CS-GY 6513: Project Report
Antarctica: The Last Frontier

May 17, 2019

Prof. Juan Rodriguez

Meghan Bongartz (mb7052); Michelle La (mal928);
Varun Mathur (vsm277); Yequ Sun (ys3278)

Introduction

In 2019, climate change poses the number-one threat to humanity's continued existence on Earth, and its well-documented effects are becoming increasingly frightening. One such consequence is the melt of Antarctic ice, contributing to rising sea levels and disrupting Antarctica's native wildlife. While we tend to think of this in the context of penguins losing their homes, it is important to also remember that there is a huge continent land mass underneath the ice. Although we certainly hope that the impacts of climate change will be slowed or reversed, the current warming trends for the near future indicate that while hot equatorial areas may become uninhabitable, the bedrock beneath the ice in Antarctica could simultaneously become exposed. Meanwhile, a second issue facing the planet is that of overpopulation. Given the human drive to explore new frontiers and the temperature rise of currently habitable regions, humans would likely begin colonization of a newly habitable continent. Relevant predictions regarding location and ice level will be vital to future infrastructure, environmental policies, and international politics.

We expect that humans would first settle Antarctica in areas with no ice since it would be difficult to build permanent structures, grow food, or travel in ice-covered areas. While it is by no means impossible for people to live in tundra biomes, we must draw a distinction between a few people surviving and the formation of full settlements similar to U.S. cities. In an effort to understand what human colonization of Antarctica might look like, we analyzed ice thickness levels and ablation rates, or ice mass loss due to melting and sublimation, in Antarctica to find out when areas of bedrock would be exposed. Then, we analyzed global population as a function of elevation, and mapped this onto Antarctic topography exposed from the ice at different time points.

Ablation Rate Analysis

By first examining rates of ablation, we wanted to predict which areas could be habitable at certain points in the future. With a dataset from Taylor Glacier, Antarctica, sourced from the U.S. Antarctic Program Data Center, we worked with samples including location, elevation, time, and ablation rate.

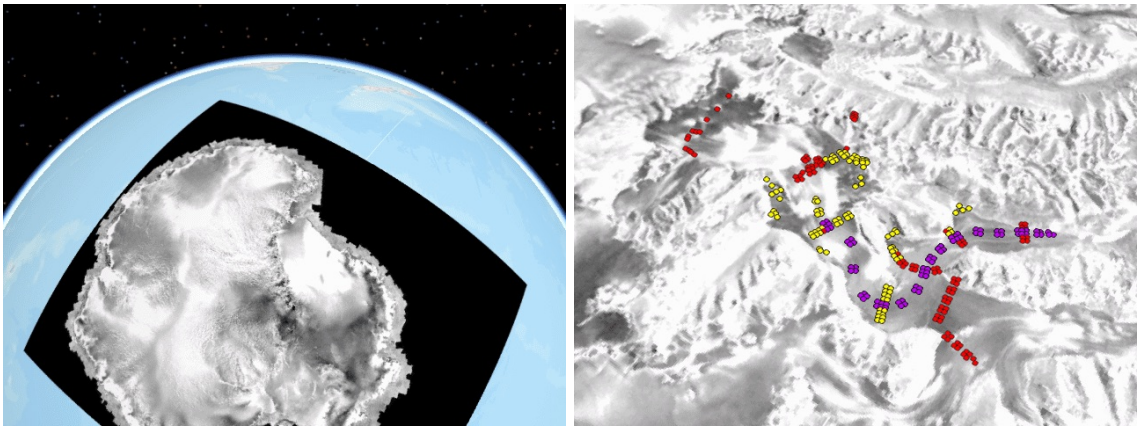
We utilized PySpark and NYU's High-Performance-Computing (HPC) clusters. PySpark enables handling of large dataset with DataFrames, while HPC clusters utilize parallel processing capacities to speed data processing. The PySpark console on the HPC Dumbo cluster was perfect for the task since it shares most of the code with our local machines. There are only two small differences in practice. First, to run jobs on Dumbo Pyspark, all files must be on the HDFS. Second, declaring a SparkContext instance is not needed on Dumbo cluster.

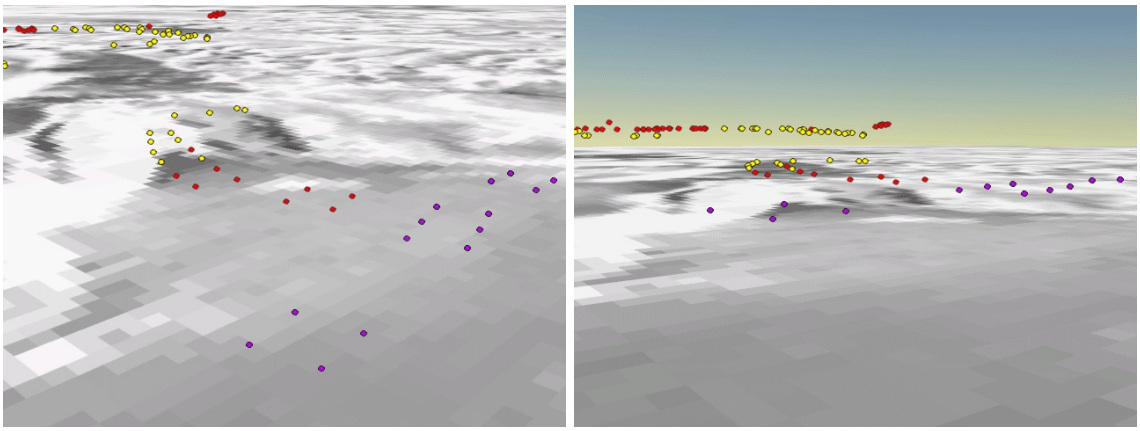
We used window functions to determine minimum and maximum elevation and ablation rate.

	Elevation	Ablation Rate
Minimum	247.270004	-0.57172
Maximum	1710.35999	0.85015

We also visualized the ablation dataset using ArcGIS Pro, mapped over a different geoTIFF image dataset of Antarctica. ArcGIS Pro's 3D layering interface enabled us to observe how our data samples were physically arranged and oriented relative to one another, as well as visualize multiple geographic datasets together and create interactions between them.

ArcGIS Pro visualizations of ablation dataset





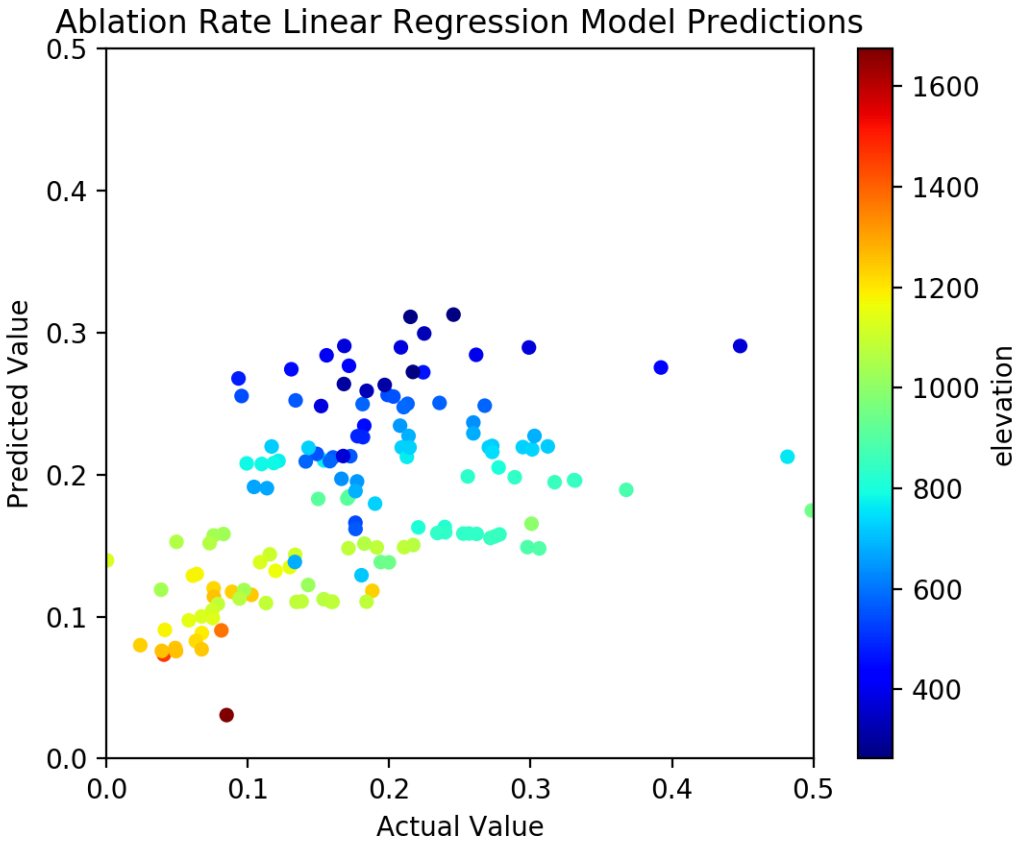
These images demonstrate the process of zooming to the layer of data samples, with colors differentiating between samples taken from each of six input files (only three colors visible due to overlap). As we reach the surface, we can rotate and tilt the visualization to observe differences in elevation for the samples.

Multivariate Linear Regression

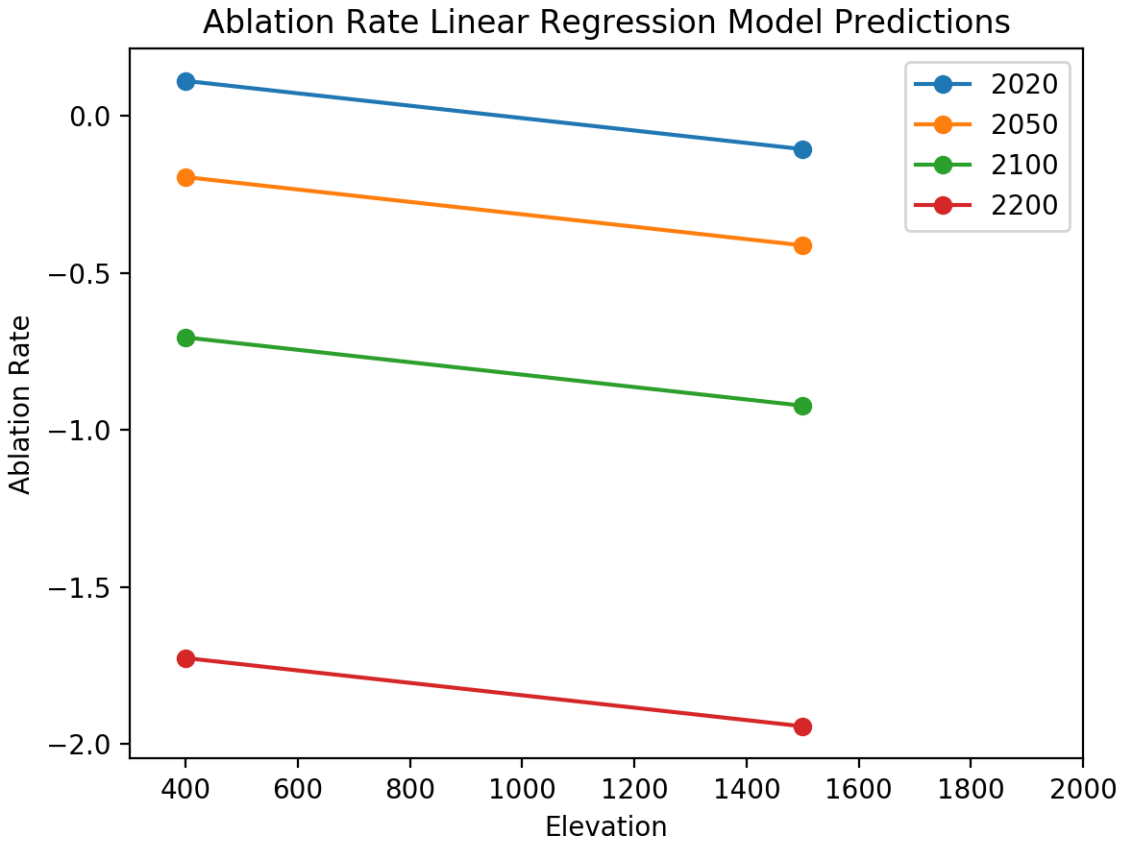
We then performed multivariate linear regression on the ablation dataset with PySpark to obtain a linear relationship between the independent variables elevation and time, and the dependent variable ablation rate. This enables us to predict the ablation rate given elevation at certain time points. We generated the following values corresponding to our model.

Coefficients	-0.000197696, -0.0102181
Intercept	20.83185
RMSE	0.10348
r^2	0.24973

The graph below shows the actual values in the test set compared with our predicted values. This is a relatively high root mean squared error (RMSE) given the standard deviation of 0.11963.



While we were able to obtain a positive correlation between these values, the model struggled with the data's inconsistencies. Our simple model predicted generally that each elevation was linked to a certain value, leading to the horizontal bands of color in the graph, but the relationship is ultimately not so simple.



The above image shows the predicted ablation rates at different elevations over time based on our model. Each line represents the predicted ablation rates for a given year. We expected that ablation rates would be higher at lower elevations - that is, ablation occurs at faster rates in valleys instead of mountains - and found this to be the case. However, we also found that for this particular data set, ice actually accumulates at faster rates (negative ablation) as time goes on.

This isn't what we expected, of course, because it's well established that Antarctic ice is melting in response to climate change. We experimented with multivariate polynomial regression to see whether the linear model was too simplistic for this dataset, but that approach yielded similar results with negative ablation rates.

Discussion

There are a few reasons why we believe this is happening in our data set with this model. This data is for a single glacier and was collected over a relatively short time period, only 8 years, when the National Oceanic and Atmospheric Administration defines climate as the average of weather over a 30 year period. This means that the scope of the data itself may be insufficient for predicting ablation rates up to 200 years into the

future, especially given that climate change can have unexpected short term effects such as the "deep-freeze that parts of the U.S. experienced this past winter. Additionally, the glacier that these rates were measured at is located far inland in Antarctica, and may not be experiencing the same rapid ice melting and iceberg calving that is occurring in more coastal regions.

We were unfortunately unable to find a sufficient ablation data set to make accurate predictions for the entire continent, so in order to move forward with our primary focus of mapping global population patterns onto Antarctica, we used a non-elevation-specific, literature-sourced rate to determine general ice reduction.

With data collected from 2009-2017, Rignot et. al. calculated average Antarctic ice loss to be 252 Gigatons annually. With the U.S. Antarctic Program's estimate of ice density at 0.9 $\frac{\text{kg}}{\text{m}^3}$ and a total Antarctic area of 14,200,000 square kilometres, our average ice ablation rate is 19.7 m w.e./year. We continued our overall analysis using this rate.

Global Population and Elevation Correlation

The next step in our eventual goal of predicting population at time points in Antarctica is to evaluate current data correlating global population with elevation.

To explore the relationship between bedrock elevation and population density, we used two huge datasets. The first one is the global elevation dataset from NASA. It has coordinates and elevation data of the corresponding location. The other one is the population center dataset from CIESIN. This dataset contains millions of coordinates of administration points and its area, population and population density.

In order to prepare the data for a Spark machine learning module to do a linear regression, the two datasets needed to be joined. Because these two datasets use very different levels of precision for the longitude and latitude, this proved a challenging task; it is impossible to find an exact match of location between the two datasets. In the end, we rounded the coordinates digit by digit until there were matches. It was the equivalent of searching the area around the coordinates to find a match. Then, the joined table was vectorized and modeled with the Spark machine learning linear regression.

In the end, the linear regression module discovered a negative correlation between bedrock elevation and population density.

Coefficient	-1.73769
Intercept	2483.69485
RMSE	3529.48720
r^2	0.02707

We used linear regression in PySpark again to model global population according to elevation. Our RMSE was quite high at 3529.48720 relative to our standard deviation of 3579.73276, due to our simple linear model which only accounted for elevation. With the correlation between elevation and population density, combined with our projections of when the bedrock in Antarctica will be exposed, we were able to predict the hypothetical population density in Antarctica for various points in the future.

Mapping Population Onto Antarctica

NOAA's ETOPO1 global relief model, or topographic model of the world, contains bedrock and ice height data across Antarctica from 2008, from which we were able to extract the ice thickness across Antarctica using PySpark.

By applying our constant ablation rate, we subtracted the meters of ice lost over time from the ice thickness data for certain time periods. Once again using window functions, we determined minimum and maximum ice thicknesses for predictions at each time point.

	2008	2050	2100	2200
Minimum	0.0	0.0	0.0	0.0
Maximum	4253.0	3425.6	2440.6	470.6

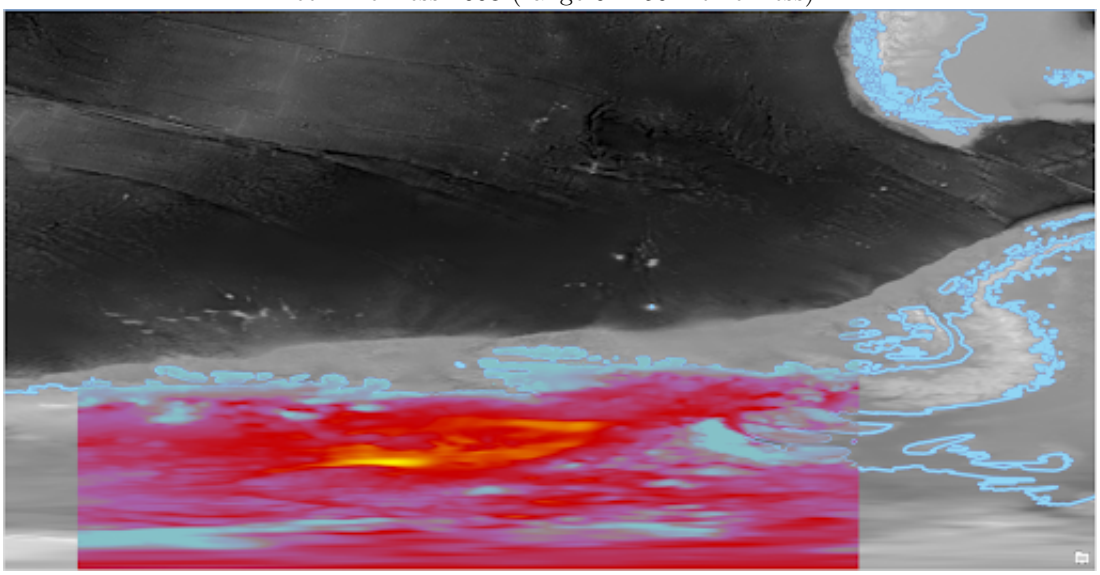
Using our constant ablation rate, we also found that by year 2224, the ice melted completely, leaving only bedrock with a minimum elevation of -5897.0 and a maximum elevation of 4217.0.

$$(4253 \text{ meters}) \,/\, \left(19.7 \,\frac{\text{meters}}{\text{year}}\right) \approx 215 \text{ years}$$
$$2008 + 216 = 2224 \text{ years}$$

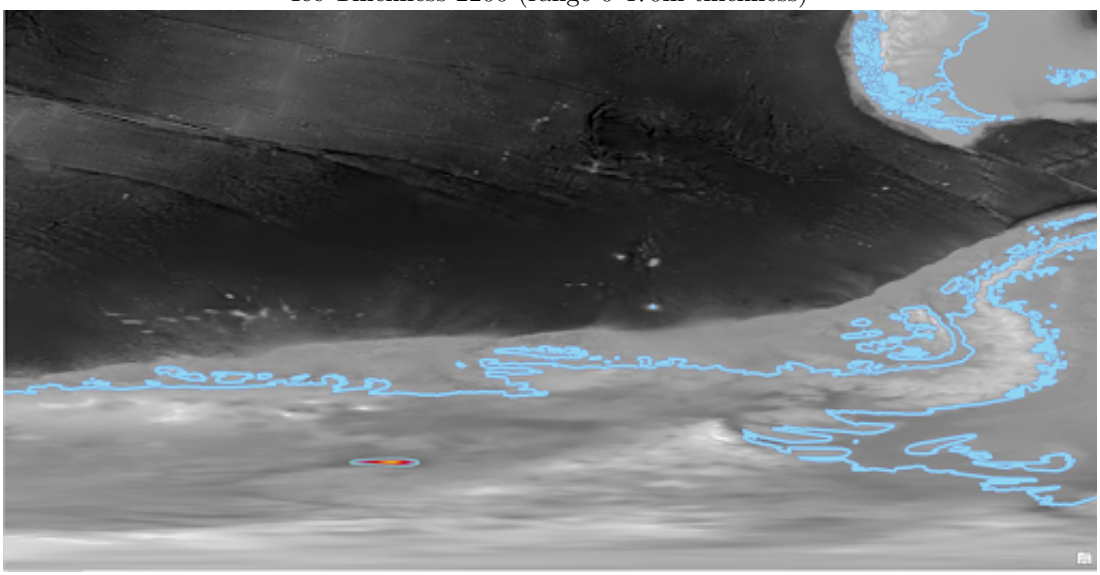
Ice Thickness Visualization with ArcGIS Pro

We visualized this in ArcGIS Pro, and the results are striking. In the images below, a color gradient was applied to ice thickness with the blue area representing the thinnest existing ice cover, and yellow representing the thickest ice cover. Note that we visualized only a sample area, so the areas to the right and left where no coloring exists in the 2008 image are not actually bare; in fact, at this scale we cannot see any patches that are entirely bare. By 2200, only a tiny patch of ice remains. This is even more notable given the scale of the imagery: ArcGIS does not allow manual manipulation of color scale ranges, so the yellow area in the 2008 image represents an ice thickness of up to 4250 meters. By 2200, the maximum ice thickness has decreased to only 470 meters.

Ice Thickness 2008 (range 0-4250m thickness)



Ice Thickness 2200 (range 0-470m thickness)



With the linear regression model ready, the hypothetical population density of Antarctica could be projected. There were two more huge datasets for this job: Antarctic bedrock elevation data and ice elevation data with more than 5 million data points combined. Firstly, the two datasets were merged into one. This time, the coordinates matched each other perfectly. Next, the ice thickness of certain coordinates could be calculated by simply subtracting the bedrock elevation from the ice elevation. Then, with the help of the constant ice ablation rate (19.7 meters per year), the future ice thickness could be calculated accordingly. Spark's RDD map function and DataFrame column-wise operation helped significantly with processing. At last, for each coordinate in the dataset, when the ice thickness of a certain year drops to 0, we can predict the hypothetical population density based on the bedrock elevation thanks to the previously-found population-elevation linear

regression model.

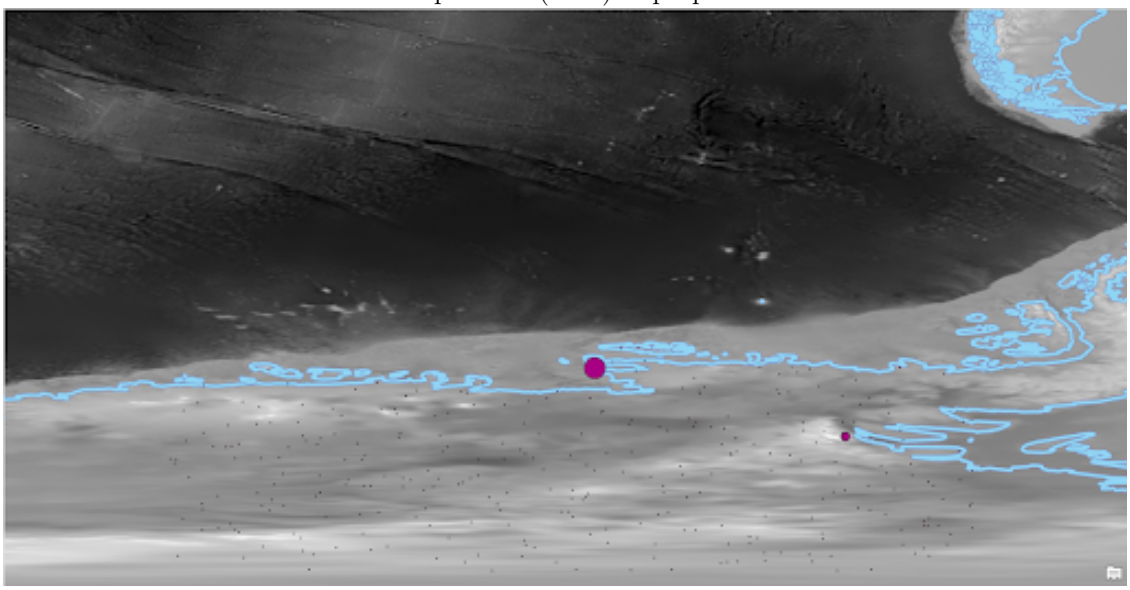
Due to the sheer size of our datasets, it was extremely hard to explore operations on full scale. Most of the work was done on the NYU HPC Dumbo cluster first before being transferred to the Jupyter PySpark notebook on our personal computers. The output .csv file for GIS with predicted population density data is over 700 MB and takes hours to write on our personal computers. Dumbo finishes the job within several minutes. The runtime of mappings and aggregations is also much, much faster on the cluster.

Population Mapping Visualization with ArcGIS Pro

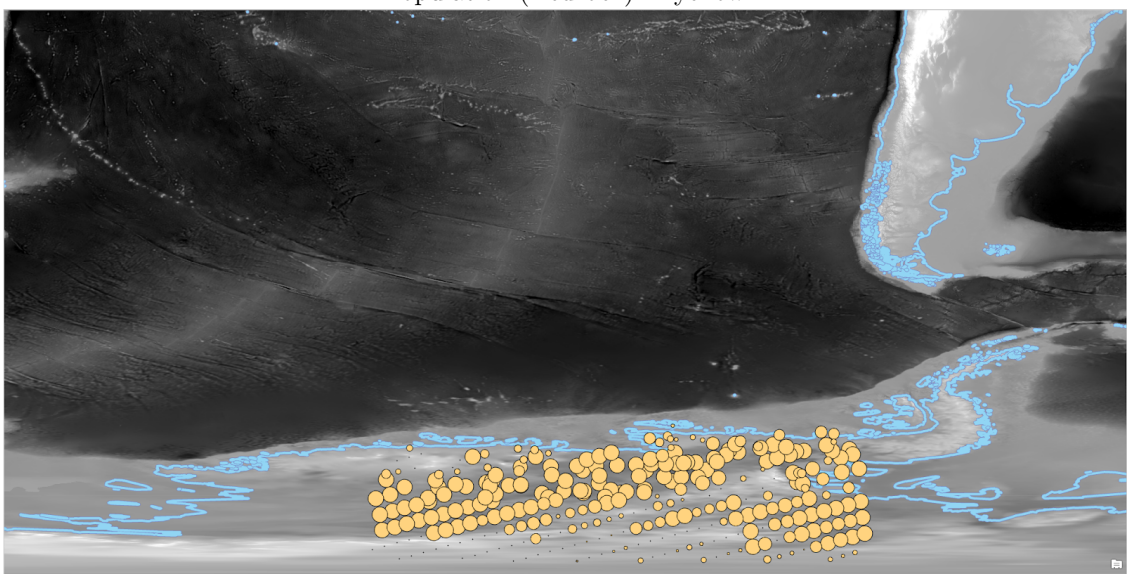
Some exploratory visualization confirmed our theory that population centers occur at lower elevations and tend to be clustered in coastal areas. We used ArcGIS Pro to visualize population on exposed bedrock for years 2008, 2050, 2100, and 2200.

This involved converting the .csv outputs from our PySpark analysis into points and clipping them to fit a shape file of the Antarctic continent to ensure that we didn't map people living in the ocean trenches. Our maps include a coastline shapefile that we converted from a vector using QGIS in order to aid viewers in understanding what they are looking at. As with our ice thickness maps, we have visualized a sample set of our data. When viewing these maps, it is important to remember that this is a zoomed in view of an entire continent, so the population centers look far closer together than in reality. Because our analysis provided us with a projected population density for every GPS coordinate on a very fine grain, when sampling the data, we selected to show one in every ten thousand points in order to simulate the appearance of cities.

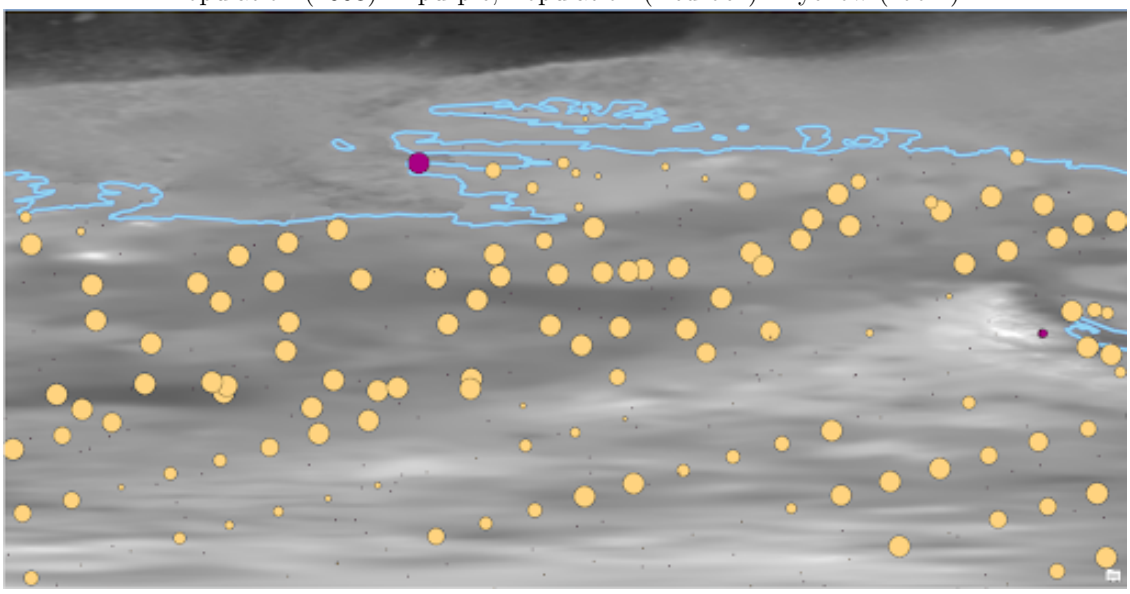
Population (2008) in purple



Population (Bedrock) in yellow

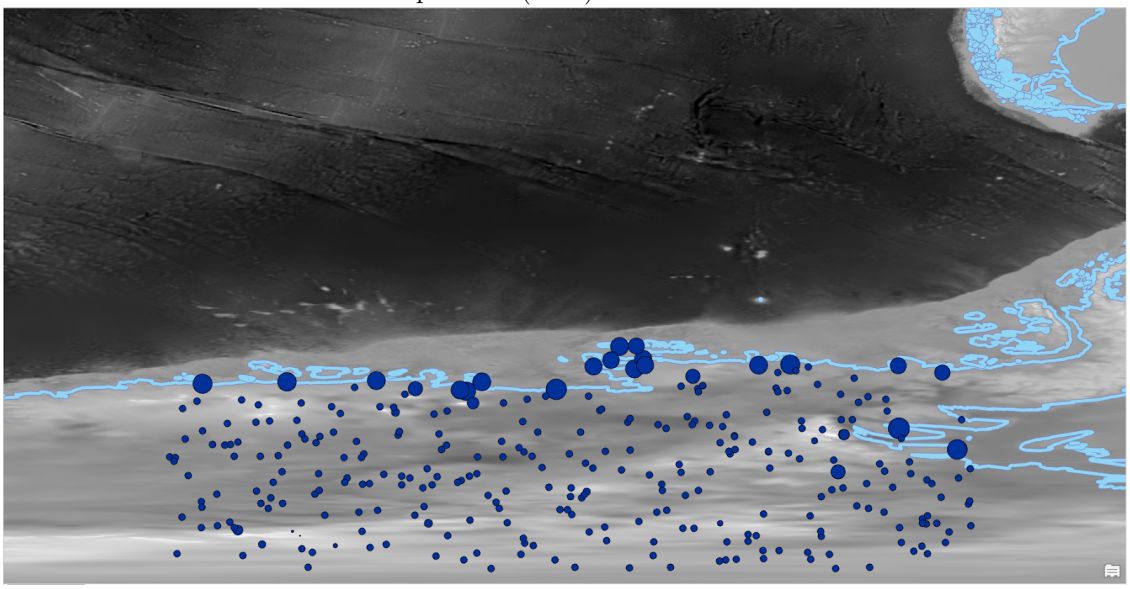


Population (2008) in purple; Population (Bedrock) in yellow (zoom)

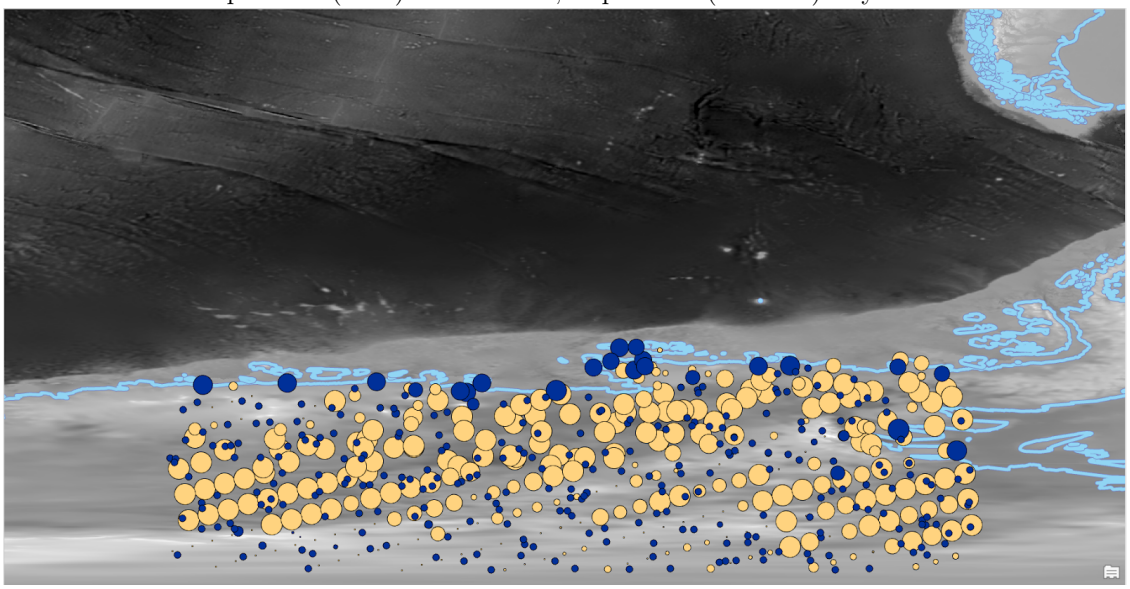


We were surprised to see that by 2050, Antarctica should already be clear enough to support a large number of population centers - though perhaps this should not be quite so shocking given recent news that the global impacts of climate change will likely be quite devastating by that not-so-far-off point.

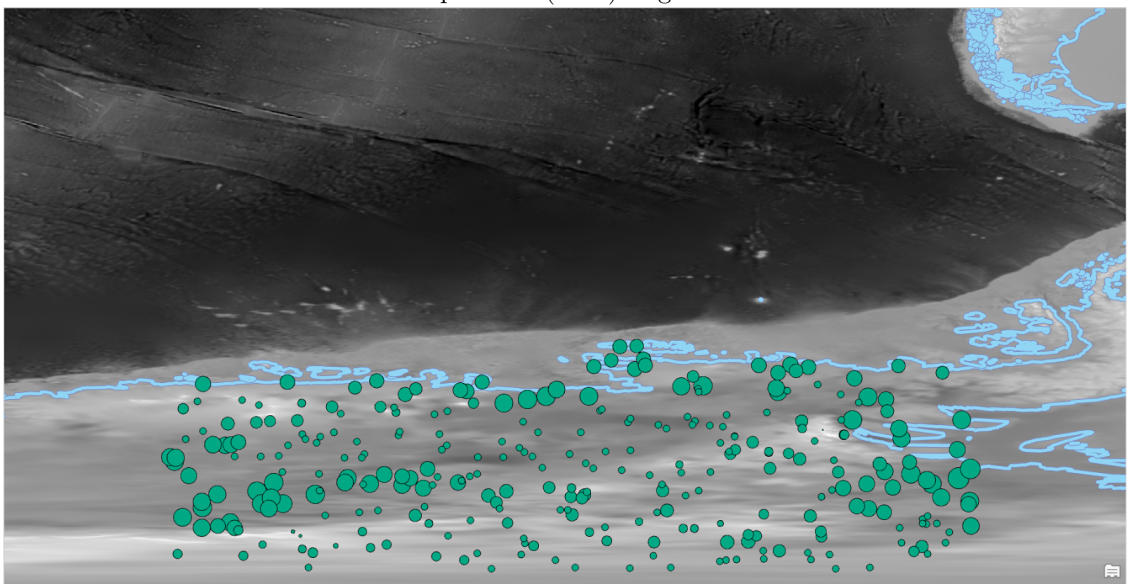
Population (2050) in dark blue



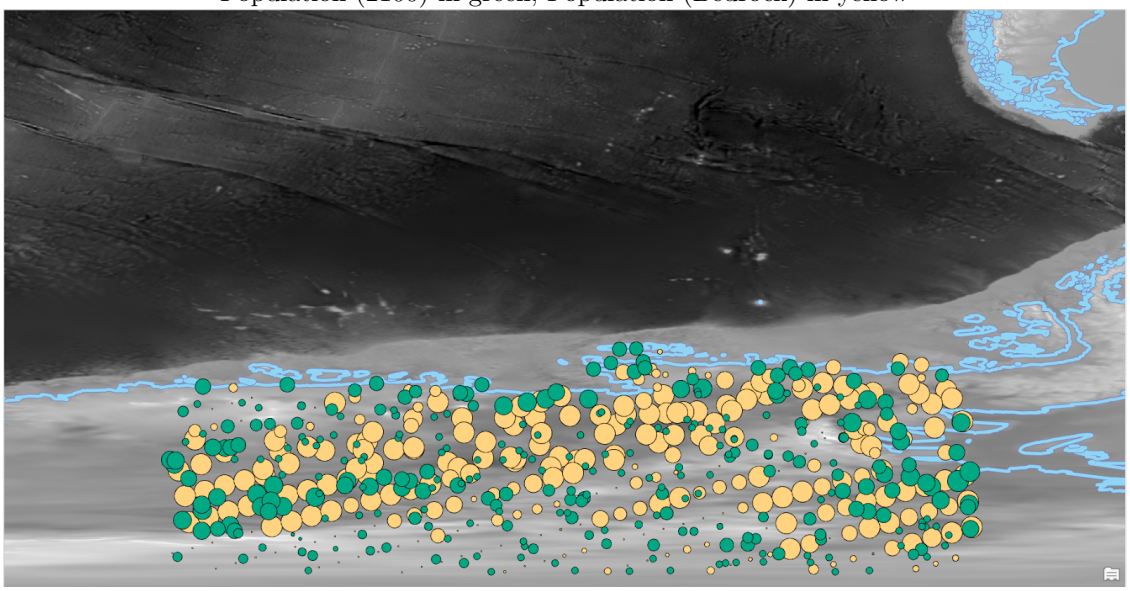
Population (2050) in dark blue; Population (Bedrock) in yellow



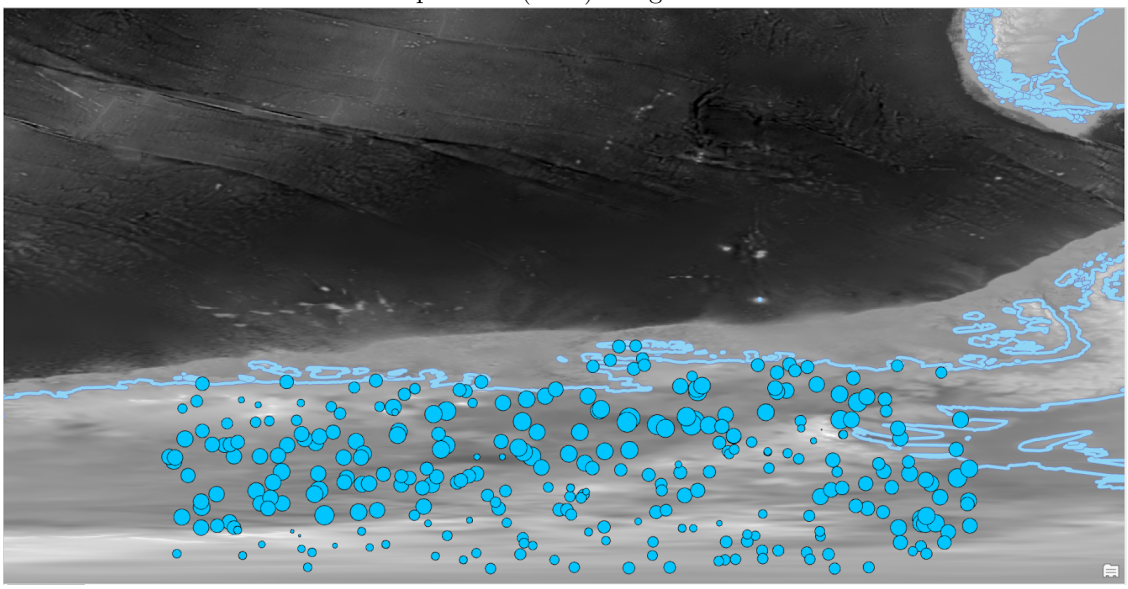
Population (2100) in green



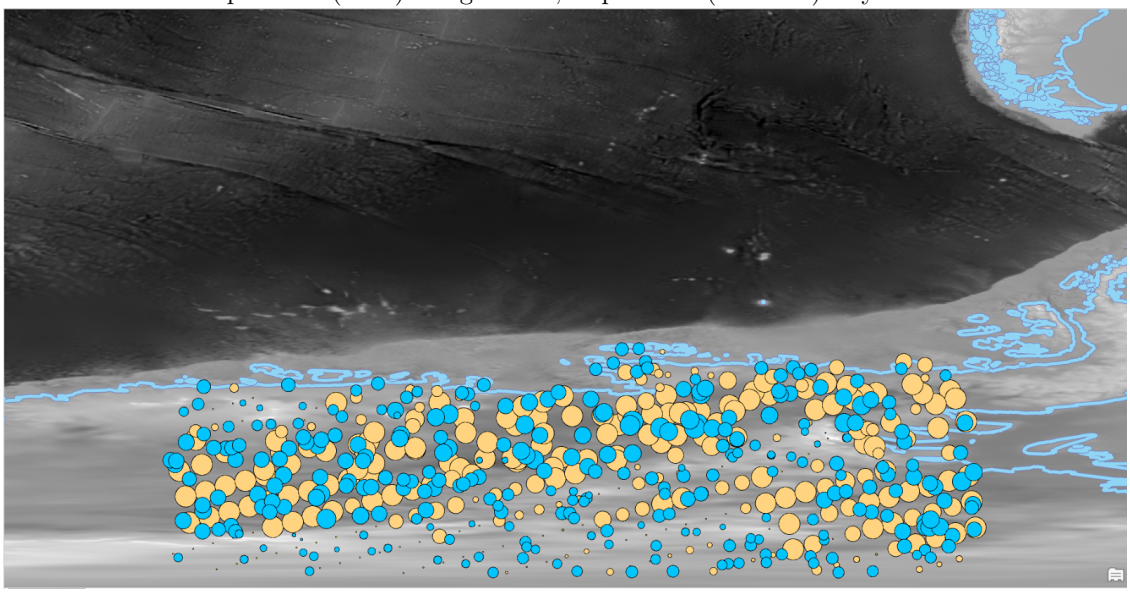
Population (2100) in green; Population (Bedrock) in yellow



Population (2200) in light blue



Population (2200) in light blue; Population (Bedrock) in yellow



Conclusion

We used multiple datasets to create novel prediction models of population based on elevation. This analysis could easily be made more useful and accurate by considering some additional factors. As the ice is melting, the sea level is also rising, so the coastlines that we used here many no longer be the coastlines of the future, and coastal areas may become less appealing to live in. Biome development/changes and weather patterns such as severe storms may make some areas that would otherwise promising settlement locations uninhabitable, driving populations into areas that they would normally be less likely to live in. Population growth rate may change.

This work is significant in beginning to address what will soon be a new reality for our planet. Analyses similar to the one we have performed here could assist governments and industries in adapting to the consequences of climate change on earth, and are not only applicable to Antarctica. Our analysis could be expanded to model future populations in other currently unpopulated areas, predict changing populations in areas that people live in today, or visualize population distributions on the moon or Mars if humans move beyond Earth. This information will be vital to these missions long before launching the first ship - or the first rocket.

Works Cited

Cuffey, K. M. et al. (2007) "Ablation Rates of Taylor Glacier, Antarctica" U.S. Antarctic Program (USAP) Data Center. doi: 10.7265/N5N29TW8.

RAMP AMM-1 SAR Image Mosaic of Antarctica, Version 2. Version 2. Archived by National Aeronautics and Space Administration, U.S. Government, National Snow and Ice Data Center. https://doi.org/10.5067/8AF4ZRPULS4H.

Amante, C. and B.W. Eakins, 2009. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center, NOAA. doi:10.7289/V5C8276M [May 9 2019].

https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-coastline/

Center for International Earth Science Information Network - CIESIN - Columbia University, United Nations Food and Agriculture Programme - FAO, and Centro Internacional de Agricultura Tropical - CIAT. 2017. Gridded Population of the World, Version 4 (GPWv4): Population Count Grid. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Accessed 9 MAY 2019.