2017

Quantitative Estimation of Causality and Predictive Modeling for Precipitation Observation Sites and River Gage Sensors

Tri Vu Nguyen
Louisiana State University and Agricultural and Mechanical College

QUANTITATIVE ESTIMATION OF CAUSALITY AND PREDICTIVE MODELING FOR PRECIPITATION OBSERVATION SITES AND RIVER GAGE SENSORS

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

in

The Department of Electrical Engineering

by
Tri V. Nguyen
B.S., Computer Science, University of Texas at Austin, 2013
August 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF LISTINGS

Abstract

This project seeks to investigate two questions: correlations from precipitation measurement sensors to river gage sensors, and predictive modeling of peak river gage heights during precipitation events. First, if correlations can be quantified, then a predictive model can be explored to predict peak water levels at river gage sensors, in response to precipitation inputs. Answering both research questions can provide early flood detection benefits and provide quantitative time assessments for flood risks. An extensive data-driven study was conducted across a geographical area of the U.S, spanning the time period 2008-2016 to identify river gage sensors that are closely correlated to nearby rainfall events. More than 1000 precipitation observation sites were identified and for each precipitation site, nearby river gage stations/sensors were ranked using a cross correlation measure. The cross correlation measures provide information such as which river gage sensors are most sensitive to nearby precipitation inputs. Predictive machine learning models were also developed around each rainfall-river gage pair to learn from historical rainfall and river gage levels, and then predict peak river gage heights. The predictive models generated were accurate and verified a strong causality between precipitation events and river gages that were sensitive to such events. A web-based and map-based decision support and visualization tool was also developed to depict the causality between precipitation and river gage sites and to graphically display the results of the predictive models. This study found about 3500 strongly correlated rain station and river gage pairs. Machine Learning models for these pairs yield high accuracy - 80 percent and above.

Chapter 1
Introduction

1.1 Time-Series and Growth of Sensor Data

A time-series is a continuous series at a certain time-interval resolution[15]. The study of time-series provides pattern identifications, trend suggestions and forecasts [15]. In this project, the domain of time-series is limited to hydrological data with observations provided by sensors. An example is the time-series of temperature in Baton Rouge from 2008 to 2016, measured by temperature sensors at Baton Rouge Airport Weather Station (KBTR). This temperature time-series could indicate the trend of temperatures through seasons in Baton Rouge, and perhaps climate-changes, if any, through the years.

As sensors technologies advance, their time-intervals become shorter, providing higher resolution time-series. Before 2008, the majority of The United States Geological Survey (USGS) river gages were reporting with a daily time-interval [38]. Whereas by the time of this project in 2017, most of those gage sensors are reporting every 15 minutes [38]. The rapid growth in numbers of sensors and the data they provide introduce new challenges to manage and analyze this new influx of information.

1.2 Project Work Outline

In this project, the two subjected time-series are precipitation and river water-level. The goal is to verify the correlation between these two time-series; then, apply machine learning models to forecast the short-term maximum of river water-level.

The first phase is to quantify the correlation level of precipitation and river gages time-series. This causality relationship of rain event and rising river water-

level may seem intuitive to the human mind. However, this may not be obvious to a computational system, which can only react to numerical signals. Thus, a score of how correlated a river is to a certain rain-event location is essential for a computational modeling. In addition, due to irrigation and urban structures, this correlation may not occur, or take an unusually long or short amount of time to observe. For example, river water level behind a levee is humanly monitored, and thus not influenced by precipitation. Whereas a creek in the city may rise immediately during the rain, as the water drainage system allows very fast draining to river channels [7]. A statistical approach will not only signify the correlation of raining and rising water, but also will derive the amount of time it takes for this correlation to occur. This is also a very key insight to flood forecasting and warning.

The next phase is to predict the peak of river water level in a number of hours after a rain-event. Since the amount of precipitation is an essential factor, prediction is only possible after the total precipitation amount has been recorded. This peak water level prediction within a time boundary is a potential enhancement to the existing flood warning system. As this modeling is purely machine learning based on historical events, it has the potential to work independently as an automated signaling system.

1.2.1 Cross-Domains Time-Series Correlations

This project first uses Cross-Correlation Function (CCF) to validate the causality relationship of local river gages in Baton Rouge area with precipitation readings recorded at KBTR. The similar CCF statistical procedure conducted by Ayuso [3] is followed. However, this work was done in River Arnoia in Spain [3] that is likely to have a different basin and weather pattern than that of Baton Rouge, LA. Therefore, this project's verification will further demonstrate the potential of

CCF statistical method in finding levels of correlations between river water gages and rain events. If CCF is applicable in both rivers in Spain and in Baton Rouge, LA, it is possible this statistical procedure is applicable in every other river in the U.S - or perhaps even worldwide.

After verifying correlations of rivers in Baton Rouge using CCF, this project takes a step further: ranking correlations of every river and local rain stations in the U.S. The goal is to run CCF computational analysis for every precipitation station and river-gage. This will give analytical insights on specific closely related river-gages and near-by precipitation stations. The hypothesis is that only a fraction of nearby river-gages are closely related to a precipitation station. In doing a systematic ranking, a sensitivity score will help finding rivers that are very likely to be influenced by precipitation inputs from a specific station.

1.2.2 A Machine-Learning Based Water-Level Prediction

Next, Machine Learning models are applied to suggest peak water levels at water gages in response a rain event recorded at weather stations. As Cross Covariance Function (CCF) shows high level of correlation between precipitation and water-level in Section 4.1, it is very likely that the amount of precipitation is a strong indicator of the peak river water level. Machine Learning (ML) models utilizes historical data as training data to adjust underlining functions or data structures, ultimately providing a best fit to the output training data. Generally the more training data given, the higher accuracy a machine learning will give. This implies higher accuracy in the future as more and more data with higher time resolution is recorded.

A number of ML models from different categories are applied in this project. The majority of these models are Regression Models, which are based on statistical

regression functions. Training regression models consequently adjust and optimize underlining weights and coefficients of the contained functions, which could be linear or polynomial [43]. Regression models are designed to predict a quantity, and work well with small sample size (less than 100 thousands records) [27]. In addition, regression models perform well when only a few features are significant, by assigning higher weight values [43].

For the training process, the selected features are limited to only climate hydrological elements, excluding physical structure elements. Specifically, the selected features are precipitation, temperature, river flow rate, and number of dry days. This is a very moderate number of features and input signal in compare to modern hydrological models [19]. A river water-level is influenced by various different attributes ranging from its river-basin size and type of soil, its upstream and downstream water level, to its area's precipitation amount and frequency [3]. In this project, the approach is to not take every attribute into account, but rather create an optimized agnostic statistical model for each river gage and rain station pair. The end-goal is to train ML models for every pair of river gage and rain station, that accounts for various river basin structures by adjusting internal coefficients.

1.2.3 Data Source and APIs

One of the major steps was to prepare hourly climate data required for this project. Analytical works require data, and in this case a large quantity of time-series weather and river records. These historical records need to be well-organized, allowing rapid query to get data from different time periods. At the time of this project, there was no publicly available application programming interface (API) for climate data in hourly resolution. Therefore, the first task was to acquire archive historical hourly weather records from public resources, and ingest these records to

a time-series database. Next, an API was created to provide simple programable data in weather station's local time. This API was later made publicly available to other climate researchers at *http://hrly.lsu.edu* [32]. The procedure of how this climate API is made is specified in Chapter 3.

Overall, this project found approximately 3500 highly correlated pairs of rain station and river gages using CCF. Machine Learning models built for each of these pairs produce accuracy averaged about 80%. This result shows that statistical approach is able to both accurately find rain-river correlations, and predict river water level.

Chapter 2
Related Works and Problem Description

2.1 A Brief History of Hydrologic Modeling

Hydrologic modeling takes a sharp innovative turn during the computational age in 1960s. Prior to this time, models were largely based on empirical experience and physical analysis [30]. Hydrologic and climate data had been very limited and hardly accessible, until the growth of computational power and the invention of the Internet. Great computing revolution around 1960s brought the necessary power for statistical modeling to work. As early as 1966, Crawford and Linsley were working on one of the first watershed model using computers [7]. This early success brought attention to applying statistical methods in hydrologic modeling.

Dramatic computing growth during the digital era have enabled tremendous opportunity for advanced Hydrologic modeling. More data is available now than ever before [3]. According to the National Climatic Data Center (NCDC) documentation, there are approximately 33 thousands registered weather stations in the U.S alone [21]. These weather stations report hourly with meteorological measurements such as temperature, humidity, precipitation and other climate elements. Regarding river water gages, the United States Geological Survey (USGS) monitors about 36,500 sensors in the U.S [38]. Each of these sensors report with a resolution as high as every 15 minutes [38]. In addition, the Internet enables access to these data very conveniently. In the U.S, these data are provided via public APIs [37], or accessible portals [22]; both of which are free of charge. This vast amount of data greatly improves simple statistical models's accuracy [26]. Along with increasing availability of data, the computing capabilities has immensely increased over the

past decades. This trend of greater computing power and capability will likely continue, which in turn allows statistical methods to be done with more data in less amount of time.

Modeling river water-level is a great application of Hydrology computing [30]. There are primarily two well-developed approaches to hydrologic modeling: deterministic and stochastic [15]. The main difference between deterministic and stochastic is the requirement for physical simulations. Deterministic modeling simulates physical structures to give estimations, whereas Stochastic modeling does not [15]. This project uses the stochastic modeling approach. Thus, this direction will be explained in greater details in Section 2.1.2.

2.1.1 Deterministic

Deterministic Modeling applies mathematical concepts and physical based equations to estimate a response given an input [39]. These conceptual equations often aim to emulate the physical realities, in addition to variables of momentum, mass and energy [30].

There are a number of attributes that may be used in a deterministic model. However, the level of influence of these attributes may vary from river to river, and season to season [26]. The following factors are significant to a deterministic hydrologic model [15]:

- Dry Periods: the period from the last rainfall. The longer the period, the more likely the ground is dry, leading to higher absorption rate and amount of rain [15].

- River Basin size and shape: the larger the basin, the more likely a rainfall would influence water [15].

- Current Water Level: the higher the water-level is, the more amount of water is required for it to rise [15].
- Soil Moisture: this is similar to dry period. A low moisture in soil suggests a higher absorption rate to rainfall. [5]
- Evapotranspiration: high temperature leads to some evaporation. More importantly, long period of high temperature leads to drier ground, and lower soil moisture [15].
- Terrain: river terrain and slopes strongly influence the speed of river rising [30].
- Location: the distance between a weather station and a river gage [7].

The attributes that are not physically simulated are considered to be included in this project's modeling in Section 4.2. Specifically, *Dry Periods* and *Current Water Level* are included, as these two attributes are easily calculated given weather station time-series data.

2.1.2 Stochastic

Stochastic model lean solely on historical data to predict a future outcome. These models operate without taking into considerations of physical attributes that may vary from river to river. In essence, this is a probabilistic black-box approach [15]. An input is given; then based-on parameters derived from the past, an output is provided. This black-box is usually a statistical method with adjusted coefficients to have the best-fit for historical events. Using these trained coefficients, the statistical method will give a prediction, given a new input.

The great advantage of statistical methods is ignoring engineering structures and other specific local physical attributes [30]. By focusing on numerical changes and

trends, those physical attributes of a river that influence the speed and the peak of water-level are already included. Thus, these statistical methods can be applied to any rivers or water streams, if enough data is provided. Whereas, deterministic approaches require domain expertise inputs, physical reasonings and simulations.

Although statistical applications were proposed as far back as 1914 [16], only in the past decade were these techniques practically utilized [34]. Before 1996, collected data was very limited in size. The accuracy, frequency and quality of instruments were not sufficient to gather enough insights on weather events. In addition, without the aid of the Internet, accessing hydrological data was a challenge. Nowadays, these hydrological data are easily accessible with accurate readings and a high spatial resolution. In addition, computational resources have become more accessible, along with numerous statistical and numerical libraries. This has revolutionized the potential of applying statistical model in hydrology. In fact, this project is only possible with open-source libraries and public weather APIs. For open-source library, this project utilizes Machine Learning library *scikit-learn* [28], and statistical library *statsmodels* [33]. Both libraries are very stable, and well-developed.

Stochastic modeling is on a rising-trend. This is as a result of increasing availability of hydrological data with higher time and spatial resolution, gradually boosting up this statistical approach accuracy. In a recent hydrological research done in [3], a stochastic approach was able to deliver high correlation accuracy and prediction [3]. This is a very good result, as river basins and channels along with other physical attributes were disregarded. Thus, this finding in [3] is potentially applicable to other flowing surface water area without dam, irrigations and other human physical structures. The question now is whether this stochastic research approach in [3] could be successfully applied to more rivers.

2.1.3 Current Hydrological Modelling

Latest hydrological models utilize a structure of sub-models that take into accounts of various factors: precipitation, temperature, river basin size, river slope, GIS and others hydrological elements [11]. A great challenge of building such a sophisticated model is optimizing parameters for each of these attributes. Every river is different and may react differently to a precipitation event. Even the same river may rise differently in a different season of the year. Thus, sub-models are built to account for these factors. In addition, a great challenge to hydrological modeling is the ability to forecast and simulate river water-levels continuously. This requires sophisticated modeling to predict water levels hourly in the next few days, up to a week. In this project, a different approach is taken. The prediction output is the peak of the water level, not necessarily when that will occur. This helps eliminating the need to consider various physical attributes of the river water basins, which are essential to the speed of water rising. The details of this peak water level output are explained in Section 4.2.

According to the Advanced Hydrologic Prediction Service (AHPS), the direction of building pathways for better science in water forecasting have been proposed to include different areas of science as part of its modeling system including: probabilistic hydrologic application, distributed hydrologic and calibration [19]. Calibration is a significant part of [19] as it largely determines the error level of a model. The calibration system generally includes soil-moisture [5], snow accumulation and ablation [2], streamflow routing [14], and reservoir simulation models [19]. These calibration is largely a labor-intensive process, as it requires experts' field knowledge. Although automatic calibration methods have been proposed. The process still generally relies on manual expertise inputs and field inspections. A historic

flood may change the basin landscape and vegetation. Thus, only through manual labor and inspections can this change be accounted for.

By learning from the work in [19], some key observations are included in this project. These essential observations to modeling accuracy are: precipitation, temperature and evaporations [19]. According to AHPS, rainfalls projections are regarded as a required input for hydrologic forecast models [19]. In this project, rainfalls forecast estimations are not included, thus leading to higher error in continuous raining events.

2.2 Project Description

This project aims to answer that question of extending and scaling the experiment in [3], by applying stochastic modeling to rivers across the U.S. Effectively, the pioneering work in [3] will be extended to every river and weather station in the U.S. Section 4.1 will put in detail the process of applying cross-correlation search across every weather station and water gages in the United States. Section 4.2 shows how a stochastic approach is used with Machine Learning to forecast peak water-level in short-terms across U.S rivers.

Chapter 3
Data Access and Storage

3.1 Precipitation and Water Level data

This project relies primarily on two types of hydrological data: weather (precipitation, temperature and humidity) and water-level. Weather data is acquired from both: National Oceanic and Atmospheric Administration (NOAA) and Aviation Weather Center (METAR). NOAA provides monitored and corrected data in a daily basis [22]. METAR data is broadcasted hourly over a hydrological satellite network [20]. NOAA ISH data is quality-controlled (QC) and near real-time accessible, whereas METAR data is non-QC but real-time accessible. Therefore, NOAA ISH weather data is used for training, as this dataset is quality-controlled, providing higher consistency and accuracy. METAR data is potentially utilized for real-time prediction, as a future extension for this project. On the other hand, real-time water-level at gages is obtained from an API provided by the United States Geological Survey (USGS) [37].

3.2 Files storage and Database

An important part to this whole project is selecting a storage structure that fits the climate informatics growing needs and delivers a rapid performance. Storing and managing weather data is a challenging task, as the data goes back as far as 1950 from thousands of stations in the United States. As the long-term goal is to deliver real-time analytics, the speed of underlying storage structure is key to both prediction abilities and visualization. The good news is time-series storage has received great industry interest recently. This is as a result of strong growth in the area of Internet of Things (IoT) devices [1]. This project considered two prominent

time-series storage options for great performance and management: HDF5 and InfluxDB time-series database [10][17].

3.2.1 HDF5 Hierarchical Data Format

HDF5 stands for Hierarchical Data Format. Using HDF5 provides the flexibility of defining a structure that best suites users' reading and writing pattern [13]. The advantage of HDF5 is a simple file format that allows very fast reading throughput [13]. This project initially used HDF5 as an experimental storage. The file structure is that each weather station is a separate file, grouped by year. Each year has all datasets corresponding to climate measurements: temperature, humidity and other observations. Each dataset is pre-allocated the hourly space for the whole year at the time of creation. The time granulation is therefore at the hour level. A later observation in the same hour will overwrite the previous one. The data type and corresponding sizes of each dataset is pre-defined and fixed.

HDF5 has a number of advantages. This file format is popular among scientific community, which provides more technical resources for support and performance tuning [13]. HDF5 also has high reading throughput within a group and dataset. In addition, there is built-in support for in-memory reading, which works as a cache layer to provide faster access to frequently viewed data.

Nevertheless, HDF5 has a number of significant drawbacks that led to a search for a more stable storage. First and foremost, HDF5 lacks built-in support for time-series data. Every timestamp is converted to an index to map to a measurement dataset array [41]. HDF5 also requires pre-defined storage, effectively a substantial initial allocated storage. Many datasets that stores infrequent events such as precipitation and snow are very sparse, leading to large chunks of unused of storage. Using HDF5, the total storage space required for all US stations in 7 years (from

2010 to 2016) is 385GB. Although storage space tuning could lower this figure, it was out-of-scope for this project. In addition, HDF5 lacks support for basic queries across groups and datasets, such as filtering by specific value or value range in a single dataset. Every query requires a custom programming function to handle data querying and filtering. HDF5 requires knowledge of reading and writing patterns to define an optimized file structure. This may limit the potential of doing more complex and advanced query in the future. HDF5 requires predefined time-resolution for initial storage space allocation. This time resolution varies among weather stations, as some report once an hour, and some report multiple times an hour. In addition, during extreme events such as hurricanes, weather stations may report as frequent as every 7 minutes. A rigid storage structure would not provide the flexibility for these sudden increase in storage allocations. Therefore, using HDF5 would require extensive additional work to provide a stable API for data query and management. As a result, a time-series database like InfluxDB is required to meet the need of this project.

3.2.2 InfluxDB Time-series database

InfluxDB has a predefined query set and API that any programming languages can interact. In addition, data consistency in a database is guaranteed. These advantages allow this project to focus on analytics, as supposed to data query and management. As a time-series database, InfluxDB has built-in support for storing data with a timestamp, and therefore allows high-performance query based on time-ranges [4]. Among time-series databases, InfluxDB has been the most recommended solution in the past couple of years [9]. In addition, InfluxDB has been highly tested and used in industry across disciplines from climate informatics

to financial analysis [36]. Using InfluxDB helps this project manage data through a well-defined SQL structure, that provides both great performance and consistency.

InfluxDB does provide a significant storage compression efficiency [24]. By using InfluxDB, the storage space required for all US stations METAR data over 17 years (from 2000 to 2016) is 19 GB, as opposed to more than 385 GB going with HDF5. This is a 20x saving in storage. This saving is in line with the 45x storage compression ratio benchmark, using InfluxDB time-structured merge tree compression storage [24]. Therefore, InfluxDB proves to be a better option to move forward in term of data storage and API. There are lots of other options for time-series databases (TSDBs) that is out of scope of this study. InfluxDB is clearly a sufficient solution for this project.

3.3 Data Sources and Ingester

Three ingesters were implemented for three different data sources: NOAA, METAR and USGS. The frequency of ingestion is daily for NOAA, hourly for METAR and daily for USGS. Both NOAA and METAR data feed are in special meteorological format, and required extensive work to create human-readable format. The extracted data is then ingested into InfluxDB. The ingestion work is optimized in batch and parallel processing to provide more up-to-date data.

3.3.1 ISH Ingester

NOAA ISH data is provided per station as a text file. Each line is the data feed for every hour. The information is encoded in a positional format to minimize the amount of text. NOAA provides a manual text to decode each line. For example, timestamp is from position 10 to 14. Temperature information is from 20 to 24. Following this instruction, a program was written to digest each line into a programming object. NOAA I.T staff provided a basic functional JAVA program to

digest each line. In this project, this program was extended to provide better data format. A separate ingester written in Java was added to read each line, decode and save into InfluxDB. This ISH data is provided and updated daily from NOAA FTP repository. Thus, an automated scheduler (cron) was created to run daily a program that would download data from NOAA FTP service, decompress the text data, and ingest into InfluxDB. The storage dedicated for NOAA decompressed text file is about 2TB, whereas the size of InfluxDB ISH data is only about 30GB. This is a clear advantage of utilizing a time series database, which has built-in support for storage compression and rapid time query.

3.3.2 METAR Ingester

Ingesting METAR is processed in 2 stages: filtering METAR data from satellite feed, and ingesting decoded METAR feed to InfluxDB. Since METAR operates in a meteorological satellite network, a lot of received information is not hydrologically related. In addition, data feed could come in multiple lines with leading and ending spaces. Thus, a filter program was written in Python to filter only hydrological feed, sanitize and put in proper format in each line. Hydrological feeds begin with either *METAR* or *SPECI*. Code *SPECIY* is to specify corrections [20]. Unit tests were written to ensure the filter performs correctly, not leaving out correct feeds, and not taking the unrelated feeds. The outcome of this filter is a file with extension *.metar*, which only contains METAR related information. Every hour, a scheduler (cron) runs a BASH script, that filters the satellite feeds into a METAR file, which will then be parsed and ingested to InfluxDB.

3.3.3 USGS River Ingester

USGS River data is provided daily by USGS API. A Python library called *ulmo* is used to access USGS data [37]. In this project, [37] is used to gather river gages

sensors in the U.S. These river gages and its information are stored in a high performance document database called MongoDB. Next, each river gage and its historical records of water-level is queried via *ulmo*. Since this project repeatedly requires querying a large time-range, it becomes a burden to USGS API services overtime. Thus, historical river water-level records along with their timestamps are also ingested into InfluxDB for high performance query.

3.4 META Data

META data is static data about hydrological stations such as coordinates, time-zone, name and international ID. These information need to be stored locally in a MongoDB for faster query and more stability to web services API. MongoDB also allows spatial queries to find nearby river-gages to a weather station.

3.4.1 Weather Station META

Worldwide weather station listing is publicly provided through the NOAA data portal. Each station contains detailed information of its quality, location, country, city, region, international id reference, timezone and coordinates. For example, KBTR weather station in Baton Rouge, LA has the following information:

```
{
    "country" : "United States",
    "region" : "LA",
    "subregion" : "East Baton Rouge Parish",
    "city" : "Baton Rouge|Liberty Farms",
    "station_name_current" : "Metropolitan Airport | Ryan Field Airport | Harding
        AAF",
    "icao" : "KBTR",
    "icao_quality" : "A",
    "national_quality" : "A",
    "lat_prp" : 30.532,
    "lon_prp" : -91.149,
    "tz" : "America/Chicago",
}
```

Listing 3.1. Weather Station META

There are totally 43757 listed weather stations in the whole world [22]. However, many of the listed stations are no longer active. Thus, only stations that have

reported data in 2016 are considered. Out of 43757 stations, only 4820 stations are considered consistent and highly active. These 4820 active stations have been reporting data for the past 1 year.

3.4.2 USGS River Gages META

Information regarding a river gage META data is acquired from USGS API via *ulmo* python library [37]. An example is the META data for the Comite river's water gages near Baker, LA:

```
{
    "code" : "07377754",
    "site_type" : "ST",
    "huc" : "08070202",
    "state_code" : "22",
    "agency" : "USGS",
    "location" : {
        "latitude" : "30.596",
        "srs" : "EPSG:4326",
        "longitude" : "-91.094"
    },
    "name" : "Comite River near Baker, LA",
    "county" : "22033",
    "timezone_info" : {
        "default_tz" : {
            "abbreviation" : "CST",
            "offset" : "-06:00"
        }
    },
    "network" : "NWIS"
}
```

Listing 3.2. USGS River Gages META

The *site_type* field is *ST*, which means this a streaming flow of water. This is a key field to filter rivers and creek gages, as rivers and creeks are considered flowing water streams. The locations in latitude and longitude allow MongoDB spatial search to find these water-gages in radius of 20 miles around every weather station.

3.5 Web Service API

A Tornado (Python) web service was written to provide both weather data and station META data in a programmatic format [32]. Tornado is a simple and powerful framework to build modern web APIs and applications [12]. This project's

Web API is predefined with instructions available on the SRCC Hourly website: *http://hrly.lsu.edu* [32]. Web APIs provide stable data feeds, which is beneficial not only for this project, but also for other research groups.

There are two APIs written: META data and Weather data. META data API provides META data regarding weather station. This includes the weather station classification, quality, region, timezone, and its geographical coordinates. Weather Data Service API joins the data from NOAA and METAR queried from the InfluxDB to provide a simple continuous data interface. This web service is setup as an API with declarative formats and requirements that can support a variety of needs. When a request comes in, NOAA data is first queried, METAR data is then queried to fill in gaps that NOAA data may have had. This dynamic data routing guarantees a higher data quality and greater data coverage to end users. Since NOAA data is updated daily, the latest data feed is likely drawn from METAR database. Whereas, more historical data is likely provided from NOAA ISH, which is monitored and corrected by NOAA staffs. The available measurements for queries are as below with the programming code and its explanation.

The Weather Data Service API Allowed Variables is specified as followings:

- `cloud_ceiling: the observed code for cloud ceiling`
- `dewp: dew point, which could be converted to relative humidity`
- `max_temp: maximum temperature in every hour`
- `min_temp: minimum temperature in every hour`
- `precip_1hr: precipitation amount in one hour`
- `precip_6hr: total precipitation in 6 hours`

- precip_12hr: total precipitation in 12 hours
- precip_24hr: total precipitation in 24 hours
- press_sea_level: pressure at sea level
- sky: observed sky condition
- snow_depth: depth of snow felt
- temp: temperature in Fahrenheit
- vis: visibility range in miles
- wind_dir: direction of wind blow
- wind_gust: sudden brief increase of wind
- wind_speed: speed of wind in miles per hour

Chapter 4
Computational Analysis

The focus of this project is to analyze the effects of precipitation rainfall and river stages (measured using river gage height). Portions of this work use ideas in [3], which shows that large amounts of rainfall can cause river water-levels to rise. However, the amount of rain and the type of soil in each river basin may vary the amount of water level increase from one river to another. In addition, irrigation systems in metropolitan areas may influence the water-flow paths and behaviors [15]. In this computational analysis, statistical methods are applied, which would take into account only physical phenomenal attributes such as: precipitation, temperature and humidity. Human intervention such as levees, canals and other irrigation systems are out of the scope of this analysis.

4.1 Cross Correlation of Precipitation and River Water-Level Time-Series

The initial challenge of this project was to show the correlation between precipitation events and river gage height increases. This correlation can usually be lagged by a number of hours, as it can take time for the rain water to flow downstream along rivers and canals. The challenge is - more specifically - seeking that number of lagging hours, when two time-series' event (upward trend) align. It turns out this is a fundamental problem in time-series analysis [8]. A number of statistical functions have already been developed to address this. Among those functions, Cross Correlation Function (CCF) was selected to use for this project, as it was a standardized method and had already been implemented in public statistical libraries such as R and Statsmodels [33].

4.1.1 Cross Correlation Function (CCF)

Cross Correlation function (CCF) provides a quantitative assessment of two time-series similarity at different time-shifts [8]. Statistically, CCF is a linear function of lagging time, defined as a convolution integral of two time-series as in Equation 4.1 [25]:

$$CCF(\tau) = \int_{-\infty}^{\infty} X(t+\tau)Y(t)dt \tag{4.1}$$

In the case of this project, the X(t) function is the precipitation time-series, and the Y(t) is the river gage height time-series. The number of hours shifting is the τ. CCF(τ) measures the linear association of X(t+τ) and Y(t). The higher CCF(τ) represents the higher associations [8]. A range for τ is limited from 0 to 48 hours to search for the best CCF(τ) for every precipitation time-series and water level time-series.

The programming language used for this study was Python. The library used in this project to calculate Equation 4.1 is Python Statsmodels package *statsmodels.tsa.stattools.ccf* [33]. This cross-correlation function for one-dimensional array is implemented using Python Numpy library *numpy.correlate* that calculates the convolution of two time-series X(t+τ) and Y(t) [29]. The result from *statsmodels.ts.stattools.ccf* is an array of correlations values with indexes are the lagging number of hours τ specified in Equation 4.1. Based on this result, the first CCF(τ) local maxima is the selected point. This point represents the lag hour and correlation value of two time-series X(t) and Y(t).

4.1.2 CCF General Statistical Procedure

The procedure of calculation is as following, and is depicted in Figure 4.1.

1. Find a *Rain Event*, and locate the moment that precipitation begins (*Rain Begins*), and when it ends (*Rain Ends*) (Section 4.1.2.3).

2. Add an *Extended Lag Period* number of hours to the *Rain Ends* to find the *CCF Ends* hour (Section 4.1.2.4)

3. Calculate *River Level Rise* series from River gage height series, using difference function (Section 4.1.3)

4. Create two time-series of precipitation (*X(t)* and river gage height (*Y(t)*) from time-range: *Rain Begins* to *CCF Ends*.

5. Use CCF to compute *Correlation Score* from *X(t)* and *Y(t)*

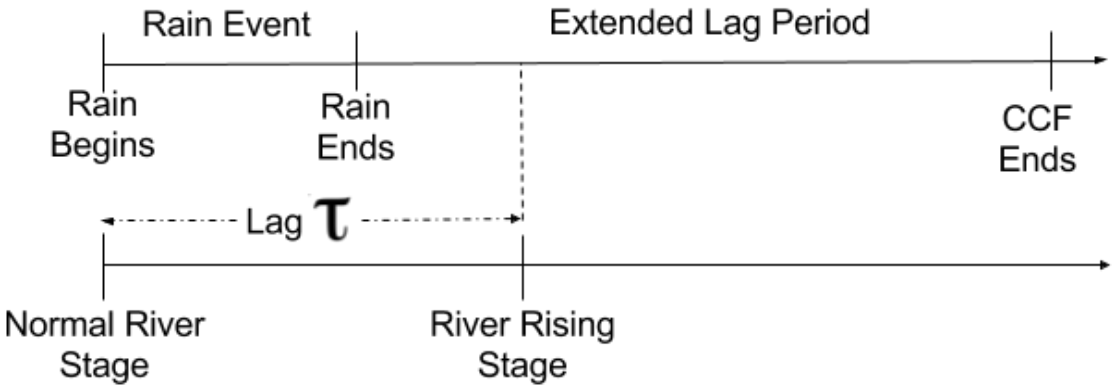

FIGURE 4.1. CCF Lag Period

- **4.1.2.1 Correlation Score**

Correlation Score is defined as the output of CCF given two time-series inputs X(t) and Y(t), and a lag value (time-shift) τ. This *Correlation Score* ranges from 0 to 1 with no correlation having a score of 0 and highly correlated having a score 1.

- **4.1.2.2 Lag Period**

Lag Period is the number of time-shift hours before an increase in precipitation indicates an increase in river gages. As shown in Figure 4.2, a time-shift of one

hour indicates that river water-level begins to rise one hour after a rain event. In this project, a range of time-shift from 0 to 48 hours is used with CCF. The first CCF output's local maxima is selected, and the corresponding lag is the *Lag Period* of the two input time-series. The reasoning is that the first significant water rise event (first local maxima), is likely to correlate with the most recently occurred rain-event. On the other hand, a global maxima can be potentially attributed to precipitations that could have occurred upstream.

- **4.1.2.3 Rain Event**

  A *Rain Event* is a continuous occurrence of rain, regardless of length of time and precipitation amount. A rain-event may last a few minutes or a few hours. A *Rain Event* may contain precipitation of 2mm, which is an extreme rainfall event, or 0.2 mm, which is a regular rain amount [6]. In Figure 4.1, *Rain Begins* is the hour when rain begins. *Rain Ends* is the last hour that rain has stopped, and recorded precipitation is zero.

- **4.1.2.4 Extended Lag Period**

  Extended Lag Period is an additional period of time after a *Rain Event*. After a rain event, it may take several hours before a rise in river height occurs. This additional period of hours is called *Extended Lag Period*, and it may or may not intersect with another *Rain Event*. In [3], 80 hours is chosen as the max *Extended Lag Period*. In this project, *Extended Lag Period* is limited to 48 hours. This is to avoid the likelihood that a river-gage is influenced by unrelated weather events occurring further upstream, or outside the river basin area. The extended lag period number of 48 hour was derived from empirical studies done in Baton Rouge and other areas in Louisiana. These experimental results as listed in Table 4.2 and 4.3 show that a lagging period is most likely within the first 24 hours. Therefore, a double of that timespan (48 hours) was chosen to capture rising events in a river after rain event.

4.1.3 CCF with River Level Rise Series

River Level Rise is the 1st discrete difference series of the river gage height series [29]. The calculation is according to Equation 4.1.3. In which, RLR is the *River Level Rise* series, and R is the river gage height series.

$$RLR(t) = R(t+1) - R(t) \tag{4.2}$$

This is a differing point from what has been done in [3]. Essentially, this project replaces water gage height time-series with its *River Level Rise* series. As the time-interval of measurement is fixed at hourly, *River Level Rise* effectively calculates the speed of water rising or receding. As a result, *Rive Level Rise* is more sensitive to rain-events than the regular river gage height series. The regular CCF procedure in [3] produces the lagging correlation around 6-8 hours or more. Whereas in this study, the mean lagging hours is around 1 to 2 hours, as shown later in Section 4.1.7. In reality, river water-level could be influenced by various numbers of reasons, from rain in other area further upstream, to the water flow-speed, soil moisture and absorption, and river-basin physical structures (levees) [15]. The longer the lagging period, the more likely river water-level is influenced by non-accounted factors. Using *River Level Rise*, CCF overall suggests a smaller and more consistent lagging period that agrees with actual occurrences. Therefore, *River Level Rise* was used instead of direct water gage height.

Table 4.1 is a comparison of CCF score using regular river water level series and the difference of that series (*River Level Rise*). The columns in this table are:

- *Time*: the day that the precipitation event occurred
- *Rise Lag*: Lag Period in hours using *River Level Rise* series

TABLE 4.1. CCF Comparison River Difference at KBTR and Comite River in 2016

Time	Rise Lag	Rise CCF	Direct Lag	Direct CCF	Precipitation
12/04/16	1	0.84	2	0.94	1.6
11/28/16	3	0.62	7	0.44	1.2
08/17/16	1	0.67	**17**	**0.39**	3.6
08/12/16	7	0.45	8	0.83	4
08/12/16	0	0.60	2	**0.11**	8.7
06/03/16	2	0.84	9	**0.38**	2.2
05/01/16	1	0.72	3	0.91	1.4
04/13/16	2	0.94	**21**	0.51	1.5
03/10/16	1	0.76	7	0.64	3
02/02/16	2	0.79	**45**	0.83	1.7
01/21/16	2	0.87	11	**0.37**	1.2
01/09/16	3	0.84	**26**	0.54	1.2

- *Rise CCF*: CCF score using *River Level Rise* series
- *Direct Lag*: Lag Period in hours using direct river gage-height series
- *Direct CCF*: CCF score using river gage-height series
- *Precipitation*: The amount of rain in inches

In this comparative study, the selected weather station is KBTR and the selected river-gage is Comite river at Comite. The number in bold in Table 4.1 indicates the instances where direct CCF calculations with river gage height have poor results: high lagging hours and very low CCF values. Table 4.1 shows that using *River Level Rise* provides a better quantitative output for lagging hour and CCF score value.

4.1.4 Verify CCF in Baton Rouge area

In this verification study, the selected weather station is KBTR (Baton Rouge Airport) as this is the most consistent station in the local area, with a sufficiently large history of records. In addition, Baton Rouge has experienced a number of extreme rain events in the past decade, mostly during hurricanes and thunderstorms. These

events could potentially provide more insights with high correlations and distinctive visualizations.

To make this analysis easier to follow, KBTR is selected as the weather station. The date of study is December 4th 2016, as there was totally 1.87mm of rain. The selected water gage in Baton Rouge area is at the Comite River, near Comite, Louisiana. First, time-series data of precipitation amount and water-gages height is graphed to visualize the correlation. Then, CCF is applied to find the optimal length of lagging period. This lagging period should match with the correlation found in the precipitation and water gage height time-series in Figure 4.2.

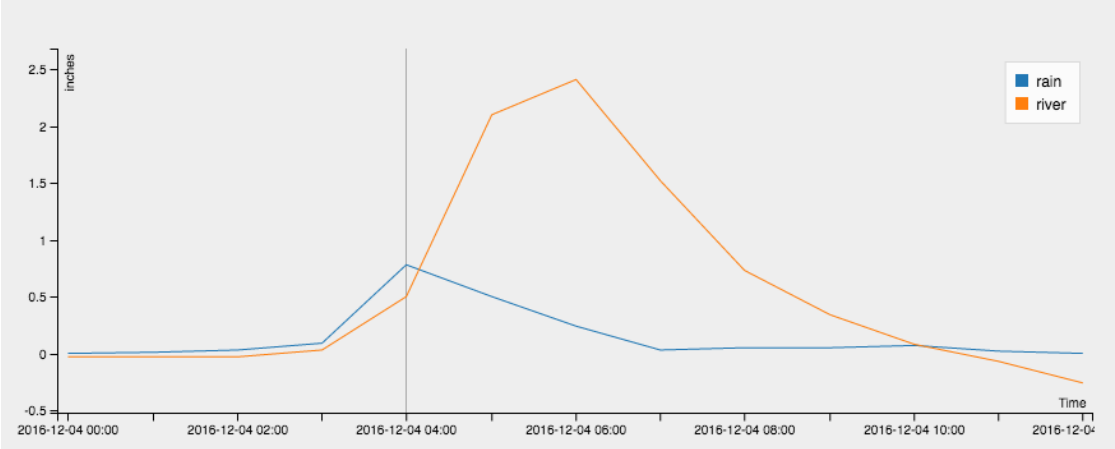

FIGURE 4.2. KBTR Precipitation & Comite River Water Level - December 4th 2016

In Figure 4.2, the blue line is precipitation amount at the given hour. The orange line is the *River Level Rise*, which shows that the river begins to rise as soon as rain event occurs. The river reaches its local maximal about 1 hour after rain event peaks.

In the CCF correlation function in Figure 4.3, the X axis shows the lag hours, and the Y axis shows the corresponding CCF correlation values. The Figure shows that correlations begin at lag 0-hour with value = 0.4. The correlations peak at lag 1 with value = 0.9. The negative section begins at 4-hour because *River Level*

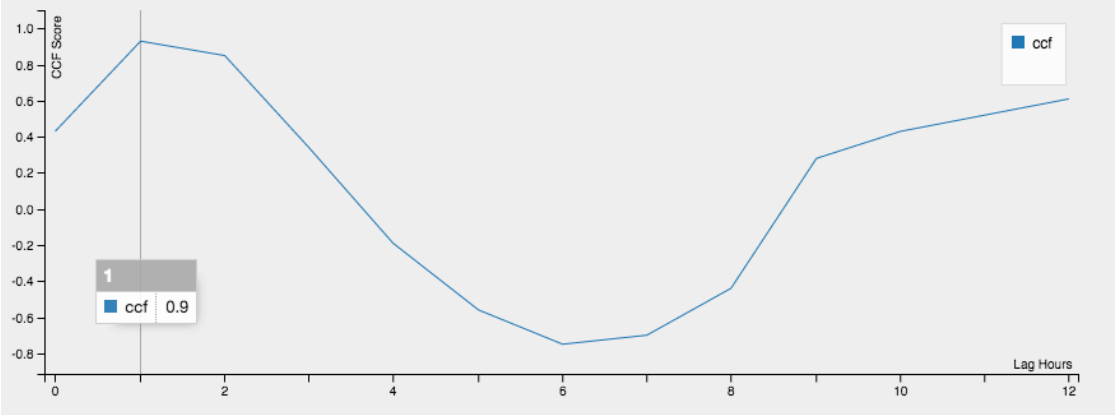

FIGURE 4.3. CCF Correlation - KBTR & Comite River - December 4th 2016

Rise series begins to recede, which is contradicting to a rain event. The positive section begins later at 9-hour can be ignored as this suggest correlation when no rain occurs, and the water is likely increasing due to other factors than rain.

From Figures 4.2 and 4.3, it is observed that CCF correlation peaks at 1-hour, which matches the river and rain plot timelines. These two figures suggest the strong correlation of KBTR weather station and Comite river-gage in a rain-event. Thus, Figures 4.2 and 4.3 verify that CCF correlation procedure is applicable in any rivers without human interventions and physical structures [3].

4.1.5 CCF Mean Score and CCF Mean Lag Hour

As the scope of this study is every rain event, which may be in hundreds over the historical years, a general score is required to represent all these events regarding the correlation level of one weather station and one river gage. Therefore, *CCF Mean Score* and *CCF Mean Lag Hour* are introduced to give a general sense of how a river gage and a weather station are correlated. Here are some used terms and their definitions:

- *CCF Mean Score*: the mean of CCF *Correlation Scores* for every rain-event
- *CCF Mean Lag Hour*: the mean of CCF *Lag Periods* for every rain-event

TABLE 4.2. CCF Score and Lag Hour - KBTR & Comite River

Time	Lag Hour	CCF Score	Precipitation
12/04/16	1	0.93	1.87
11/28/16	3	0.63	1.23
08/17/16	1	0.67	3.6
08/12/16	0	0.48	13.39
08/11/16	0	0.45	2.14
08/10/16	4	0.98	1.26
06/12/16	1	0.93	1.01
06/03/16	2	0.84	2.12
05/19/16	0	0.79	1.9
05/01/16	1	0.83	1.63

For every rain event recorded by a selected weather station, a *CCF Mean score* and *CCF Mean Lag Hour* are calculated for a selected water-gage sensor. These CCF scores vary from one rain event to another, due to the precipitation amount and potentially other hydrological factor such as temperature and soil moisture. Thus, a mean score of all these CCF scores can capture the general number of lagging hours of a river-gage in response to a rain event. This calculated mean score is stored in a MongoDB database instance for every weather station and its nearby river-gages. This score is then used for ranking correlations level as specified Section 4.1.6.

Table 4.2 lists the CCF scores between KBTR weather station and Comite River near Comite LA river gage. The score of every rain event is calculated and stored as a list. Then, a mean score is calculated from these. Accordingly, the calculations are carried out according to Equations 4.3 and 4.4:

$$\begin{aligned} CCFMeanScore =& (0.93 + 0.63 + 0.67 + 0.48 + 0.45 + 0.98 \\ & + 0.93 + 0.84 + 0.79 + 0.83)/10 = 0.75 \end{aligned} \tag{4.3}$$

$$CCFMeanLag = (1 + 3 + 1 + 0 + 0 + 4 + 1 + 2 + 0 + 1)/10 = 1.3 \tag{4.4}$$

TABLE 4.3. CCF Mean Score at KBTR

River	CCF Mean	CCF Lag Mean
Comite River, Baton Rouge, LA	0.76	2
Comite River, Comite, LA	0.75	1.92
Grays Creek, Port Vincent, LA	0.72	2.35
Beaver Bayou, Baton Rouge, LA	0.71	1.89
Comite River, Baker, LA	0.7	2.5
North Creek, Baton Rouge, LA	0.69	0.41
Amite Rive, Denham Springs, LA	0.67	3.58
Bayou Manchac, Kleinpeter, LA	0.65	2.95
Ward Creek at Essen Lane	0.63	1.4
Ward Creek at Government	0.62	0.47

4.1.6 Ranking river-gages and precipitation correlation

Pre-calculated *CCF Mean Score* is used to rank every weather station and its nearby river-gages correlation level. For faster search of nearby river-gages, these gages' locations are queried from USGS API and stored locally in a local MongoDB database. The locations and meta data of weather stations are also locally stored in the MongoDB instance. As explained in Section 4.1.5, CCF analysis are applied to every weather station and near-by water gages, to aggregate the *CCF Mean Scores* and *CCF Mean Lag Hour*. Again, *CCF Mean Score* reflects the correlation level of a river water level, in response to a rain event. Therefore, *CCF Mean Score* can also serve as a ranking measure of river-gages correlation level to a selected weather station.

For example, Table 4.3 is the ranking for river-gages around KBTR weather station. The result from Table 4.3 reflects the intuition that nearby gages are more likely to respond to a rain event. Comite River water-gages are located very near to Weather station KBTR, which is at the Baton Rouge airport. In addition, Comite River drainage basin is in the East Baton Rouge area. Thus, the score and ranking above match the intuition of how Comite River water-gages are likely to be more responsive than other water gages in the area. The ranking in Table 4.3 also shows

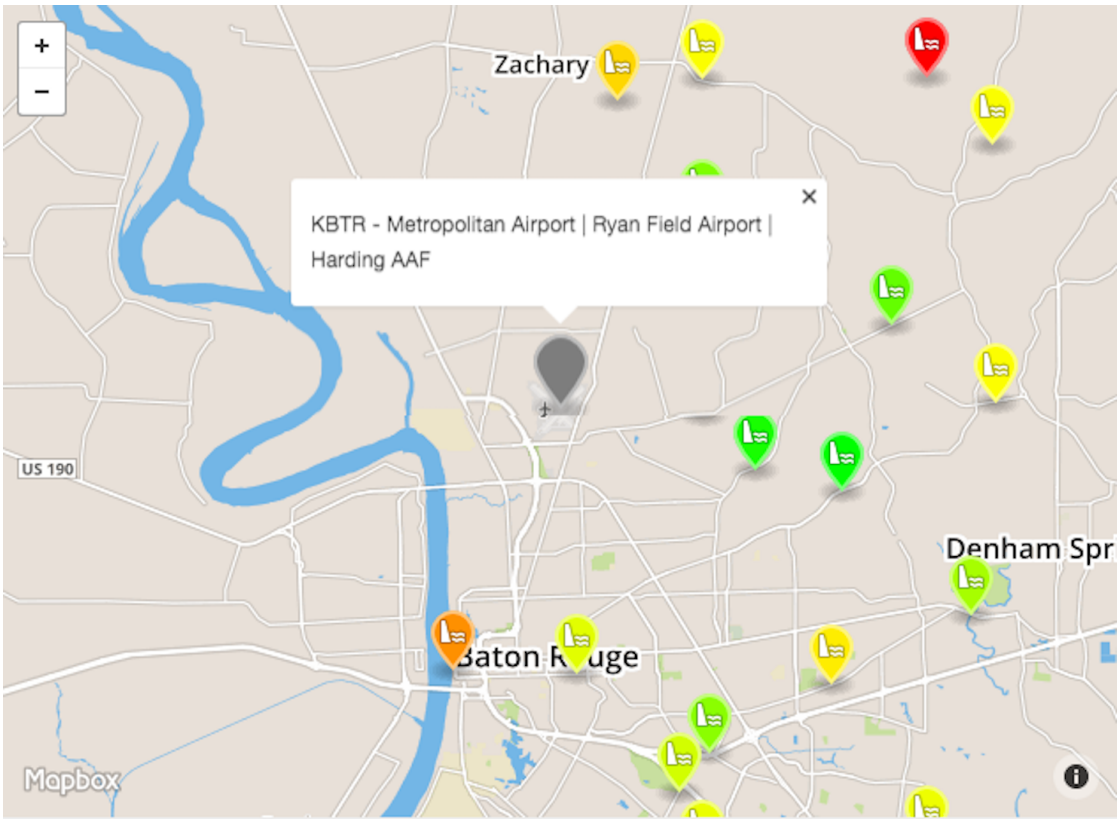

FIGURE 4.4. KBTR and local River Gages

that the Mississippi river gage is not responsive to rain-events in Baton Rouge. This is correct because of the artificial levee system that acts as a barrier on the Mississippi river. Using this ranking scheme, one can draw a cluster of responsive water-gages around a weather station. Figure 4.4 shows the locations of river gages around KBTR weather station. A web-based visualization was developed for this purpose as shown in Section 4.1.8.

4.1.7 CCF Analytics Result

CCF Analysis does not function well in the case of continuous raining, which frequently occurs in the case of extreme rain events such as during hurricanes or tropical storms. During these events, precipitation would rise and go down to a non-zero value, and then rise again. This pattern confuses CCF functions of where the peak precipitation value is, to match with the rising of river water level. Figure 4.5 is the graph of rain and river values during the historic heavy rain during 08/12/2016. In this particular event, rain was very heavy at some hours, and light at other hours. However, it was continuous as one event.

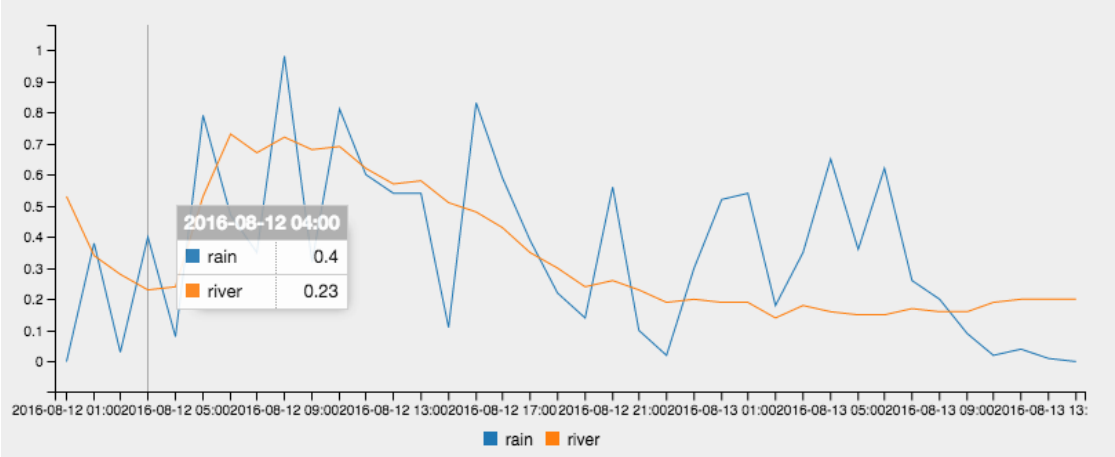

FIGURE 4.5. KBTR Precipitation & Comite Water Level - 12th August 2016

According to Figure 4.6, the CCF values were mostly negative with small lagging hours, suggesting the wrong signal of no coordinations. The chosen CCF *Lag Period* is 0, as this is the first local maxima. However, 0.46 (rounded to 0.5) is an exceptionally low *Correlation Score* for water gage at Comite River, which is geographically close to KBTR weather station and has *CCF Mean Score* at 0.75. Table 4.2 is the CCF scores for Comite River at every extreme rain events (above 1mm) in 2016 [6]. This table shows a general high correlations score and low lag period hours for the rain events. The exception was during the historical flood around 08/12/2016. During this period, rain occurred continuously in hours and water level stayed high for a long period of time, leading to a low score for finding the causality lag hours.

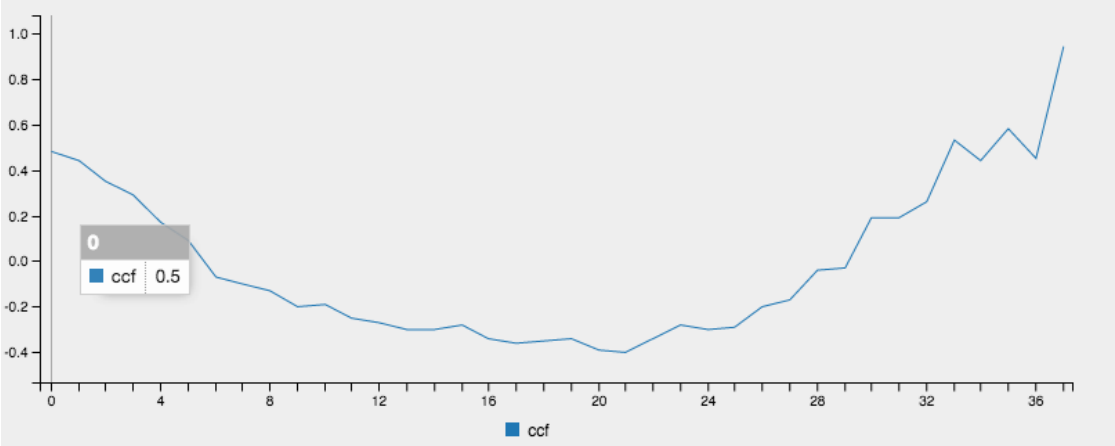

FIGURE 4.6. CCF KBTR & Comite River - 12th August 2016

4.1.8 Visualization on CCF Analysis Findings

A web application has been built to visualize the CCF analysis correlations and Machine Learning Prediction (Section 4.2). This web application is available for public use at: *http://rainriver.lsu.edu* [31]. Each of the visualization tool is in a separate page with navigation on top as: CCF, River Prediction. As shown in Figure 4.7, clicking on the navigation tab will lead users to a visualization page

that allows more in-depth view of the study between a weather station and a river gage. This section will explain in details the visualization tools built for CCF Analytics Results.

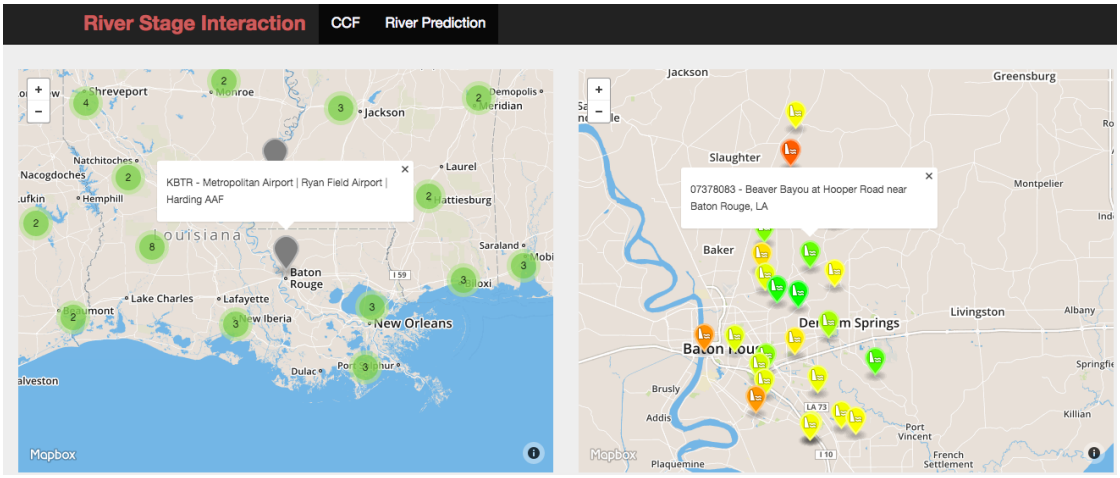

FIGURE 4.7. River State Interactive Visualization

- **4.1.8.1 Weather Station and River Gage Selection**

The first view has two maps positioned on left and right. In Figure 4.8, the left map is all active weather stations for initial selection. The right map is the river gages associated to a weather station, that must be selected on the left map. For easier and faster view in the left map, these weather stations are clustered in number. The clustering label number and size represent the number of stations under that cluster. Users can click on clusters (green icon) to zoom in and eventually find a weather station of interest. Upon selected, each station will display its name and its national code name. In addition, the right map will coordinate position and display the river water gages around that selected weather station.

Each water gage marker is colored to signify the correlation level with the selected weather station. The color distribution ranges from: red (almost no correlation) to orange (very low correlation), to yellow (low correlation) and finally to green (high correlation).

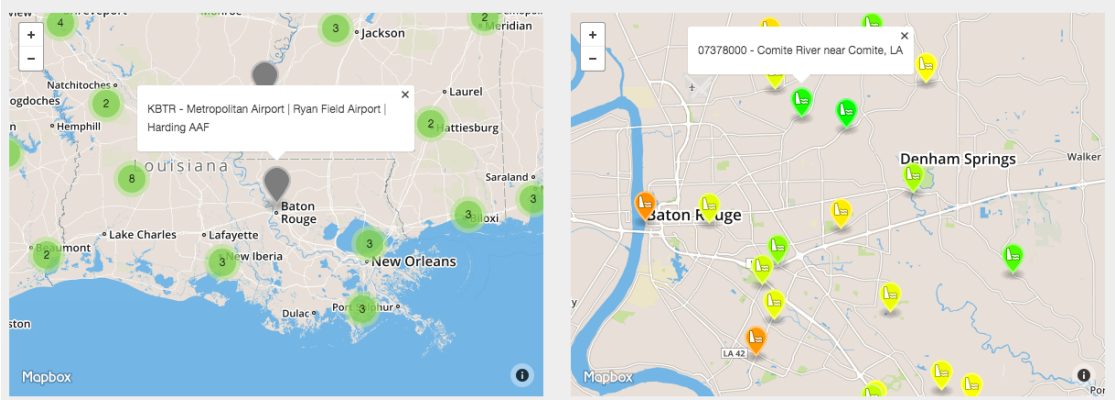

FIGURE 4.8. Two Maps - Selection Weather Station & River Gages

Figure 4.9 shows the markers of river gages around KBTR weather station. For example, Mississippi water gage is in orange color, as the correlation level is very low due to a levee on the river. Whereas Comite Rivers in Baker shows green color, indicating a high correlation and perhaps a higher predictability as later found in Section 4.2. As a river water-gage is selected, a graph is shown visualizing the model accuracy and prediction of future water-gage level.

- **4.1.8.2 CCF Correlation Graphs**

CCF visualization tables and graphs illustrate and support the scores provided by the CCF function. The table shows all the analyzed rain-event along with its recorded precipitation, maximum CCF score and lag hour.

Upon selecting a rain-event in a table list (Figure 4.10), two graphs are displayed. The top graph (Figure 4.11) draws the rain and river recorded measurements during the selected rain-event. The bottom graph (Figure 4.12) shows the CCF score in different lag hours.

Figure 4.11 and 4.12 are both visualizing the rain event in December 4th 2016 at Comite River, Baker LA. In Figure 4.11, the blue line represents the rain measurement, whereas the river slope is in the orange line. During this event, the rain

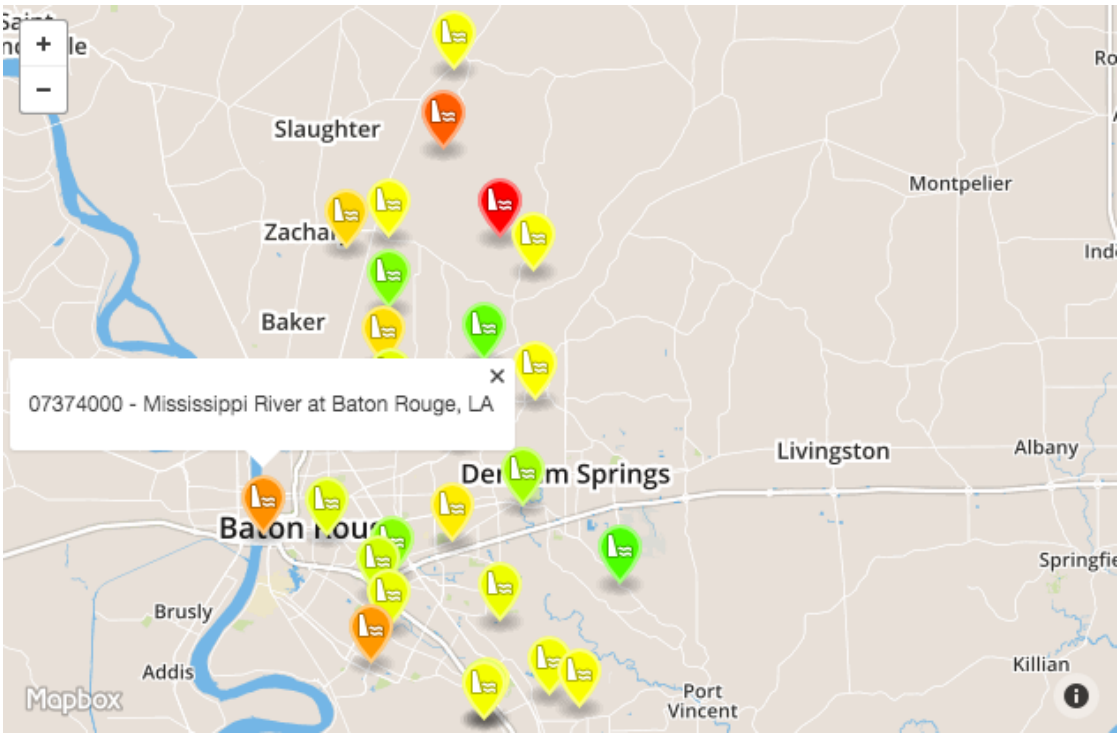

FIGURE 4.9. CCF Analytics Water Gage Colors Distribution around Baton Rouge

	Time	Lag Hour	CCF Score	Precipitation
☑	2016-12-04 00:00	2	0.94	1.87
☐	2016-11-28 22:00	0	0.92	1.23
☐	2016-05-01 07:00	1	0.95	1.63
☐	2016-04-13 04:00	14	0.76	1.54
☐	2016-03-10 21:00	1	0.61	2.87
☐	2016-03-10 11:00	2	0.82	2.98
☐	2015-05-26 02:00	1	0.49	1.83
☐	2015-01-03 19:00	3	0.81	1.57

FIGURE 4.10. Rain River Normalized Measurement

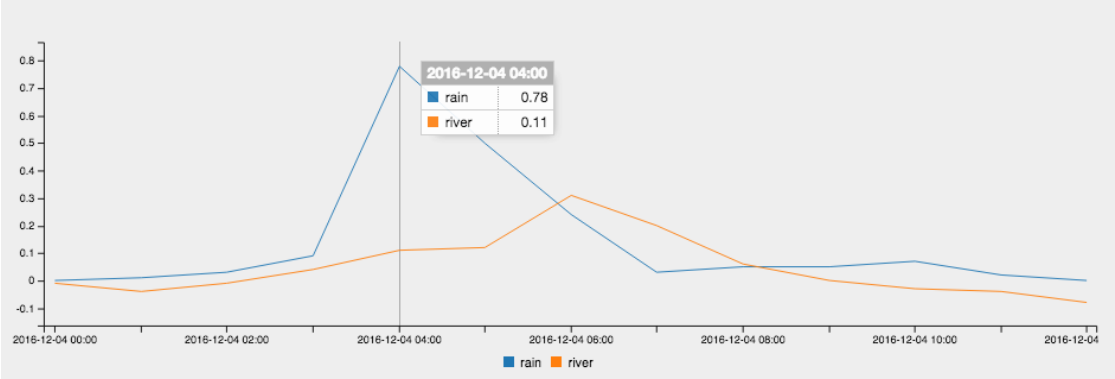

FIGURE 4.11. Rain River Normalized Measurement

poured heaviest at 4A.M. The river began to rise during the rain event, but rose most sharply at 6A.M. This is the correlation that CCF analysis expects to find.

Figure 4.12 shows a graph of CCF score value at different lag hours from 0 to 48. This CCF score reaches its maximum at 2, which is the same number of hours of difference from 4AM to 6AM. Therefore, this particular rain-event approves the CCF Analysis and its score. This two-graphs visualization comparison allows end-users to easily verify the CCF Analysis study across all weather stations and river gages.

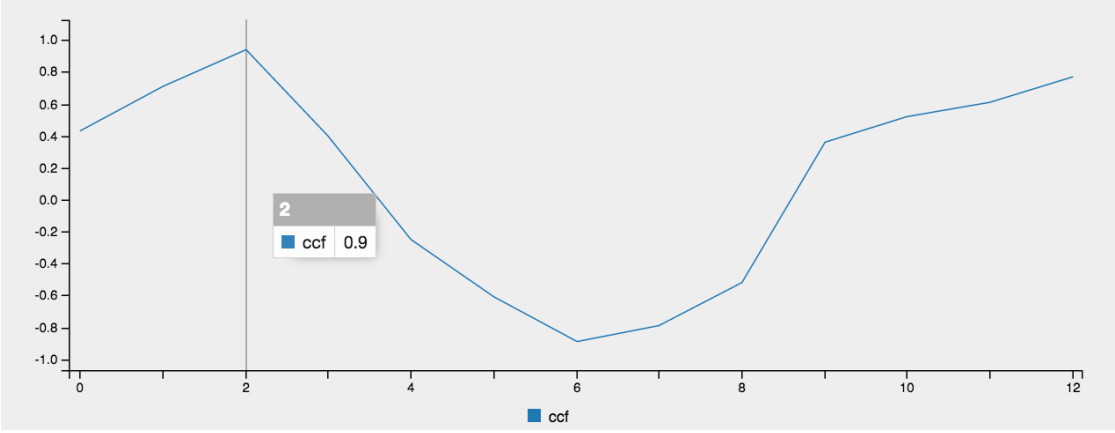

FIGURE 4.12. CCF Lagging Hour

4.2 Predictive Analytics using Machine Learning

The next question that this project attempts to answer is forecasting river water peak after a rain event. The CCF study in Section 4.1 provides a statistical proof that rain events that are very likely to cause water-rising events. How high will the water rise is a different question that machine learning models will be a great fit to answer.

4.2.1 Machine Learning: Procedure and Data Preparation

Machine Learning (ML) is a broad study that encompasses numerous areas such as classification, clustering, dimension reduction and regression. ML Regressor models are the main focus of this project, as the goal is predicting a quantitative value: peak amount of rain. The accuracy output of these models are stored in database for further analytics and optimizations.

Applying Machine Learning (ML) models follow two stages: training and testing. A ML model is first trained with one set of data, and later tested with a separate set of data to observe its level of accuracy [28]. In the training process, a model will be given a dataset of multiple features, optimizing to best predict a column of values. In this study, the predicting value is the maximum water-level that a river will reach during a period of P hours. Following Figure 4.13, this P hours includes an extended period after rain (X hours), ranging from 4 to 48 hours. This X hour is used to study which time-period is most predictable. The training dataset in this project has five features: total precipitation amount, beginning river water level, last rain period, average temperature and river flow rate. The details of these features are listed in Section 4.2.1.1

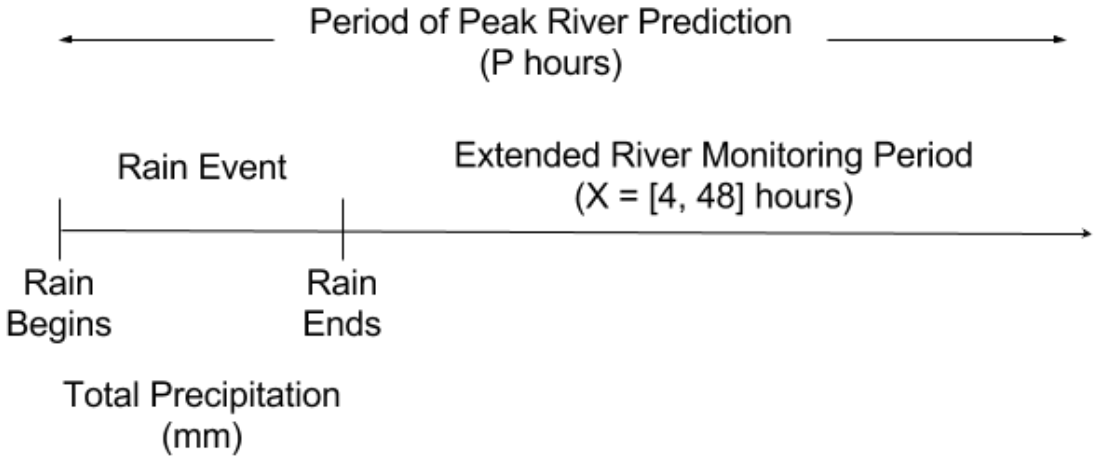

FIGURE 4.13. Peak River Prediction

- **4.2.1.1 Training Features**

ML rely on a set of features (columns) that influence the output column to train and produce a best fit model. Previous studies done in [19] and [3] already suggested a number of influential attributes to water-level. These verified attributes are listed in Section 2.1.1 and 2.1.2. This project only includes a subset of those recommended attributes, which are time-series and made available online. These selected attributes are put in details as below:

- Total Precipitation: this is the total precipitation input of a continuous rain event. This is likely the most essential feature as the overall model relies on the CCF hypothesis: precipitation input will eventually cause an increase in river water gage level, the predicting column value.
- Current Water Level: this is the water level at the same hour before it begins to rain.
- River Flow Rate: this is the flow speed of water current. The higher the flow rate, the more likely water will escape in pouring downstream, potentially leading to a lower peak water level.

- Dry Periods: this is the period in number of hours from the last rain fall. This feature suggests the ground condition as longer dry periods may lead to drier ground. Figure 4.14 explains how this period is calculated.

- Mean Temperature: this is the mean temperature of the whole dry period. A higher temperature will likely lead to higher evaporation rate and lower soil moisture.

- Mean Dew Point: this is the mean dew point temperature of the whole dry period. Dew point is the temperature at which water vapor forms liquid [40]. Thus, dew point is a related measurement for air humidity. A higher dew point indicates higher soil moisture.

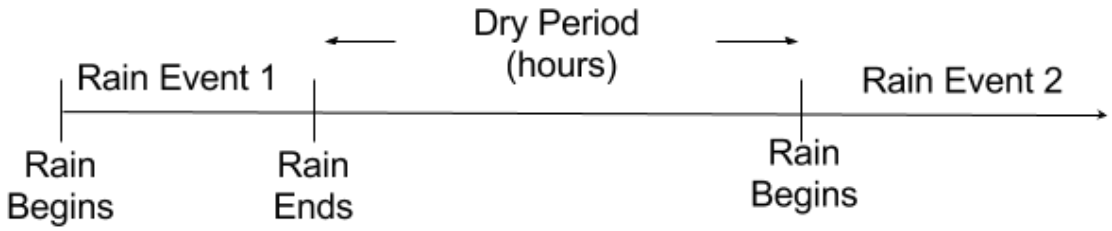

FIGURE 4.14. Last Rain Period

- **4.2.1.2 Linear Regression**

Linear Regression Model optimizes coefficients of a linear function. For every feature column, a coefficient *Wi* is assigned. *Linear Regression* seeks to minimize this difference function

$$||Xw - Y||^2 \tag{4.5}$$

The output that *Linear Regression* model gives is the minimized output of function 4.5. In addition, the coefficient shows the significance of each feature column, based on which some features can be disregarded to improve the model precision.

TABLE 4.4. KBTR Linear Regression Coefficients and Results

River	Rain	River	Dry	Temp	Dewpt	Flow	Score
Mississippi River	0.09	1.00	0.0000	-0.01	0.00	0.0000	1.00
Amite, Magnolia	0.87	1.01	0.0002	-0.06	0.05		0.94
Bluff Swamp	0.20	0.97	-0.0001	-0.01	0.01		0.93
Bayou Manchac	1.04	0.91	-0.0003	-0.04	0.03		0.91
Amite, Denham	2.27	1.02	0.0003	-0.11	0.10	-0.0001	0.89
Comite, Baton Rouge,	2.75	1.01	0.0003	-0.04	0.01		0.88
Comite, Comite	3.86	1.17	0.0006	-0.14	0.11	-0.0006	0.79
Comite, Olive	0.94	1.25	0.0002	-0.05	0.03	-0.0003	0.78
Grays Creek	1.45	0.87	-0.0001	-0.06	0.03		0.74
Comite, Greenwell	1.02	0.98	-0.0002	-0.13	0.17		0.73
Comite, Comite Dr	1.38	0.99	-0.0002	-0.04	0.05		0.70
Little Sandy Creek	0.34	1.14	0.0026	-0.11	0.18		0.65
North Branch Ward Creek	1.65	0.77	0.0016	-0.03	0.02		0.56
Bayou Fountain	0.34	0.95	-0.0001	-0.04	0.05		0.55
Sandy Creek	1.69	1.01	0.0004	-0.13	0.09		0.55
Beaver Bayou	2.08	0.85	-0.0001	-0.08	0.06		0.48
Comite, Baker	1.59	1.01	0.0000	-0.04	0.03		0.38
Ward Creek, Essen	1.05	0.76	0.0018	-0.06	0.12		0.36
Ward Creek, Government	1.57	0.66	0.0000	-0.02	0.02		0.35

Table 4.4 shows the *Linear Regression* prediction accuracy of different rivers around Baton Rouge area. The table is sorted in descending score order. There are six column features, in which the values are coefficient in the Linear Regression model:

- Rain: precipitation input.
- River: beginning river water level.
- Dry: dry period in hours.
- Temp: average temperature over the dry period.
- Dewpt: average dew point over the dry period.
- Flow: flow rate.

The first and second water gages at Mississippi and Alligator Bayou have perfect scores due to levees. As a result, precipitation does not immediately affect the water level, leading to un-change water level. *Linear Regression* fits this by adjusting a very high coefficient to *Beginning River* water level, leading to a perfect fit of 1. Nonetheless, other rivers are subjected to water rising, and thus the prediction scores are more relevant. The sign value of each feature shows how the feature may affect the water peak level.

- Rain: the precipitation input is always a plus feature, as this determines the amount of water being added in to the river.
- River: the beginning water level helps projecting the final peak value of the river in the near future. It is a plus feature as water is almost always going to rise.
- Dry: the dry period should be a negative feature. The longer the dry period, the drier the ground, leading to more soil absorption of water and lower amount of water drained to the rivers.
- Temp: the average temperature is also a negative feature, with similar explanation to the dry period. Higher temperature leads to lower soil moisture and higher evaporation rate. These two factors lead to lower peak water level.
- Dewpt: the average dew point is a positive feature. A higher dew point means more moisture in the air [40]. Thus, a higher dew point means higher soil moisture, leading to higher amount of water pouring to rivers.
- Flow: this is a negative feature, as a higher flow rate leads to more water able to flow downstream, and a lower peak water level. Flow rate is potentially an influential index, as it is directly related to river physical condition. It is

TABLE 4.5. Features Comparison - Linear Regression - KBTR

River	Dry	Temp	Dewpt	Flow	All
Mississippi River	1.00	1.00	1.00	1.00	1.00
Amite, Magnolia	0.94	0.94	0.94	0.94	0.94
Bluff Swamp	0.90	0.94	0.94	0.91	0.93
Bayou Manchac	0.91	0.91	0.91	0.91	0.91
Amite, Denham	0.92	0.92	0.92	0.88	0.89
Comite, Baton Rouge,	0.87	0.87	0.87	0.85	0.88
Comite, Comite	0.81	0.81	0.81	0.79	0.79
Comite, Olive	0.82	0.82	0.81	0.78	0.78
Grays Creek	0.70	0.74	0.73	0.71	0.74
Comite, Greenwell	**0.78**	0.71	0.69	**0.78**	0.73
Comite, Comite Dr	0.66	0.56	0.58	0.65	**0.70**
Little Sandy Creek	0.64	0.55	0.63	0.67	0.65
North Branch Ward Creek	0.56	0.56	0.56	0.56	0.56
Bayou Fountain	0.50	0.39	0.40	0.47	0.55
Sandy Creek	**0.71**	0.60	0.64	**0.70**	0.55
Beaver Bayou	0.46	0.48	0.47	0.46	0.48
Comite, Baker	0.37	0.34	0.34	0.37	0.38
Ward Creek, Essen	0.40	0.31	0.36	0.40	0.36
Ward Creek, Government	0.35	0.34	0.34	0.34	0.35

unfortunate that only a fraction of river gages have sensors for this measurement. With flow rate data, the accuracy may significantly improve, such as in the case of Comite River at Comite, LA (Table 4.5)

Among the above six features, the first two (*Rain* and *River*) are primary features, contributing to the majority of model accuracy. The subsequent features are optional and subjected to further study for their influential level. A further comparative study is done among the optional features to see the improvements that each feature can add in.

Table 4.5 shows the comparison of *Linear Regression* optional features. The default features are *River* (beginning river water level) and *Rain* (precipitation input). The subsequent columns is the score using that additional column feature. For example, column *Dry* is included in the model using three features: *River*, *Rain*

and *Dry* (dry period in hours). A number significant individual feature improvements are observed and marked as bold. For example Comite, Greenwell shows 18.5% improvement using either *Dry* or *Flow*. This shows that some rivers are influenced more by *Flow* and *Dry* than others.

- **4.2.1.3 Lasso Regression**

  Lasso is a linear model that enhances solutions with few feature columns [28]. Using Lasso gives the advantage of built-in preference over a small subset of features [35]. In this project, this small subset of features are: initial river level and precipitation amount. The other features such as temperature, flow rate and dew point will largely be ignored. Thus, using Lasso model brings evaluation of primary features importance.

- **4.2.1.4 Kernel Ridge Regression**

  Kernel Ridge Regression is based on Ridge Regression, which supports data regularization [42]. The *Kernel* used in this project is a Radial Basis Function (RBF). Regularization in *Ridge Regression* uses a decaying weight coefficient to avoid over-fitting for the function [42]. Using *Kernel Ridge* gives a point of reference of how important over-fitting prevention is to modeling water gage height.

- **4.2.1.5 Random Forest**

  Random Forest is an ensemble method, which contains a number of *Decision Trees* [18]. Each decision tree is trained with a randomized subset of training data. The output of *Random Forest* is the average of these sub *Decision Tree Regressions* [18]. *Decision Tree Regression* operates very similar to *Decision Tree*. It builds a tree of output values (decisions) based on ranges of feature column values. Unlike *Linear Regression*, *Decision Tree* is non-linear and the output is a cumulative average of possible decision paths [18].

4.2.2 Results from the Machine-Learning Models

Machine Learning models such as Random Forest, Linear Regression and Kernel Ridge Regression are applied to predict the peak water-level in response to a rain event. The next section is general outlines regarding each machine learning method and how it fits to use in this study. Every model is first trained with the defined above set of features: total amount of rain, the starting amount of water, along with other features such as temperature and water flow rate. Then, these models perform predictions on the peak water level in the next X hours after rain has stopped.

- #### 4.2.2.1 Models Comparison at KBTR Baton Rouge

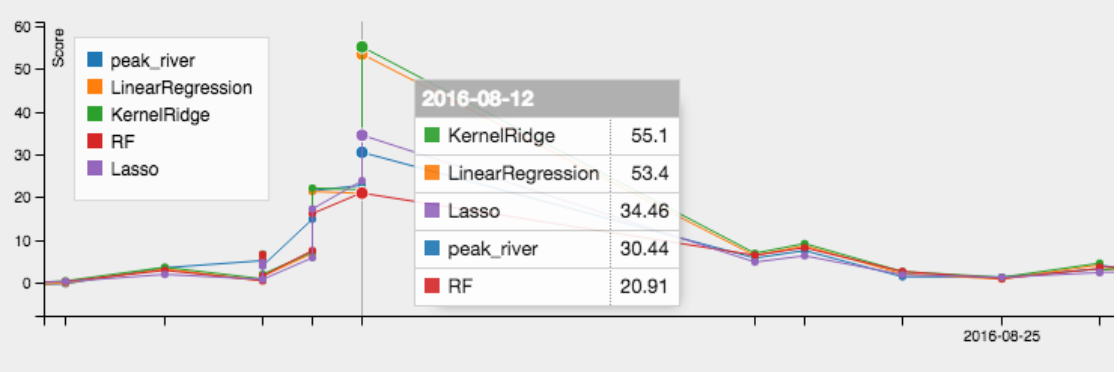

FIGURE 4.15. KBTR Comite - ML Models - Aug 12th 2016

A comparative study is done across various types of machine learning models. These models are trained and tested with the same dataset. The training data is 80% of rain events, and the testing data is the rest 20%. The training column features for all the models are 6 features: *Rain*, *River*, *Dry*, *Temp*, *Dewpt* and *Flow*. Table 4.6 is the table of prediction accuracy at KBTR weather station. According to this result, *Linear Regression* is a sufficiently good model for the majority of rivers. However, *Linear Regression* is outperformed in a few rivers using certain models. These cases are marked with bold: Comite at Baton Rouge using *Lasso*,

TABLE 4.6. ML Models Comparison at KBTR

River	Linear Regression	Kernel Ridge	Random Forest	Lasso
Mississippi River	1.00	1.00	1.00	0.99
Amite, Magnolia	0.94	0.94	0.81	0.84
Bluff Swamp	0.93	0.92	0.88	0.54
Bayou Manchac	0.91	0.91	0.91	0.85
Amite, Denham	0.89	0.88	0.88	0.79
Comite, Baton Rouge	0.88	0.87	0.81	**0.90**
Comite, Comite	0.79	0.78	**0.84**	0.77
Comite, Olive	0.78	0.78	0.67	0.53
Grays Creek	0.74	0.68	0.76	0.67
Comite, Greenwell	0.73	0.73	0.76	0.66
Comite, Comite Dr	0.70	**0.73**	0.48	0.72
Little Sandy Creek	0.65	0.50	**0.78**	0.67
North Branch Ward Creek	0.56	0.54	0.54	0.31
Bayou Fountain	0.55	0.54	0.35	0.55
Sandy Creek	0.55	**0.61**	0.49	0.42
Beaver Bayou	0.48	0.43	**0.71**	0.66
Comite, Baker	0.38	**0.42**	0.07	0.12
Ward Creek, Essen	0.36	0.38	0.16	0.31
Ward Creek, Government	0.35	0.33	0.32	0.15

Comite at Comite using *Random Forest* and a few other cases. This shows that using the right ML model for a river could bring significant accuracy improvement, as these models consider input attributes differently. It will be beneficial to verify how this model improvement in efficiency aligns with physical properties of the river basin and areas. For example *Random Forest* has the accuracy of 0.71 with *Beaver Bayou*, whereas other models were having lower accuracies around 0.5. This 50% difference in accuracy is an indicator of a hydrological physical attribute, which *Random Forest* gives a higher considerations than others. Very similar to how *Mississippi* river accuracy is always 100%, these anomalies in prediction outcomes are potential keys to a physical factors that are primarily contributing to river water level behavior.

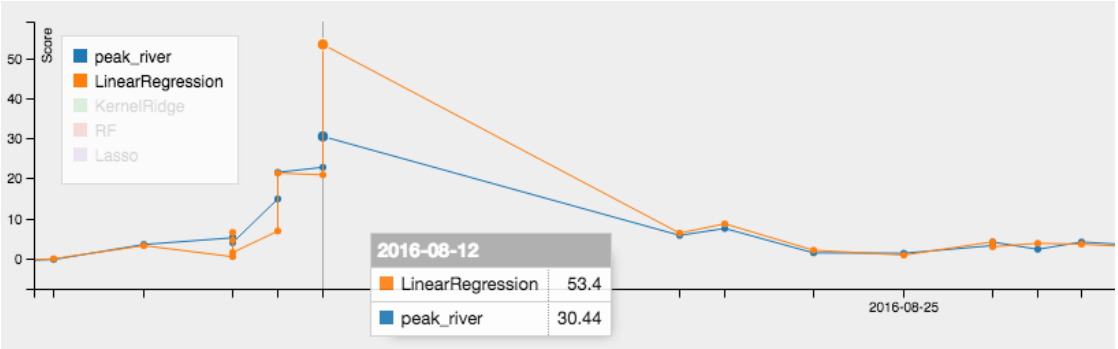

FIGURE 4.16. KBTR Comite - Linear Regression - Aug 12th 2016

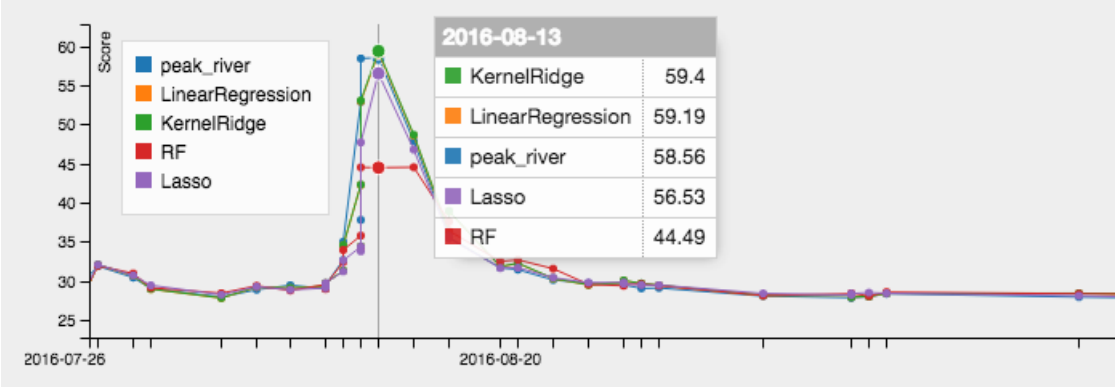

FIGURE 4.17. KBTR Comite - ML Models - Aug 12th 2016

Figures 4.16 and 4.15 show ML predictions of River Gage Height around the historical flood period in August 12th 2016 at Comite River. This is a precipitation event that CCF also gives low correlation score. Figure 4.16 shows that *Linear Regression* was able to accurately predict other rain events before and after the flood periods, but not so much during the flood. *Linear Regression* as well as other models shown in Figure 4.15 over-predicts the peak water level in flood event on August 12th by a wide margin.

However, rivers with high overall prediction accuracy are more predictable during extreme events. For instance, ML models perform well on predicting Amite river, even during flood periods. Amite river has an average accuracy of 0.94 whereas Comite is 0.79. Figure 4.17 depicts Amite River, Magnolia during the same historical flood period. Unlike the outcome at Comite river, at Amite river ML models were able to predict river gage height with low error.

- **4.2.2.2 Models Accuracy in San Antonio, TX and Nashville, TN**

Two other cities outside Baton Rouge, LA were selected to observe ML models's performance: San Antonio, TX and Nashville, TN. These two cities are picked as they have numerous local rivers and creeks, and have had a sufficient historical climate records . For each city, the accuracy table is listed, followed by graphs of ML models predictions.

Table 4.7 lists the prediction accuracy using ML for rivers around San Antonio. This table shows that ML models do generally well in predicting river gage height in San Antonio area rivers. Figure 4.19 shows the Linear Regression predictions for San Antonio River in late 2016, including an extreme rainfall (3.13 mm) on Dec 3rd 2016 [6]. The figure approves that ML models was giving good predictions, even in the extreme rain event.

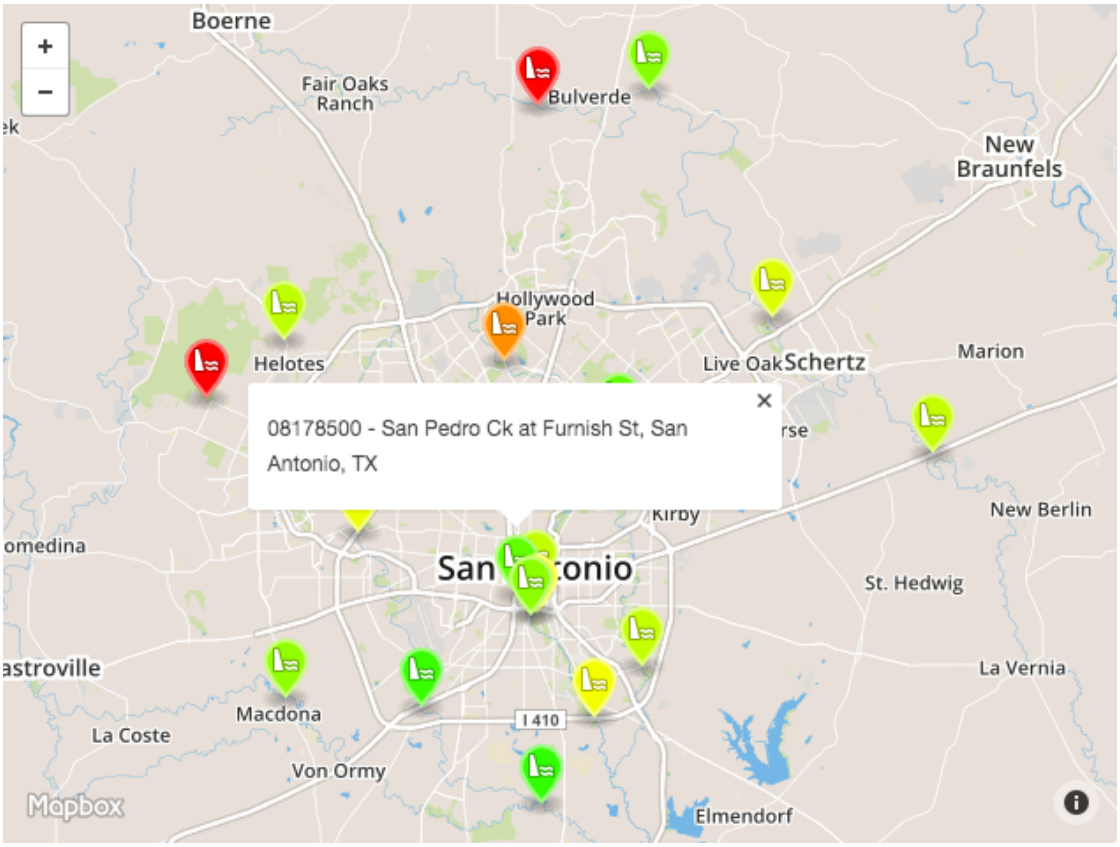

FIGURE 4.18. San Antonio River Gages

TABLE 4.7. ML Models Accuracy at San Antonio Rivers

River	Linear Regression	Kernel Ridge	Random Forest	Lasso
Medina River	0.95	0.95	0.84	0.68
Leon Creek	0.94	0.93	0.67	0.47
Salado Creek	0.89	0.88	0.73	0.18
San Pedro Creek	0.89	0.89	0.44	0.07
Cibolo Creek	0.84	0.87	0.83	-0.16
San Pedro Creek, Probandt	0.84	0.53	0.92	0.01
Olmos Creek	0.83	0.85	0.41	0.02
Medina River, Macdona	0.83	0.82	0.42	0.3
Helotes Creek	0.75	0.75	0.77	0.04
San Antonio River	0.74	0.74	0.62	0.21
Salado Creek	0.74	0.75	0.68	0.43
Cibolo Creek	0.74	0.74	0.7	0.38
Cibolo Creek, Selma	0.68	0.69	0.59	0.24
Leon Creek, Loop 410	0.63	0.58	0.53	0.34
San Antonio River, Loop 410	0.6	0.62	0.49	0.05
San Antonio River, Mitchell	0.58	0.58	0.5	0.27

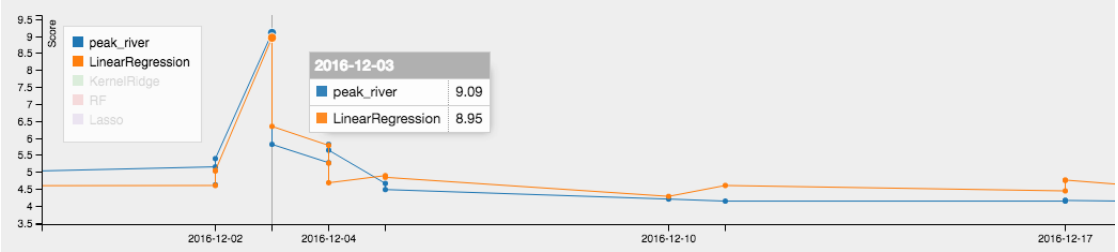

FIGURE 4.19. San Antonio River ML Accuracy

TABLE 4.8. ML Models Accuracy at Nashville, TN Rivers

River	Linear Regression	Kernel Ridge	Random Forest	Lasso
Cumberland, Hermitage	0.99	0.99	0.98	0.98
Cumberland, Nashville	0.98	0.98	0.98	0.94
Cumberland, Omohundro	0.98	0.98	0.98	0.97
Cumberland, Briley	0.98	0.98	0.97	0.98
Cumberland, Bordeaux	0.98	0.98	0.98	0.96
Cumberland, Cockrill	0.98	0.98	0.97	0.95
Cumberland, Edenwold	0.97	0.97	0.96	0.97
Cumberland, Woodland	0.93	0.93	0.81	0.87
Stones River	0.9	0.89	0.87	0.88
Harpeth River, Franklin	0.78	0.78	0.74	0.66

Table 4.8 lists the ML models accuracy at rivers in Nashville, TN area. Since it rains more frequently in Nashville, ML models have more training data; thus these Nashville rivers generally have higher accuracies than those at San Antonio.

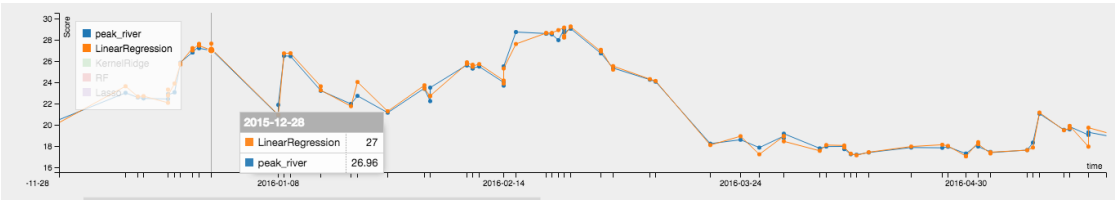

FIGURE 4.20. Nashville River ML Accuracy

Figure 4.20 displays the high accuracy of predicting Cumberland River using Linear Regression. The Cumberland river runs through the city of Nashville. Therefore, a precise prediction of river water level has a significant impact on flood prevention and forecasting for the city. Nashville and San Antonio are a good ex-

amples of how cities can use ML Models to gain quick and accurate predictions of river gages.

- **4.2.2.3 States Comparison**

Average scores are aggregated across 50 states to find the generally best model as shown in Table 4.9. The *Count* column is the number of weather station and river gage pairs that were included for the score aggregation process. *Linear Regression* and other columns show the average scores of corresponding ML models in each state. These averaged scores are aggregated using MongoDB Aggregation pipeline as detailed in Listing 4.1

```
aggregate([
  {$match: {'LinearRegression.score':{'$gt': 0, '$lt': 0.99}}},
  {$group: {
        _id: "$state",
        LinearRegression: {$avg: "$LinearRegression.score"}}}
])
```

Listing 4.1. MongoDB Aggregation for Model Score

The negative and perfect scores are considered invalid and filtered out. Negative score may occur due to lack of data or wrong input data. Perfect scores usually occur due to physical structures that cause water to always remain stable. For example, the levee in Mississippi river allow prediction accuracy to be perfectly 100%. However, this score, if included, does not fairly measure the performance of a ML model, as it simply gives the output the same as beginning river water level. Statistics in Table 4.9 give a general summary of ML Models performance. In addition, this table reveals states with more river gages and weather stations such as Florida (FL), California (CA), Illinois (IL) and Texas (TX).

4.2.3 Visualization ML Prediction

The weather station and river gage selection procedure and interaction are identical to that of the CCF Visualization [31]. However, the river gages coloring is now indicating the prediction accuracy using *Linear Regression* model. Again, the

TABLE 4.9. ML Average Score across U.S.A States

State	Count	Linear Regression	Kernel Ridge	Random Forest	Lasso
AK	19	0.82	0.83	0.86	0.62
AL	108	0.76	0.75	0.77	0.62
AR	100	0.74	0.73	0.67	0.51
AZ	74	0.65	0.66	0.55	0.35
CA	483	0.66	0.66	0.61	0.37
CO	81	0.64	0.69	0.48	0.47
CT	95	0.81	0.82	0.79	0.48
DC	22	0.65	0.59	0.53	0.37
DE	27	0.66	0.64	0.55	0.37
FL	601	0.83	0.83	0.77	0.57
GA	207	0.69	0.71	0.62	0.47
HI	58	0.50	0.48	0.50	0.35
IA	47	0.72	0.73	0.68	0.62
ID	46	0.74	0.70	0.47	0.57
IL	321	0.77	0.77	0.73	0.55
IN	115	0.82	0.80	0.77	0.62
KS	134	0.77	0.76	0.74	0.57
KY	102	0.66	0.64	0.64	0.53
LA	121	0.77	0.79	0.77	0.67
MA	221	0.84	0.85	0.81	0.61
MD	120	0.59	0.61	0.56	0.36
ME	8	0.83	0.83	0.83	0.61
MI	53	0.74	0.75	0.56	0.58
MN	46	0.85	0.76	0.77	0.76
MO	167	0.69	0.70	0.67	0.48
MS	56	0.75	0.71	0.67	0.61
MT	16	0.81	0.88	0.63	0.67
NC	347	0.70	0.71	0.62	0.41
ND	18	0.85	0.77	0.89	0.66
NE	32	0.84	0.77	0.51	0.58
NH	52	0.85	0.88	0.85	0.64
NJ	234	0.76	0.76	0.69	0.43
NM	124	0.77	0.77	0.70	0.45
NV	43	0.66	0.64	0.58	0.40
NY	84	0.67	0.63	0.57	0.41
OH	223	0.60	0.54	0.48	0.38
OK	51	0.77	0.77	0.68	0.53
OR	119	0.75	0.73	0.58	0.51
PA	241	0.71	0.69	0.63	0.46
RI	34	0.84	0.84	0.81	0.46
SC	128	0.70	0.74	0.66	0.52
SD	31	0.69	0.69	0.59	0.55
TN	56	0.73	0.68	0.68	0.59
TX	362	0.62	0.62	0.58	0.45
UT	40	0.76	0.77	0.71	0.58
VA	128	0.62	0.58	0.56	0.43
VT	28	0.81	0.79	0.74	0.53
WA	184	0.82	0.84	0.79	0.61
WI	71	0.73	0.74	0.59	0.70
WV	77	0.69	0.68	0.58	0.48
WY	11	0.88	0.77	0.56	0.69

web application interactive visualization can be accessed at *http://rainriver.lsu.edu* [31].

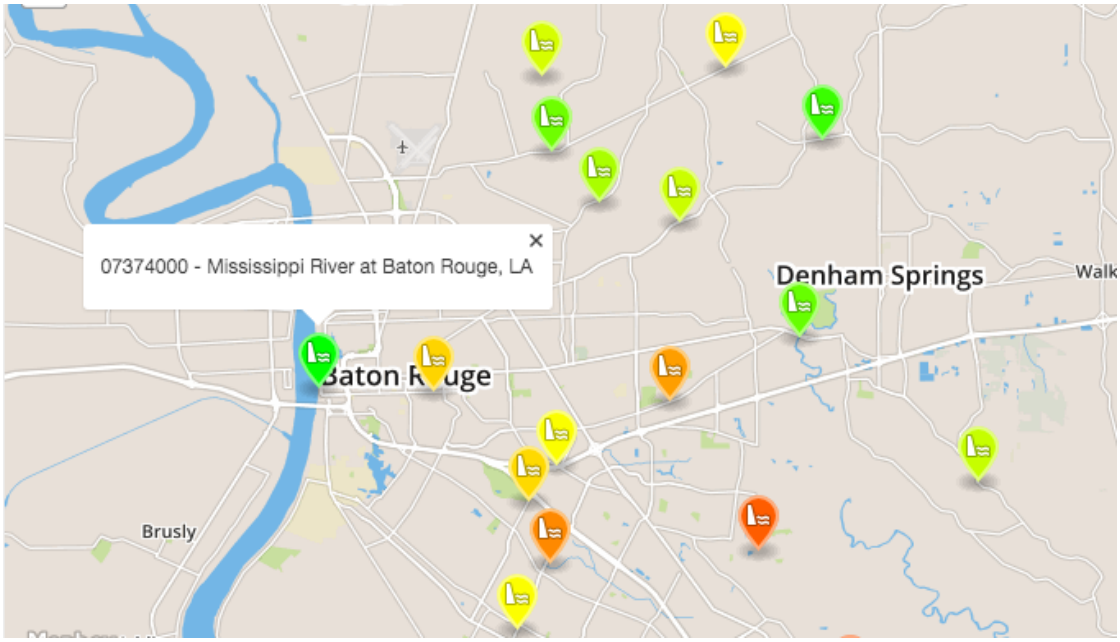

FIGURE 4.21. Machine Learning Visualization Gage Colors

- **4.2.3.1 Weather Station and River Gages Selection**

Figure 4.21 is an example of river gages markers around KBTR weather station. Upon selecting a river gage marker, a time-series graph is displayed showing the predictions using different Machine Learning models, against the actual recorded measurements.

- **4.2.3.2 Predictive Analysis Graph**

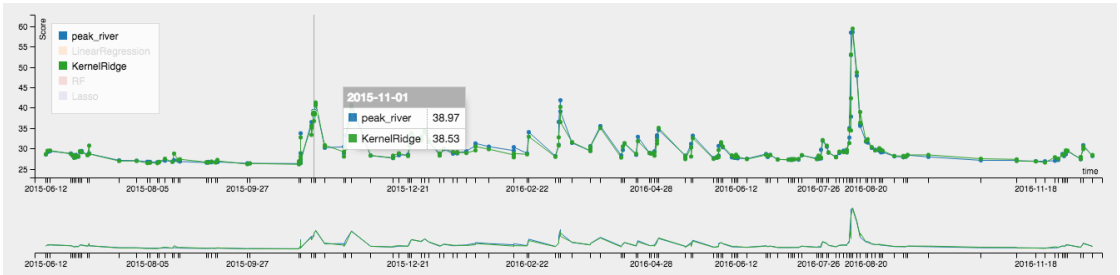

FIGURE 4.22. Kernel Ridge Prediction

The Prediction Analysis Graph generally has 3 lines: rain-amount, water-level and predicted water-level. This graph allow visualization of the causal effect of rain and water-rising. In addition, it shows the accuracy level of our predictive model. It works as a verification tools, as some river-gages provides incorrect or missing data. The visualization system allows end-users to truly see if rain events affected the river-gage water level in a repeated, predictable pattern.

Figure 4.22 shows the prediction using Kernel Ridge machine learning (green line), against the actual peak river water level (blue line). The river gage is Amite River at Magnolia, LA, and the graph represents all the rain-events during 2015 and 2016.

Chapter 5
Conclusions

Overall, this project has answered two questions. The first answered question is finding if there is a causality relationship between a precipitation recorded at a weather station, and a rise in water level as measured at a nearby river gage sensor. If this correlation exists, the second answered question is whether Machine Learning is applicable to predict the peak water gage-height, given a precipitation event.

5.1 Cross Correlation and Machine Learning Results

Across the U.S, Cross Correlation Function (CCF) shows the causality relationship of weather stations precipitation and their local river water-level time-series. A correlation score and lagging period were derived for every pair of weather station and its local river. The correlation score suggest the level of responsiveness a river gage height to a rain event. Whereas, the lagging period indicates an average number of hours that the river's gage height will rise as a result.

ML then takes advantage of the CCF study to use precipitation as a primary feature input for building models to predict peak river's gage-height. A number of additional elements are also included in training ML models such as: river flow-rate, beginning river gage-heigh, to name but a few. Finally, standard ML models such as *Linear Regression* and *Random Forest* are applied across the U.S to find accuracy scores. It is found that some rivers are more predictable than others, and some column-features are more significant than others. There is also a variations

of features significance from one river to another river. Overall, the result shows great potential in forecasting short-term peak of river gage-heights.

5.2 Future Development and Study

In this project, ML models is able to provide a relatively high accuracy to only about half of the rivers and other surface water streams. The rest with low (less than 80%) or inconsistent ML prediction accuracy needs further in-depth studies. There are various possible issues that can fail ML models ranging from failed sensors' readings to physical structure interventions. This requires additional manual field data and possibly expertise inputs. Therefore, a features that include manual inputs is suggested in Section 5.2.1. In addition, a real-time signal system is outlined in Section 5.2.2 to bring this project closer to a productional environment that may benefit communities in river basin areas.

5.2.1 User Optimization model feedback

This is a feature that allows users to adjust the numerical values in our model and visually see if there is a better outcome. For flooding model, it is better to over-predict, rather than under-predict, which happens quite often in regression models [23]. The users can then save the different set of values for future use.

5.2.2 Real-time Prediction and Signal System

A potential application is to integrate with river flooding records to give signal on the visualization map. The weather data system is ingested hourly. As a rain event occur, this signal can send the prediction system to work. The system then use existing optimized model to predict the future water-gage heights, and the number of hours before water reaches its peak. If this peak level water is above the flooding stage, icons of rain-stations and river-gages will be blinking in warning colors on map to indicate an alert for the local areas. As this is automated, hydrologists can

quickly use this as a signal to narrow down the area of study and deliver more accurate results.

Another possible application for this project is to support and enhance a flood warning system. In the United States, flood warning services are provided by the Advanced Hydrologic Prediction System (AHPS) [19]. According to AHPS's published white paper in 2009, the present flood warning system applies complex mathematical models using various inputs ranging from: river basin structure, land moisture to automated sensor and manual inputs [19]. Such a sophisticated approach require experts of different input-areas and lots of computing power. AHPS model allows forecasting of water-level by hour in both short-range (few hours) and up to long-range (few weeks). Although this project is limited to finding the peak water level in short-range time period (48 hours), it relies solely on machine-learning models to calculate the peak water-level that will occur in a few hours. As a result, this project could contribute to AHPS modeling as a automated signal input.

References

[1] Ala Al-Fuqaha, Mohsen Guizani, Mehdi Mohammadi, Mohammed Aledhari, and Moussa Ayyash. Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials*, 17(4):2347–2376, 2015.

[2] EA Anderson. Techniques for predicting snow cover runoff. *The role of Snow and Ice in Hydrology*, pages 840–863, 1973.

[3] Fernando Ayuso. *Analysis of correlation between rainfall and river water level in the short term.* PhD thesis, Dublin, National College of Ireland, 2014.

[4] Andreas Bader. Comparison of time series databases. 2016.

[5] Nathan Buras. *Reflections in Hydrology: Science and Practice.* American Geophysical Union, 1997.

[6] Climdex. Datasets for indices of climate extremes, 2013. [Online; Last updated 29-January-2013].

[7] Norman H Crawford and Ray K Linsley. Digital simulation in hydrology'stanford watershed model 4. 1966.

[8] Jonathan D. Cryer and Kung-sik Chan. *Time series analysis. [electronic resource] : with applications in R.* Springer texts in statistics. New York : Springer, c2008., 2008.

[9] db engines.com. Trend of influxdb popularity, 2017. [Online; accessed 13-February-2017].

[10] Francesco De Carlo, Doga Gürsoy, Federica Marone, Mark Rivers, Dilworth Y Parkinson, Faisal Khan, Nicholas Schwarz, David J Vine, Stefan Vogt, S-C Gleber, et al. Scientific data exchange: a schema for hdf5-based storage of raw and analyzed data. *Journal of synchrotron radiation*, 21(6):1224–1230, 2014.

[11] Julie Demargne, Limin Wu, Satish K Regonda, James D Brown, Haksu Lee, Minxue He, Dong-Jun Seo, Robert Hartman, Henry D Herr, Mark Fresch, et al. The science of noaa's operational hydrologic ensemble forecast service. *Bulletin of the American Meteorological Society*, 95(1):79–98, 2014.

[12] Michael Dory, Allison Parrish, and Brendan Berg. *Introduction to Tornado: Modern Web Applications with Python.* " O'Reilly Media, Inc.", 2012.

[13] Mike Folk, Gerd Heber, Quincey Koziol, Elena Pourmal, and Dana Robinson. An overview of the hdf5 technology suite and its applications. In *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*, pages 36–47. ACM, 2011.

[14] DL Fread. Technique for implicit dynamic routing in rivers with tributaries. *Water Resources Research*, 9(4):918–926, 1973.

[15] Sushil Kumar Gupta. *Modern hydrology and sustainable water development*. John Wiley & Sons, 2011.

[16] Allen Hazen. Storage to be provided impounding reservoirs for municipal water supply. In *Proceedings of the American Society of Civil Engineers*, volume 39, pages 1943–2044. ASCE, 1913.

[17] Benjamin Leighton, Simon JD Cox, Nicholas J Car, Matthew P Stenson, Jamie Vleeshouwer, and Jonathan Hodge. A best of both worlds approach to complex, efficient, time series data delivery. In *International Symposium on Environmental Software Systems*, pages 371–379. Springer, 2015.

[18] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[19] John McEnery, John Ingram, Qingyun Duan, Thomas Adams, and Lee Anderson. Noaa's advanced hydrologic prediction service: Building pathways for better science in water forecasting. *Bulletin of the American Meteorological Society*, 86(3):375–385, 2005.

[20] MeteoCentre. Aviation routine weather report (metar), 2017. [Online; accessed 27-January-2017].

[21] NCEI. Historical observing metadata repository, 2017. [Online; accessed 13-February-2017].

[22] NOAA. Noaa ish ftp portal, 2017. [Online; accessed 07-February-2017].

[23] GR Pandey and V-T-V Nguyen. A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*, 225(1):92–101, 1999.

[24] Dix Paul. The new influxdb storage engine: Time structured merge tree, 2015. [Online; posted 7-October-2015].

[25] Bradley M. Peterson. Time series analysis in studies of agn variability, 2010. Gamma Ray-Burst Workshop Presentations.

[26] A Ramachandra Rao, Khaled H Hamed, and Huey-Long Chen. *Nonstationarities in hydrologic and environmental time series*, volume 45. Springer Science & Business Media, 2012.

[27] scikit learn. Choosing the right estimator, 2017. [Online; accessed 13-February-2017].

[28] scikit learn. An introduction to machine learning with scikit-learn, 2017. [Online; accessed 13-February-2017].

[29] Scipy. Scipy documentation, 2017. [Online; accessed 08-March-2017].

[30] Vijay Singh and Mauro Fiorentino. *Geographical information systems in hydrology*, volume 26. Springer Science & Business Media, 2013.

[31] SRCC. Rain river interactive visualization, 2017. [Online; accessed 13-February-2017; http://rainriver.lsu.edu].

[32] SRCC. Srcc's hourly climate data lister, 2017. [Online; accessed 13-February-2017; http://hrly.lsu.edu].

[33] Statsmodels. Welcome to statsmodels documentation, 2017. [Online; accessed 08-March-2017].

[34] Charles E Sudler. Storage required for the regulation of stream flow. *Transactions of the American Society of Civil Engineers*, 91(2):622–660, 1927.

[35] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[36] Persen Todd and Winslow Robert. Benchmarking influxdb vs. open tsdb for time-series data, metrics and management. 2016.

[37] Ulmo. Ulmo library, 2017. [Online; accessed 07-February-2017].

[38] USGS. Usgs water surface water data, 2017. [Online; accessed 07-February-2017].

[39] Baxter E Vieux. Distributed hydrologic modeling using gis. In *Distributed Hydrologic Modeling Using GIS*, pages 1–17. Springer, 2001.

[40] John M Wallace and Peter V Hobbs. *Atmospheric science: an introductory survey*, volume 92. Academic press, 2006.

[41] Yi Wang, Yu Su, and Gagan Agrawal. Supporting a light-weight data management layer over hdf5. In *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*, pages 335–342. IEEE, 2013.

[42] Max Welling. Kernel ridge regression. *Max Welling's Classnotes in Machine Learning*, pages 1–3, 2013.

[43] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

Vita

Tri Nguyen was born in Hanoi, Vietnam. He attended University of Texas at Austin for his BS degree in Computer Science. Tri is interested in visualization and database technologies. He works at the NOAA Southern Climate Region Center (SRCC) as a graduate research assistant.