Machine Learning Algorithm Development for Quality Assurance of Lutz Catchment Runoff Data

Project Proposal & Statement of Work

Gillian McGinnis, Project Manager

ADVISORS:
Dr. Sriram Iyengar (*The University of Arizona College of Medicine, Phoenix*),
Steven Paton (*Smithsonian Tropical Research Institute*)

Date: September 23, 2025

Version	Summary of Changes	Date
0.1	Converted template to LaTeX; started References, §1	09/13/25
0.2	Began §2; transcribed §4; completed §5	09/16/25
0.3	Continued §2; updated §4	09/18/25
0.4	Switch to research template; updated §2	09/19/25
0.5	Updated §4, expanded §3; added Gantt chart	09/22/25
1.0	First draft complete	09/22/25
"	Submitted for signatures	09/23/25
2.0	Additional map; expanded §2 with BCI info; rephrased portions of §1, §3; typo fix	09/23/25
"	Resubmitted for signatures	09/23/25

Contents

1 Executive Summary

The primary goal of this project is to create models that can automate quality assurance of temporal runoff data from the Lutz catchment of Barro Colorado Island (BCI), Republic of Panama (Figure 1). This will remove the need for manual data corrections, which not only use the valuable time and energy of researchers, but also have the potential to be imprecise or inconsistent. Rainfall and soil moisture measurement data will be integrated to produce a more well-informed model.

Runoff data has been continually collected from the weir since 1972, at first through hand-written records (which have since been digitized) in intervals ranging from three minutes to three hours, and electronically measured every five minutes since 1989. Quality issues in the data come about due to environmental factors (such as partial blockages of the outflow channel), sensor failures, and calibration issues, to name a few. Currently, quality assurance is conducted manually using a customized program in Visual FoxPro to allow for visual inspection and correction of the data. Not only does this approach run the risk of imprecise or inconsistent corrections, but the deprecated status of the language and interface leaves it vulnerable to inaccessibility. Past attempts at automated quality assurance using traditional stochastic methods have failed.

An outline of anticipated project responsibilities is provided in Table 1. Relevant data sets are available on the Smithsonian Research Data Repository with `CC BY 4.0` licenses.[1] Steven Paton—director of the Physical Monitoring Program at the Smithsonian Tropical Research Institute (STRI)—will be the primary contact regarding background information of the data, knowledge of exceptional data points, and any hydrological questions that come about. Project work will be conducted remotely, with regular advisor check-ins via teleconference.

The goals by the end of the project are to have a system for multiclass classification to flag periods of time with failure mode(s), and models to correct flagged data points. Measured success of the models can provide information into the feasibility of applying machine learning (ML) models in the context of micro-catchment runoff data quality assurance, as well as identify the most significant difficulties which can be addressed in future refinements of the models. To maintain a feasibility & realistic goal for the semester, it is anticipated that accurate correction models can be made for at least two failure modes.

Team Member	Feature responsibility
Gillian McGinnis	Literature review & research
"	Exploratory data analysis (EDA)
"	Failure mode classification model
"	Failure mode correction models
"	Model quality evaluations
"	Reproducibility documentation
"	Course deliverables and presentation preparations

Table 1: Preliminary Subsystem Responsibilities

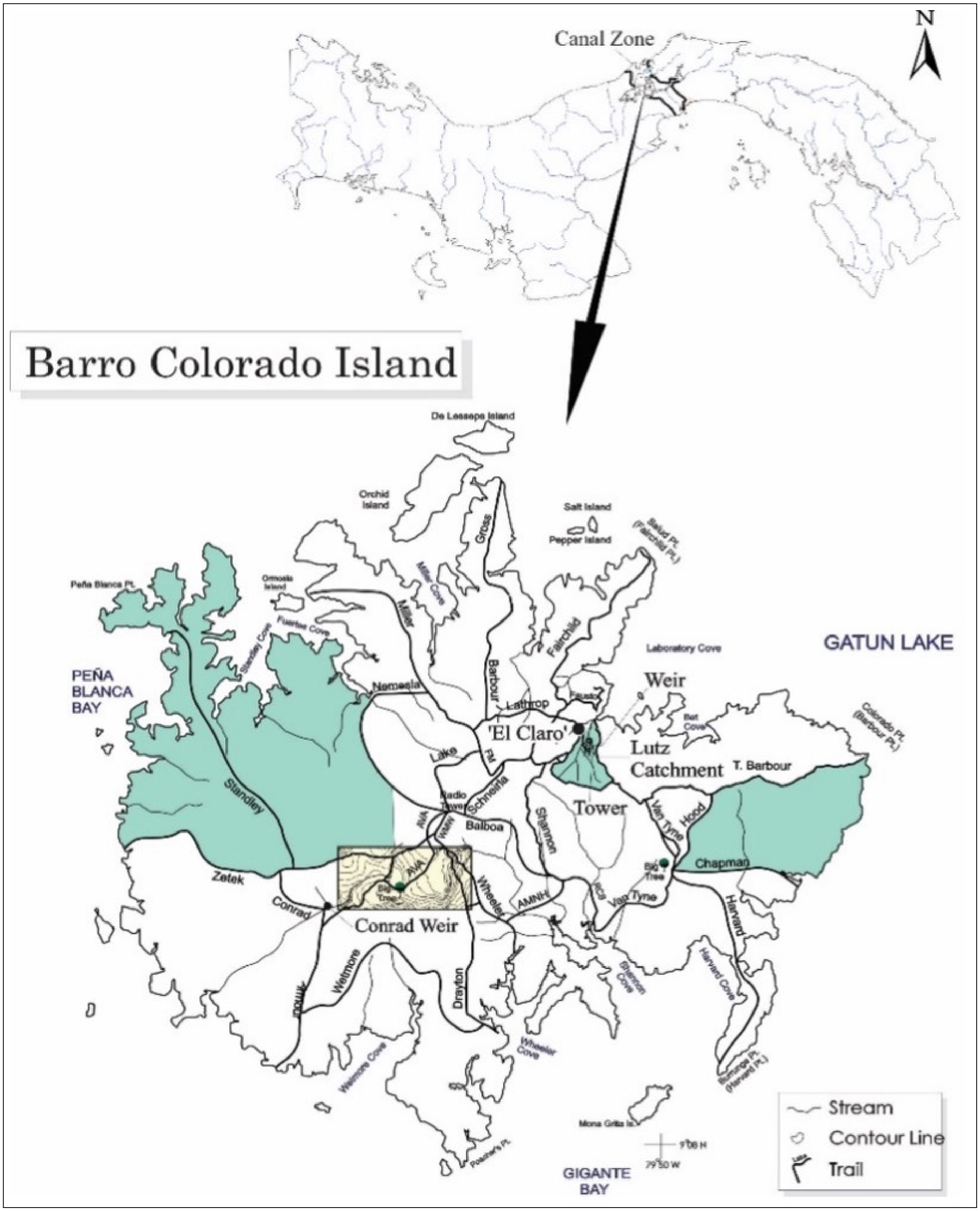

Figure 1: The location of BCI, Republic of Panama, showing the locations of the Lutz catchment, weir, and tower.[2]

2 Literature Review/Market Research

Overall research landscape

ML has been used in hydrology for applications such as predictive runoff modelling[3] and modelling agricultural drought.[4] Rainfall runoff data collected by citizen scientists in Ethiopia has been visually (via graphical inspections) and quantitatively (via statistical methods) analyzed for quality compared to nearby professional references.[5] In the field of remote sensing and geospatial analysis, recurrent neural networks—especially long short-term memory (LSTM)—are successful for analyzing multi-temporal data, however scalability is limited and they can be prone to overfitting.[6] In environmental ML, improper missing data management techniques can also worsen data leakage.[7]

The Lutz catchment (Figure 2) is a 9.73 ha area located in the secondary low land tropical rainforest on BCI, a 15 km^2 island located in Lake Gatun in the Republic of Panama.[2] BCI is operated by STRI, a branch of the Smithsonian Institution.[8] Runoff at the Lutz catchment—as well as other meteorological and hydrological data—have been continually monitored since 1972, making it one longest, continually monitored micro-catchment datasets for the neotropics.[2]

External variables of rainfall and soil moisture content are necessary for a well-informed model. In the context of modelling expected droughts, Houmma et al.[4] highlights how incorporation of local and microclimate variables are becoming increasingly necessary due to climate change and "exceptional situations" which otherwise make extreme peaks of variables & interdependencies no longer able to be statistically modelable.

User insights

Paton currently uses Visual FoxPro to manipulate the data. While the interface has been sufficient for completing the tasks at hand, the software is vulnerable due to its depreciation. Furthermore, current quality assurance methods rely on subjective decisions by the individual analyzing the data, meaning that results are potentially biased to some unknown degree.[2] Additionally, it is not currently set up to automatically detect problematic portions of the data. Because data monitoring is ongoing, new data must be regularly checked for occurrence of any type of failure mode.

According to Paton, there have been few—if any—attempts to apply ML algorithms to singular catchment runoff data quality assurance. Previously, stochastic methods were attempted for automation of the Lutz catchment data quality assurance, however the team concluded it was unsuccessful in being more accurate than Paton's manual corrections. Although water flow is a complex phenomenon, it is anticipated that the quantity of raw and manually-corrected data as well as inclusion of additional variables will enable an algorithmic approach to correct the data to a reasonably-accurate degree.

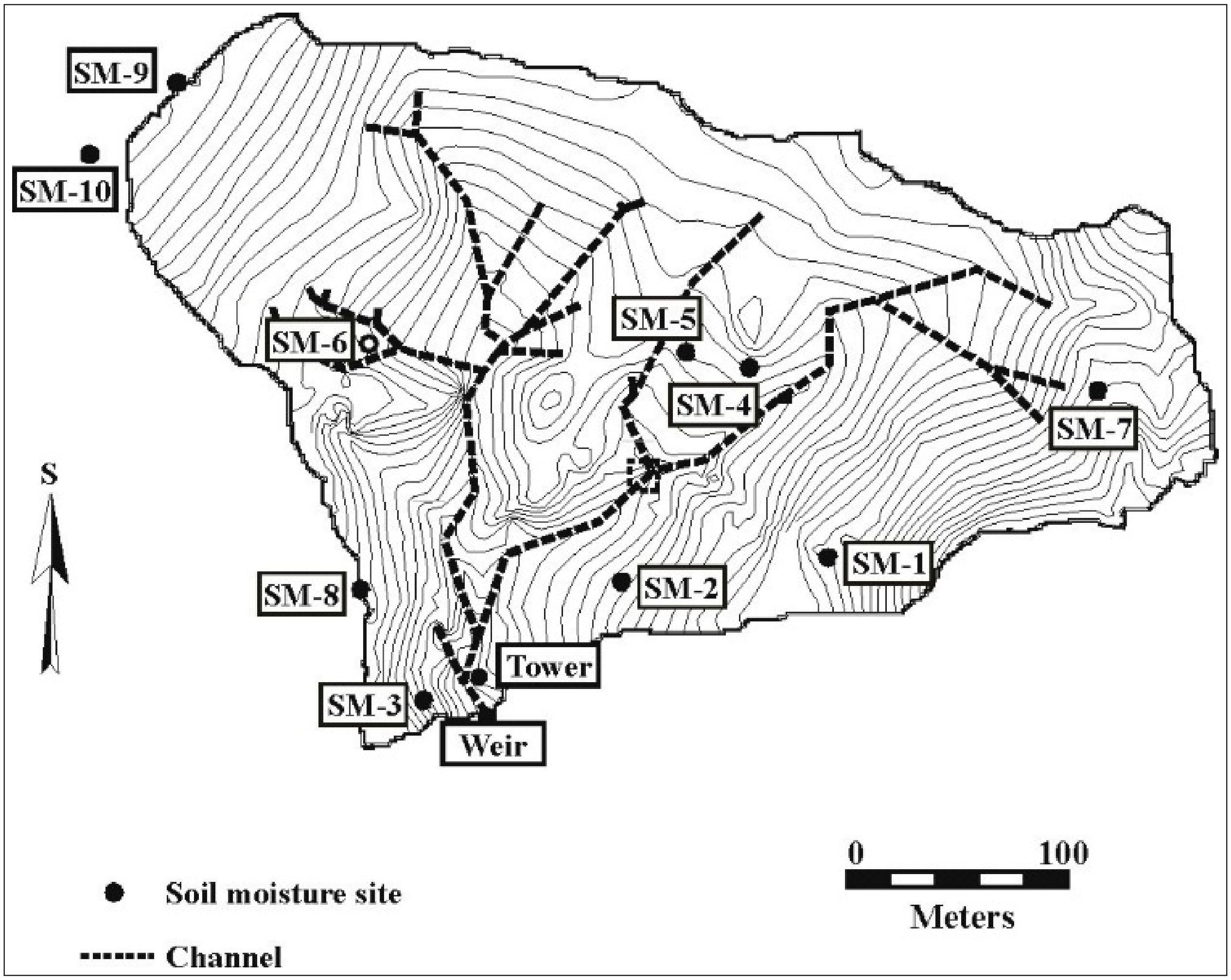

Figure 2: A fine-scale topographic map of the Lutz catchment showing the locations of the weir, tower, and ten soil moisture sampling sites.[9]

3 Research Project Deliverables

Final Presentation Format

The final project will be submitted in the form of a GitHub repository containing code that is relevant, readable, and thoroughly annotated to maximize accessibility and reproducibility. In addition to code files, a written report with background information, methods/approach, and final model performance statistics & visualizations will also be created.

What Analysis Is Being Run?

Classification of Failure Mode

Periods of time in the data can be flagged as containing “failure modes”, as explained in Table 2. The system will determine where there are readings containing a particular failure mode. This also requires the system to determine the time range in which a failure mode is occurring. A challenge for this feature is that some failures occur in extremely short intervals of time (even as few as one reading), while others can spoil data spanning multiple weeks.

To flag data points, supervised classification can utilize rolling statistics and the external environmental variables (which provide some seasonal insights). Gradient boosted trees or LSTM can be used, but more literature review and exploratory data analysis will need to be conducted prior to finalizing which model(s) to utilize.

Failure Mode	Description	Complicating Factors
Calibration	Standard checks that require baseline correction or slight data pivot.	Recovery after blockage clearing can render calibration points ineffective.
Spike	Short and abrupt changes in level typically caused by equipment issues.	Extremely short-term issue that could be skipped over by random sampling.
Sub-zero	The stream runs dry, or the pond is being drained for cleaning.	Rain during a pond draining can render new data unrecoverable.
Signal noise	Equipment failure results in impossible variability of reported values.	Impossible to manually fix.
Blockage	Debris blocks the weir’s ‘V’, resulting in gradual increases in water level, and later abrupt decreases following blockage removal.	Rainfall events before, during, or shortly after a blockage can interrupt the base flow and decay curves.

Table 2: Overview of Weir Failure Modes[9]

Quality Assurance on Periods of Failure

Based on the classification result(s), additional models will attempt to adjust the raw outputs to the appropriate values had one of the above-mentioned failure modes not occurred. This will involve creation of multiple separate models, as the method in which the data is corrected differs based on the classification of failure type—for example, an interrupted drainage caused by a blockage in the weir can be fixed by the expected decay curve, while a spike is simply “flattened” to the level of adjacent data points using a simple linear regression model.

What Accuracy Is Expected?

Accuracy for certain failure mode classifications are expected to be high—for instance, "sub-zero" values are simple to flag. Others, such as "spikes", are also expected to be easily found. Accuracy for multiple failure modes may be more difficult, especially in the dry season when there are non-erroneous fluctuations in flow caused by evapotranspiration. Multiple, large rainfall events also pose particular challenges, as well as overlapping failure modes (e.g., simultaneous blockage and calibration problems).

Similarly, correction models for simpler failure modes are expected to be highly accurate and precise, since they are often simple linear gap-fills. Corrections for the failure mode of a blockage in combination with a rain event may have reasonable accuracy but lower precision.

What if the Analysis doesn't work?

If the error type categorization has poor performance or is unable to consistently perform, and the needs of the correction models become more pressing as the semester continues, the data set does include the manually-flagged error types. The results of the ML-assisted corrections will be compared against the results of the manual corrections to data not included in the training sets. Large deviations from the expected values will be considered "failures".[2] If the correction models do not work, it will still provide valuable insights into which failure modes are most difficult to automate or most "resistant" to certain algorithms.

4 Project Timeline & Gantt Chart

A tentative schedule of major project milestones is provided in Table 3, while a detailed Gantt chart draft is shown in Figure 3. It is anticipated that integrating the correction models in combination with managing computational performance requirements will take the greatest amount of time for the project.

Milestone	Date
Mentor finalization	09/12/25
Project data meeting(s)	09/19/25
Bibliography building / lit review	09/22/25
Signed proposal	09/23/25
EDA	09/29/25
Failure mode multiclass categorization model	10/27/25
Individual failure mode QA models	11/24/25
Final performance measurements	11/27/25
iShowcase poster	12/01/25
iShowcase presentation practice	12/04/25
iShowcase	12/05/25
Documentation/repository cleanup	12/09/25
Final capstone	12/10/25

Table 3: Milestone Schedule

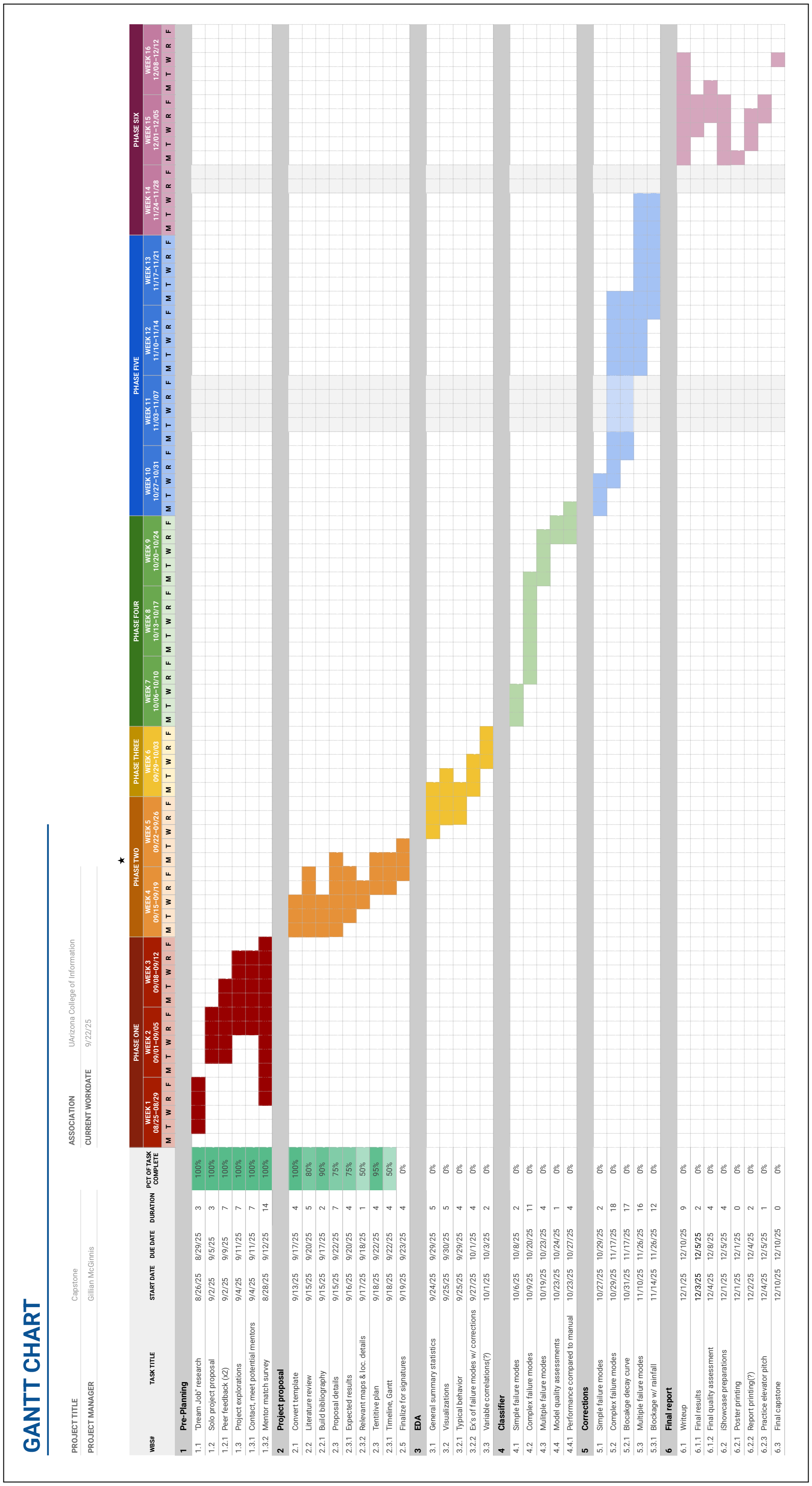

Figure 3: Gantt Chart Draft

5 Ethics

#	Question	Generally	Data Breach
1	Could a user sell drugs or other illegal items on your platform?	No	No
2	Could a user of your platform engage in sex trafficking?	No	No
3	Could a user sell class notes or cheat on their homework on your platform?	No	No
4	Could a stalker use your project to find someone?	No	No
5	Could your app be used to spy on or track individuals?	No	No
6	Could your app/software access the camera or microphone and record things without users being aware?	No	No
7	If someone uses your platform, could they be re-traumatized or have their mental health impacted in some way?	No	No
8	Could your algorithm promote material that would traumatize or upset individuals?	No	No
9	Would your users be upset if the data you collect was given to someone else?	No	No
10	Could a data leak potentially lead to identity theft?	No	No
11	If your site was hacked, would users of that product potentially lose their job, spouse, or family?	No	No
12	Should there be an age limitation on your product?	No	No
13	Could someone use your product to find, contact, and potentially commit elder abuse?	No	No
14	If the data on your platform was breached, could it be used to blackmail the users?	No	No
15	Does the existence of your project imply that a particular racial group, gender, religion or other protected category is inherently bad, gross, or unwanted?	No	No
16	Could your product be used to commit hate crimes against a specific group?	No	No
17	Does the primary content of your game or algorithm focus on something considered deeply unethical?	No	No
18	Does your game or software contain race, gender, or other stereotypes?	No	No
19	Could users of your app scam other individuals?	No	No
20	Is your particular algorithm biased towards predicting correctly only for one race, gender, or other group?	No	No
21	Are the users of your project, players of your game, or those being surveyed for your data aware of how their data will be used?	No	No
22	What are the possible misinterpretations of your results? For example: would a white supremacist or misogynist be stoked about your results if they misinterpreted it?	No	No
23	Does the use or purchase of your data potentially contribute to a dangerous group or regime?	No	No
24	Could your virtual reality environment cause injury to the user?	No	No
25	Are your study participants or game players aware that their data will be collected and used?	No	No
26	Does your game or app contain addictive design elements without benefit to the user?	No	No
27	Does your survey contain an aspect of compulsion or unusually large incentive, that would command users to take it even if it was to their detriment?	No	No
28	Could your research outcomes harm an individual or entity?	No	No

6 Approvals

This document is based upon and supersedes the "First Draft" Version 1.0. Deviations, (versus clarifications), from the PDR have been clearly noted. For any requirements not listed in this SOW, the PRD requirements shall remain in effect.

Approver Name	Title	Signature	Date
Gillian McGinnis	Project Manager		09/23/25
Sriram Iyengar	Primary Advisor		09/24/25
Steven Paton	Data Advisor		09/24/25
Nitika Sharma	Course Instructor		

	Section	Author	Word Count
1.	Executive Summary	Gillian McGinnis	389
2.	Literature Review/Market Research	"	408
3.	Research Project Deliverables	"	491
4.	Project Timeline & Gantt Chart	"	N/A
5.	Ethics	"	N/A

7 Appendix

A. Advisor Engagement

1) Project Team Responsibilities

- The Project Manager will set up and facilitate a weekly call/meeting with the Faculty Advisor. The Project Team will provide weekly status updates to the Faculty Advisor including upcoming deliverables, critical issues, and any adjustments to the Project Plan.
- Documents will be provided to the Faculty Advisor with adequate time for review and signature. The time necessary for review will be agreed with the Advisor. The minimum review time will be 3 days prior to the document due date.
- Design files will be provided to the Faculty Advisor as requested in a format agreed to with the Advisor.
- Support requirements will be clearly requested from the Faculty Advisor with the dates required and an adequate time for fulfilling the request.
- Modifications requests to the Project Plan by Faculty Advisor will be reviewed and agreed to within 1 week of the request.

2) Faculty Advisor Responsibilities

- The Faculty Advisor will provide knowledge and expertise to help the group stretch their skills.
- The Faculty Advisor will participate in a weekly or bi-weekly call/meeting with the Project Team to review the project status, upcoming deliverables, priorities, issues, and progress to the agreed Project Plan.
- The Faculty Advisor will provide document review, feedback and approval, rejection, approval with contingencies with adequate time for the Project Team to meet the course due dates.
- The Faculty Advisor will provide feedback to requested support requirements from the Project Team. This includes feedback and guidance on design implementations decisions, design files, test plans, test procedures and test results.
- The Faculty Advisor shall provide technical advice and guidance to the Project Team answering inquiries approximately 1 hour per week.
- Modifications to the Project Plan by the Project Team will be resolved and documented within 1 week of the request.
- Grade the finalized project using a skill-based rubric
- Attend iShowcase in May.

B. Ground Rules

As a team and as individual team members, we agree to:

1. **Stay focused on our objectives and goals.**
   Each time the team meets, we will clearly define our objectives and desired outcomes at the beginning of the meeting. We will politely remind team members if we are getting off track.

2. **"Sidebar" any issues that are relevant but not consistent with the immediate objectives.**
   Occasionally, important matters are raised that are not relevant to the immediate goals of the meeting. To keep the group on track, but avoid losing the issue, create a "sidebar" where these topics can be listed and discussed later.

3. **Listen when others are speaking.**
   We will listen and consider others' input before adding our own comments.

4. **All viewpoints will have an opportunity to be heard.**
   We understand that some team members may be quieter than others. We will make an effort to get each team member's viewpoint and that no one dominates the discussion.

5. **Differences of opinion will be discussed respectfully.**
   We will identify areas of agreement before assessing areas of disagreement. We will encourage each other to look beyond our own point of view. We will discuss different ideas respectfully. As a team, we will weigh the merits of different opinions and agree on a process for choosing a direction. All team members will respect and follow the decision or direction.

6. **Look for the good points in new ideas.**
   We will endeavor to explore the value in each idea as we assess and select our path forward.

7. **Focus on the future, not the past.**
   We will use our past experience to inform our decisions, but focus the discussion on the future objectives. Blame for past performance is counterproductive, we will focus on finding solutions.

8. **Agree upon specific action items and next steps.**
   At the end of each meeting and discussion, we will summarize and agree on specific next steps, action items and assignments.

9. **Accountability**
   As team members, we will each be responsible for our individual assignments and contribution to achieving the team objectives and goals. We will honor our responsibilities and not let our team members down.

References

(1) Smithsonian Institution Smithsonian Research Data Repository, https://smithsonian.dataone.org/ (accessed 09/19/2025).

(2) Paton, S. Runoff proposal suggestions, personal communication, 2025.

(3) Mohammadi, B. A review on the applications of machine learning for runoff modeling. *Sustainable Water Resources Management* **2021**, *7*, 98, DOI: 10.1007/s40899-021-00584-y.

(4) Houmma, I. H.; Mansouri, L. E.; Gadal, S.; Garba, M.; Hadria, R. Modelling agricultural drought: a review of latest advances in big data technologies. *Geomatics, Natural Hazards and Risk* **2022**, *13*, 2737–2776, DOI: 10.1080/19475705.2022.2131471.

(5) Kebede Mengistie, G.; Demissie Wondimagegnehu, K.; Walker, D. W.; Tamiru Haile, A. Value of quality controlled citizen science data for rainfall-runoff characterization in a rapidly urbanizing catchment. *Journal of Hydrology* **2024**, *629*, DOI: 10.1016/j.jhydrol.2024.130639.

(6) Dritsas, E.; Trigka, M. Remote Sensing and Geospatial Analysis in the Big Data Era: A Survey. *Remote Sensing* **2025**, *17*, 550, DOI: 10.3390/rs17030550.

(7) Zhu, J.-J.; Yang, M.; Ren, Z. J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environmental Science & Technology* **2023**, *57*, PMID: 37384597, 17671–17689, DOI: 10.1021/acs.est.3c00026.

(8) Smithsonian Tropical Research Institute Barro Colorado (Clearing, Lutz, Conrad weir), 2025, https://striresearch.si.edu/physical-monitoring/barro-colorado/ (accessed 09/13/2025).

(9) Paton, S. BCI Hydrology Introduction, personal communication, 2025.