Predicting & Visualizing Vulnerability to Flooding for Upstate NY in Near Real-time with Google Earth Engine API

Concept Note for NYS RISE (Resiliency Institute for Storms and Emergencies)

Bessie Schwarz | Bessie.Schwarz@Yale.edu | Beth Tellman | Elizabeth.Tellman@Yale.edu
Yale School of Forestry and Environmental Studies | 195 Prospect Street (G11), New Haven, CT, 06511

In partnership with Dr. Keith Tidball | kgtidball@cornell.edu

Summary: This concept note represents initial efforts to down scale a global flood vulnerability model developed in a cloud based computing tool Google Earth Engine for the noncoastal "upstate areas" of the State of New York. This customized New York application of the model is the result of collaboration with our colleague at Cornell University. The model analyzes social and physical vulnerability to riverine flooding based on multiple data inputs, outputs the high risk areas for flooding, and runs statistics on the population living in the flooded zone. Initial results examine the ability for the model to predict risk for a specific storm area, county, or watershed in 1-30 seconds. Future work requires further testing and validation of the model, a more advanced algorithm, and dynamic user-friendly interface for public risk communication of both underlying vulnerability and an early warning system. A zoomable web map of baseline vulnerability data is available for exploration here.

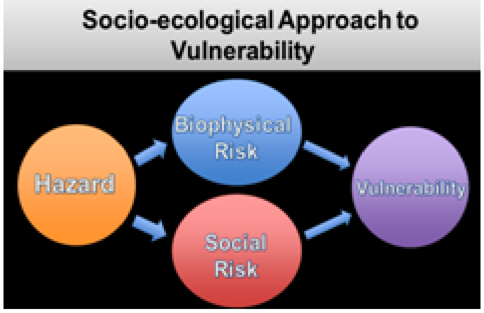
I. Introduction:

We are developing an application in the new Google Earth Engine API that uses highly parallelized cloud computing to model social-ecological vulnerability to flooding at high spatial resolution. The model is currently designed to produce coarse results at a global scale. For this project, we will refine the model to run at higher resolution on watersheds in New York State.

The proposed activity draws from modeling data readily available on the Google cloud platform, including elevation, satellite imagery, and census data to dynamically refine a surface of risk within a flood prediction zone produced by weather services (i.e. NOAA). In this proposed model, the algorithm first finds the flood zone using a set of biophysical indicators of vulnerability and then analyzes social vulnerability within the flood risk zone to find the overall area of risk. The current application finds the physical area of highest risk, and the number of people living in the risk area.

Model Framework and Context

"To foster resilience and sustainability within a system, an understanding of adaptive cycles within the coupled human-environmental system, and the scale at which they occur, is necessary"
-from "A place-based model for understanding community resilience to natural disasters," Cutter et al 2008



This models draws from a social-ecological systems approach that assumes that flood vulnerability is the product of both biophysical and social risk. The model computes an underlying vulnerability index for riverine flooding (NOTE: not coastal flooding/ storm surge) based on basic assumptions in physical and social sciences, *e.g.* that flooding occurs in areas that are low and flat, pools at the bottom of larger watersheds, occurs in watersheds that have a lot of impervious surfaces and less capacity to infiltrate rainfall, and that people who are very poor, very old, and very young, live in fragile communities and are less likely to have the means to evacuate themselves, or to be evacuated by their neighbors (see Cutter et al 2003 and Cutter et al 2008). The model scores each pixel by adding up each of these biophysical and social indicators.

This pilot research focused on application of the vulnerability model to the entire state of New York. Most attention in the post Sandy environment has been on building disaster resilience and coastal flood surge modeling capabilities for New York City. Little attention has been focused on building disaster resilience, risk communication, and forecasting for the rest of the state. While FEMA flood maps exist for some areas (and were recently updated for NY), many counties remain without flood maps, or have maps as much as 30 years out of date. Furthermore, when a storm arises, flood predictions are often too late (pers. Communication, NY Office of Emergency Management) or poorly communicated (pers. Communication, Dr. Keith Tidball director of NY EDEN (Extension Disaster Education Network). An interactive, live updating, publicly available flood vulnerability model in the cloud would allow citizens to zoom into their local area and see risk areas for themselves. Furthermore, integration with other early warning text messaging systems, such as Ushahidi and Frontline SMS, could be spatially targeted to citizens living in predicted risk areas days in advance. For communities in flood prone watersheds, such as the Oswego, Mohawk, and Upper Hudson, a live flood vulnerability model could be a valuable tool in an extreme event. Cloud computing could allow for rapidly updating predictions in multiple watersheds. Furthermore, one built, the flood modeling architecture could be used as a scenario-modeling tool for future prediction of strong storms under climate change scenarios.

In addition to opportunities presented by the NY RISE initiative, developing a cloud computed flood vulnerability model for testing is ideal in a place like New York. New York (like most locations in the USA) has access to data from stream gages and historical flood maps that will aid in calibrating and testing the vulnerability model. Recent storms (i.e. Ike, Irene, and Lee), and even flooding last July, provide recent and reliable data for calibration and testing. Testing the viability of Earth Engine flood vulnerability in an accessible location like New York State will provide insight into the potential for global model expansion that might encompass data poor countries.

II. Biophysical Indicators of Vulnerability:

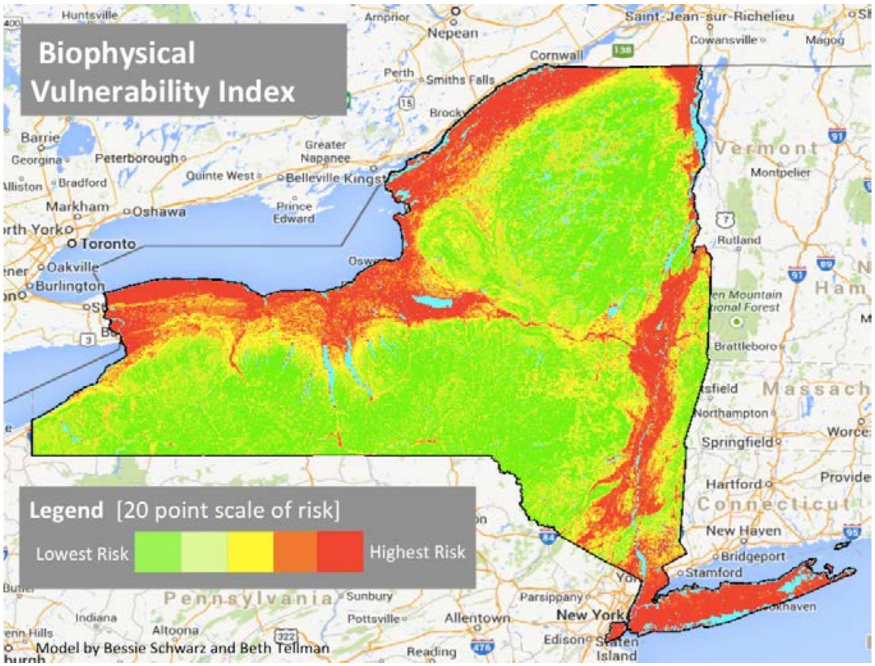
The biophysical risk prediction area refines an already existing weather prediction of general "flood risk." This is not a physically process based watershed hydrologic and hydraulic model that "flows" quantitative precipitation estimates over a surface in real time. Examples of complex flood mapping models with 2 and 3-d differential and numerical equations include LISFLOOD, MIKE-SHE, and TOPKAPI that involve flow routing routines that are not possible in the parallelized computing environment in Google Earth Engine. The parallelization process sends individual pixels (or potentially groups a region of pixels) in a grid to different parallel servers to process each step of an algorithm, and then puts them back together again. For hydrologic and hydraulic modeling that requires complex routing algorithms where pixels must communicate in one another in sequential fashion in multiple dimensions (i.e. the D8 flow accumulation algorithm used to calculate contributing area used to calculate topographic index in TOPMODEL), parallel computing is problematic (Qin and Zhan 2012). Our model simplifies the assumptions made in these more complex models and computer factors with simple algebra on a pixel by pixel basis. This model takes advantage of Google's cloud computing and is thus tailored towards a rapid assessment evaluation in an oncoming storm; it is not meant to replace detailed 3-D hydrologic modeling. The four biophysical parameters are explained below.

1. *Elevation:* uses 10m pixel resolution DEM from the National Elevation Dataset for analysis in US. Elevation is rescaled from 0-5, with low elevations (below 125 meters above sea level receiving a score of 5, elevations between 140 and 125 a score of 4, etc.).

2. *Slope:* calculated in degrees using the elevation layer. Slope is rescaled from 0-5, with slopes of .00001 to 0 gaining a score of 5, slopes of .5 to .00001 a score of 4, etc.

3. *Impervious Surface*: calculated from vegetation data using an NDVI index with a 2012 global composite of Landsat 7. A 2012 composite of Landsat 7, representing the average pixel radiance for the year, was pulled into the earth engine API. We selected the near infrared and infrared banks to calculate NDVI, the normalized vegetation index. This equation is infrared-red bands/infrared+red bands. The NDVI output is a relative index of vegetative cover from -1 to 1. Thus, less vegetation represents lower infiltration rates, and higher risk areas in a flood. Low vegetation areas, such as cemented urban surfaces, have a low NDVI. The NDVI was rescaled from 0-5, such that NDVI less than zero receives a score of 5, and .2 to 0 a score of 4, and so on.

4. *Topographic Index:* describes the spatial distribution of the soil moisture and related landscape processes (equation in Moore et al 1991). This final indicator is a function of the contributing area of each pixel in its watershed and slope gradient. As catchment area increases and slope gradient decreases, topographic index and soil moisture content increase. Topographic index is used in flood models such as TOPMODEL (variable contributing area conceptual flood mode- see Beven et al 1984l), in which the major factors affecting runoff generation are the catchment topography, and the soil transmissivity that diminishes with depth. Topographic index controls flow accumulation, soil moisture, distribution of saturation zones, depth of water table, evapotranspiration, thickness of soil horizons, organic matter, pH, silt and sand content, and plant cover distribution.



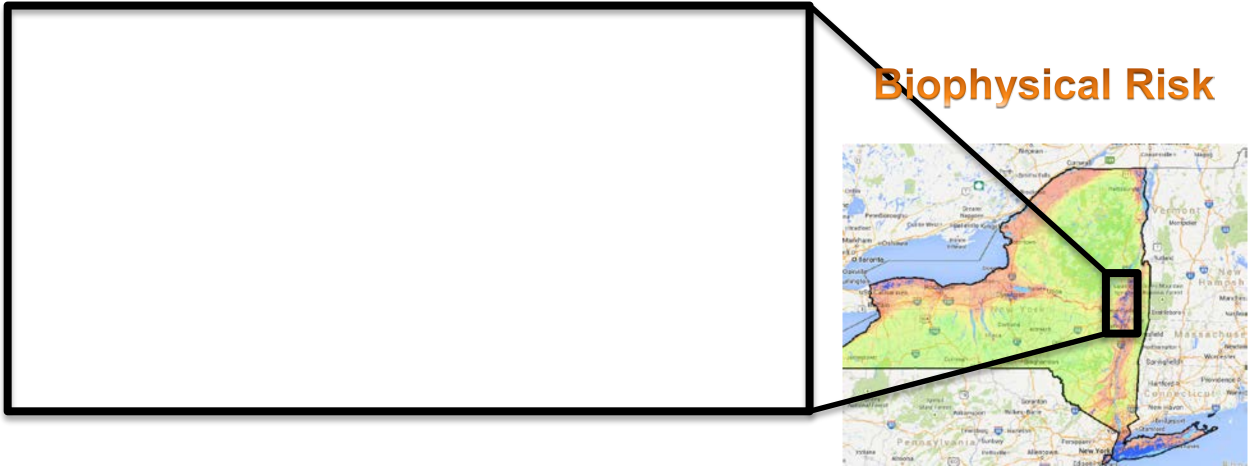
The equation we use comes from Moore et al (1991), and is:

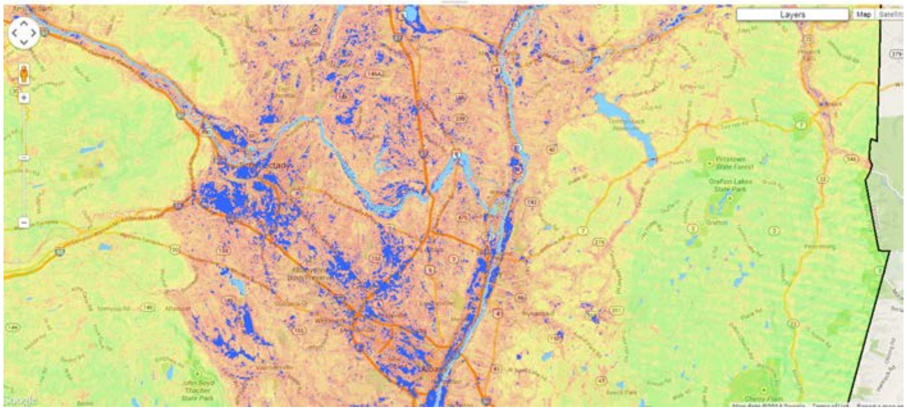
Topographic Index = ln [[(flow accumulation+1)*cell area]/tan(slope)]]

Flow accumulation raster data comes from a 30 arc second resolution raster by USGS, now made publicly available in the Google Earth Engine library per our request. Flow accumulation is a function of how many cells flow into each pixel, and represents the “upslope” contributing area for each point. Flow accumulation, plus one, is multiplied by the cell area to get the contributing catchment area for each cell. This is divided by the slope gradient at each point. Slope is calculated from the USGS 10m DEM National Elevation Dataset. .000001 is added to each slope, so that the natural log of zero slope data points (like many rivers and floodplains) are not removed from the topographic index. All other mathematical functions are available in Google Earth Engine.

All four biophysical variables are combined (added) to yield a composite score of 1-20. The risk surface displayed here shows increasing redness as the risk score increases. Note that existing standing water (lakes, ponds, and rivers, shown in bright blue) have been masked out of the biophysical risk index based on the “open waters” categories of the National Land Cover data set. Open water thus receives a default risk score of “0”

The predicted flood area includes all pixels in the 95th percentile of the biophysical risk surface. This results in variable thresholds for what “level” of biophysical risk denotes a truly “flooded” area, shown in medium blue, here for the Albany/Schenectady area.





III. Social Indicators of Vulnerability:

The model predicts social vulnerability to disaster by adapting techniques and scholarship by Dr. Susan Cutter at the Hazard and Vulnerability Institute at the University of South Carolina (Cutter et al 2003, Cutter et al 2008). The model scaled five of Cutter's indicators to NY social data using US Census tract level data from 2010. Each indicator is converted into a five-point scale of risk based on its absolute numbers. Thus the combined social risk index is on a twenty five point scale.

1. *(Age) Children:* calculated based on tracts with a large number of children under 5. Areas with a large number of young children are more vulnerable because they require assistance during evacuation, potentially endangering themselves and their caretakers.

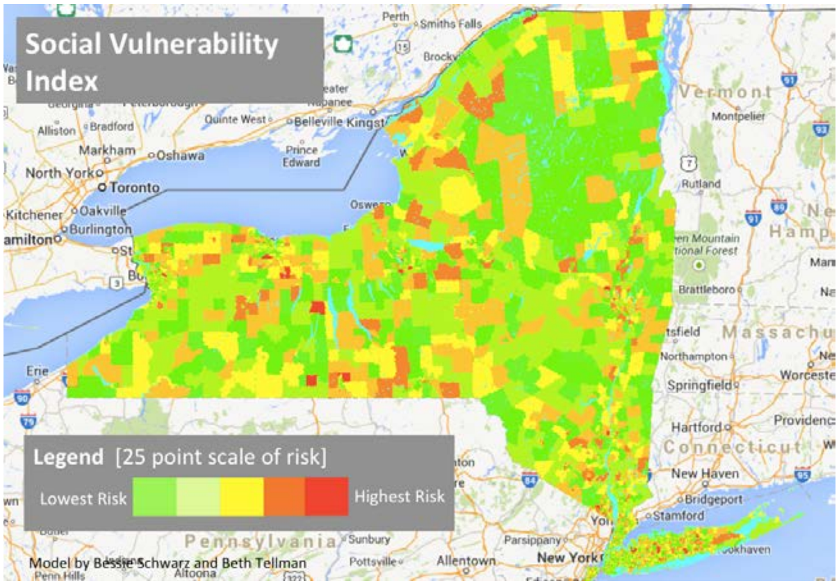
2. *(Age) Elderly:* calculated based on tracts with a large number of people over 85

3. *Community Cohesion*: calculated based on percentage of change in the population between 2000 and 2010 as a proxy for communities with increasing or decreasing populations. The higher the change the less "cohesive" the community is deemed to be.

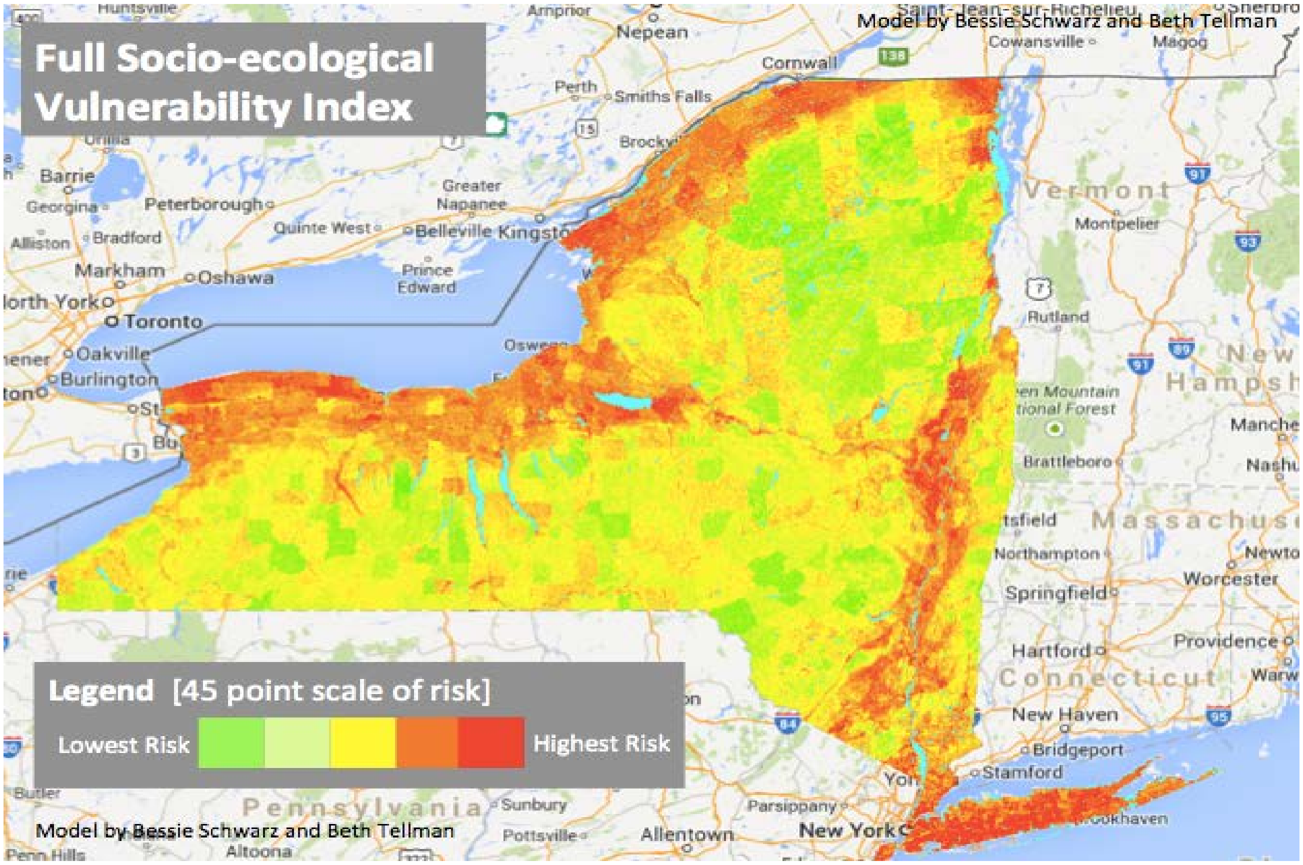
4. *Density:* calculated based on tracts with a high population per square Kilometer

5. *Poverty:* calculated based on tracts with a large number of individuals below the poverty line.

The social risk index in the model is run solely on the predicted flood zone by applying the 25 level social risk surfaces to the flood zone. The threshold for the high risk zone is dynamically determined based on the overall risk for the area of interested input into the model, examining the 95th percentile of risk for the area of interest.


Social Vulnerability Index
Legend [25 point scale of risk]
Lowest Risk
Highest Risk
Model by Bessie Schwarz and Beth Tellman
Toronto
Oshawa
Oswego
Vermont
Massachu
Connecticut
Pennsylvania
New York

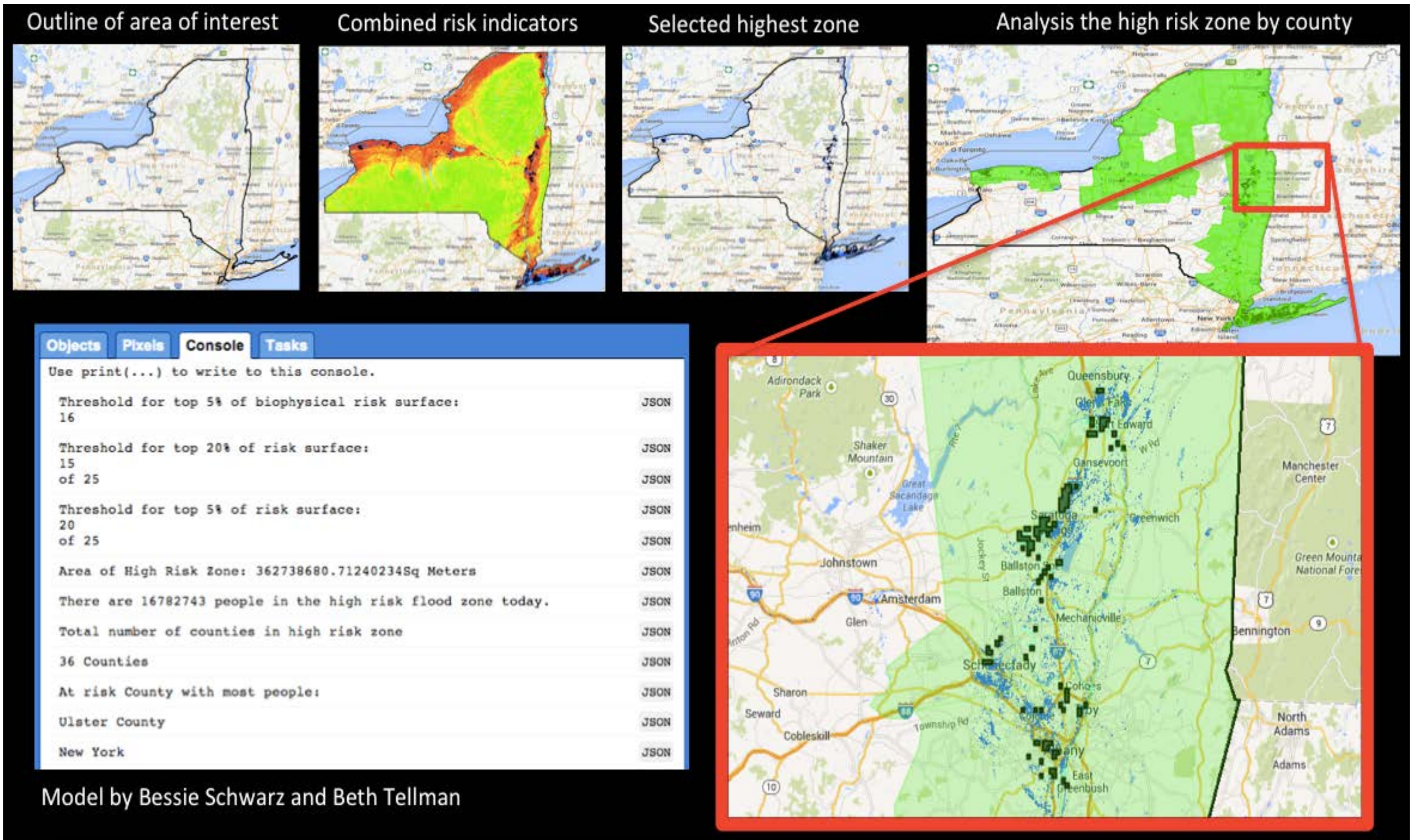
Therefore, the model naturally favors the physical indicators of risk, predicting results only within the flood. The desired percentile of risk for the index can be selected (the sample images provided here use a variety of thresholds), and the highest risk zone will then be used for further analysis.



IV. Outputs:

This link leads to an interactive web map with several of the risk indexes and the watersheds of NY (full link: https://mapsengine.google.com/08039105425737821391-10002667888828033184-4/mapview/?authuser=1)

After determining physical and social risk indexes and identifying the highest risk zones within the area of interest, the model analyzes areas most at risk to determine key attributes including: physical size, number of residents within, and the county with most number of people at high risk.

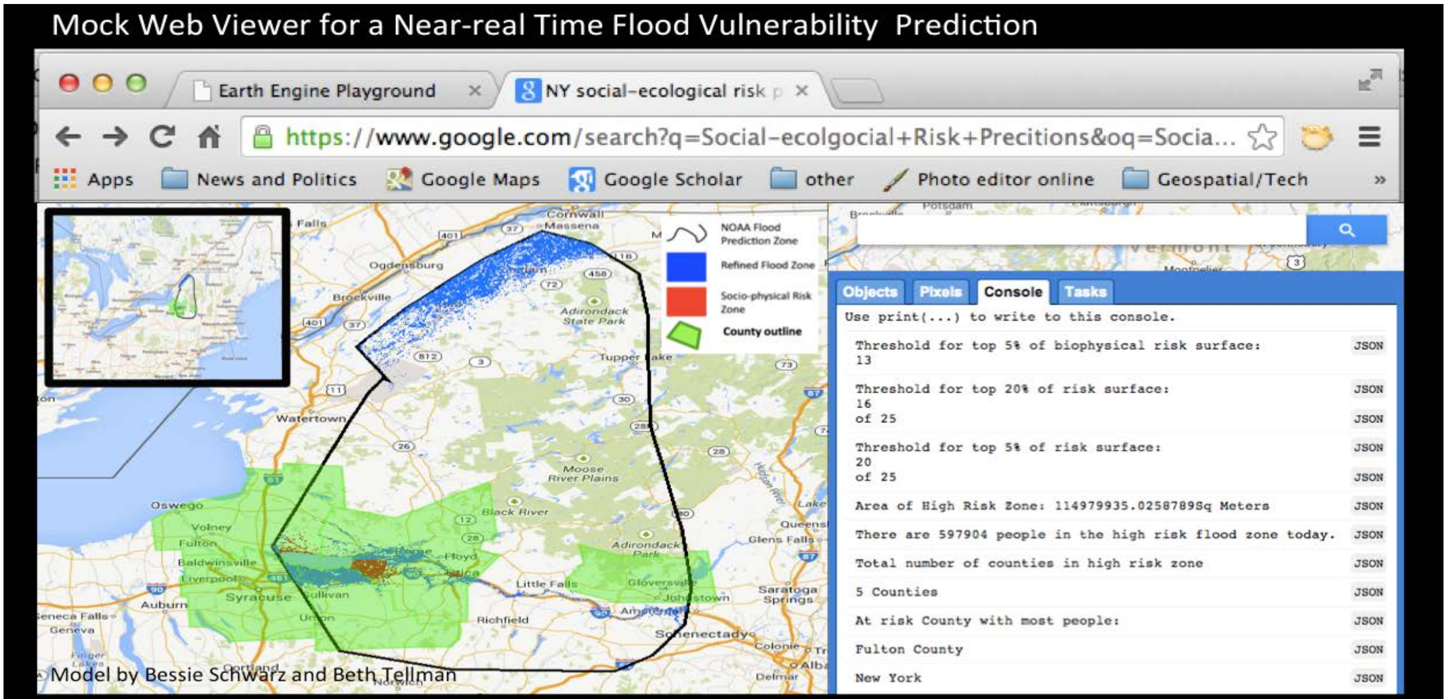


```
Use print(...) to write to this console.
Threshold for top 5% of biophysical risk surface:
16
Threshold for top 20% of risk surface:
15
of 25
Threshold for top 5% of risk surface:
20
of 25
Area of High Risk Zone: 362738680.71240234Sq Meters
There are 16782743 people in the high risk flood zone today.
Total number of counties in high risk zone
36 Counties
At risk County with most people:
Ulster County
New York
```

Model by Bessie Schwarz and Beth Tellman

Full process flow of model. The model takes an area of interest (predicted storm zone, watershed, county, city etc.), determines its highest risk area by adapted the combined risk index, and output the demographic and other data about risk for the area. This process takes 1-30 seconds and can be automated.

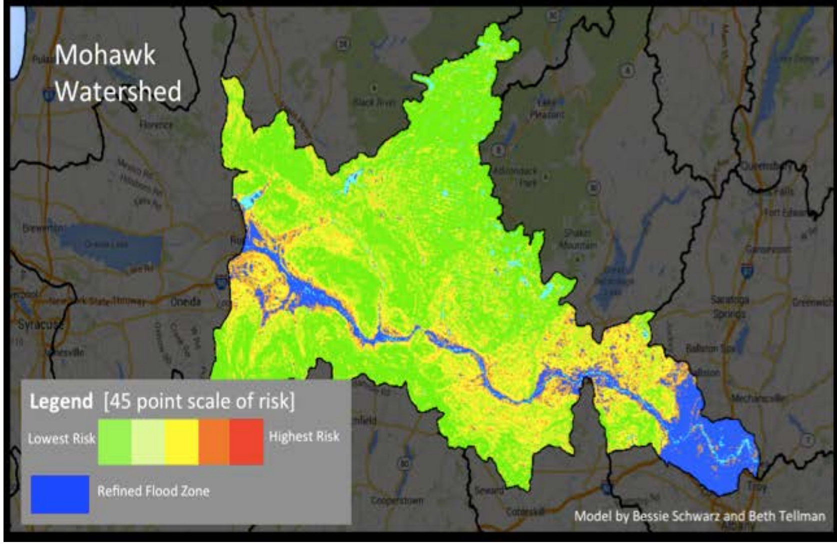
Mock Web Viewer for a Near-real Time Flood Vulnerability Prediction



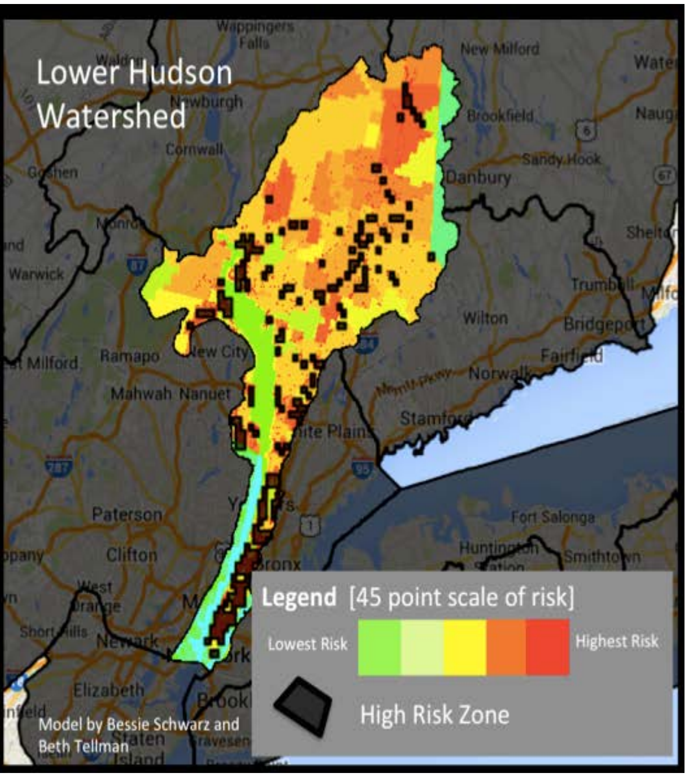
```
Use print(...) to write to this console.
  Threshold for top 5% of biophysical risk surface:        JSON
  13
  Threshold for top 20% of risk surface:                   JSON
  16
  of 25                                                    JSON
  Threshold for top 5% of risk surface:                    JSON
  20
  of 25                                                    JSON
  Area of High Risk Zone: 114979935.0258789Sq Meters       JSON
  There are 597904 people in the high risk flood zone today. JSON
  Total number of counties in high risk zone               JSON
  5 Counties                                               JSON
  At risk County with most people:                         JSON
  Fulton County                                            JSON
  New York                                                 JSON
```

Since the model is cloud based in Google Earth Engine a web application can could be build to run the analysis on a the most up to date storm predictions (as available for download on the web *e.g.* NOAA 5 day flood prediction). This mock up shows a example storm for upstate NY.

Prediction by Watershed: the model can be run on a variety of areas of interest, generating customized and actionable information for a given decision-maker. These images demonstrate risk predictions for different NY watersheds.



Model by Bessie Schwarz and Beth Tellman



Model by Bessie Schwarz and Beth Tellman

V. Future Directions:

There are many exciting directions for this project based on initial research. These include a more advanced regression based flooding algorithm, partnering with Azavea to create web based communication vulnerability tools, higher resolution social data, and additional analysis on proximity to hospitals and safe evacuation routes out of the flood zone.

More Advanced Flood Algorithm

The current biophysical risk algorithm is a simplification of major physical processes that play a role in flooding. However, this could be further calibrated and validated with historic flood data to generate a locally specific algorithm that takes into account not only general watershed topographic characteristics, but also flood control management structures such as levees and dams, and their ability to mitigate flood risk in different storms for each watershed. Not only would this capture the reality of flooding better, but could result in a more dynamic algorithm that inputs quantitative precipitation data predictions (in a raster data format such as NEXRAD). This approach would require a multiple logistic regression.

Partner with Azavea

We have begun initial talks with Azavea, a web-GIS company based out of Philadelphia. Azavea has pioneered geo cloud computing for hydrology in its GeoTrellis platform. Azavea has tested initial research on live watershed modeling in Philadelphia, Delaware, and Southeastern Pennsylvania funded by the National Science Foundation (NSF). Azavea's efforts in this Wiki-Watershed project enable live user interaction with models by drawing areas of land use change and analyzing effects on runoff. In initial conversations with developers and the Azavea CEO, this tech company is very interested in collaborating with use to help develop an exciting user interface, communication tool, and advanced real time computing.

Refine Social Data

To refine the social risk index we could increase the resolution from the already fine US Census tract level to the block level and include more social indicators like race and percentage rental homes.

Tailor output

There are many more aspects of the high risk zone that the model can analyze including evacuability of the high risk areas, more refined number of residents in the high risk zone, number of emergency relief stations in or near the high risk zone.

VI. Disclaimer:

This script for and right to this model are property of Bessie Schwarz and Beth Tellman. Any internal distribution of the document's content or images must acknowledge them, and their permission is necessary for external distribution.

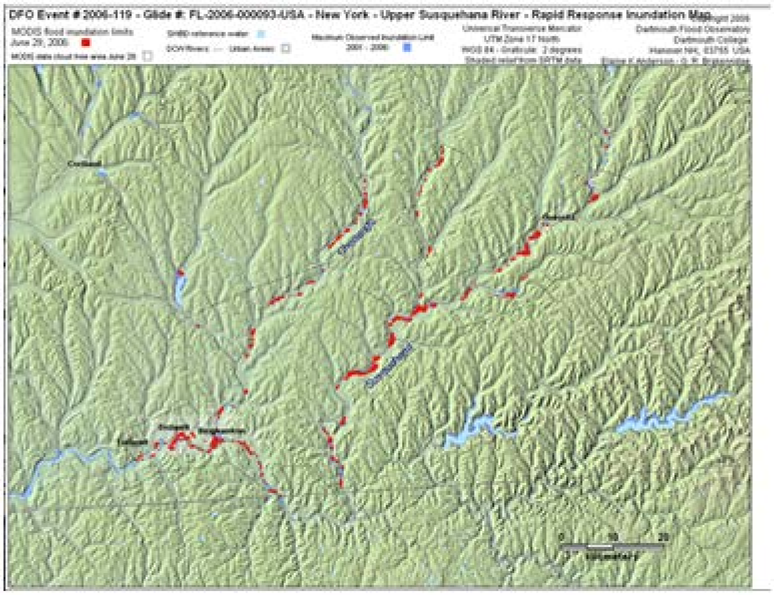
Acknowledgements:
Our institutional sponsor is The Yale Project for Climate Change Communication at the Yale School of Forestry and Environmental Studies. Funding for this project comes from a sub-award from Cornell University's College of Agriculture and Life Sciences, Department of Natural Resources, a member and partner in the New York Resiliency Institute for Storms and Emergencies (NYS RISE)

Appendix:

More on advanced flood algorithms

This approach would require a multiple logistic regression that uses historic flood data to predict the contribution of flood risk factors on a pixel by pixel basis (the area of impervious surface upstream, the number of dams upstream, local topographic index, soil type, watershed size, etc.) and the amount of precipitation necessary to "flood" each pixel. This type of data driven approach (see Pradhan 2010 for a review and example) builds a more complex model that respects local flood characteristics, with an algorithm that avoids computationally intensive flow routing.

The multiple logistic regression will use the Dartmouth Flood Observatory (DFO) database, which has spatially explicit rasters of flood extent. These are commonly used in flood vulnerability modeling as flood observations for validation (Bates and Roo 2000). A quick search of the DFO database shows at least 4 storm event observations for upstate New York from 2000-2008. The image below is an example of flood extent for June 29th 2006 in the upper Susquehanna River Basin. In addition to this data, the NY RISE Initiative may have access to more localized high-resolution data for recent events (such as Hurricane Irene in 2011, or flooding Summer 2013). Some events will be reserved for manual model calibration (i.e. decisions to remove the impervious surface parameter in the index, add in other considerations such as location of levees, bridges, reservoirs, and other hydraulic structures), while others will be reserved to test the model and quantify uncertainty. Part of the reason for selecting DFO data for model validation is that the entire database will be included in Earth Engine by May 2014. Thinking ahead to global application of the model and model validation, devising a method to use DFO data is appropriate.


DFO Event # 2006-119 - Glide #: FL-2006-000093-USA - New York - Upper Susquehana River - Rapid Response Inundation Map
MODIS flood inundation limits June 29, 2006
MODIS data cloud free area June 29
Universal Transverse Mercator
UTM Zone 17 North
WGS 84 - Graticule: 2 degrees
Shaded relief from SRTM data
Dartmouth Flood Observatory
Dartmouth College
Hanover NH, 03755 USA
0
10
20
Kilometers

Works Cited:

Cutter, Barnes, Berry, Burton, Tate, and Webb. 2008 A place-based model for understanding community resilience to natural disasters. *Global Environmental Change.* 18:598-606.

Cutter, Boruff, and Shirley.2003. Social Vulnerability to Environmental Hazards. *Social Science Quarterly.* 84:2.

Bates, P.. & De Roo, a. P.. (2000) A simple raster-based model for flood inundation simulation. *Journal of Hydrology*, **236**, 54–77.

Hapuarachchi, H. a. P., Wang, Q.J. & Pagano, T.C. (2011) A review of advances in flash flood forecasting. *Hydrological Processes*, **25**, 2771–2784.

Pradhan, Biswajeet. 2009. Flood susceptible mapping and risk area delineation using logistic regression, GIS, and remote sensing. *Journal of Spatial Hydrology.* 9:2 p 1-18.

Bates and Roo. 2000. A simple raster-based model for flood inundation. *Journal of Hydrology.* 236:54-77.

Moore, I.D., Grayson, R.B. and Ladson, A.R. (1991). Digital Terrain Modeling: A Review of Hydrological, Geomorphological, and Biological Applications. Hydrological Processes, 5:3-30.

Beven, Kirkby, Schofield, and Tagg. 1984. Testing a physically-based flood forecasting model (TOPMODEL) for three U.K. catchments. *Journal of Hydrology.* 69:10. Pp119-143