Project Matsu:
Flood Prediction using Topological and Climate Data Cross Correlated with Water Identification from Satellite Imagery over MapReduce

Open Cloud Consortium
Open Data Group

February 23, 2013
Version 2.0

Materials for Matsu

Matsu is a public project. Currently, code and documentation are maintained on Github. The Open Cloud Consortium has an account on Github where projects may be released from time-to-time. The official Open Cloud Consortium account is `github.com/opencloudconsortium`. The Matsu project page is `github.com/opencloudconsortium/matsu-project`.

Information about the computing resources available from the OCC can be found at `https://www.opensciencedatacloud.org/`.

Overview

A set of analytics for correlating satellite imagery, topological data, and climate data, have been developed to yield a simple model for flood prediction. Despite simplifications in the science of the physical model, the approach yields intuitive results and demonstrates that the utility of a common approach that leverages the computing environment and its particular appropriateness to rasterized data, independent of the specific nature of the data sets themselves.

Contents

1 Introduction **5**

2 Modeling surface topology and derived topological features. **5**

3 Characterizing Terrain by Flood Basins and Modeling Water Coverage **10**

4 Correlating Water Volumes with Historical Rainfall Data **11**

5 About Project Matsu **17**

1 Introduction

One class of analytics relevant to cloud-based infrastructures and map-reduce type processing capabilities consists of those appropriate to multiple rasterized data samples. Flood prediction applied to rasterized terrains associated with topological features, climate data, and other geo-located data is an example application of such analytics. In the following, a prototype sequence of model development for this example is reviewed.

The overarching goal of the model is to relate rainfall to accumulation of water at the level of geo-located pixels presented as geoTIFF images. To achieve this, the following steps are undertaken:

1. Modeling surface topology at the pixel level by calculating derived features based on 'raw' elevation data. The collection of raw and derived features constitutes a topological dataset.

2. From the topological data, calculate large scale structures, *basins*, wherein water will tend to accumulate.

3. By comparison with satellite imagery with identifiable standing water, relate water within basins to specific dates.

4. Correlate recent rainfall within basins to dated water content within basins.

As discussed in section 4, a simple analytic representing the correlation between rainfall and water accumulation within a basin can be built, for example, as a linear regression between the two. In a predictive sense, known or anticipated rainfall can be used to calculate water content for a given basin and rendered as an image which shows individual pixels and the associated water depth (if any) at their location.

2 Modeling surface topology and derived topological features.

Elevation data is available in both vectorized and rasterized forms and at multiple resolutions from many sources. For this exercise, topological data was obtained as a set of $1^\circ \times 1^\circ$ granules derived from satellite-born infra-red spectrometry. The granules are presented as geoTIFF files with pixel sizes of approximately 30 m $\times$ 30 m (3600 x 3600 pixels with a small latitude-dependent variation in linear dimensions). The value associated with each

pixel indicates the average height of the terrain at that location. Figure 1 shows an example geoTIFF for the elevation levels of a single granule.

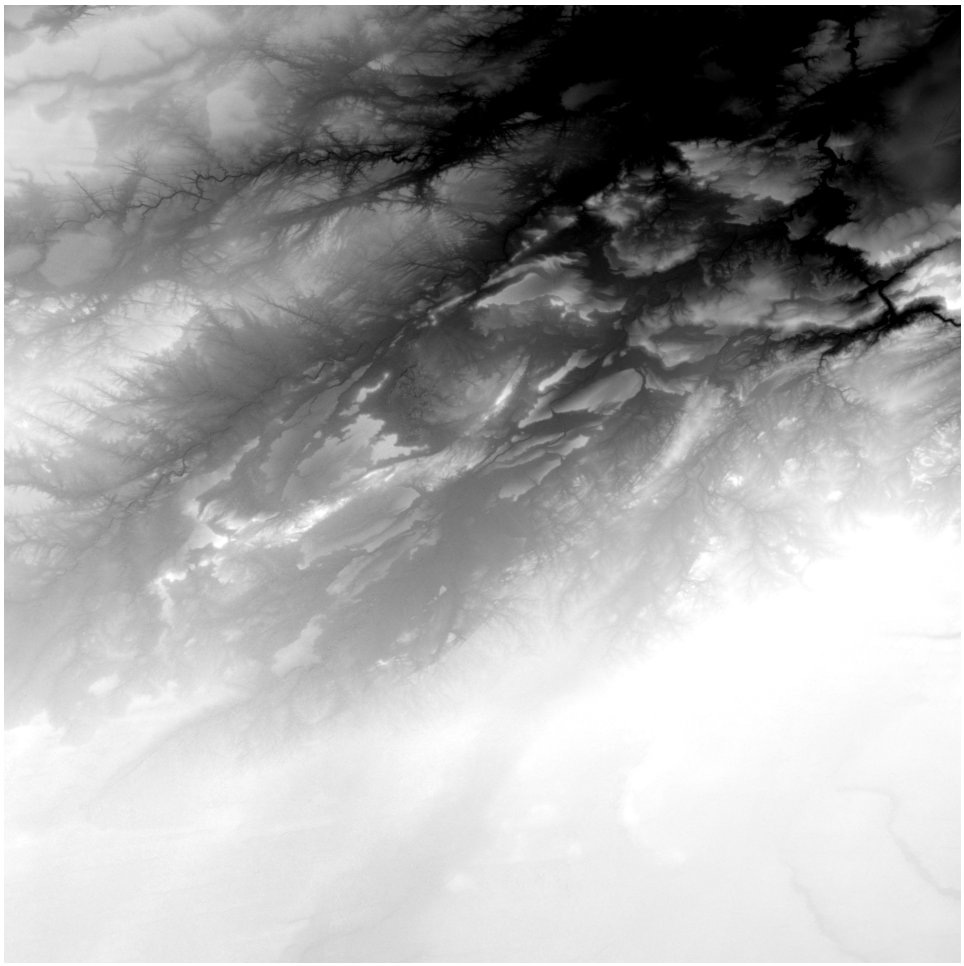


Figure 1: Raw elevation data for 1° by 1° granule. White indicates a higher elevation.

A topological data sample is created by augmenting the elevation data for each granule of pixels with a set of derived features. The derived features describe quantities such as gradients and may be rendered as separate geoTIFF images for each. Calculation of these features may be accomplished as part of a map-reduce job and the results can be persisted as either tables in an Accumulo database or as files in a cloud-based system such as HDFS. Each feature is also available to visually inspect through a WMS interface.

Persistence in Accumulo and in HDFS were both implemented and tested for this exercise, using the set of granules covering southern Africa below 1° South.

The derived features included gradient components and magnitude, curvature components and magnitude and directions of maximal gradient and curvatures. Calculation of these quantities proceeds in the usual way for estimations with finite differences. The gradient in the longitudinal and latitudinal directions, for example, are calculated, respectively, as follows:

$$g_{ij}^{+\phi} = \frac{h_{ij} - h_{ij+1}}{d\phi} \tag{1}$$

$$g_{ij}^{+\lambda} = \frac{h_{ij} - h_{i+1j}}{d\lambda}, \tag{2}$$

where h_{ij} is the elevation for the pixel at row i and column j of the geoTIFF, ϕ is the longitude, λ is the latitude, and $d\phi$ and $d\lambda$ the pixel size in degrees longitude and latitude.

Figures 2, and 3 show examples of topological features rendered as geoTIFFs.

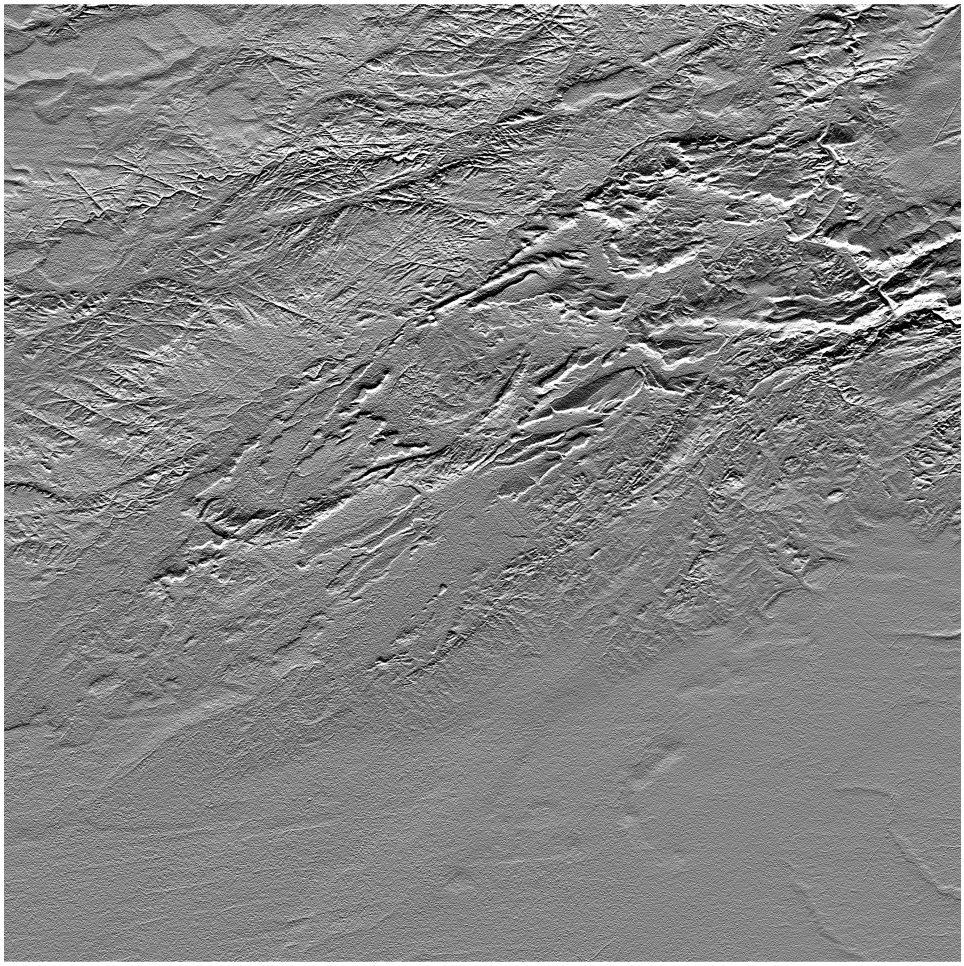


Figure 2: Gradient in the Latitudinal direction. As one moves from higher to lower latitudes, a bright color indicates a rising slope and darker color indicates a falling slope. Ridges and valleys are thus paired rising and falling slopes. The gradient is calculated using a simple finite-difference calculation.

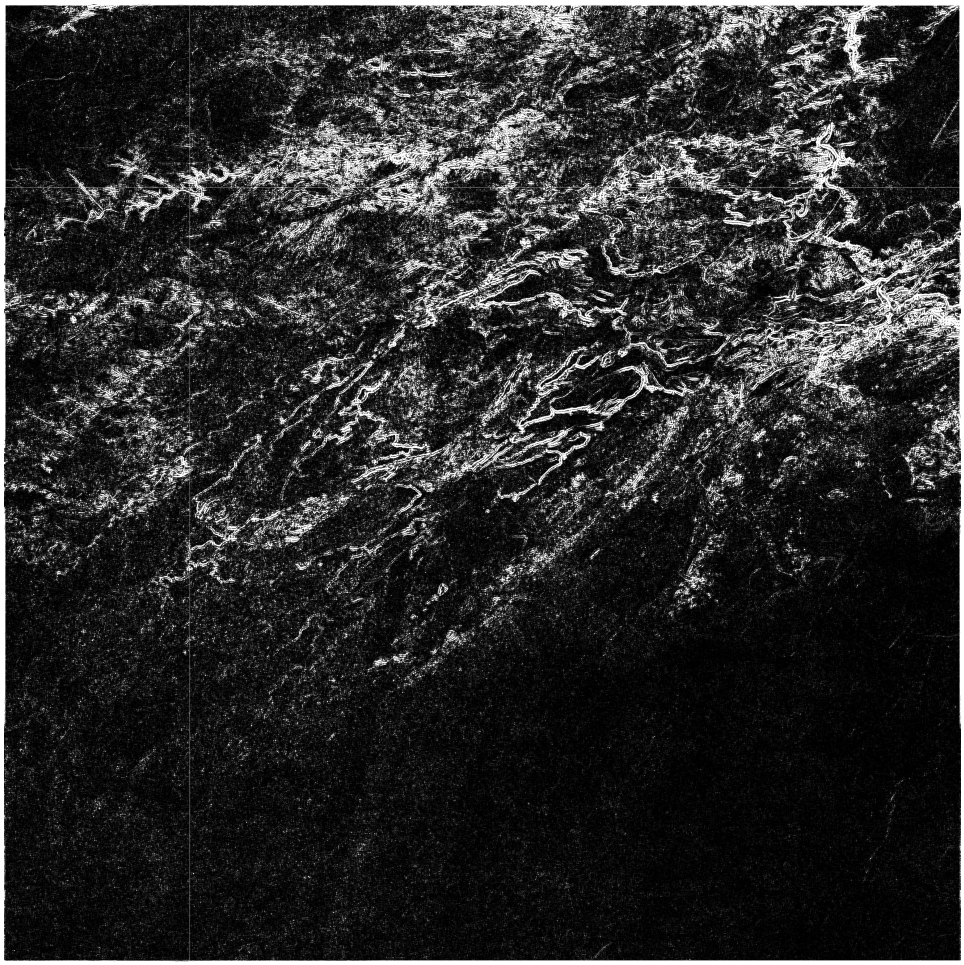


Figure 3: Magnitude of elevation curvature over a 1° by 1° ASTER granule. Curvature calculated as the magnitude of the finite-difference calculation of the second derivative in the latitudinal and longitudinal directions.

3 Characterizing Terrain by Flood Basins and Modeling Water Coverage

A flood basin is a contiguous set of pixels for which each pixel has at least some path by which to reach the lowest lying set of pixels within the bin. The lowest lying set of pixels are required to be contiguous with each other for each flood basin. A simple algorithm that identifies such basins was developed for this effort and relies on the gradient data calculated earlier to characterize the surface terrain. The algorithm is as follows:

1. Find the first row which includes a pixel with elevation equal to the global minimum for the image.

2. Tag all pixels in the first row with a global minimum up to the first column on either side for which the gradient begins to decrease (possibly indicative of a new local minimum). This step is referred to as 'stretching'.

3. Tag all pixels in the next row which are either level with or uphill from, and are subtended by the tagged pixels in the first row. Stretch each of these pixels along the row. These stretches are also terminated if a previously tagged pixel is reached.

4. When a row is reached for which no pixels in the next row are tagged or when the last row is reached, repeat the process moving back *up*.

5. Repeat the iterations up to a maximum number of iterations.

6. For each set of iterations, the elevation of tagged pixels is artificially changed to be 1 m larger than the highest elevation in the image prior to the iterations. These pixels constitute a *basin*.

The procedure is repeated a fixed number of times. Many of the iterations will terminate quickly with a very small number of pixels tagged. The majority include 1 to 3 basins of sufficient size to capture some 80% to 90% of the area of a single granule.

Each basin can be independently associated with a function that correlates a volume of water with water coverage per pixel. Let n_i represent the distribution of the height of pixels above the local minimum of a basin. For example, n_1 indicates the number of pixels within a basin that have an elevation of 1 meter above the bottom of the basin. If h is the height of the

water above the basin floor, then the volume of water within the basin, V, is given by finding the value k which maximizes the following:

$$\frac{V}{A_{pixel}} = h \sum_{0}^{k} n_i - 1 \times \sum_{0}^{k} i n_i. \tag{3}$$

Here, A_{pixel} is the area of a single pixel, approximately 900 m^2. The height of water within a basin at each pixel is then given by $h - h_{ij}$, where h_{ij} is the height of the pixel at coordinates i, j relative to the basin floor. Pixels for which this quantity is positive are under water. For any basin, one can use this calculation to tabulate a series of hypothesized values of h and the associated volume of water, V. A pair of functions, $\nu(V)$ and $\kappa(h)$, are then interpolated from the table of values to simulate water coverage as a function of water volume and vice-versa.

Figure 4 shows simulated introduction of progressively more water into an identified basin where one can see the increasing flood coverage as more water is introduced.

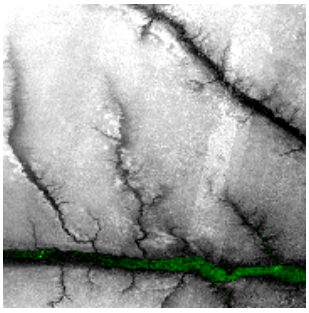
4 Correlating Water Volumes with Historical Rainfall Data

An existing analytic which identifies water coverage was used to visually validate identified basins and to correlate water coverage (and therefore water volume) to specific dates. After applying the water identification analytic to a satellite image, the extent of water coverage was compared to a sequence of images showing the simulated water coverage for different inputs of water volumes. The water coverage which most closely resembled the analyzed satellite image was used to infer a volume of water associated with the date of the satellite image. Figure 5 shows an example of such a comparison.

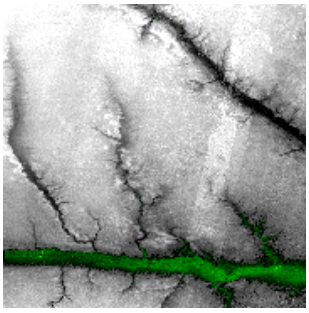
The USGS also makes available climate data such as rainfall estimates. This data is available in a rasterized format which may be rendered as geoTIFF images (see figure 6. Calculation of rainfall accumulation within basins is a straightforward computation between the two geoTIFFs as they are both simple arrays differing solely in their dimensions. Table 1 summarizes the results of analyzing each available image in the data sample considered.

Figure 7 depicts the relationship between recent rainfall and basin water volumes. The two variables show a positive correlation ($r = 0.43$). The correlation can be exploited to devise an analytic relating observed, expected,

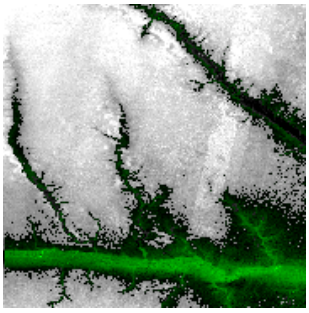
or hypothesized rainfall and potential or predicted flood levels at any area of interest. As an example, linear fits between prior ten-day rainfall and basin water volume were calculated and are superimposed on Figure 7. One of the two fits includes a constraint of zero water volume when there was no prior rainfall.



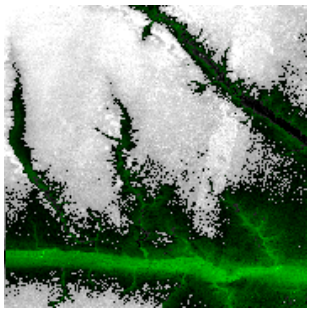
(a) Flood Sequence 1



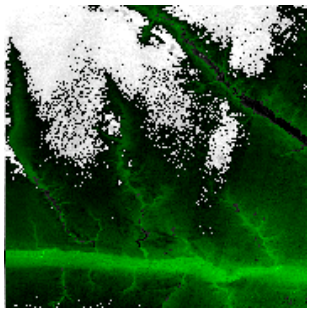
(b) Flood Sequence 2



(c) Flood Sequence 3



(d) Flood Sequence 4



(e) Flood Sequence 5

Figure 4: Simulated flooding sequence for a single basin. In each figure, green indicates water coverage with the brightness proportional to the water depth.

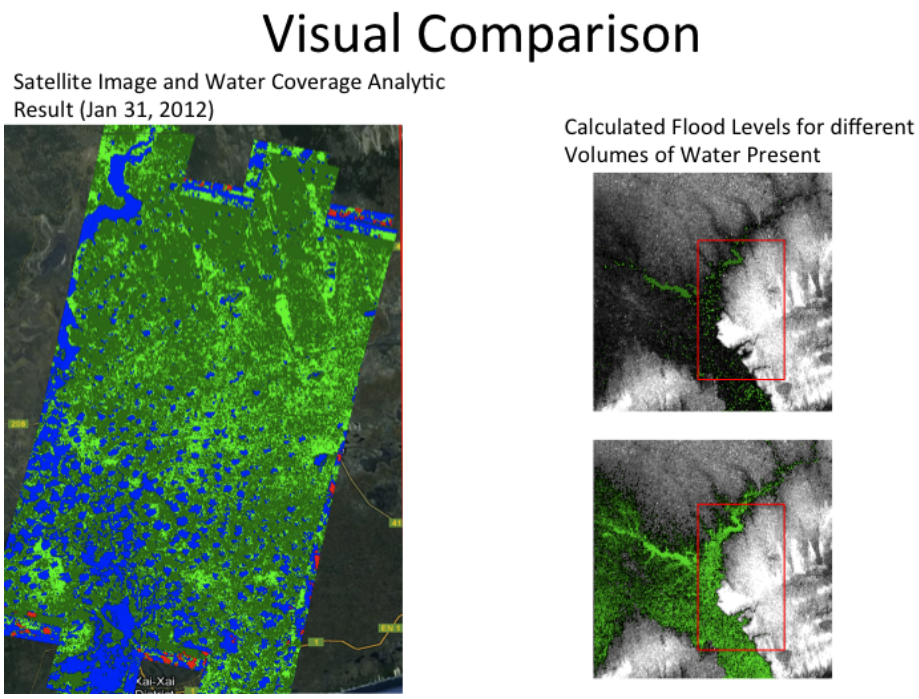


Figure 5: Satellite image (left) with results of analysis for water coverage superimposed; blue regions are consistent with water. Simulated flooding of basin in the region (right) for two differing volumes of water introduced. By comparing the simulated level of water with the water coverage deduced from the satellite image, the volume of water in a flood basin at a given date may be estimated.

ASTER granule	Basin	Image Date (Date)	Volume ($\times 10^3$ m^3)	RFE (mm)	Dekad	Location	Notes
11e017	1451	5-11-2012	586	0	12051	Angola	
11e017	1452	5-11-2012	0	5.3	12051	Angola	
11e017	1451	5-19-2012	1075	18.4	12051	Angola	
11e017	1451	5-19-2012	17639	577.7	12051	Angola	
11e017	1451	6-14-2012	14111	178	12051	Angola	
11e017	1451	7-18-2012	8550	245.8	12051	Angola	
11e017	1451	8-21-2012	0	0	12051	Angola	
16e023	1101	3-06-2012	14800	1400	12023,12031	Zambia	
16e023	1101	3-27-2012	15381	651	12032	Zambia	
16e023	1101	4-01-2012	16565	4690.1	12033	Zambia	
16e023	1101	4-22-2012	8252	3.4	12042	Zambia	clouds
16e023	1101	5-31-2012	9465	1.0	12053	Zambia	clouds
16e023	1101	7-12-2012	0	0	12071	Zambia	clouds
18e019	1201	2-16-2012	5100	10000	12021,12022	Namibia,Angola	
18e016	1151	2-20-2012	27700	9946	12022	Namibia,Angola	
18e016	1151	3-07-2012	19033	5382	12031	Namibia,Angola	
18e021	1103	1-09-2012	750	32.2	12011	Caprivi	
19e021	1322	1-09-2012	394	105	12011	Caprivi	
19e021	1301	1-09-2012	1933	3101	12011	Caprivi	
18e021	1103	1-19-2012	6087	3430	12012	Caprivi	
19e021	1301	1-19-2012	13537	16427	12012	Caprivi	
18e021	1103	2-01-2012	0	2038	12013	Caprivi	
19e021	1303	2-01-2012	6700	8198	12013	Caprivi	
26e028	1651	2-24-2012	22702	4900	12022,12023	South Africa	
26e027	1851	2-24-2012	19147	5100	12022,12023	South Africa	
27e027	1801	2-24-2012	11000	6359	12022,1202	South Africa	

Table 1: Evaluation of simulated flood basin flooding with estimated water coverage and estimated rainfall. Rainfall is estimated for the basin indicated based on a ten-day period prior to the date for which a satellite image was obtained and analyzed for water coverage. Ten-day time periods, *dekads*, are fixed so that the first dekad, 120101, begins with January 1, 2012. In some cases where the satellite image is taken in the middle of a dekad, the average of the dekad containing the acquisition date and the prior dekad are used.

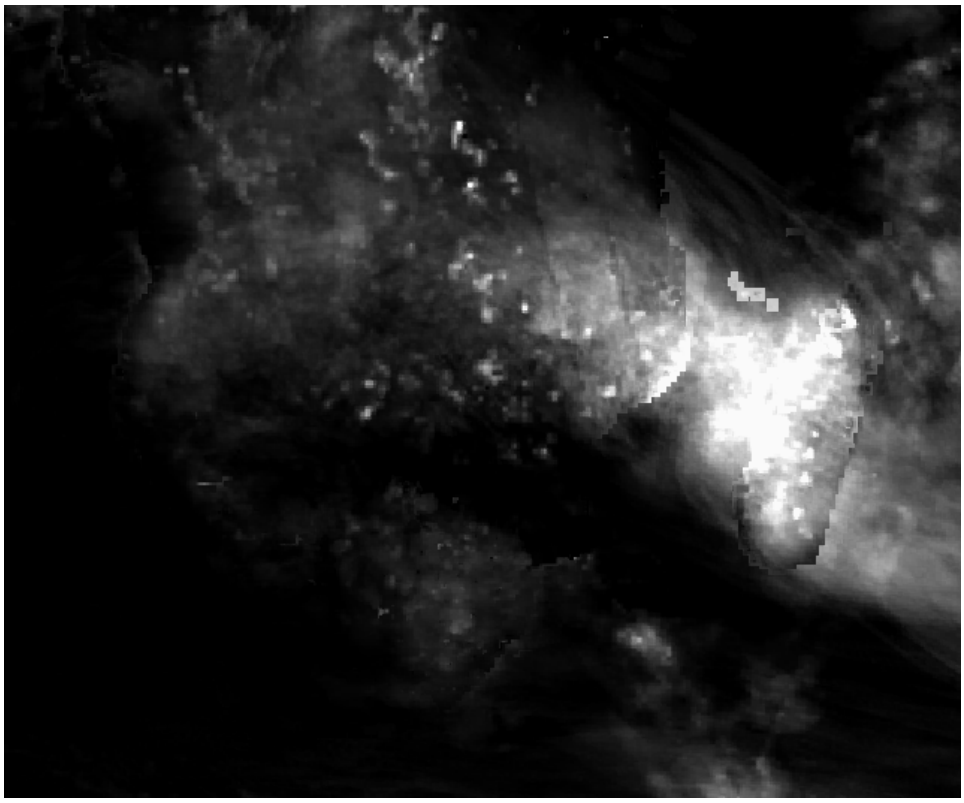


Figure 6: Ten-day estimated rainfall for souther Africa, obtained from USGS.

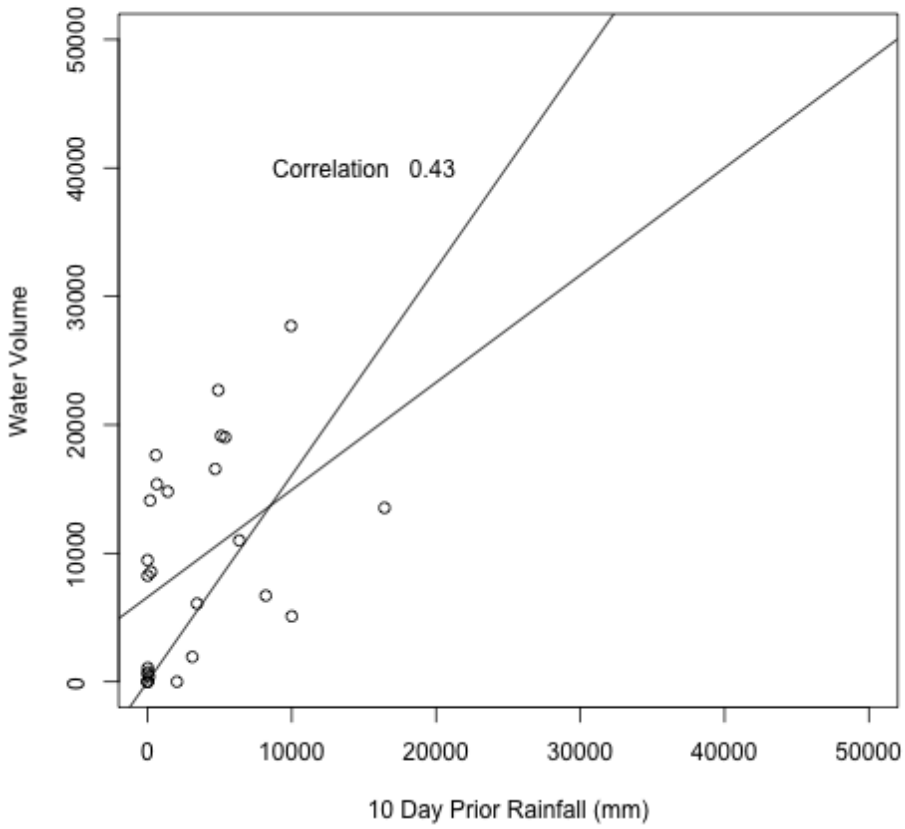


Figure 7: Volume of water within flood basins compared with the total rainfall within the basin for the prior ten-day period. The lines represent linear fits with and without a constraint of zero volume corresponding to zero rainfall. The (Pearson's) correlation for this sample was 0.43.

5 About Project Matsu

Matsu is an Open Cloud Consortium (OCC)-sponsored project. The source code and documentation will be made available on GitHub under Open Cloud Consortium (`https://github.com/opencloudconsortium`).

The OCC is a not for profit that manages and operates cloud computing infrastructure to support scientific, environmental, medical and health care research. The OCC is focused on using this technology to make scientific advances by working with scientists in a variety of disciplines. Visit us at `http://opencloudconsortium.org` or, for more information, email `info {at} opencloudconsortium {dot} org`.

Open Data is a member of the OCC. Open Data began operations in 2001, specializes in building predictive models over big data, and is one of the pioneers using technologies such as Hadoop and NoSQL databases so that companies can build predictive models efficiently over all of their data. More information can be found at `http://opendatagroup.com` or by emailing `info {at} opendatagroup {dot} com`.