EOSC 510 - Term Project: Assessing Similarity of Western Canadian River Basins Using Runoff and Spatial Distribution of Precipitation

Dan Kovacek (35402767)

2020-04-06

Contents

1 Abstract **1**

2 Introduction **1**

3 Data **1**
3.1 Historical Weather Radar 1
3.2 Daily Average Streamflow (Runoff) 2
3.3 Catchment Boundaries 3

4 Methodology **4**
4.1 Use of Anomaly Detection 4
4.1.1 AD Algorithm 5
4.2 Radar Image Preprocessing 6
4.3 Hydrograph Reconstruction 7
4.4 Clustering Spatial Precipitation Distribution 7

5 Results and Discussion **8**

6 Conclusions **8**

References **8**

1 Abstract

Hydrological analysis is often undertaken in ungauged river basins on the basis that historical records from a reasonably similar basin can serve as a proxy. Runoff characteristics in mountainous regions across British Columbia and Alberta have a high amount of variability at the local level, confounding efforts to develop relationships that apply accurately at the regional level. In this study, precipitation data is captured for 144 basins in British Columbia and Alberta using historical weather radar imagery. Spatial distribution of precipitation is estimated using a sample of precipitation events identified in summer and fall between 2007 and 2018. Self organizing maps (SOM) are used to identify similarity in spatial distribution of precipitation. SOMs are applied independently to measured runoff at the same catchments to demonstrate similarity in runoff patterns. The precipitation and runoff based SOMs are compared. The results show...

2 Introduction

The characterization of water resources is a critical step in support of natural resource project proposals in Canada. Hydrological analysis is critical to the planning, design, permitting, and operation of mines and hydropower facilities in particular. The standard of resource engineering practice in assessing precipitation and runoff is often limited to quantification at seasonal or annual levels due to the limitations in the resolution of available data. Precipitation gauges measure a single point in space, while spatial distribution of precipitation is highly variable. As a result, there is considerable uncertainty associated with applying precipitation measurements to an ungauged basin of interest for analytical purposes, and derivative metrics such as runoff ratio and evapotranspiration are accordingly highly uncertain. Similarly, regional relationships are applied in practice to adjust long-term estimates from one location to ungauged locations some distance away. Combining precipitation and runoff information from different sources helps develop a reasonable picture of expected values.

Despite its own inadequacies, weather radar data uses low frequency radio waves to measure the density of water droplets in the air. The calibration of radar sensing to precipitation intensity has its own limitations and uncertainty, but weather radar has the benefit of providing a spatial projection of precipitation intensity within some sensor radius. The goal of this research is to develop a probabilistic estimate of the uncertainty in using weather radar data for predicting runoff in ungauged locations. Historical weather radar data and concurrent streamflow data at 144 hydrometric stations in British Columbia (BC) and Alberta (AB) are used in conjunction with weather radar data from five stations in BC and AB to compare estimated precipitation to the resultant runoff.

-Reference the UBC PhD dissertation and describe the difference in approach

3 Data

3.1 Historical Weather Radar

Environment Canada (EC) provides free access to historical weather radar data as far back as 2007 for some radar stations, though programmatic access is not provided at the time of writing. Historical weather radar as a result was obtained by a web-crawling script on specific time periods corresponding to summer and fall runoff events. Five stations in BC and Alberta were used based on coverage of mountainous basins. Weather radar coverage encompasses a circular area described by a radius of 250km from the radar station. An example radar image is shown in Figure 1

The resolution of radar imagery corresponds to 1 pixel for every 1kmx1km, and the image is centred on the station. It is these two pieces of information that allow a reasonable projection of a coordinate system onto an unreferenced image. The algorithm behind the extraction of radar data is discussed in greater detail in Section X. Note the information layers, such as place names and concentric circles are embedded in the image and cannot be removed, causing issues for some catchments.

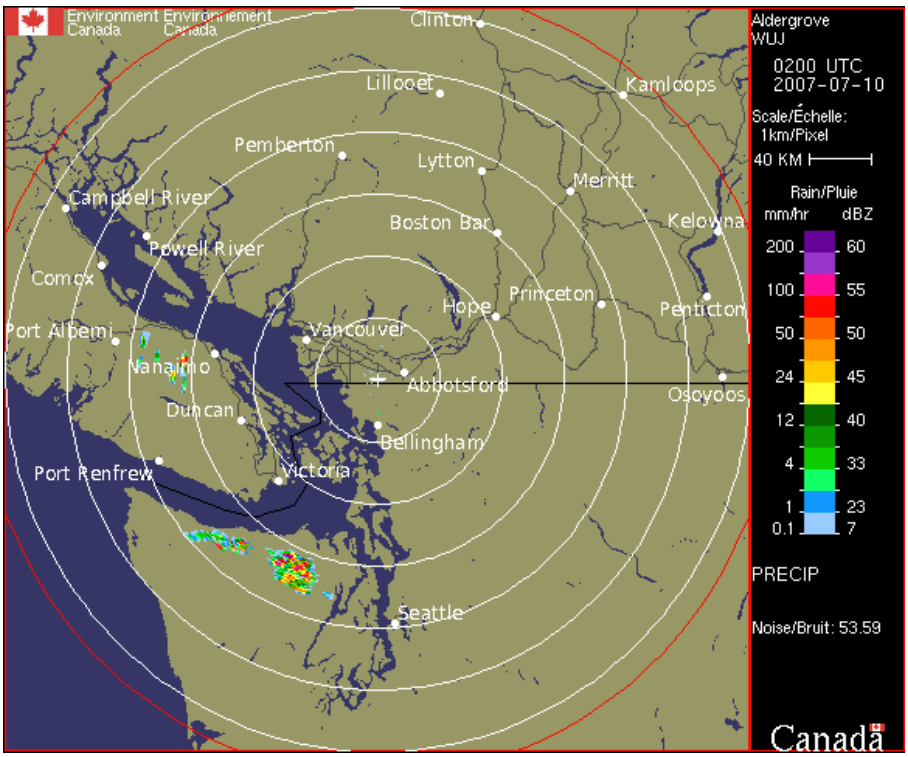

Figure 1: Example Radar Image (Aldergrove, BC Radar Station)

3.2 Daily Average Streamflow (Runoff)

The Water Survey of Canada (WSC) provides open, programmatic access to historical daily average streamflow records for over 8000 active and inactive hydrometric stations across Canada. A database file (HYDAT) containing all historical WSC streamflow data is maintained and updated quartely, and the October 2019 HYDAT database file is used in this study. An example streamflow time series is shown in Figure 2. Figure 2 and subsequent figures are plotted using the Bokeh Data Visualization library (Bokeh Development Team 2020) for the Python programming language.

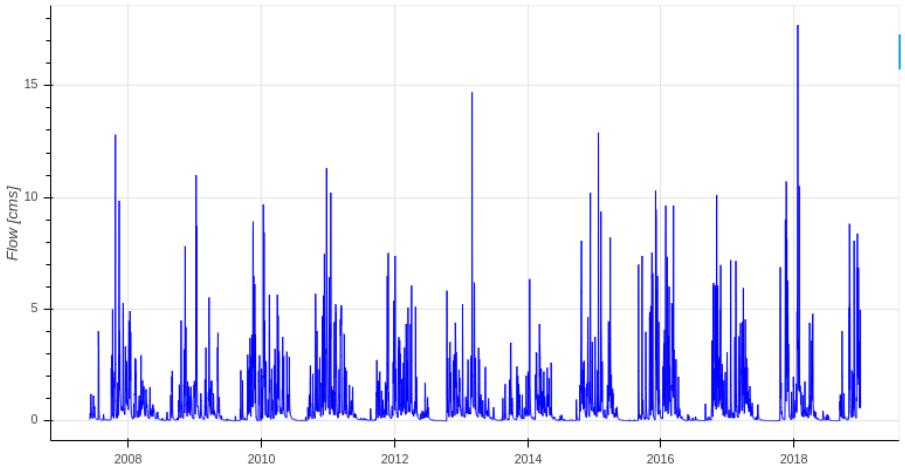

Figure 2: Example Daily Average Flow Timeseries (WSC 08HB048: Carnation Creek at the Mouth)

The HYDAT database is filtered to include stations in BC and Alberta, to include stations with historical record concurrent with the weather radar stations, and to include stations falling within the measurement radius of a radar station. Since the radius of radar measurement is limited to 250 km,

and also because information layers embedded in the radar images obstruct some areas of the images, the WSC stations were also filtered to include stations capturing a drainage area of less than $1000km^2$. A total of 141 stations were found to fit the filtering criteria. The smallest WSC catchments are in the order of $10km^2$ which correspond to a mere ~10 pixels of the radar image. As a result, the accuracy of the smallest and largest catchments is expected to be the poorest. The WSC stations satisfying the above criteria are plotted on Figure 3.

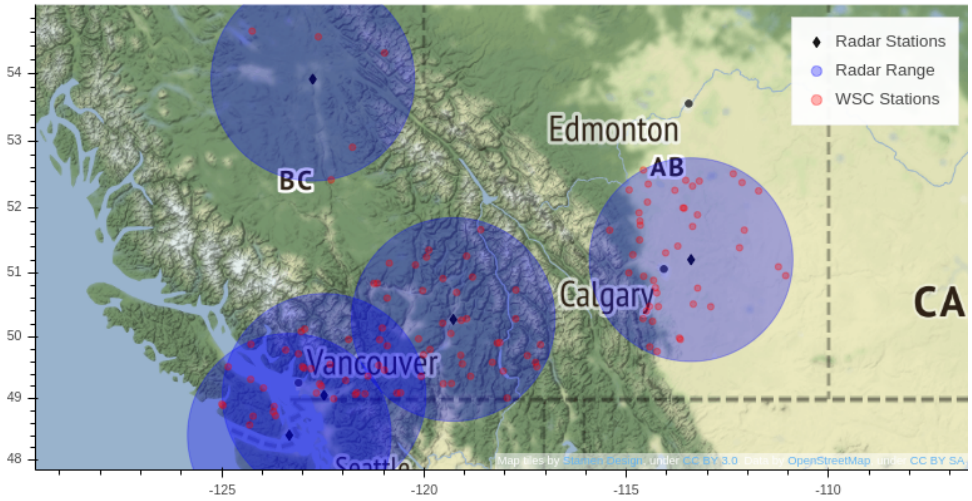

Figure 3: Radar Stations and Range (Approximate), and WSC Stations within Radar Coverage

3.3 Catchment Boundaries

Geographic polygons corresponding to most of the WSC hydrometric stations are available from the Government of Canada's Open Data Platform. Given the low resolution of the radar images, the shape files are believed to be suitable for the intended purpose of extracting from the radar images the pixels corresponding to each catchment of interest. An example catchment boundary with its corresponding station location is shown in Figure 4

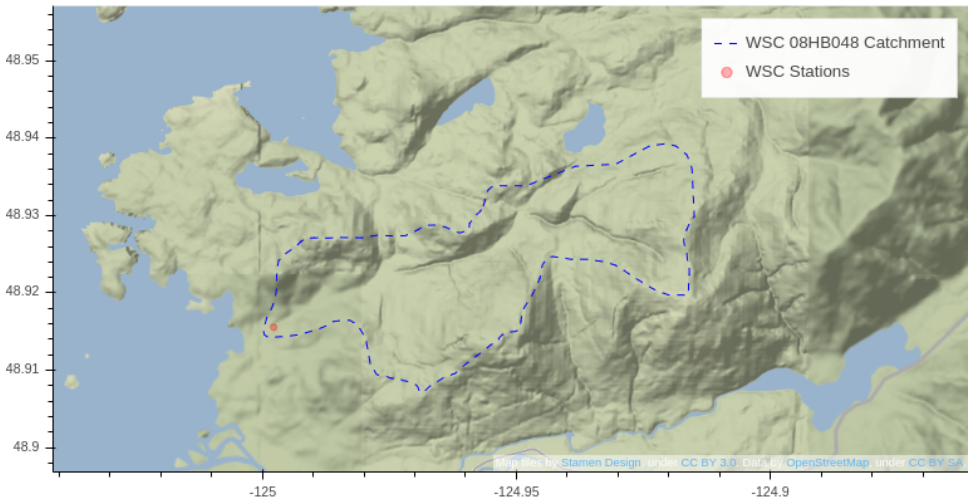

Figure 4: Catchment boundary for WSC 08HB048: Carnation Creek at the Mouth

4 Methodology

The primary research question involves the comparison of weather radar data and streamflow data to ultimately find clustering patterns of watersheds based on spatial distribution of rainfall, physiographic characteristics, and geographic location. The data acquisition process necessitated an additional analytical step which will be addressed first to set the context for the subsequent analysis. The remainder of data pre-processing is then described, including extraction of precipitation data from radar images corresponding to WSC station catchments, and the reconstruction of hydrographs. Finally, the methods used to evaluate the predictive power of radar data are presented.

4.1 Use of Anomaly Detection

A limiting step in the data acquisition process is radar imagery retrieval. The number of server requests required to capture 12 years of radar images at 10-minute intervals at 5 stations in BC and Alberta is in the order of 3E6, which is excessive for a free service and invites a ban from use. Focusing the study to summer months to simplify the interpretation of radar data (by avoiding precipitation as snow) does not alone reduce the number of requests to a viable level. As a result, an anomaly detection (AD) algorithm is used to identify isolated runoff events in summer and fall to reduce the total number of radar image requests to a reasonable number. However, the execution time of a sufficiently complex AD algorithm could negate the (time) cost savings from a reduced number of image requests, and a sufficiently sensitive AD algorithm could label every oscillation in the input signal, no matter how minute, as a 'runoff event', resulting in more image requests and longer AD algorithm execution time. A tradeoff then exists between running the anomaly detection algorithm to reduce the total radar image server calls to a viable number, the time required for the AD algorithm execution, and the number of runoff events identified (true positives).

The AD algorithm is supervised, in the sense that a training period is provided as an input. Initial testing of the AD algorithm demonstrated a high level of sensitivity to the training period selected. The search space of all years and all months is computationally intractable, so Monte Carlo (MC) simulation is used to identify variability in AD performance as a function of training period selection. To quantify the variability of AD performance (using number of events identified as a metric), a sample of 30 random input training periods for each WSC station are used as input into the AD algorithm, and the results for all stations are grouped by training year. Figure 5 shows the variability in number of events identified by the AD algorithm based on 1000 random selections of training year input.

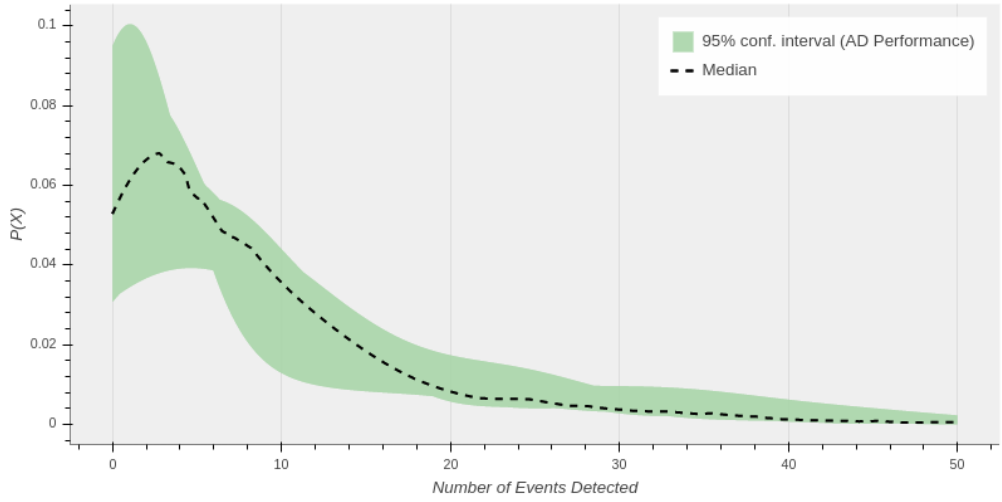

Figure 5: MC simulation: 1000 random selections of training year (KDE probability density function fit)

Combining the random selection of a single year (2007-2018) for input training with a random selection of 1-12 months (inclusive) yields a total search space of roughly 50K alternatives. The execution time

of the AD algorithm is such that computing the full search space is intractable for practical purposes. Better efficiency in code may be possible, however the main function of the AD algorithm already employs the well-optimized Tensorflow Python library (Abadi et al. 2015). To illustrate the time cost of a random search for input parameters, a random sample of 30 input parameters applied to the AD algorithm took 70 minutes for 144 stations on a six-core gpu-enabled (CUDA) machine. As shown in Figure 6 below, the number of runoff events detected by the AD algorithm using the random sample of 30 training parameter combinations per station are exponentially distributed, highlighting the opportunity for an improved search method. Since the training inputs are not continuous variables, a gradient-based search method cannot be directly used.

4.1.1 AD Algorithm

The identification of individual runoff events in nonlinear streamflow signals is a difficult problem, and generalizing the problem across catchments of widely variable characteristics adds to the complexity. For the specific use case of this study, it is important to identify a sufficient number of samples (true positives) in order to support meaningful analysis in the subsequent steps addressing the primary research question, which requires representative estimates of spatial distribution of precipitation. False negatives (missing events) are considered less important than false positives (identifying a runoff event where there isn't one) in reducing the quality of the dataset, as false positives tend to result in biased outcomes corresponding to either 0 or infinite runoff ratio.

The AD algorithm itself takes a daily average runoff time series, and builds a matrix of some number of lag periods proportional to the size of the catchment. Using up to 15 lag periods, or 15 days, is expected to be suitable for the time of concentration and hydrograph response of basins in the range of 30 to 1000 km^2. Principal Component Analysis (PCA) is then applied to reduce the number of lag series to those components describing a minimum of 90% of the variance in the data. Figure 7 shows that most of the time just 2 components are required to meet the 90% variance target, however the expected AD performance is a maximum for between 4 and 5 components, and the confidence interval highlights the large amount of variance in the data for between 2 and 6 PCA components. Much of this variance is expected to result from the runoff signals themselves, as for example few runoff events will be identified from June to September in the semi-arid climate of the BC interior.

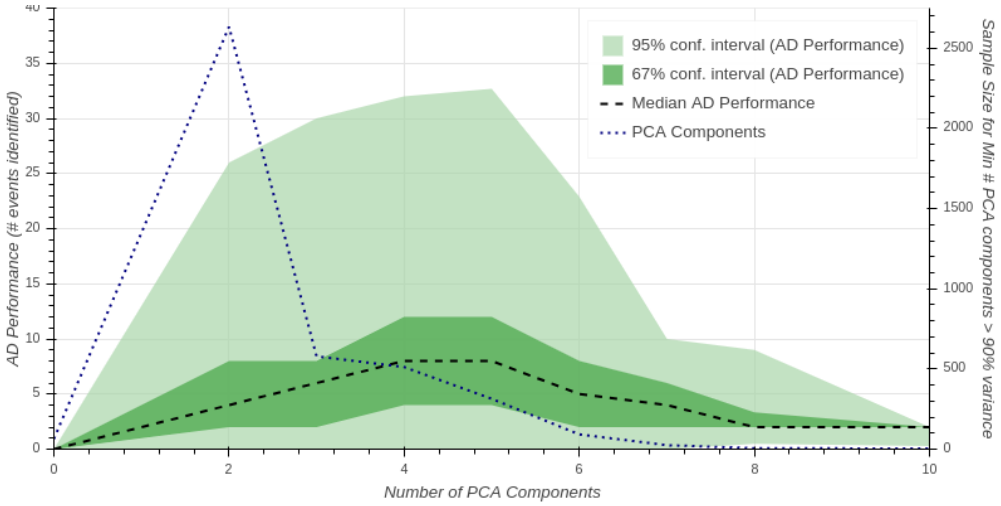

Figure 6: PCA Components and AD Performance

The principal components (PCs) comprising a minimum 90% of the variance in lagged data are then used to create a single variable time series corresponding to the Mahalanobis distance (MD) – the Euclidean distance from each row of data points (each detrended observation and its n lags) to the corresponding PCs. A threshold Mahalanobis distance then represents some magnitude of deviation from the PC within a timeframe correspondent with the number of lags (components). The runoff events identified by the AD algorithm then correspond to the timestamps where the MD crosses the threshold

(in both directions). The results are plotted in a small multiples format to facilitate verification of the sample of events for each station, an example of which is presented in Figure 8 below.

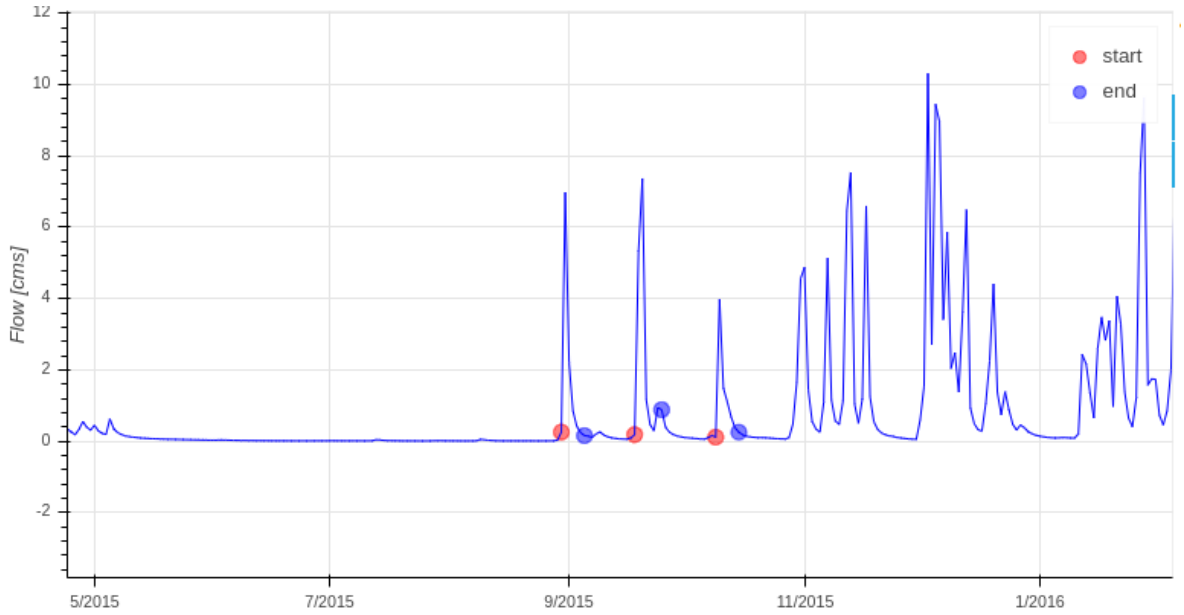

Figure 7: Example AD Results (Timeseries) for WSC 08HB048: Carnation Creek at the Mouth

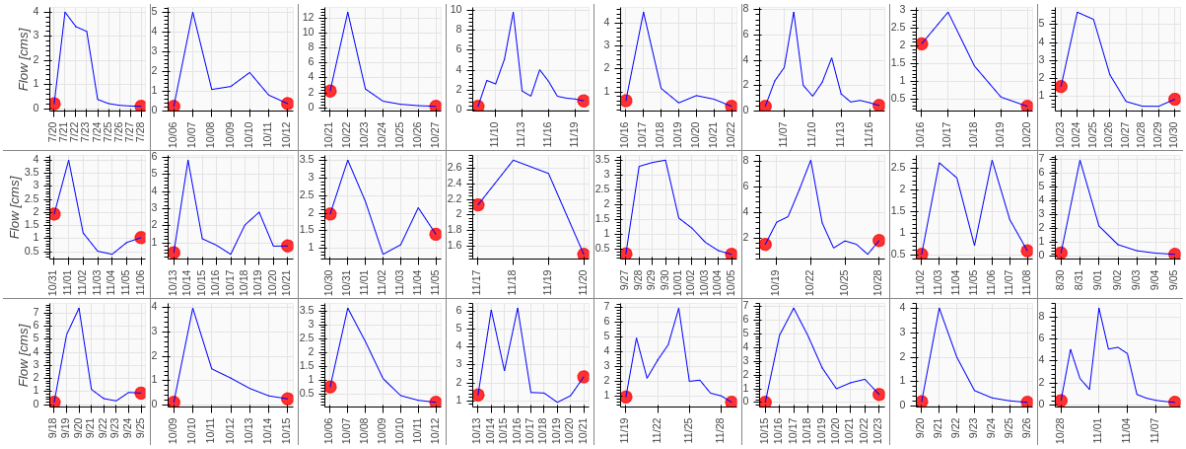

Figure 8: Example AD Results (Event Hydrographs) for WSC 08HB048: Carnation Creek at the Mouth

4.2 Radar Image Preprocessing

With a set of runoff events identified at each WSC station, the start and end times are used to compile a list of query parameters with which to retrieve the concurrent radar images from the nearest radar station. Once retrieved, a geographic projection is applied to the image based on the stated image resolution (1px = $1km^2$) and based on the known geographic coordinates of the radar station corresponding to the centre pixel. Ensuring the coordinate reference system is consistent between data sources, the catchment basin geometry is then used to create a boolean 'mask' such that the radar image pixels representing each basin can be retrieved in a batch process. The conversion of distance from the stated image resolution of $1km \times 1km$ is $\frac{1}{111.32}\frac{degree}{m}$ which propagates an error of less than 1% of the image resolution at the edges, so the error is neglected.

As shown in Figure 1, rainfall intensity is described by a unique array of 14 colours. The final step in the radar image processing is to map the colour values in the masked radar images to their corresponding

precipitation intensity. The final result is a time series of matrices representing the volume of water that fell on each cell comprising the catchment basin.

Summing the precipitation time series and applying a colour map to the normalized output volume yields a representation of the spatial distribution of precipitation for the sample events captured. Figure 10 shows a grid plot of a subsample of 100 WSC basins scaled to similar size to emphasize the differences in colour patterns, where yellow represents less precipitation and blue represents more precipitation.

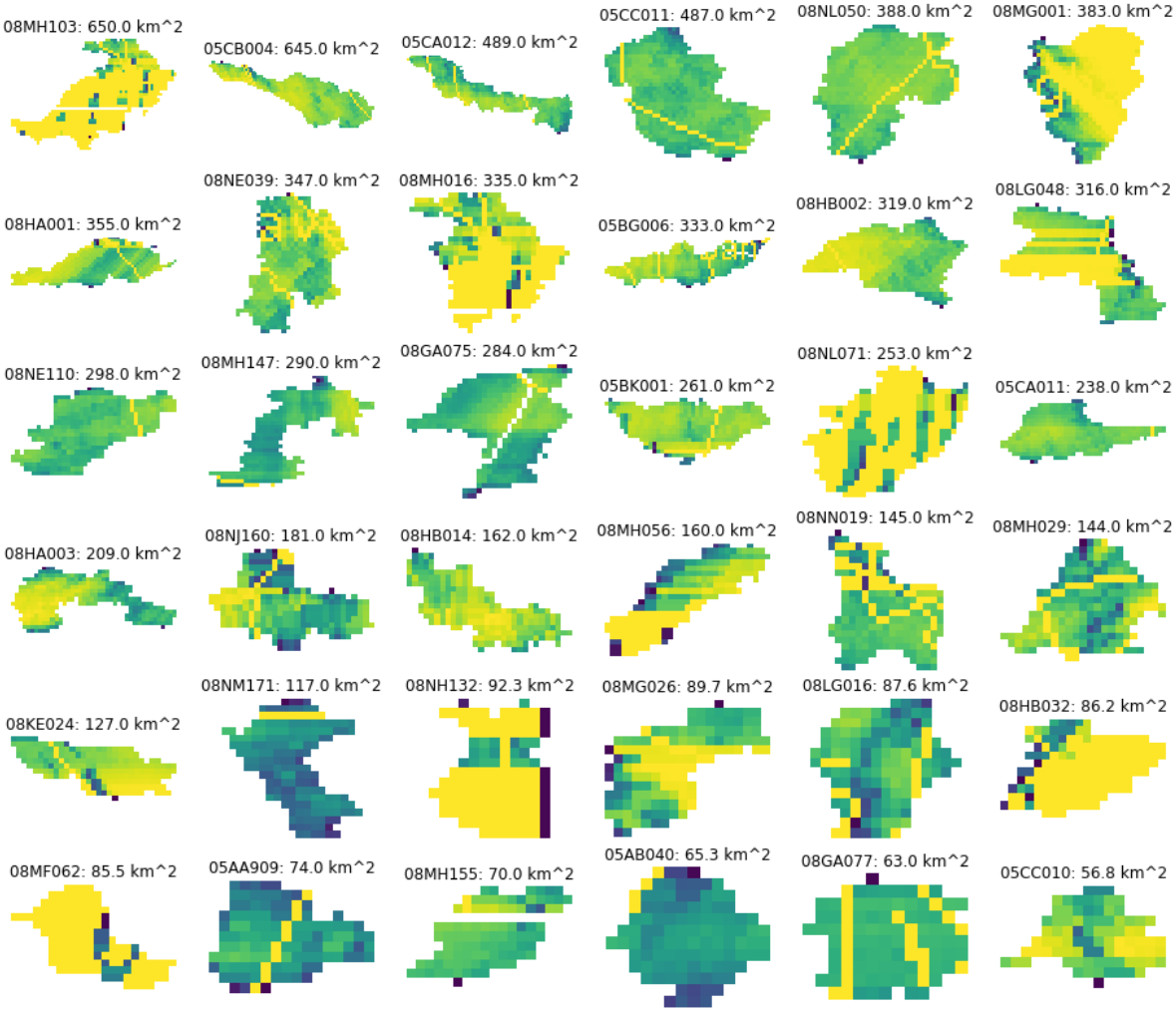

Figure 9: Spatial Distribution of Precipitation at 100 WSC Stations

Vestiges of the information layer are apparent in some of the basins in Figure 10 where parts of place names and the concentric rings from the image block the radar information. In other basins, non-uniformities are evident (08MG001, 08HA003) suggesting orographic effects. Spotted patterns (08NL071) suggest precipitation falling in convective cells in one or few events in the summer season, though this could also be attributable to noise or other interference in the radar data.

4.3 Hydrograph Reconstruction

4.4 Clustering Spatial Precipitation Distribution

Often basins of interest to a researcher are ungauged. In the case of estimating water resources for ungauged basins, information from gauged basins is projected to the location of interest based on similarities in physiography. (Obedkoff, Sustainable Resource Management, and Branch 2003) divides

the hydrology of British Columbia into many subregions on the basis of there existing some level of homogeneity in runoff characteristics at a local level. The hydrologic zones in BC are typically aligned with the coast as the source of air moisture is predominantly the Pacific Ocean. Measured runoff statistics such as long-term averages and extremes then form the basis for the region for the purpose of estimating runoff at ungauged locations.

5 Results and Discussion

-having all the data would eliminate the costly AD step, negate the issues with false negatives, and give a more complete picture of total precip.

6 Conclusions

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." http://tensorflow.org/.

Bokeh Development Team. 2020. *Bokeh: Python Library for Interactive Visualization*. https://bokeh.pydata.org/en/latest/.

Obedkoff, W., British Columbia. Ministry of Sustainable Resource Management, and British Columbia. Aquatic Information Branch. 2003. *Streamflow in the Lower Mainland and Vancouver Island*. Government of British Columbia. https://books.google.ca/books?id=II7iAAAACAAJ.