Forecast Infrastructure

2022-07-01

Near-term Ecological Forecasting Initiative Short Course

Jake Zwart (he/him)

With input and slides used from USGS co-workers

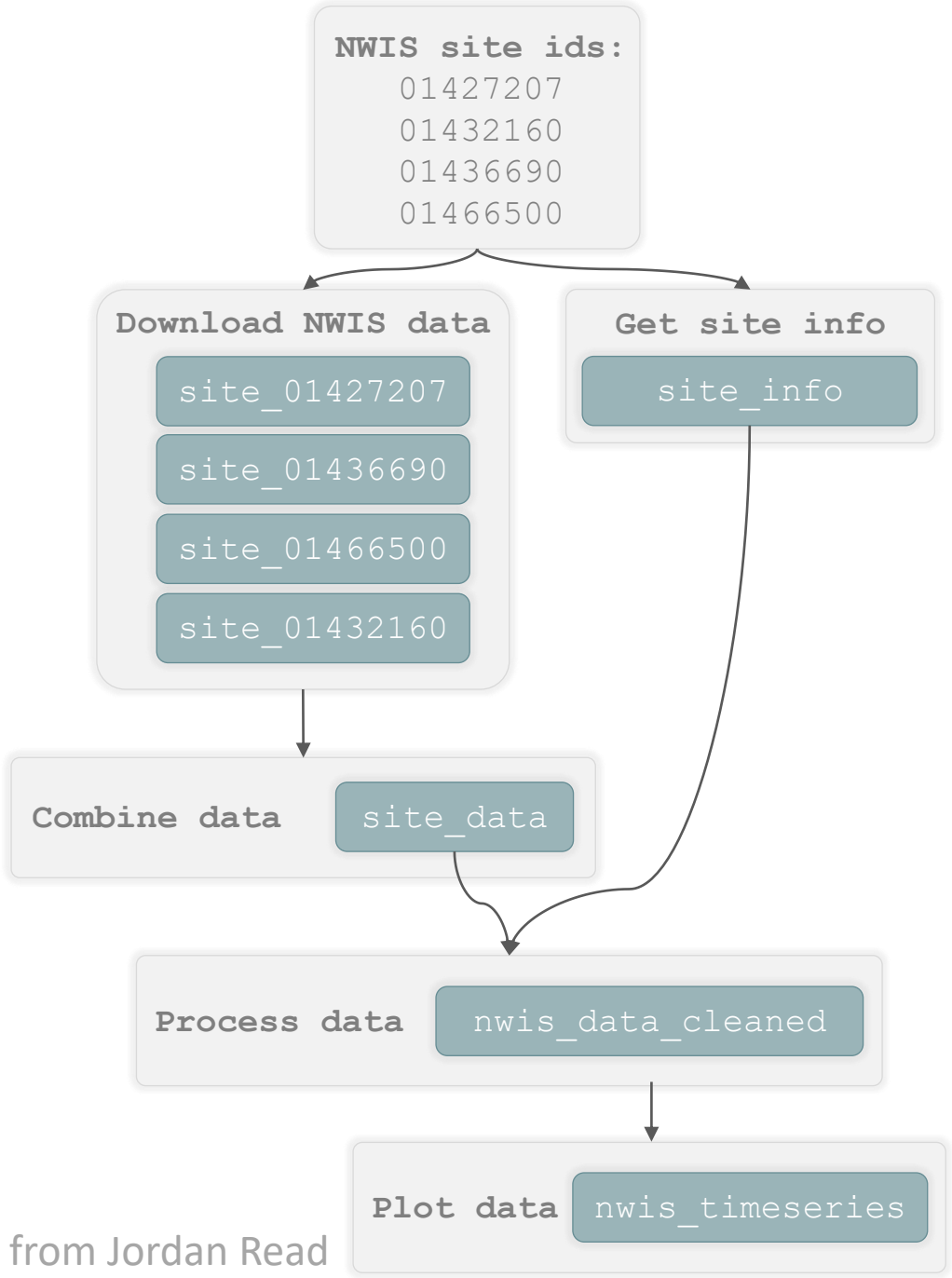
Sprague, Lori A 10/27 10:58 AM

Shoot, I know I told you we were done with the analysis and data release, but I just heard that site 01436690 had a sensor malfunction and they've since corrected the data

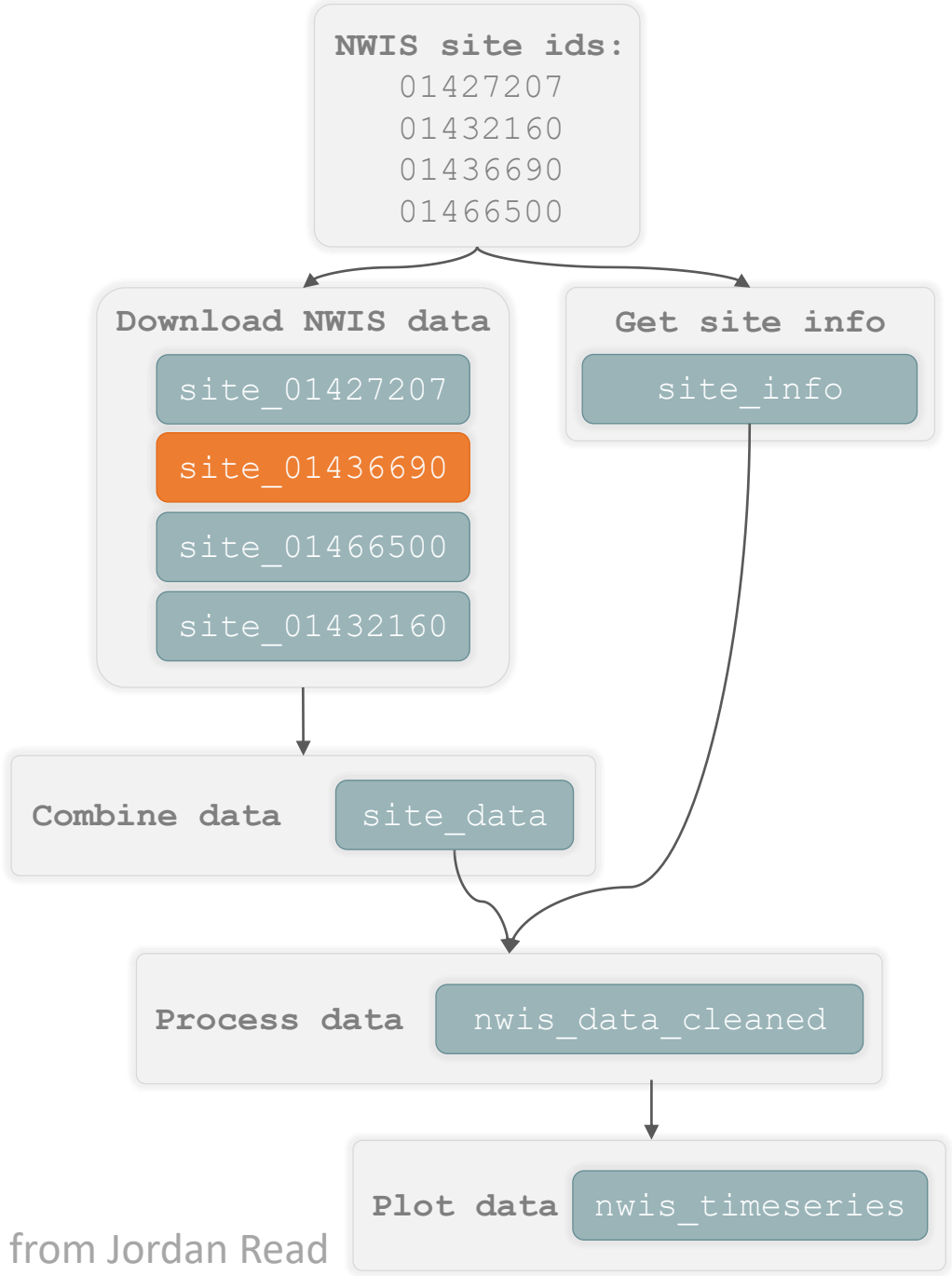
10:59 AM 👍 1

No problem. Easy fix on our end!

Slide from Jordan Read



Slide from Jordan Read



Slide from Jordan Read

R Console | C:/Users/usgs_employee/complex_analysis

```
> run_pipeline()
> Skipping site_01427207
> Building site_01436690
> Skipping site_01466500
> Skipping site_01432160
> Building site_data
> Skipping site_info
> Building nwis_data_cleaned
> Building nwis_timeseries_plot
```

Sprague, Lori A 10/27 10:58 AM

Shoot, I know I told you we were done with the analysis and data release, but I just heard that site 01436690 had a sensor malfunction and they've since corrected the data

10:59 AM 1

Umm…I *think* I can figure out which results are impacted

Slide from Jordan Read

Sprague, Lori A 10/27 10:58 AM

Shoot, I know I told you we were done with the analysis and data release, but I just heard that site 01436690 had a sensor malfunction and they've since corrected the data

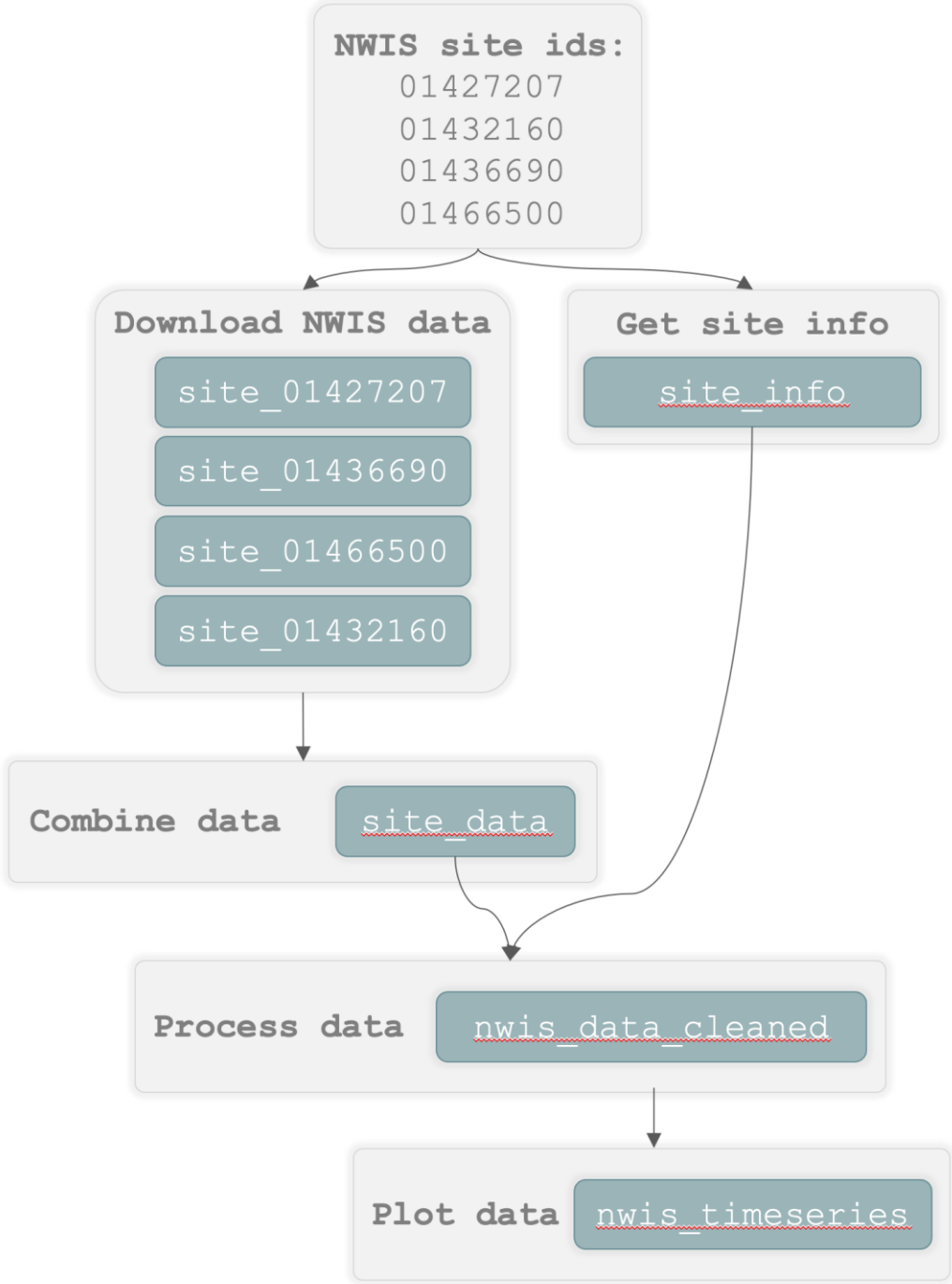
10:59 AM 1

Umm…I *think* I can figure out which results are impacted

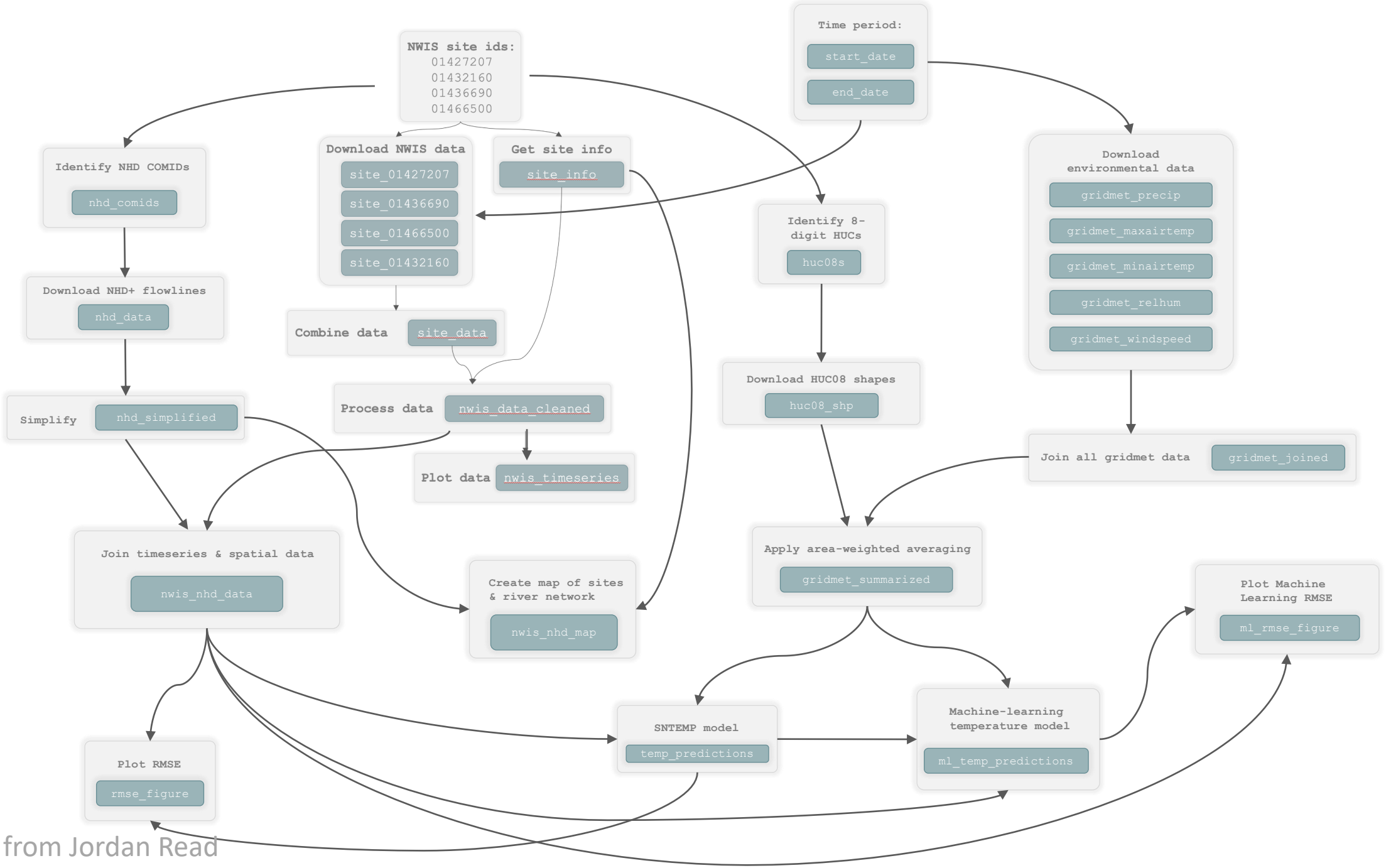
11:00 AM 1

Never mind. I can just re-run the whole thing. Might take some time though

Slide from Jordan Read



Slide from Jordan Read



Slide from Jordan Read

Tiers of reproducibility and repeatability

Document and rebuild your own analysis

Scripting

Rebuild, easily reuse and troubleshoot, plus track changes over time

Scripting

Modular functions

Version control

Efficiently rebuild a full analysis and collaborate effectively on development

Scripting

Modular functions

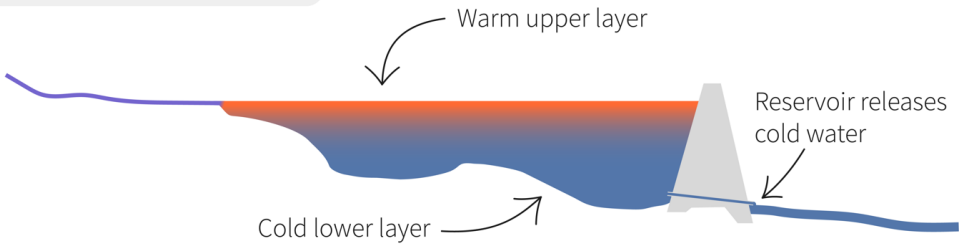
Version control

Dependency tracking

Automated Workflows

Slide from Jordan Read

Reservoir thermal bank



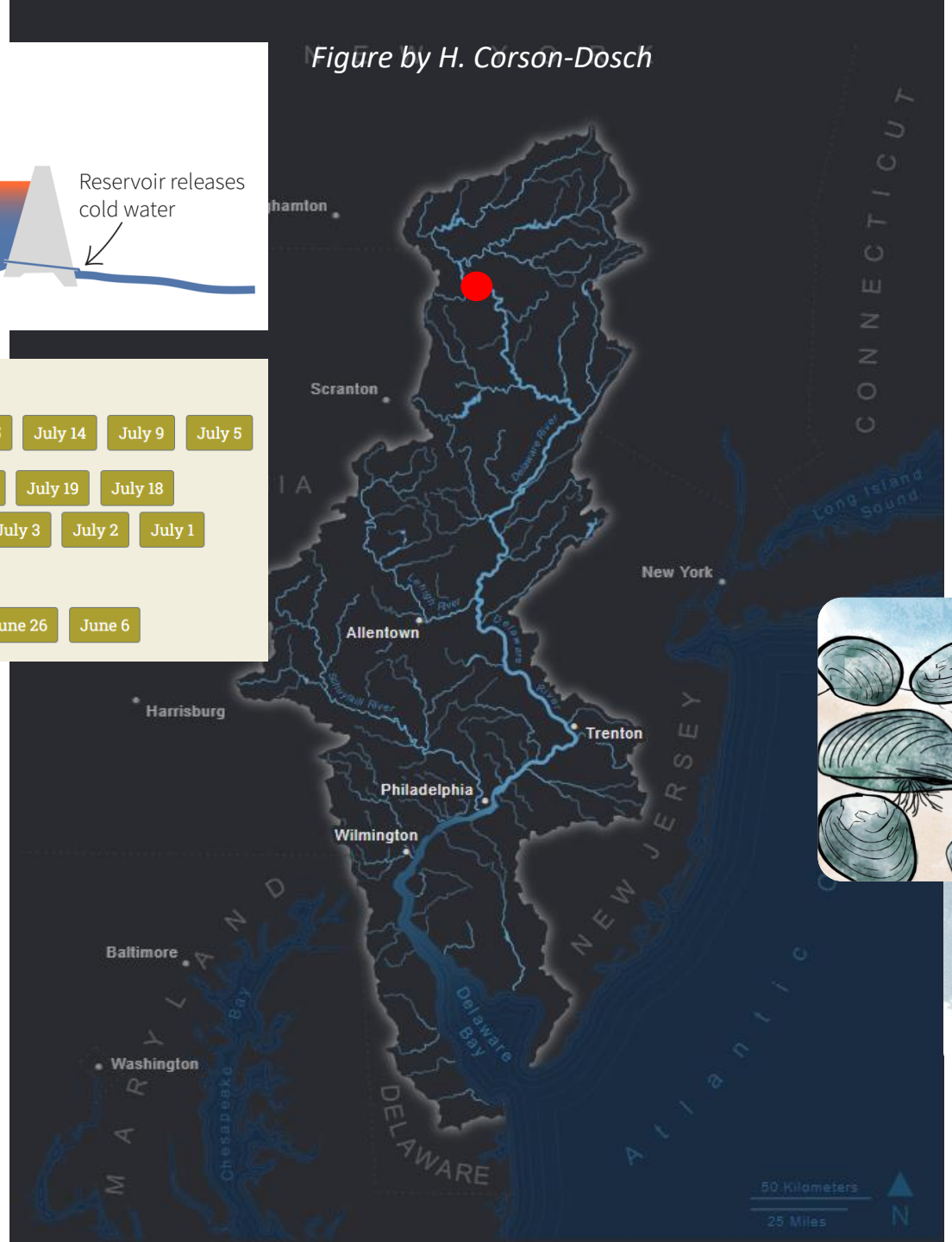
Thermal Releases

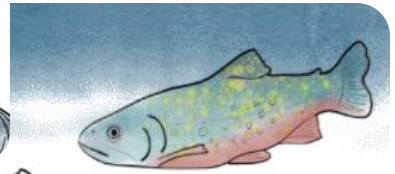
2019 August 2 | July 29 | July 27 | July 19 | July 17 | July 15 | July 14 | July 9 | July 5

2020 July 31 | July 28 | July 27 | July 26 | July 25 | July 20 | July 19 | July 18 | July 17 | July 14 | July 8 | July 7 | July 6 | July 5 | July 4 | July 3 | July 2 | July 1 | June 22 | June 20 | June 18 | June 9

2021 August 11 | August 10 | June 29 | June 28 | June 27 | June 26 | June 6

Figure by H. Corson-Dosch

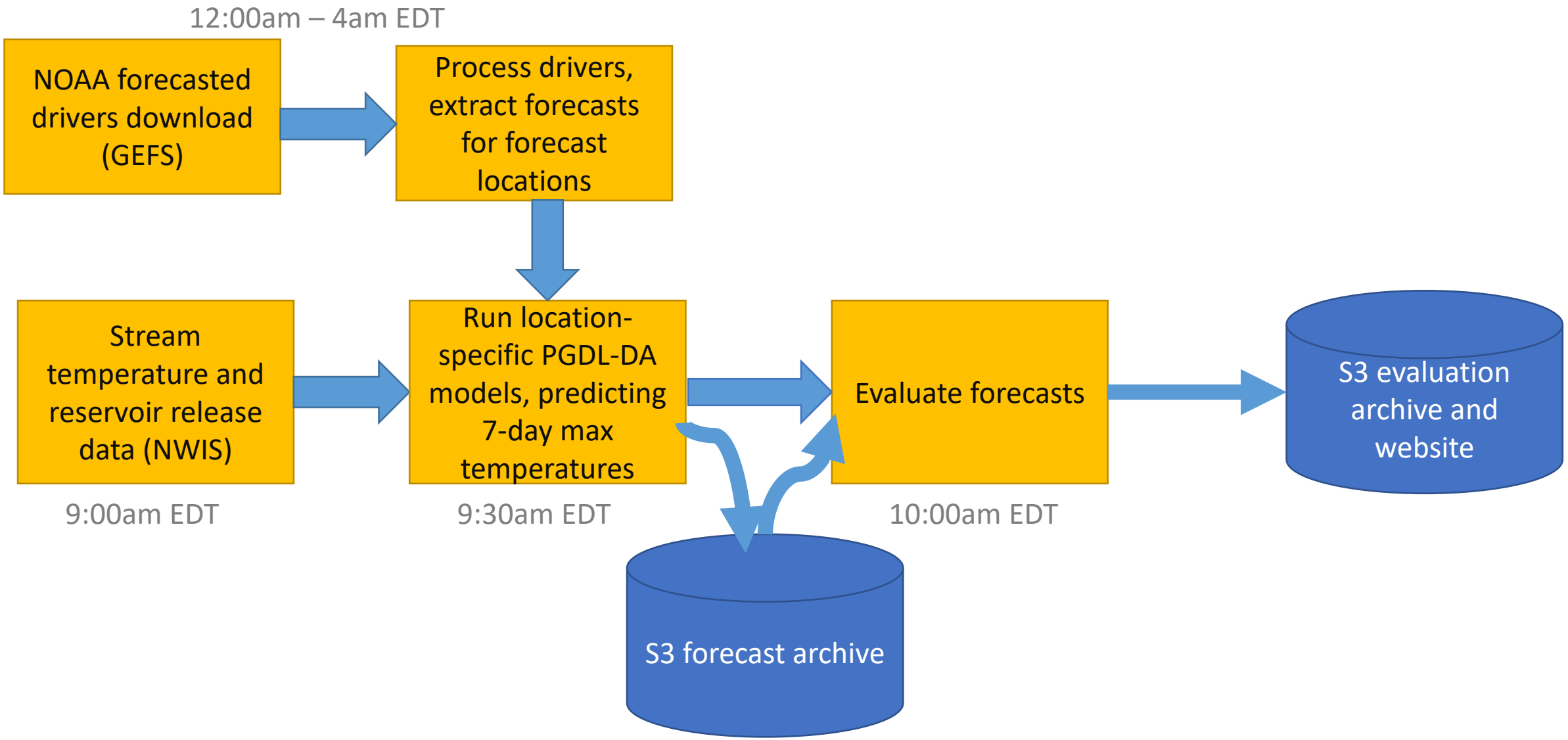


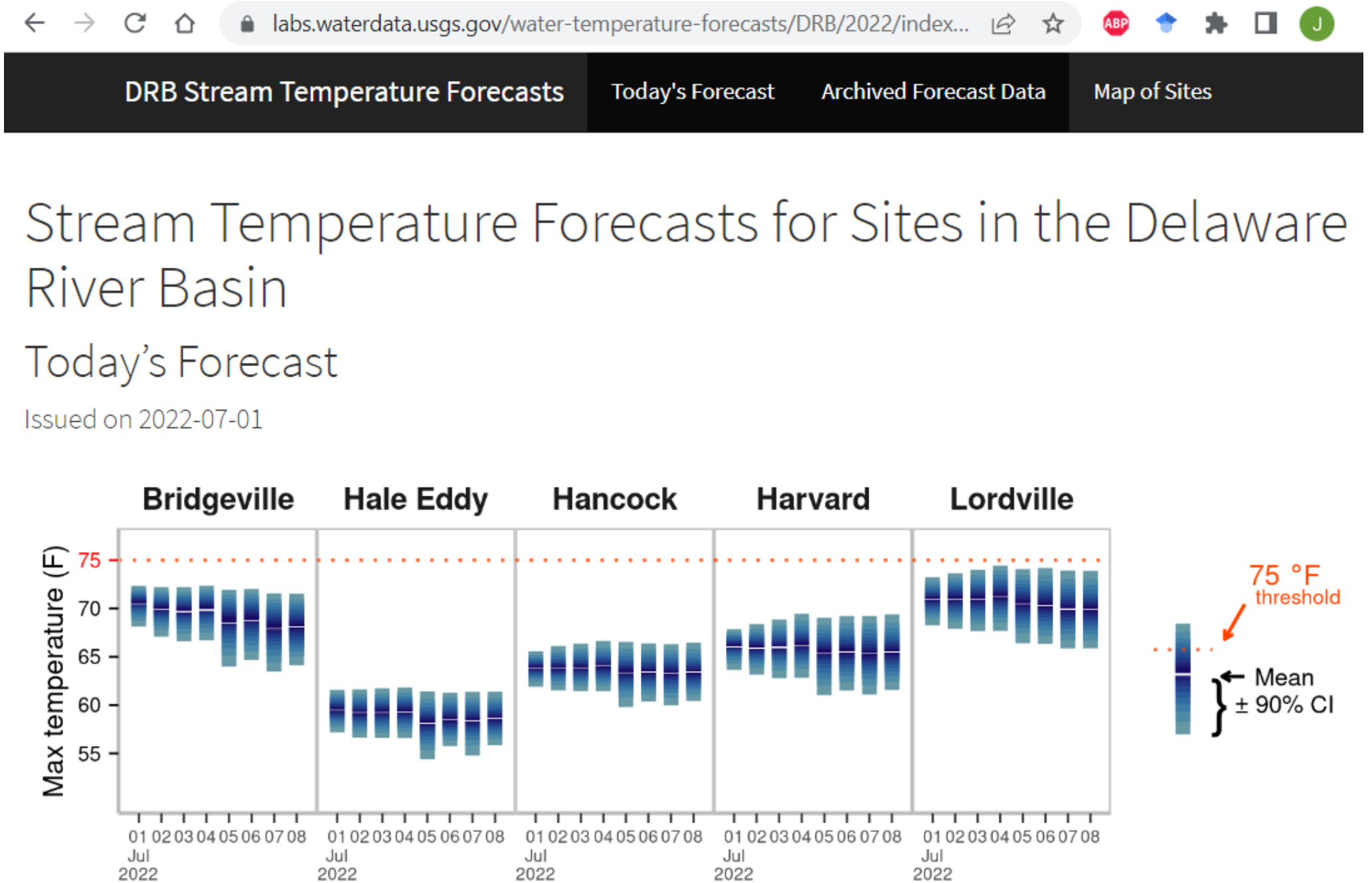


Operational Needs

- Produce daily updating 7-day forecast
- Use most recent meteorological forecasts and observation data available
- Forecasts visible to NYC reservoir operators by mid-morning EDT

Daily workflow





https://labs.waterdata.usgs.gov/water-temperature-forecasts/DRB/2022/index.html

Keys to success

Software dev best practices: Able to isolate operational workflows from code under development

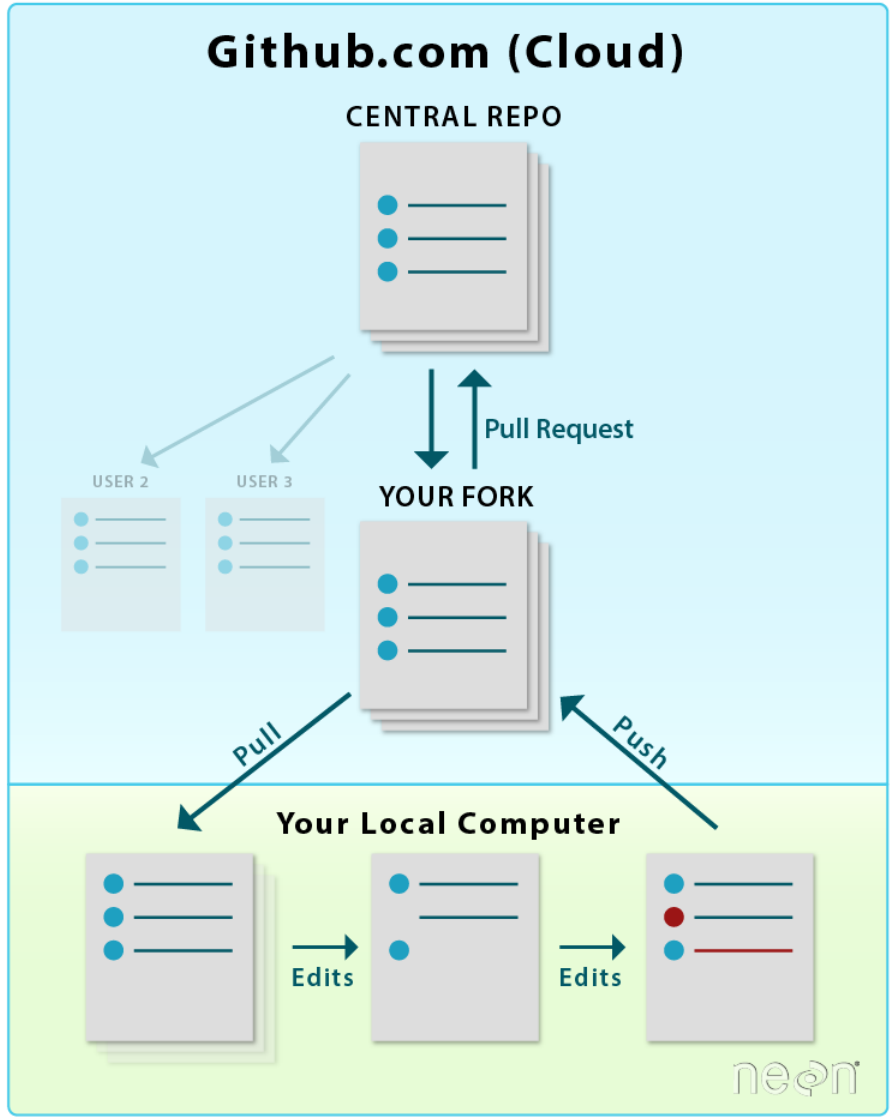
Modular design: Workflow pieces could be modified and scheduled independently

Containers: Workflows were instantly portable to cloud from local machines

Teamwork: Clearly defined roles, diversity of skills, and focused communication among staff

Software development best practices

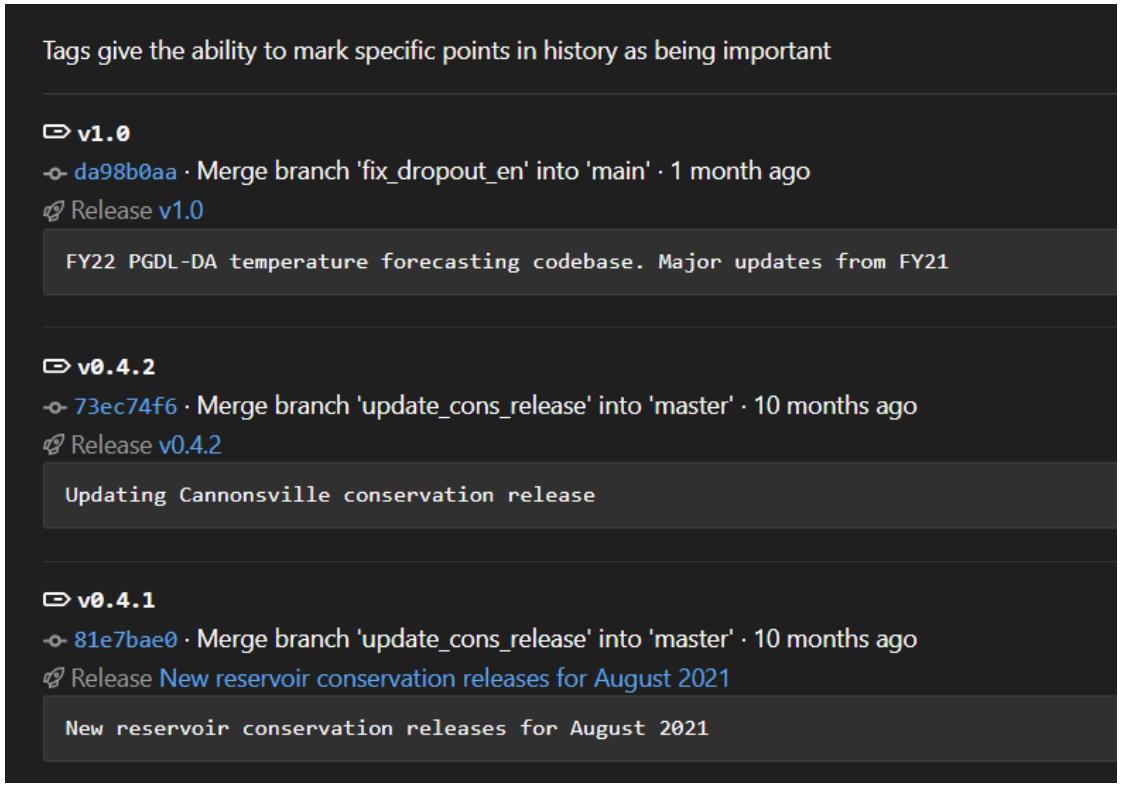
- Version control – Git and GitHub



https://www.earthdatascience.org/workshops/intro-version-control-git/pull-request/

Software development best practices

- Version control – Git and GitHub
- Use tagged releases – snapshots of codebase



Software development best practices

- Version control – Git and GitHub
- Use tagged releases – snapshots of codebase

Tags give the ability to mark specific points in history as being important

v1.0
da98b0aa · Merge branch 'fix_dropout_en' into 'main' · 1 month ago
Release v1.0

```
FY22 PGDL-DA temperature forecasting codebase. Major updates from FY21
```

v0.4.2
73ec74f6 · Merge branch 'update_cons_release' into 'master' · 10 months ago
Release v0.4.2

```
Updating Cannonsville conservation release
```

v0.4.1
81e7bae0 · Merge branch 'update_cons_release' into 'master' · 10 months ago
Release New reservoir conservation releases for August 2021

```
New reservoir conservation releases for August 2021
```

Software development best practices

- Version control – Git and GitHub
- Use tagged releases – snapshots of codebase
- Unit testing – ensure parts of code behave as expected

Software development best practices

- Version control – Git and GitHub
- Use tagged releases – snapshots of codebase
- Unit testing – ensure parts of code behave as expected
- Continuous integration – automated tests and deployment

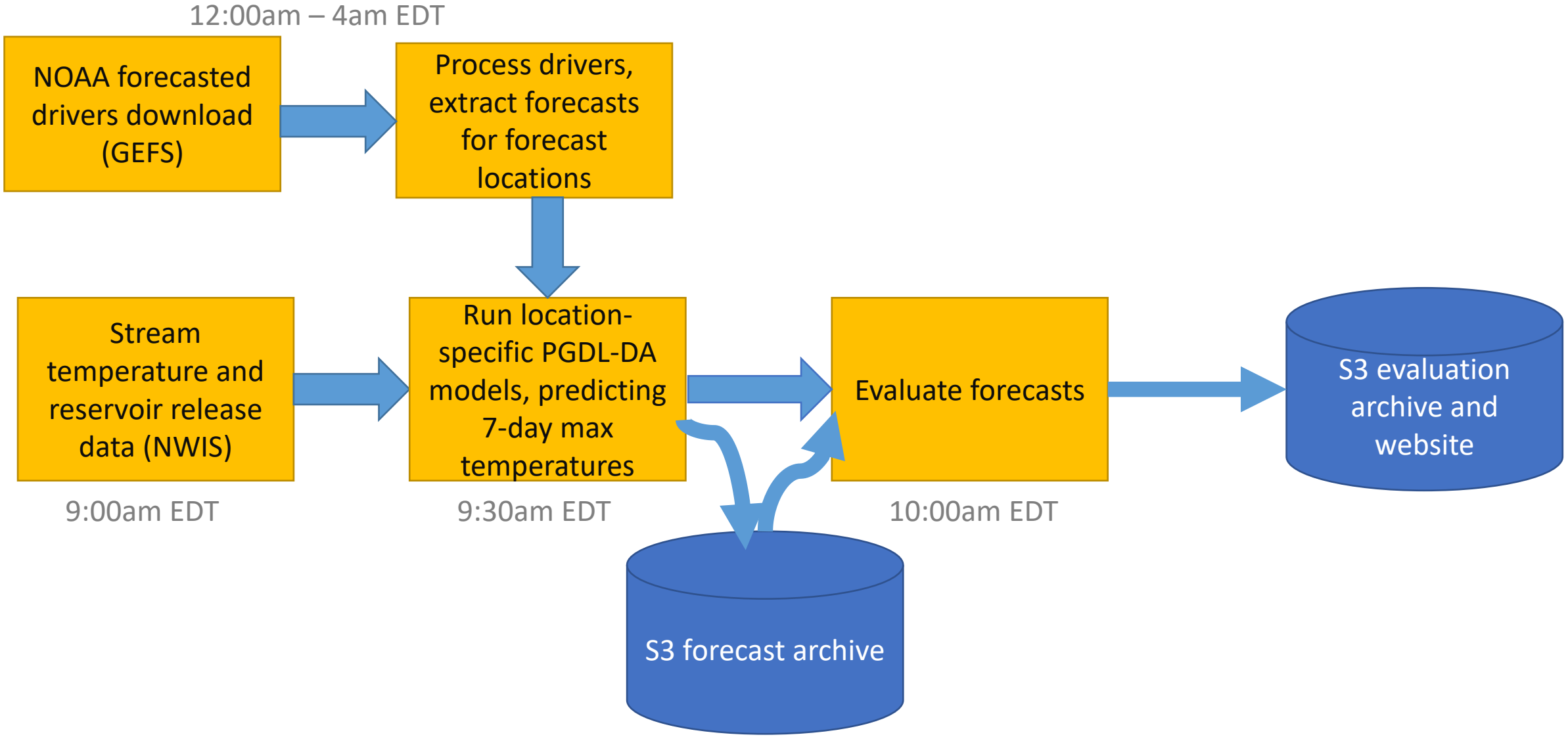
testthat

GitHub Actions

Jenkins

CI/CD

Modular design



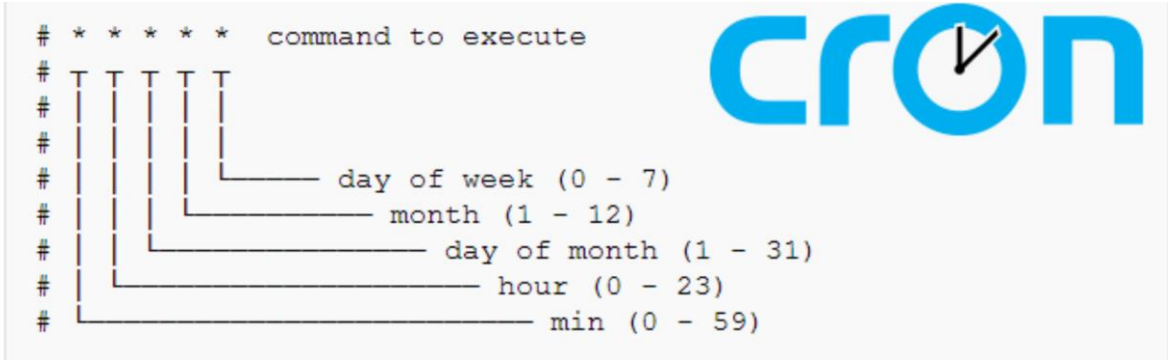
Modular design

- Separate repository for each part of the entire pipeline

12:00am – 4am EDT

NOAA forecasted drivers download (GEFS)

Process drivers, extract forecasts for forecast locations



Stream temperature and reservoir release data (NWIS)

9:00am EDT

Run location-specific PGDL-DA models, predicting 7-day max temperatures

9:30am EDT

Evaluate forecasts

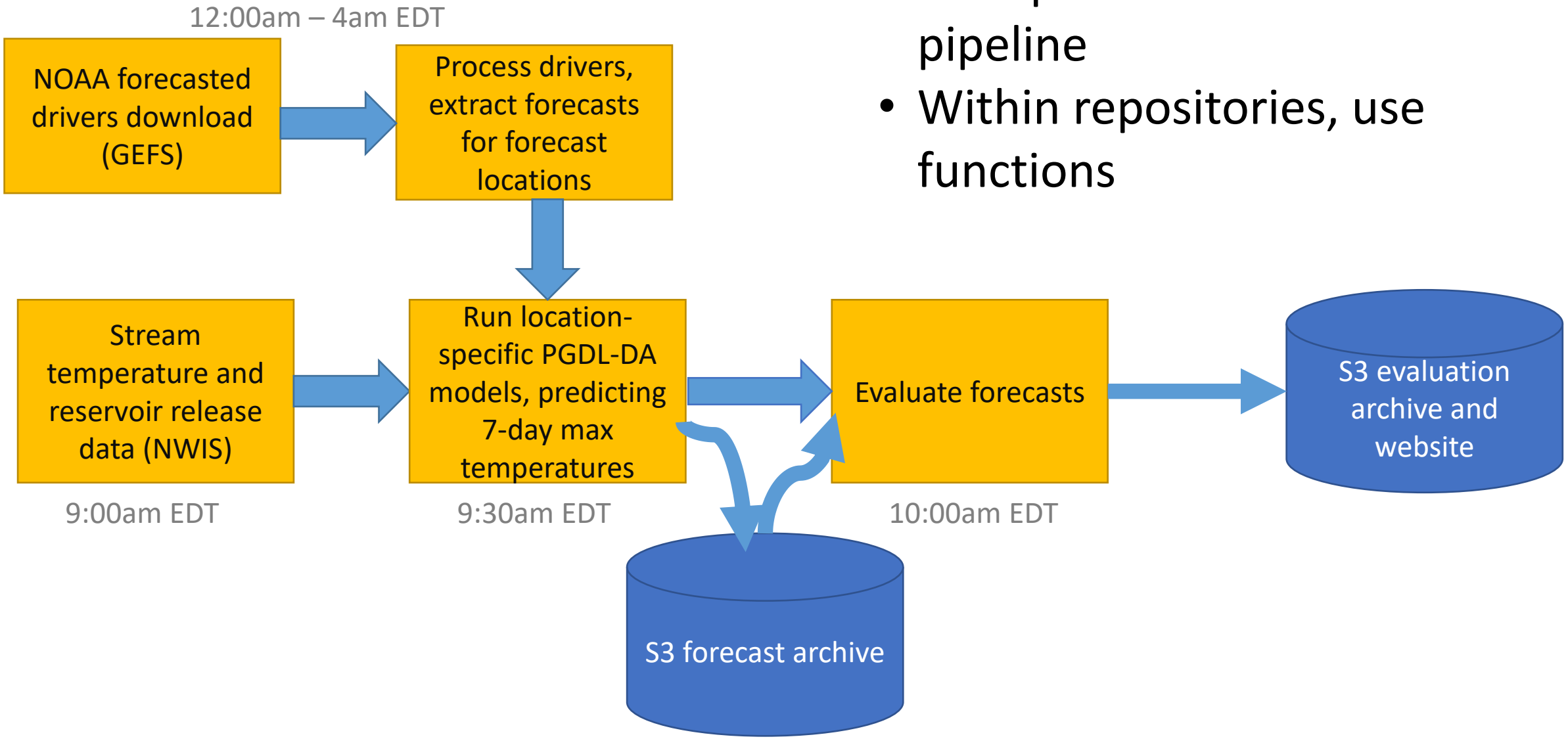
10:00am EDT

S3 evaluation archive and website

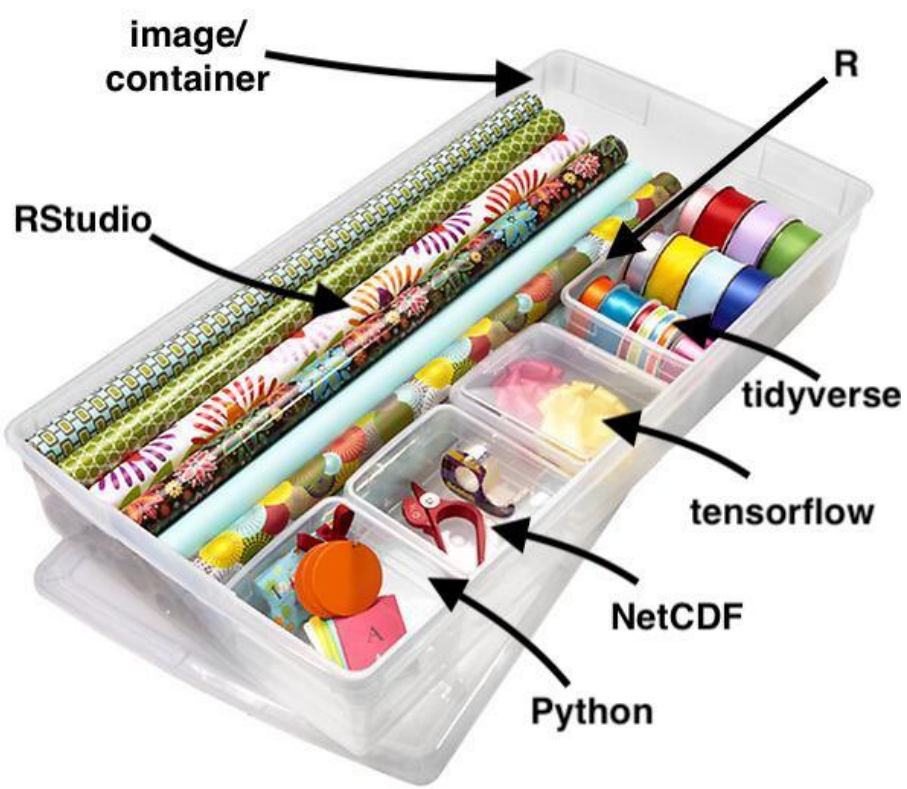
S3 forecast archive

Modular design

- Separate repository for each part of the entire pipeline
- Within repositories, use functions



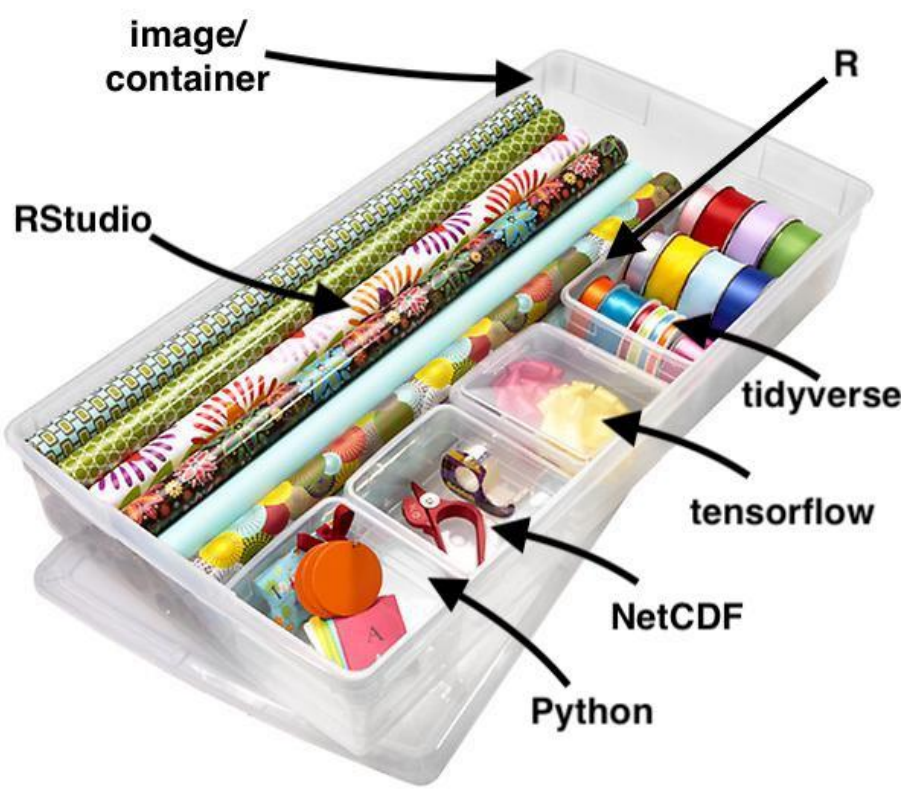
Containers



- Box that carries all the tools you need for a set of tasks

Image from David Watkins

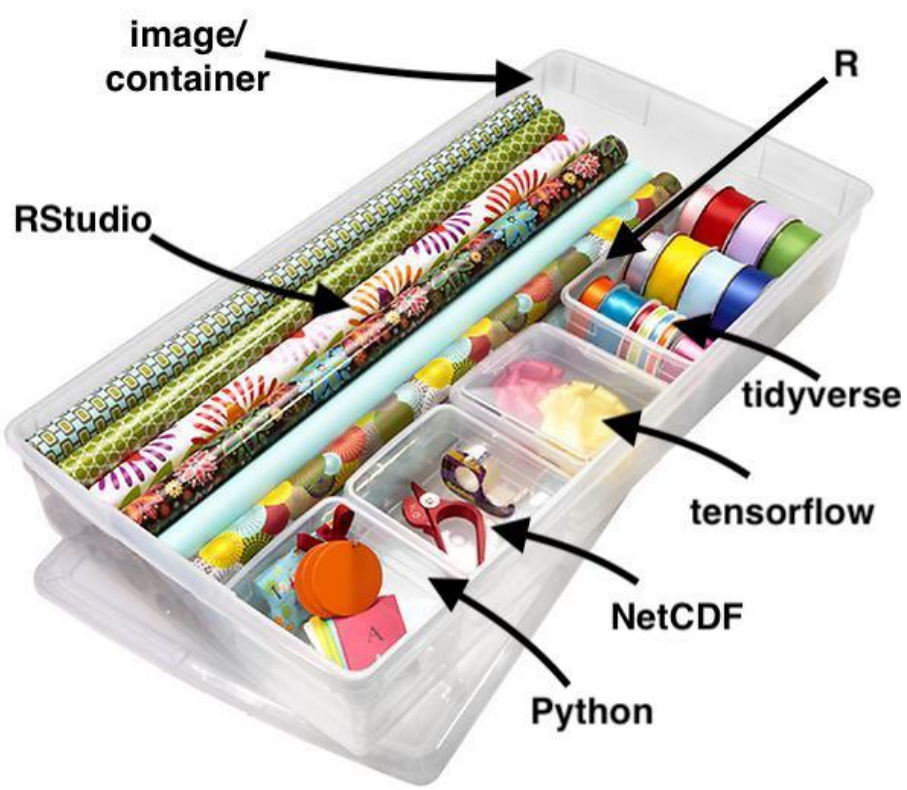
Containers



- Box that carries all the tools you need for a set of tasks
- Docker https://docs.docker.com/get-started/overview/

Image from David Watkins

Containers



- Box that carries all the tools you need for a set of tasks
- Docker https://docs.docker.com/get-started/overview/
- rocker – containers for the R environment

The Rocker Project
Docker Containers for the R Environment

Image from David Watkins

Containers

2 Images

2.1 The versioned stack

image	base image	description	pulls
rocker/r-ver	ubuntu	Install R from source and set RSPM as default CRAN mirror	docker pulls 2.4M
rocker/rstudio	rocker/r-ver	Adds RStudio Server	docker pulls 7.9M
rocker/tidyverse	rocker/rstudio	Adds tidyverse packages & devtools	docker pulls 4.8M
rocker/verse	rocker/tidyverse	Adds tex & publishing-related package	docker pulls invalid
rocker/geospatial	rocker/verse	Adds geospatial packages	docker pulls 632k
rocker/binder	rocker/geospatial	Adds requirements to run repositories on mybinder.org	
rocker/shiny	rocker/r-ver	Adds shiny server	docker pulls 1.7M
rocker/shiny-verse	rocker/shiny	Adds tidyverse packages	docker pulls invalid
rocker/cuda	rocker/r-ver	Adds CUDA support to rocker/r-ver	docker pulls invalid
rocker/ml	rocker/cuda	Adds CUDA support to rocker/tidyverse	docker pulls invalid
rocker/ml-verse	rocker/ml	Adds CUDA support to rocker/geospatial	docker pulls 21k

- Box that carries all the tools you need for a set of tasks
- Docker https://docs.docker.com/get-started/overview/
- rocker – containers for the R environment

The Rocker Project

Docker Containers for the R Environment

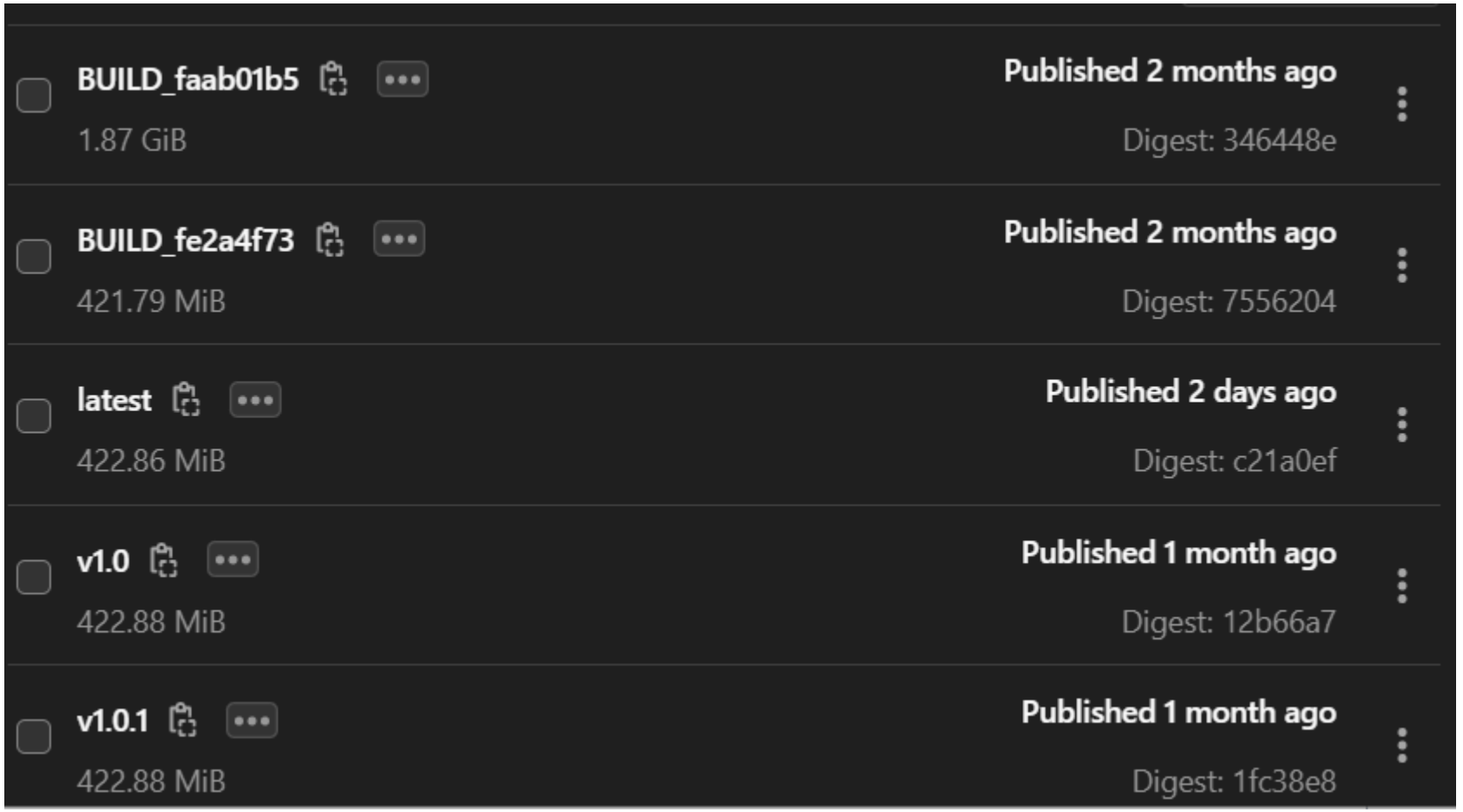
Containers + Continuous Integration

Do something if these files change

Build docker container based on our recipe

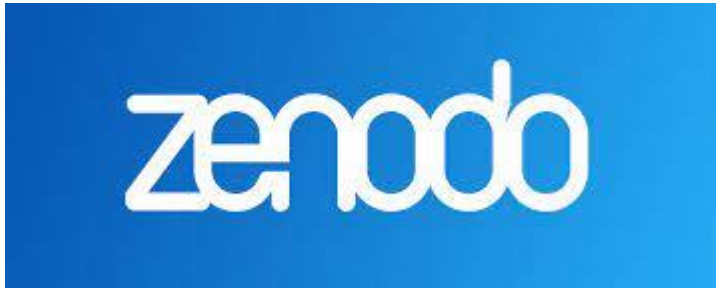
```
39    rules:
40      - changes:
41        - docker/Dockerfile
42        - docker/Makefile
43        - docker/requirements.txt
44        - docker/DOIRootCA2.crt
45        - .gitlab-ci.yml
46    script:
47      - echo "${CI_REGISTRY_PASSWORD}" | docker login -u "${CI_RE
48      - cd docker
49      - apk add make
50      - make build
51      - make push
```

Containers + Continuous Integration



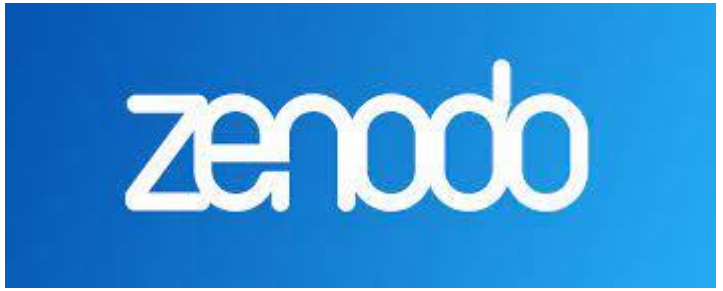
Dissemination

- Publishing code – Zenodo integrates with GitHub



Dissemination

- Publishing code – Zenodo integrates with GitHub



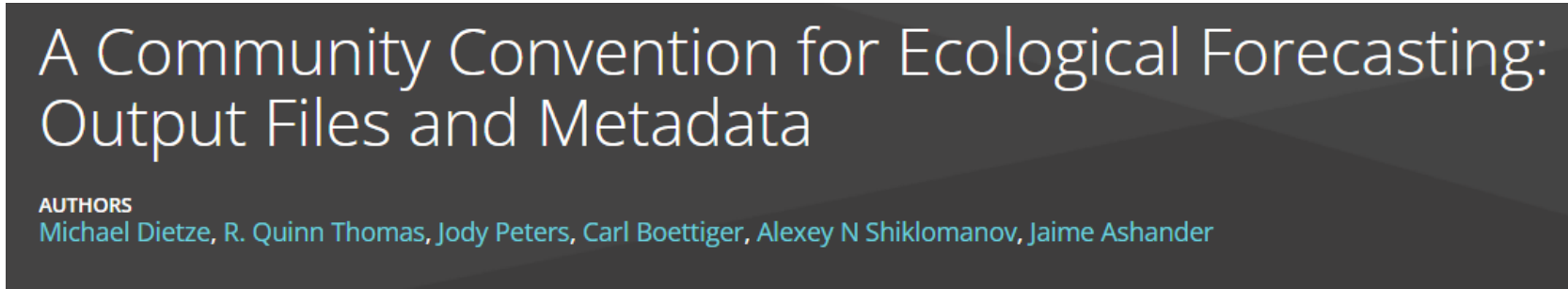
- Publishing forecasts – AWS, DataOne, Environmental Data Initiative

Dissemination

- Publishing code – Zenodo integrates with GitHub

- Publishing forecasts – AWS, DataOne, Environmental Data Initiative

- Metadata - https://ecoevorxiv.org/9dgtq/



Pipeline implementation

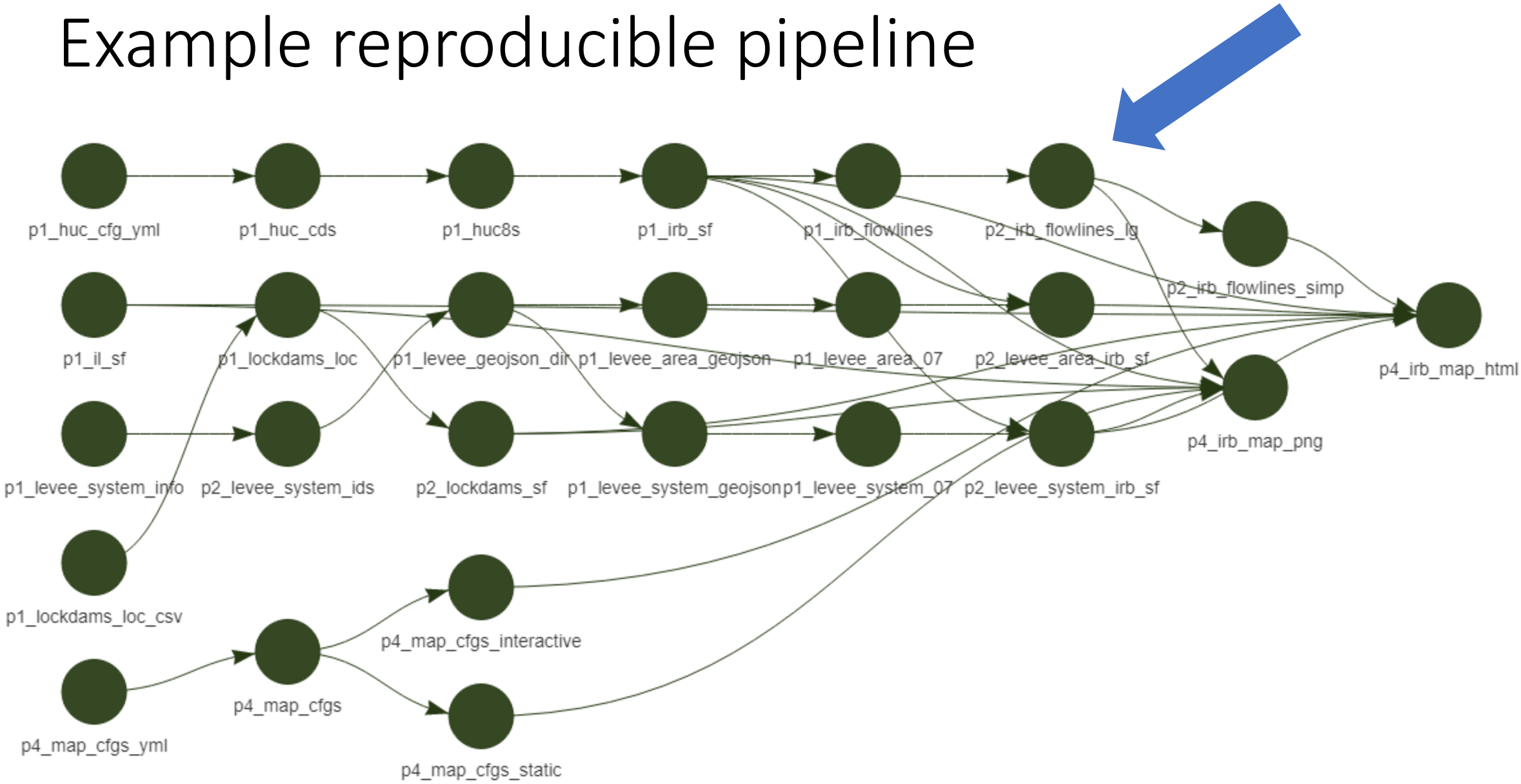
- R based
- Can integrate parts from other languages
- books.ropensci.org/targets

Slide from Lindsay Platt

Example reproducible pipeline

Slide from Lindsay Platt

Example reproducible pipeline

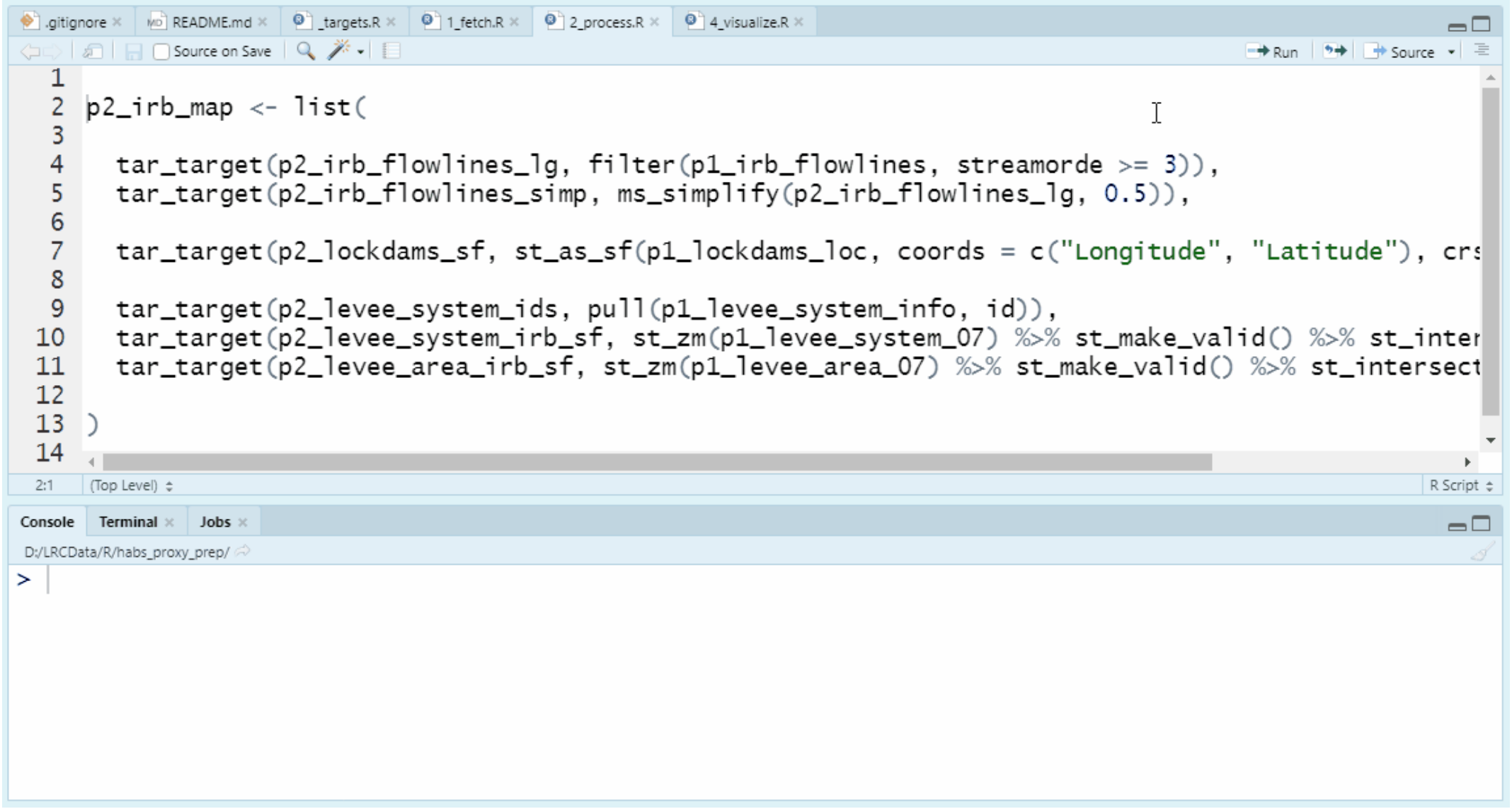


Slide from Lindsay Platt

Example reproducible pipeline

Slide from Lindsay Platt

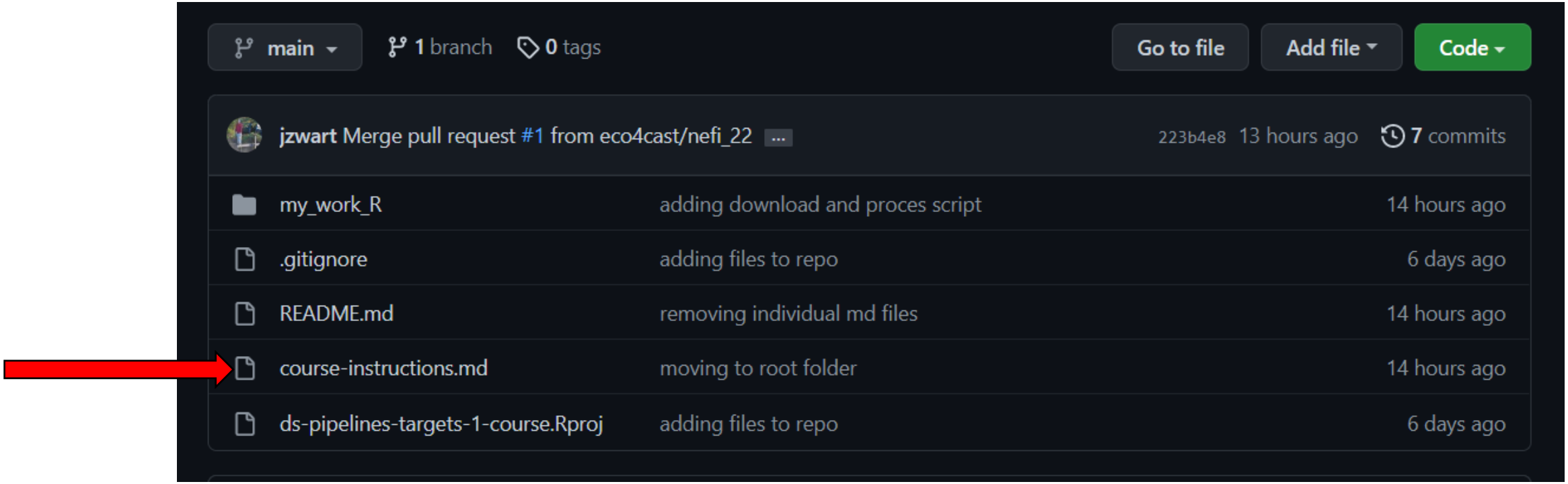
Example reproducible pipeline



jzwart@usgs.gov

Hands On Activity

- https://github.com/eco4cast/ds-pipelines-targets-1-course
- Fork to your GitHub
- Follow details in course-instruction.md



Forecast Example

```yaml
on:
  workflow_dispatch:
  #schedule:
  #- cron: "0 20 * * *"

```

- https://github.com/eco4cast/neon4cast-example