km in this study, and that the error may be modeled as a stationary stochastic field within the region for which the field is defined (the strata in this application of the methodology). The stochastic field is modeled using semivariograms, a model of spatial autocorrelation, that are defined within each strata. All sub-regions in Figure 4 except for the Valley and Ridge were used. The Valley and Ridge sub-region was not used because too few well data points were available to construct reliable semivariograms for this sub-region.

Variogram Analysis

In expectation, the difference in the geothermal field for observations that are closer to one another should be smaller than the difference in the geothermal field for observations that are more distant from each other. One metric that describes this relationship is the semivariogram: the semivariance as a function of distance between all (n*(n-1))/2 point pairs in the dataset. There are five key parameters that must be fit in order to empirically model the semivariogram. These are the nugget, range, sill, anisotropy, and shape of the semivariogram. The nugget refers to the apparent discontinuity in semivariance at infinitesimal distances between two wells. If we had perfect measuring devices, and properties of the thermal field remained constant through time, and the assumptions used to model the thermal field were perfectly accurate, the nugget would be essentially zero. Because the nugget is not zero, is represents the measurement, positioning, and modelling errors present within the dataset. The range is the distance to which spatial correlation is modeled, and the sill is the value of semivariance at the range. The shape of the semivariogram may be selected from a collection of nearly 20 classic functional forms that provide positive definite and non-singular matrices for spatial prediction. Of these options, the Gaussian, Exponential, and Spherical shapes were used in this project to model and fit semivariograms. Variograms were fit using a preferential weighting given by the number of points within a bin representing the average semivariance for all points at some distance away, divided by the distance to the bin squared. This is the traditional approach, which ensures that more weight is placed on fitting points closer to one another rather than farther away. This approach was used for all cases except when semivariograms were complex. Ordinary least squares regression (no weights) were used for complex models. A further explanation of the ordinary kriging methods used in this analysis is provided in the Appendix of Stutz et al. (2015) under the subheading Kriging Interpolation.

Anisotropy

Anisotropy refers to directional dependence in the structure of spatial autocorrelation. Figure 5 shows the anisotropy within each of the nine sub-regions identified in Figure 4. As expected, each sub-region does not exhibit the same degree of anisotropy, and some do not appear to exhibit anisotropy at all. This result supports the use of stratified kriging that can capture these differences to be used in prediction of the thermal field.

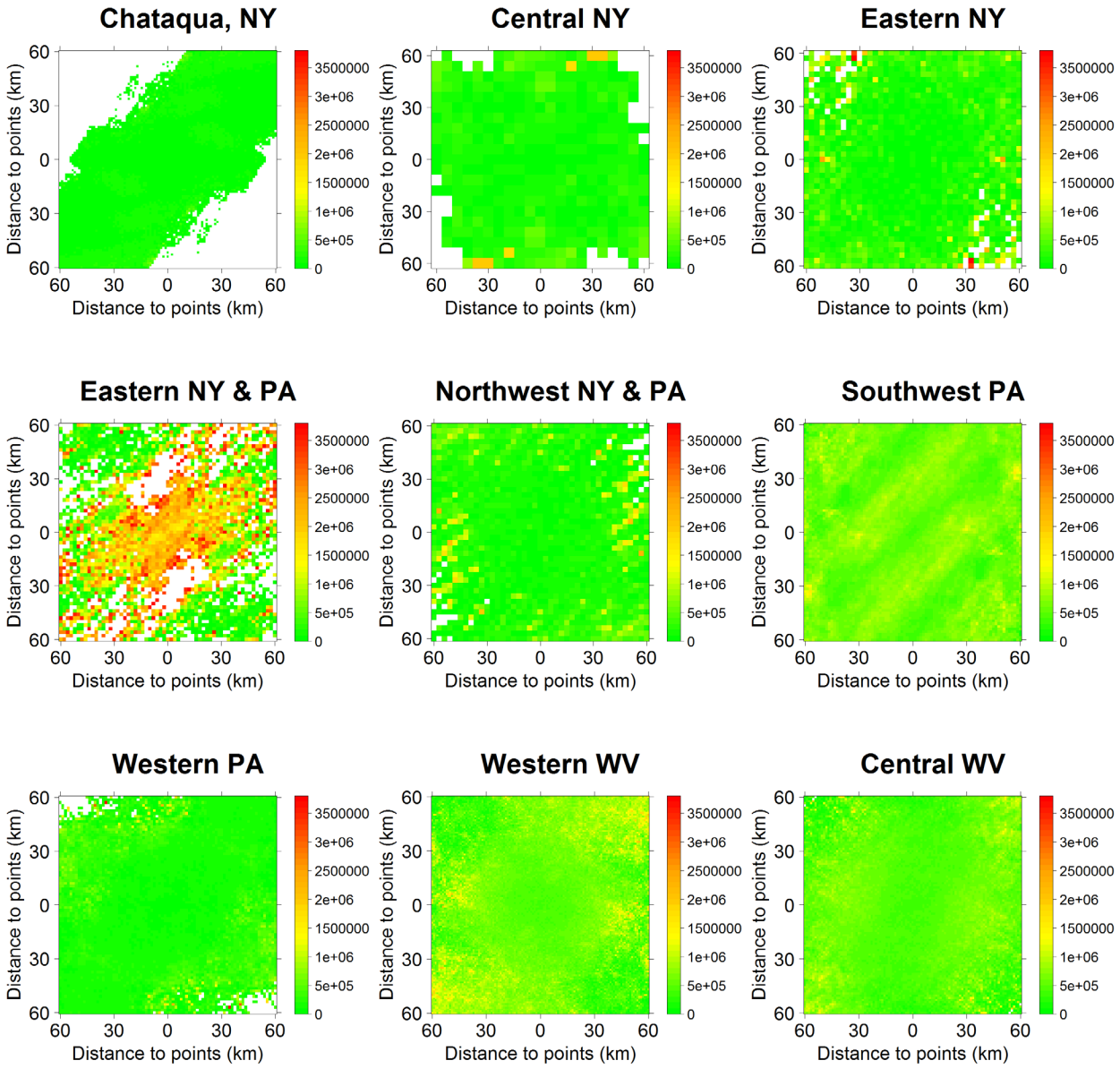

Figure 5: Contour plots of semivariance for Depth to 80 °C. All plots have the same distance and semivariance color bar. White areas are where insufficient data exist to display a value of the semivariance. Where the plots appear elliptical, there is anisotropy (directional preference) in the structure of spatial correlation. Anisotropy must be captured by spatial prediction algorithms to appropriately model the thermal field.

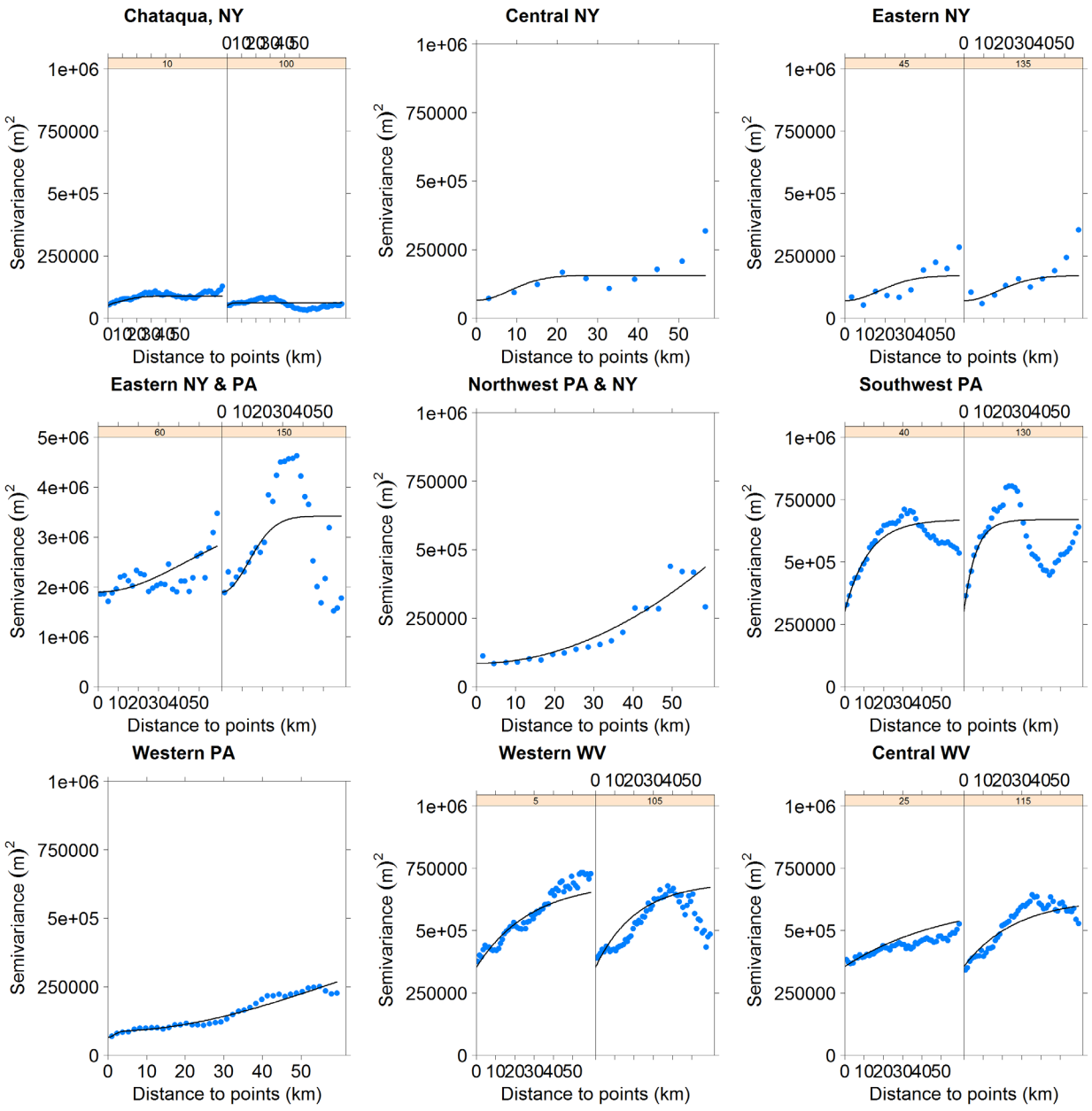

Figure 6: Fitted semivariograms used for interpolation within each of the nine provinces. Provinces that showed anisotropy in Figure 5 have variograms defined along the major and minor axes of anisotropy, labeled by the angle in the pink bar. The fitted semivariogram shape at angles between the major and minor axis angles is located somewhere between these two extremes. All plots have the same vertical axis, except Eastern New York and Pennsylvania. The distance labels are provided up until the interpolation distance of 50 km. The interpolation distance was selected because the maximum range to the sill for these regions is approximately 50 km.

Using these fitted semivariogram models the ordinary kriging interpolation was run within each of the sub-regions. Only those wells located within a sub-region were used to predict in the sub-region. Additionally, a minimum of 5 wells within 50 km of a location were needed to make a prediction. A maximum of 50 points were used for computational efficiency.

Results

The kriging algorithm provides the predicted mean and the standard error of the predicted mean. Maps for each of these are provided for the Depth to 80 °C. Uncertainty in the predicted mean is of great interest because the a large portion of the risk in a geothermal project is the first step – declaring that a resource is present in a quantity and quality that is desirable for the end uses considered. Therefore, quantifying the uncertainty in the thermal field prediction can greatly aid decision makers in determining the most certain hot spots in an area of interest.

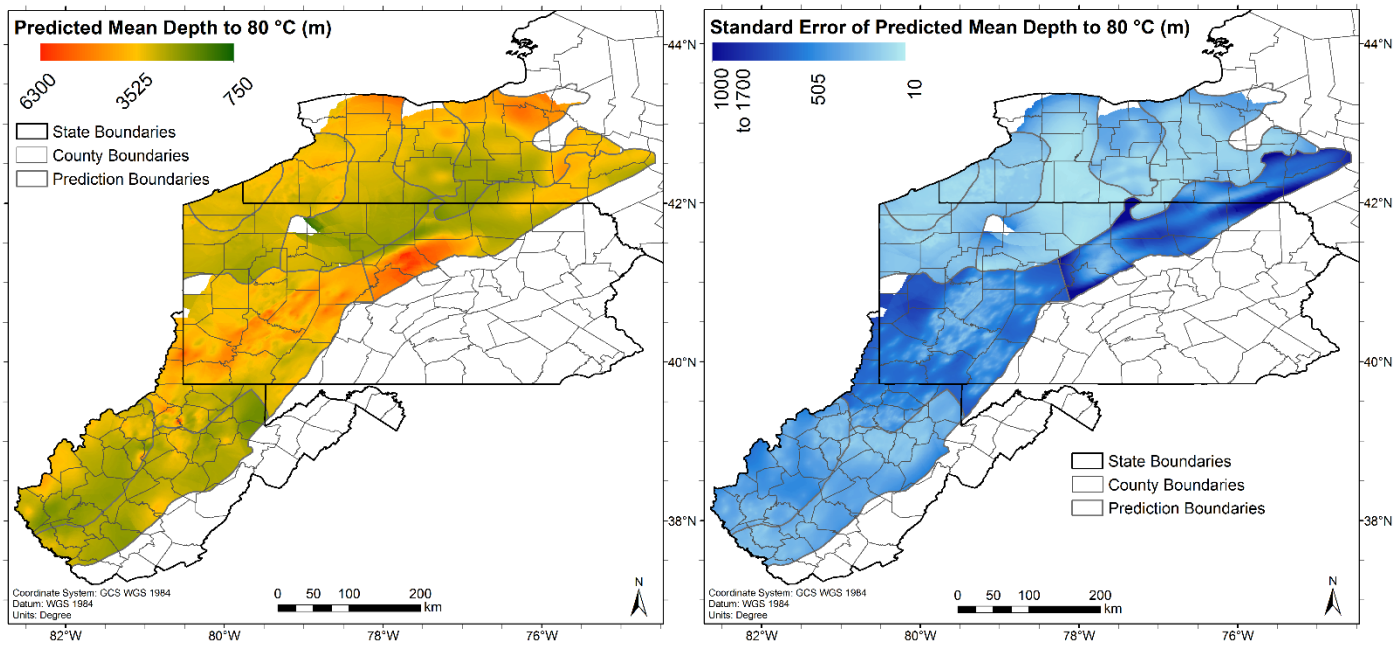

Figure 7: Predicted mean and standard error of the predicted mean depth to 80 °C. Interpolation boundaries (Figure 4) are shown in light gray lines. Note that red (deeper) is bad and green (shallower) is good on the predicted mean map. The upper end of the standard error map has a range because nearly all error between 1000 – 1700 m is contained within one section – Eastern New York and Pennsylvania.

Cross Validation

As stated above, the results of the Leave One Out cross validation are provided in the memo entitled Selection of Four Counties in Each State with the Best Thermal Resources. The cross validation using equilibrium temperature data at 1.5 km is presented in Figure 8.

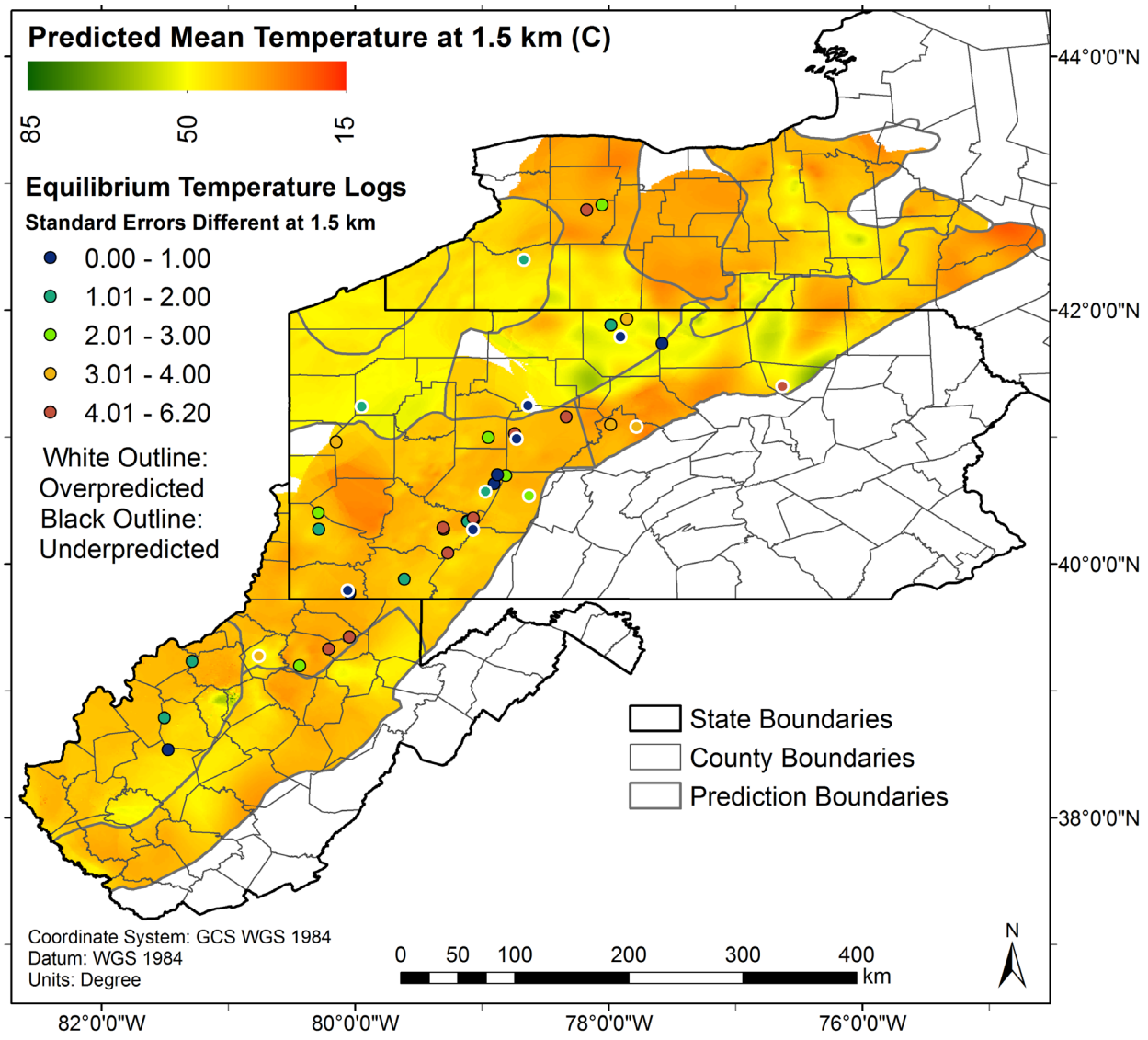

Figure 8: Wells of equilibrium or reliable temperature data are compared to predicted temperatures at 1.5 km depth. The colors of the circles show differences from measured and predicted temperature at 1.5 km.

References:

Aguirre, G.A.. (2014). Geothermal resource assessment: A case study of spatial variability and uncertainty analysis for the states of New York and Pennsylvania [M.S. thesis]: Ithaca, New York, Cornell University, 482 p.

Pebesma, E.J. and C.G. Wesseling. (1998). Gstat, a program for geostatistical modelling, prediction and simulation. Computers & Geosciences, 24(1). Pp. 17-31.

Roy, R.F., D.D. Blackwell and F. Birch. (1968). Heat generation of plutonic rocks and continental heat flow provinces. *Earth and Planetary Science Letters, 5*. Pp. 1 – 12. North Holland Publishing Comp. Amsterdam.

Smith, J. D., C. A. Whealton, & J. R. Stedinger. 2014, December. *Spatial Analysis of Geothermal Resources in New York and Pennsylvania: A Stratified Kriging Approach*. Poster presented at the Renewable Energy III Session of the 2014 American Geophysical Union Fall Meeting, San Francisco, CA.

Stutz, G.R., E. Shope, G.A. Aguirre, J. Batir, Z. Frone, M. Williams, T.J. Reber, C.A. Whealton, J.D. Smith, M.C. Richards, D.D. Blackwell, J.W. Tester, J.R. Stedinger, and T.E. Jordan. (2015). Geothermal energy characterization in the Appalachian Basin of New York and Pennsylvania. Geosphere, v. 11(5). p. 1291–304. doi: 10.1130 /GES00499.1.