

ATUR Retreat Open Data and Wiki Integration

Summer 2024 Hart Prairie August 2024

Before we begin.....You should have:

Download Zotero Desktop (or use Mendeley if you prefer)

A dataset you created .zip

A source dataset you used in creating your dataset.zip

Some literature that describes why you did what you did to make your dataset, or alternatively a list of publications which are relevant to ATUR (spreadsheet, .bib, or doc + pdfs)

For each you should have done the following:

(Title, DOI, 1-2 sentence description of the relevance to ATUR that could be understood by an 2nd year undergraduate student in Environmental science)



Agenda:

Introduction to Open Data and Single Sources of Truth (15 min)

Data Audience Exercise (5 Minutes)

Git-Wiki Exercise (10 Minutes)

Project Structure Exercise (10 Minutes)

File-naming Exercise (5 minutes)

(if time) Data Documentation Exercise (10 minutes)

(if time) Data Sharing Exercise (10 minutes)

Suitability Mapping Exercise (5 minutes)

Homework assignment explained (5 minutes)

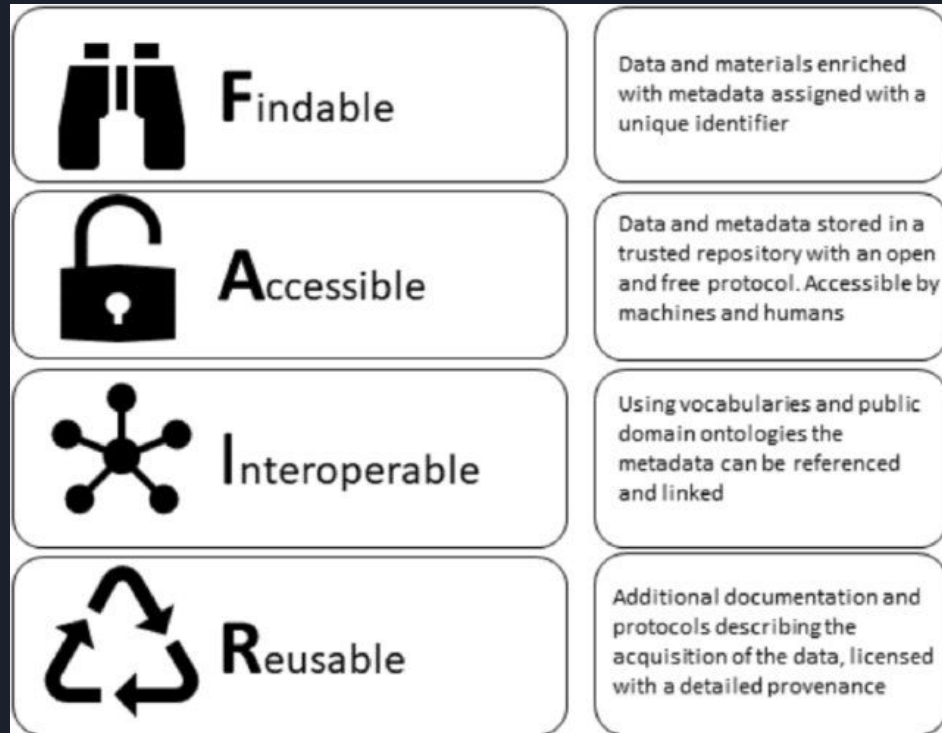
FAIR Guiding Principles for scientific data management and stewardship

Findability

Accessibility

Interoperability

Reusable





Goals of today's exercises

- To learn new ways to make our data and products FAIR
- To share knowledge and data with other folks in our project
- Start seeing ourselves as a community of collaborators!
- Think critically about the end users of our products and how we can produce science that is useful to them, and is FAIR
- Move toward a standard organizational structure and naming conventions throughout the project

Benefits

- FAIR data and products will enable folks at the AZDWR to use our data rather than allowing it to sit on a shelf.
- Organized, documented and vetted data and products make publishing manuscripts much easier
- Get a DOI for each data product you create (more citations!)
- Easier onramp for folks who are new to your labs or the project
- Re-evaluating and organizing your old work may bring up new ideas for this project and new ways to collaborate



An illustration of reasons why we should care about working reproducibly. [The Turing Way]

Single Sources of Truth SSOT

We want to employ the concept of Single Sources of Truth.

A single source of truth (SSOT) refers to the practice of structuring information models and associated data schema such that every data element is stored exactly once, ensuring consistency and accuracy across all systems and processes that rely on that data. This approach prevents duplication and discrepancies, making it easier to manage and trust the information used in decision-making.



Single Sources of Truth SSOT for ATUR

a single source of truth (SSOT) ensures that all teams use the same, consistent datasets, methodologies, and definitions throughout the project. This central repository of information prevents discrepancies between maps produced by different teams, ensures that everyone is working with the most up-to-date data, and facilitates seamless integration of results into a unified, accurate representation of recharge suitability across the state.






So things to think about and discuss later..

What items/datasets/processes should we prioritize creating SSOTs for?

Should we keep using Cyverse to hold our single sources of truth or have the data products stored somewhere else?

How will we track them and ensure we are using the same ones?

How do we ensure that everyone across the team can quickly access them?



Why Github!? I don't want to learn a new thing...uhg

1. Its free (for most things)
 - a. Public Repos are Free!!
2. Collaboration across disciplines
 - a. Version control allows multiple researchers to work on projects simultaneously without overwriting each others work
 - b. Researchers can work on their own branches and merge changes when they are ready allowing the main project to remain stable
3. Transparency and Reproducibility
 - a. You can make your work publicly available and reproducible which is key for cross-validation and further research
 - b. You can track the changes you have made and understand how your project evolved
4. Project Management
 - a. Makes it easy to manage tasks, track progress, and see individual contributions
5. Skill development and professional growth
 - a. See stats on right
6. Integration with other tools
 - a. Great integration with Jupyter, R, Python, Quarto, Bookdown and other programming languages
7. Community and Support
 - a. Theres likely someone on github who has had the same issue you are having, crowdsource solutions to problems.

Key GitHub Statistics in 2024 (Users, Employees, and Trends)

Jeremy Holcombe | Published: April 5, 2023 | Updated: October 2, 2023

- Around 100 million developers across the globe use GitHub. The majority of them are based in the US, India, and China.
- Over 90 percent of Fortune 100 companies use GitHub.
- 88 percent of developers who use GitHub Copilot say that it helps them be more productive and finish projects faster.



Wiki's

Wiki: *a website that allows collaborative editing of its content and structure by its users.*

Github Wiki's

Ryan3Lima / ATUR-WIKI

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

ATUR-WIKI Public

Pin Unwatch 1 Fork 3 Star 1

main 1 Branch 0 Tags

Go to file Add file Code

Ryan3Lima	Update README.md	0e3572d · last month	29 Commits
.ipynb_checkpoints	fix jupyter notebook	4 months ago	
Images/Readme	updates quarto	2 months ago	
Literature	updates quarto	2 months ago	
Notebooks	updates	2 months ago	
docs	updates	2 months ago	
.gitignore	updates	2 months ago	
.nojekyll	YES	2 months ago	
LICENSE	Initial commit	9 months ago	
README.md	Update README.md	last month	
]	updates	2 months ago	
_quarto.yml	updates	2 months ago	
index.qmd	updates	2 months ago	

README MIT license

About

Test Wiki For the ATUR Project (NAU Contributions)

- Readme
- MIT license
- Activity
- 1 star
- 1 watching
- 3 forks

Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Languages

- Jupyter Notebook 98.0%
- JavaScript 1.4%
- Other 0.6%

Home

Ryan Lima edited this page last week · 17 revisions

Welcome to the ATUR Test Wiki

This Wiki is intended to provide documentation, guidance, and literature used in the [Arizona Tri-University Recharge and Water Reliability Project](#) (ATUR)

Escalating drought over the past two decades has led to growing concerns regarding water quantity and quality for Arizona's communities. At the request of the Arizona Department of Water Resources, a team of researchers from the University of Arizona (UArizona), Arizona State University (ASU), and Northern Arizona University (NAU) will study locations and methods for protecting and enhancing groundwater recharge across the state.

More than 90% of precipitation that falls as rain and snow in Arizona evaporates before it enters a stream or recharges the groundwater. Capturing this water before it escapes to the atmosphere and encouraging it to percolate into the ground to replenish aquifers (enhanced recharge) can increase water supplies for communities and support ecosystems.

Project Goals:

- Identify ways to protect water supplies across the state by capturing precipitation before it evaporates or is used by plants
- Focus exclusively on water supplies that would not otherwise have reached a natural channel
- Identify locations for enhanced recharge for human and wildlife needs
- Identify land and vegetation management practices to enhance water availability
- Develop a system for prioritizing recharge sites

Meet the [Project Team](#)

Project Structure

The Arizona Tri-University Recharge and Water Reliability Project consists of 5 subteams comprised of researchers, post docs, and graduate students from the three state universities, the University of Arizona, Arizona State University, and Northern Arizona University. The project subteams are Forests and Natural Environments, Urban Environment, Recharge

[Edit](#) [New page](#)

Pages 15

Home

Welcome to the ATUR Test Wiki
Project Goals:
Project Structure

[Arizona HU8 TNM](#)[Cyverse](#)[Github](#)[Groundwater Recharge Seasonality In...](#)[Guidance: Adding Literature to Lit Ta...](#)[Hydrologic Cycle Process Diagram](#)[Karst groundwater recharge in Arizona](#)[Lit Review Sample](#)[Managed Aquifer Recharge \(MAR\)](#)[Seasonality in Ecohydrology Studies I...](#)[State-wide Karst Sink Delineation](#)[Suitability Layers](#)



Github Wiki's use Markdown language

Markdown is a lightweight markup language that you can use to add formatting elements to plaintext text documents. Created by [John Gruber](#) in 2004, Markdown is now one of the world's most popular markup languages.

Resources for Markdown:

Markdown Cheatsheet [[link](#)]

The Ultimate markdown Cheatsheet [[link](#)]



Other Markdown Resources

1. Create a Markdown table of content - [binarytree](#), [github-markdown-toc](#)
2. Create an empty Markdown table - [Tablesgenerator](#)
3. Convert Excel to Markdown table - [Tableconvert](#)
4. Markdown preview for Sublime Text 3 - [Packagecontrol](#)
5. Markdown preview Visual Studio Code - [Markdown Preview Enhanced](#)
6. A collection of awesome markdown goodies - [Awesome Markdown](#)
7. Markdownlint - [markdownlint](#), [markdownlint-cli2](#), [markdownlint-cli2-action](#), [vscode-markdownlint](#)



Data Audience Exercise (5 Minutes)

Take five minutes to think about the following questions and write down your answers somewhere, perhaps a notes app, or word doc, or notepad (be prepared to share)

- Who in this group needs your data or could benefit from it? (make a list)
 - Is it documented, named, and organized in a way they can understand?
- How might ADWR use your data or how will it enter into a final product? (describe 1-2 sentences)
 - Is it documented, named, and organized in a way that someone at ADWR would understand?



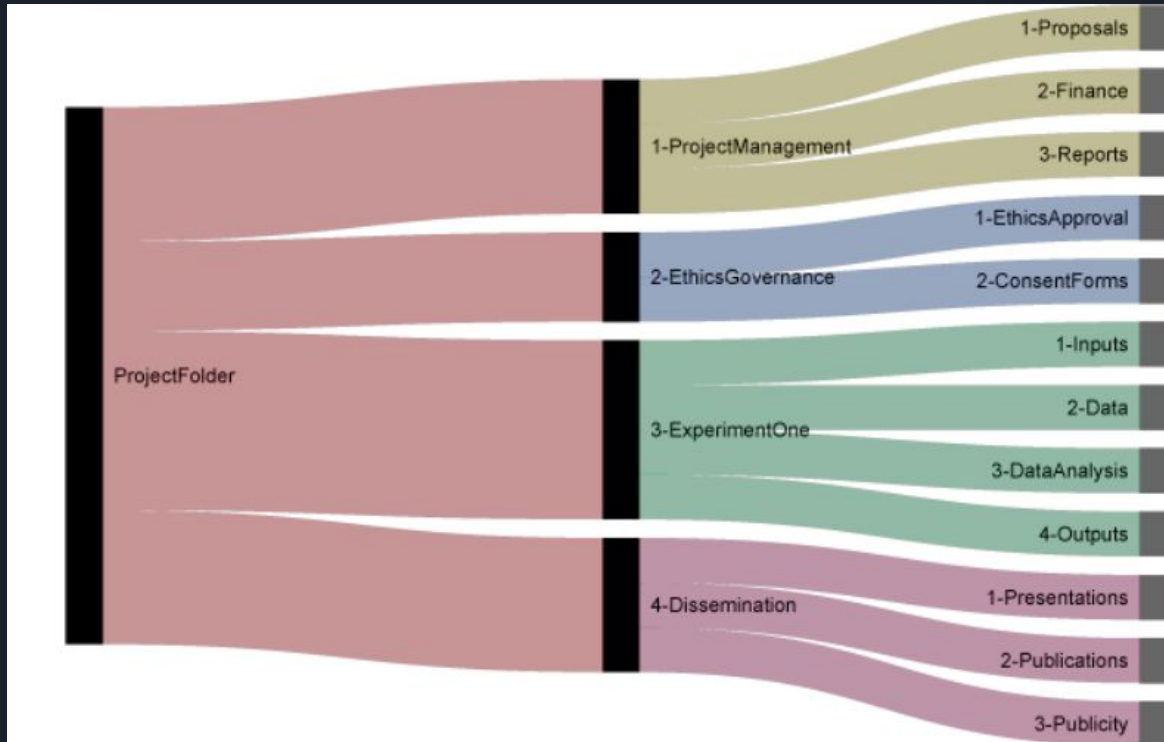
Git-Wiki Exercise (10 Minutes)

- Open Github (if you haven't create an account and download github desktop if you don't plan to use the command line exclusively)
- In a new or existing github repository - create a wiki page for your created and sourced datasets
 - Briefly describe the source data or inputs
 - Briefly describe the processing you did
 - Briefly describe the outputs, including all variables
 - Briefly describe what your dataset could be used for and why you created it
 - **SAVE YOUR CHANGES**

Project Structure Exercise (10 Minutes)

“The key is in adopting a logical and consistent project folder structure.

Having an already existing and uniform folder structure prior to starting a research project means you can work more efficiently, find your files more easily, prevent duplication, and be able to share and work collaboratively while minimising the risk of data loss.”





Project Structure Exercise (10 Minutes)

On the github page for your produced data product, think about what would need to be included in a package for someone to reproduce your work.

1. Create a package structure (file structure), treat this page like a sample read.me. Just an outline...

Ask the following questions:

Will this work for my other projects or products?

Download this folder structure by Nikoa Vukovic [[document](#)]



Project Structure Exercise - Other Resources

Download this folder structure by Nikoa Vukovic [[document](#)]

Project Structure, how to by Danielle Navarro - [[slide show](#)]

How-to-organize-your-digital-files - nytimes.com [[article](#)]

Transparent Project Management Template - [[template](#)]

MIT's recommendation for file naming and folder hierarchy [[Webpage](#)]

File-naming Exercise (5 Minutes)

1. Use logical or Chronological naming of files

Use Standard Date-Time formats

Number scripts in order of how they are run

2. Make filenames Machine Readable

Machines sometimes get confused by spaces

Easier to write code when files don't have spaces

Machines get confused by special characters

Avoid Accents, Don't assume case is meaningful

Use separator characters wisely

3. Make filename Human Readable

Short, descriptive, Logical

File names sometimes need to include dates

... but these dates don't sort in chronological order 🤔

- 1-April-2012.R
- 1-Jan-2009.R
- 1-Jan-2012.R
- 12-Jan-2012.R
- 2-Jan-2012.R
- 31-Dec-2009.R

Happily, these dates do! 🍌

- 2009-01-01.R
- 2009-12-01.R
- 2012-01-01.R
- 2012-01-02.R
- 2012-04-01.R

Key principle with dates...

- Dates should follow the YYYY-MM-DD format
- Known as the ISO 8601 standard

- ☒ what-the-cat.docx
- ☒ essay_romeo-and-juliet_draft01.docx
- ☒ what-the-cat?.docx
- ☒ essay "romeo and juliet" draft01.docx
- ☒ essay "romeo and juliet" draft01(1).docx

Love the slug 🐸

```
"analysis01_descriptive-statistics.R"  
"analysis02_preregistered-analysis.R"  
"notes01_realising-the-problem.txt"  
"analysis03_departing-from-the-plan.R"  
"notes02_tentative-write-up.docx"
```

- Concise, meaningful description
- Usually appended to the end



File-naming Exercise Resources

Project Structure, how to by Danielle Navarro - [[slide show](#)]

How-to-organize-your-digital-files - nytimes.com [[article](#)]

MIT's recommendation for file naming and folder hierarchy [[Webpage](#)]



File-naming Exercise (5 Minutes)

Write out the method behind your file-naming madness in the wiki.

As the following questions:

Is the Name understandable by humans?

Is the Name understandable by machines?

If files contain dates.....Are dates sortable?

If the answer to either of these questions is NO, then think of a new way to name your files.



Data-Documentation Exercise (10 Minutes)

Update the github wiki for your sourced and produced data product. Or better yet create ReadMe.txt for both. With a mind to the following questions:

Are there multiple versions/options of my sourced dataset?

Why did I choose the version that I chose to use?

What would someone who is not in my field need to understand to use this data product?

Can I list out all the steps in order?



Data Sharing Exercise (15 Minutes)

PART 1: Find someone in another sub-team, and talk to each other about the data product you, show them what you produced in the last few exercises and ask each other questions.

Do you understand what this data product is?

Do you understand what it could be used for?

Do my filenames make sense?

Is the project structure intuitive?

Do they know enough to reproduce it if they had to?

(7.5 minutes)



Data Sharing Exercise (15 Minutes)

PART 2: we will go around the room and talk our partners data product, explain what it is, and what its useful for to the whole group. (7.5 minutes)



Data Sharing Resources

[Openscapes.org](https://openscapes.org) - Openscapes is an approach and a movement that helps researchers and those supporting research find each other and feel empowered to conduct data-intensive science. Through a creative approach drawing inspiration and skills from many places, we provide structures for technical skill-building, collaborative teamwork, and inclusive community development.

[Making Shareable documents with Quarto](#) - It's possible to create beautiful documentation to share online with [Quarto](#) that auto-updates with [GitHub](#). This is very new and incredibly cool. This tutorial is an example of a quarto website — it is a really powerful way to create and share your work. You can communicate about science using the same reproducible workflow you and/or your colleagues use for analyses, whether or not you write code.

Creating websites with Quarto can be done without knowing R, Python or HTML, CSS, etc, and that's where we'll start. However, Quarto integrates with these tools so you can make your websites as complex and beautiful as you like as you see examples and reuse and remix from others in the open community. This tutorial borrows heavily from a lot of great tutorials and resources you should check out too – there are links throughout.

Safe data storage and backup [[video](#)]

Tool to Convert table to markdown table [[web tool](#)]

The Turing Way - a guide for open science and open collaboration [[web book](#)]



Resources

- Tool to Convert table to markdown table [[web tool](#)]
- The Turing Way - a guide for open science and open collaboration [[web book](#)]

The Turing Way



[The Turing Way](#) is an open science, open collaboration, and community driven project. A handbook for reproducible, ethical and collaborative data science

- [Guide for Reproducible Research](#)
- [Guide for Project Design](#)
- [Guide for Communication](#)
- [Guide for Collaboration](#)
- [Guide for Ethical Research](#)

Early Career Researchers

- [Guide for Project Design](#)
- [Getting Started With GitHub](#)
- [Creating Project Repositories](#)

And more...

Project Leaders

- [Open Leadership in Data Science](#)
- [Guide for Project Design](#)
- [Creating Project Repositories](#)

And more...

Research Software Engineers

- [Citing Research Objects](#)
- [Research Software Engineer: Overview](#)
- [Research Software Engineering Personal Story](#)

And more...

Software Citation

- [Steps for Making Research Objects Citable](#)
- [Citing Research Objects](#)
- [Software Citation with CITATION.cff](#)

And more...

Suitability Modeling Exercise Example

Topic: Forest thinning to increase groundwater recharge in Arizona





Suitability Modeling Steps:

1. **Define** - *Identify the problem you are trying to address by identifying goals and supporting criteria that can be modeled with data*
2. **Derive** - *obtain data that represents the model variables, convert them to rasters that can be combined into a suitability model*
3. **Transform** - *transform the data in a suitability model to a common suitability scale*
4. **Combine** - *Combine the transformed data into a model to create a single suitability surface*
5. **Locate** - *Locate sites that are suitable for the modeled phenomena by using a suitability surface or use the suitability surface to grow regions*
6. **Analyze** - *analyze the results with visual evaluation, and apply sensitivity analyses to the various components to see how model weighting, or error in measurement effects the suitable regions.*

How this might work for ATUR...

1. Define

Category	Question	Goal	Criteria
Excess Water	Where can forest management decrease atmospheric water losses in Arizona?	Find locations where forest thinning can reduce atmospheric losses of ≥ 50 acre-feet per year	Forested Land Forests over thickened Precipitation $> 400\text{mm}$ Area sufficient to generate 50+ acre feet of water with x% change in ET
Recharge Suitability	Which forested areas in the State of the highest recharge suitability?	Find forested locations where a x% of precipitation recharges	Karst Areas Forested Areas Areas which receive significant snow Areas drained by losing streams or sinkholes Low Slope or Proximity to Mountain Front Recharge Zone
Combined Suitability	Where can forest management generate excess water, which will increase recharge	Where are areas where forest thinning can reduce ET by 50 acre-feet and x% of that will recharge?	Union of the above suitability areas



Trainings for Suitability Modeling (**Highly Recommended**)

[Suitability Modeling: Introduction](#) (requires ESRI License)

Suitability modeling enables you to find where something, such as a business or habitat, should be located. In this course, you will discover the types of real-world problems that can be addressed with a suitability model. Then, you will learn how to define a problem in terms of an analysis goal and suitability criteria that can be modeled with data. You will also gain an understanding of how to prepare data for a suitability model. When you complete this course, you will be ready to create a weighted or simple suitability model by using the data that you prepared in this course as input.

[Suitability Modeling: Creating a Simple Suitability Model](#) (requires ESRI License)

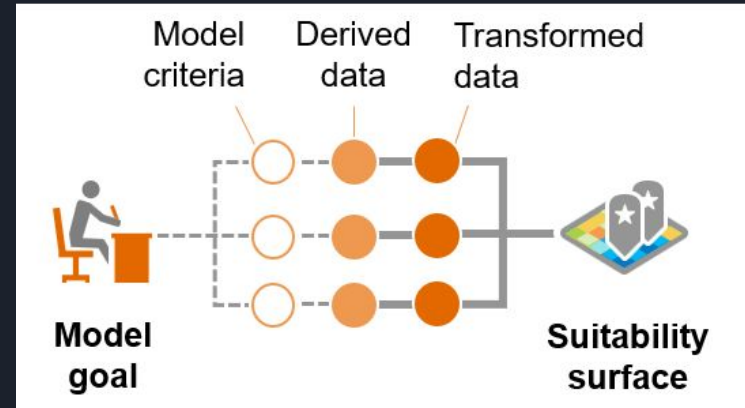
Suitability models can help you find optimal locations. But what if you merely want to identify locations or areas that meet all the criteria for your analysis—such as code requirements, permit restrictions, or minimum feasibility conditions? In this course, you will learn how to create a simple suitability model, which produces a binary result that is easy to interpret. You will also learn how to use sensitivity and error analysis to evaluate results.

[Suitability Modeling: Creating a Weighted Suitability Model](#) (requires ESRI License)

Suitability models help you find the most appropriate location for something. In this course, you will learn how to create a weighted suitability model, which adds science and expert opinion to the suitability modeling process. You will also learn how to use sensitivity and error analysis to evaluate your results.

2. Derive

- Compile necessary datasets:
 - Forest Cover Raster
 - Canopy Cover
 - Forest Type
 - Wildfire Risk
 - Predicted Land Cover Change Raster
 - Precipitation Raster 30 year normal (mean, min, max)
 - Winter precipitation
 - Avg Snow Depth/cover
 - Projected Precipitation (mean, min max)
 - Elevation Data
 - Slope
 - Aspect
 - Estimated Current ET
 - Estimated Change in ET
 - Estimated Groundwater Recharge
- Convert Data for input into model
 - Define Study Area - HUC 8 for example
 - Convert source dataset (???) to derived dataset: (for example) 8-bit unsigned raster
 - Resample to the same resolution
 - Convert to same coordinate system
 - Clip to same extent





3. Transform

1km Mean Precip mm 30-year normal (0-2000) -----> 90m MeanPrecip into Acrefeet per unit area

1km Fraction of precip in the winter (0.00 - 1.0) -----> 90m Fraction of precip in winter (0.0 - 1.0)

30m LandCover 2021 -----> 90m Forest Cover % from high to low (0.00 - 1.0)

30m Forest Type -----> 90m Forest Type Map

30m Canopy %Cover Raster (0.0 - 1.0) ----> 90m Canopy Density Raster (0.00 - 1.0)

Estimate of Change in ET with Percent Cover for Forest types (% change)



Lit Review Exercise

ON YOUR OWN

Think about all the your source dataset and your processed dataset, gather all the documentation and citations for these and **add them to your github page as a markdown table.**

Think about all the papers you have been a part of that somehow inform what we are doing here as ATUR, compile them and write 1-2 sentences about what their implications for this project.

Share the link to your Github Wiki here:

Add the 1-2 sentence summaries of your papers along with their title, authors, and DOI here:



Homework

ON YOUR OWN

Think about all the your source dataset and your processed dataset, gather all the documentation and citations for these and **add them to your github page as a markdown table**.

Think about all the papers you have been a part of that somehow inform what we are doing here as ATUR, compile them and write 1-2 sentences about what their implications for this project.

Share the link to your Github Wiki here: [LINK]

Add the 1-2 sentence summaries of your papers along with their title, authors, and DOI here: [LINK]

UPDATE FILES ON CYVERSE

If you thought up a better way to name your files, or a better structure for your data, make those changes to the data on cyverse, add your readMe.txt to the folder containing your data.

WITH 1-2 PARTNERS

Form a team of 1-3 people, develop a question related to excess water supply or recharge suitability, work with those folks over the next two weeks toL

Define: Questions, Goals

Derive: Datasets



Mapping Road-MAR Suitability in the San Pedro watershed using ArcGIS Pro Suitability Modeler



Homework

1. Bring either a table (excel or google sheets) or .bib file and the pdfs of any papers you have published which may be used in the literature reviews for any of the ATUR subteams or ATUR papers. Write a sentence or two summarizing the key points relative to our ATUR Project.
2. Bring a dataset (on a hard drive, .zip) that you are using (largely unchanged) as one of the ingredients in one of your projects. For example:
 - a. A hydrological dataset (PRISM Mean Annual Precipitation)
 - b. A thematic layer you might need (land cover, soils, geology, etc.)
3. Bring a dataset that you have altered or created as part of this project (on hard drive, zipped).
 - a. Lineament Density Map
 - b. Potential Sinkholes map
 - c. Baseflow Map
4. Read the [Guide for Collaboration](#) from the free online book [The Turing Way](#)
5. Create a github account
6. Create a [CUAHSI Jupyterhub](#) account
7. Tell us which dataset you are bringing