Subsetting Using HydroFrame and HydroData

June 2025, Princeton NJ

Reed Maxwell

The William and Edna
Macaleer Professor
Civil and Environmental
Engineering
Princeton University

Laura Condon

Associate Professor of Hydrology
and Atmospheric Sciences
University of Arizona

George Artavanis

Research Software Engineer
Princeton University

Amy Defnet

Research Software Engineer
Princeton University

It’s not just us, this has been the work of a **great team!**

BILL HASLING
Lead Software Developer
Princeton University

AMY JOHNSON
Software Developer
The University of Arizona

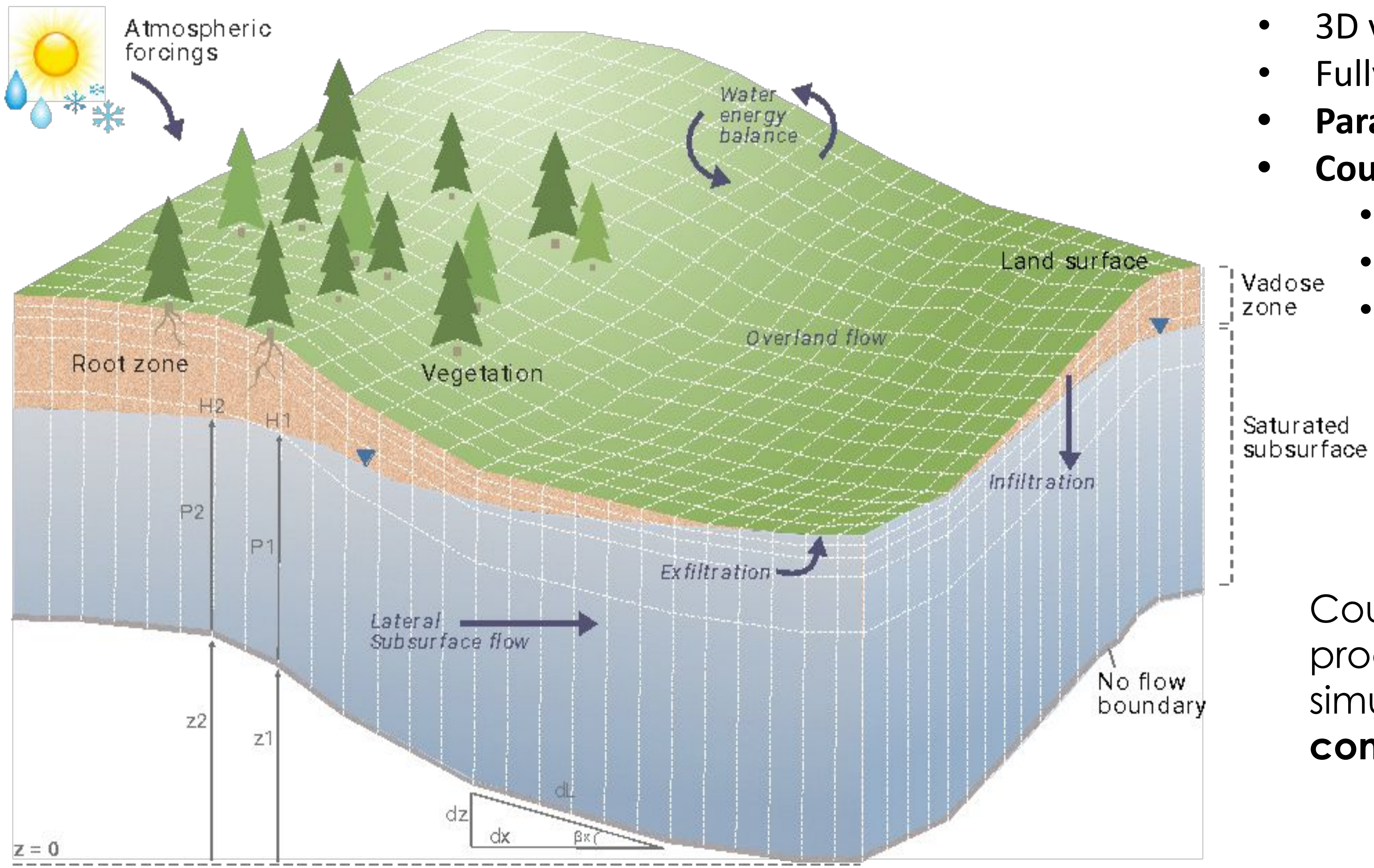
WILL LYTLE
Software Engineer
The University of Arizona

AMANDA TRIPLETT
PhD Student
The University of Arizona

Agenda

1. **Overview of the ParFlow CONUS2.0 model**
2. Overview of HydroFrame and HydroGEN toolset
3. SubsetTools Package Overview and Activity
4. hf_hydrodata Package Overview and Activity

We use the integrated hydrologic model ParFlow



- 3D variably saturated groundwater flow
- Fully integrated surface water
- **Parallel implementation**
- **Coupled land surface processes**
  - Land-energy balance
  - Snow dynamics
  - Driven by meteorology

Maxwell (2013); Kollet and Maxwell (2008); Kollet and Maxwell (2006);Maxwell and Miller (2005); Dai et al. (2003); Jones and Woodward (2001); Ashby and Falgout (1996)

Coupling with land surface processes (CLM) allows for simulation of **interactions** and **connections**

We solve the mixed form of Richards' and Shallow Water equations

$$S_S S_W(h)\frac{\partial h}{\partial t}+\theta\frac{\partial S_W(h)}{\partial t}=\nabla\cdot\mathbf{q}+q_r(\mathbf{x},z)$$

Mixed form of Richards' we solve for h

Upper boundary condition (Neumann type) combined with Shallow Water Equation (same h)

$$\mathbf{k}\cdot\left(-\mathbf{K}_S(\mathbf{x})k_r(h)\cdot\nabla(h+z)\right)=\frac{\partial\|h,0\|}{\partial t}-\nabla\cdot\|h,0\|v_{sw}+q_r(\mathbf{x})$$

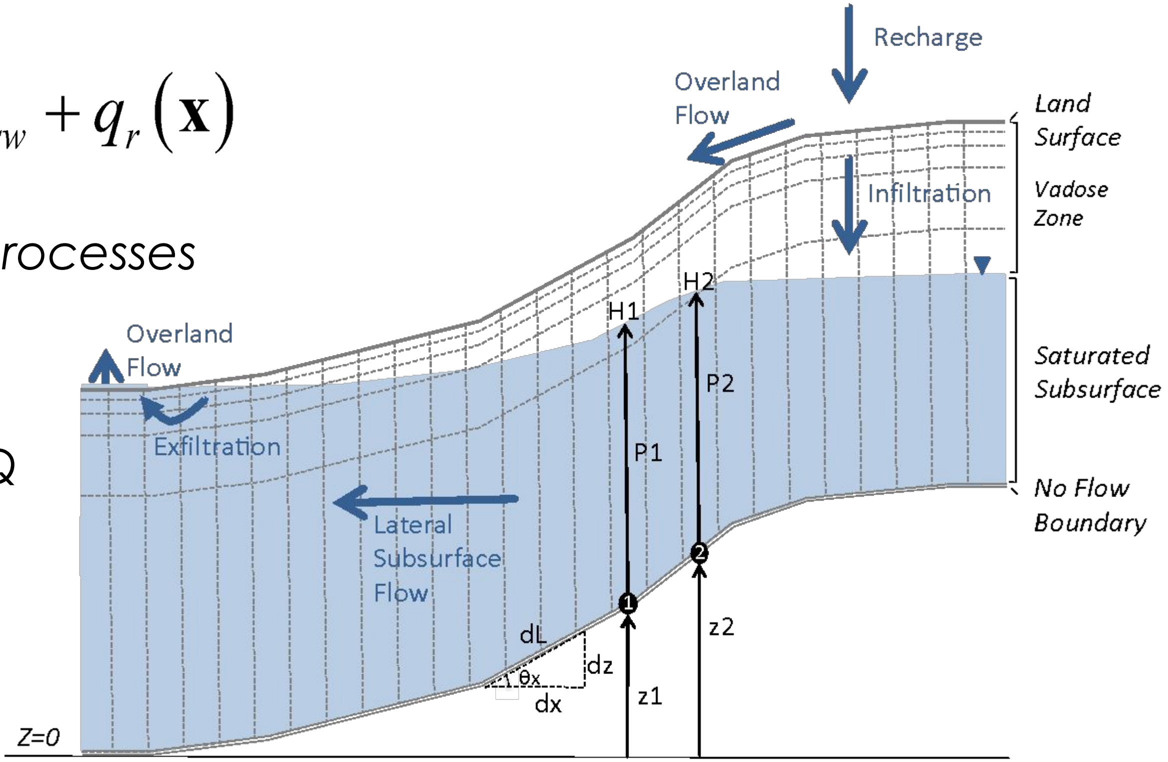
Boundary source/sink, from weather and land surface processes

$$q_r(x)=P(x)-E(x)\qquad q_r(x,z)=-E_T(x,z)$$

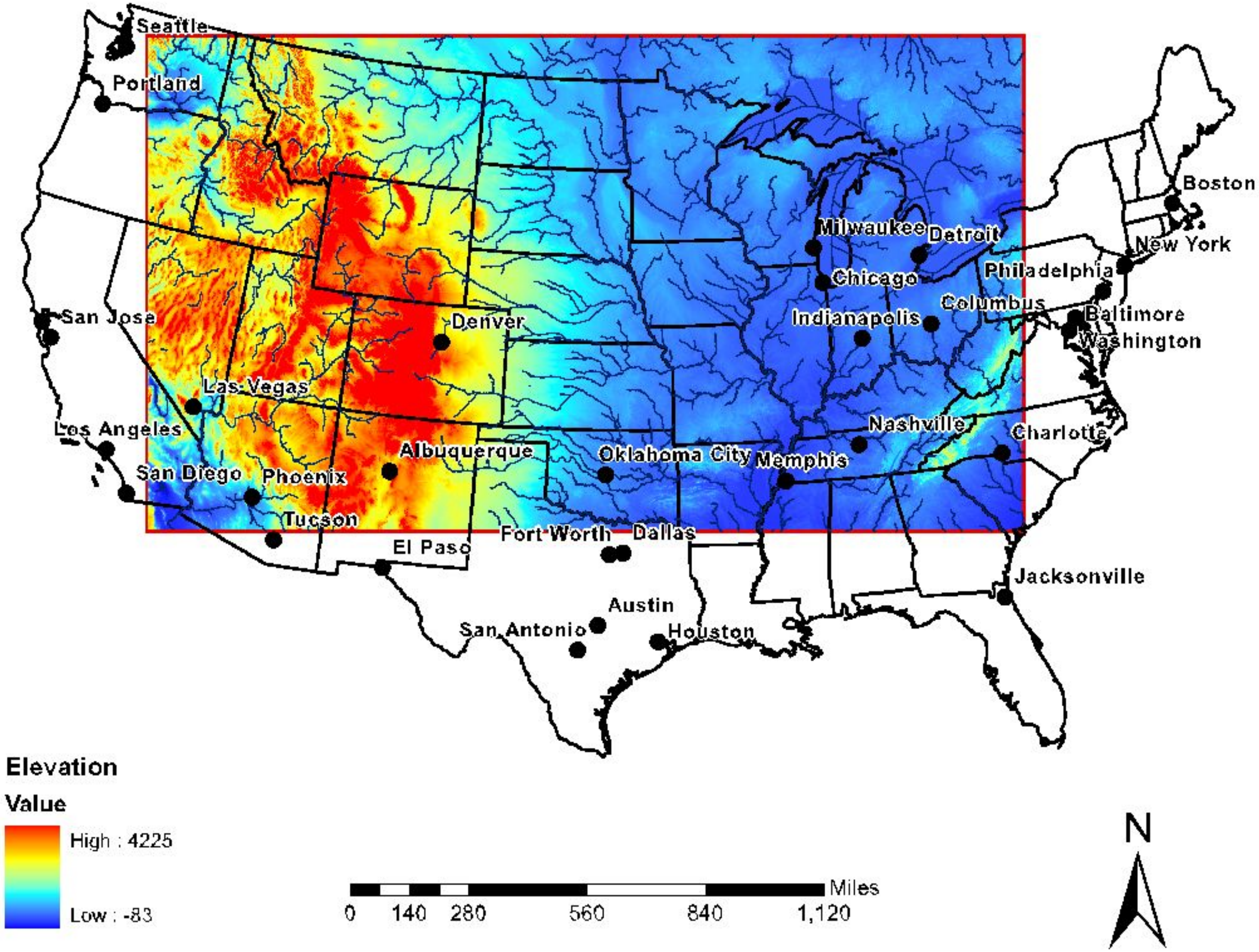
Flux Relationships from modified Darcy and Mannings EQ

$$\mathbf{q}=-\mathbf{K}_S(\mathbf{x})k_r(h)\left[\nabla(h+z)\cos\beta_x+\sin\beta_x\right]$$

$$v_x=-\frac{\sqrt{S_{f,x}}}{n}\psi_s^{\,2/3}$$



In the early teens we built a *proof-of-concept*, 6.3M km^2 domain covering much of the CONUS



- 1 km lateral resolution
- .1 – 100m vertical resolution over 102m depth
- ~32M unknowns
- The Miss and CO watersheds
- *Fully integrated*, 3D Richards' EQ, Shallow Water Equations, Land Surface Processes

Maxwell, Condon and Kollet GMD (2015)

Almost a decade ago we mapped out a path where model, scenario and observations evolve.

This became our road map for CONUS model development

Maxwell and Condon
Science (2016)

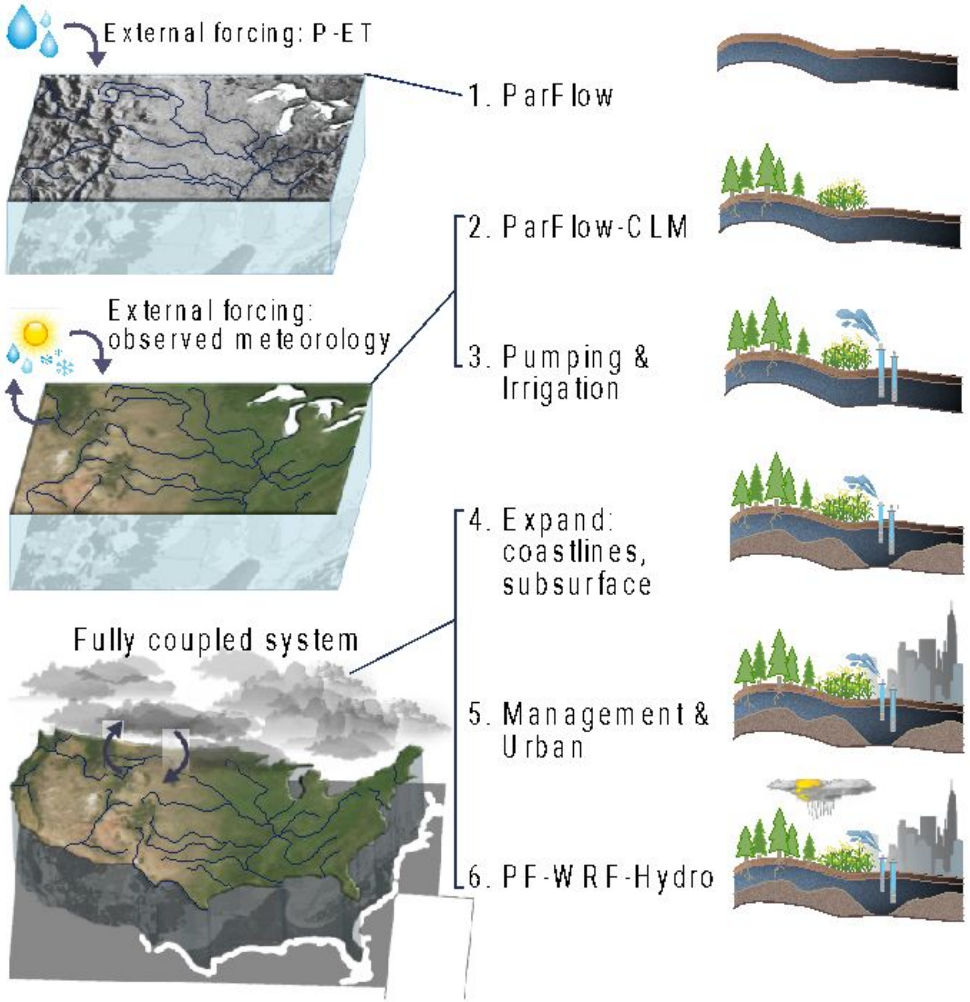


Fig. S6
We envision an evolving roadmap for model development, where complexity increases, anthropogenic influences are added sequentially, model boundaries extend and resolution increases. The starting point for this simulation (1) is a steady-state model driven by P-ET *(20)*. The current model as presented in this work (2) includes a full land surface model but no anthropogenic impacts (included in 3). The progression will evolve to include a full treatment of the atmosphere, coastlines, groundwater pumping, irrigation and water management (5-6). Eventually this could inform and merge into the National Water Model. As we develop successive model generations, models are rigorously compared to all available observations.

This roadmap is the trajectory we are following

Maxwell et al *GMD* (2015), Condon and Maxwell *WRR* (2015), Condon et al *AWR* (2015)

Maxwell and Condon *Science* (2016), Condon and Maxwell *HESS* (2017)

Condon and Maxwell (2019)

Condon et al (2020)

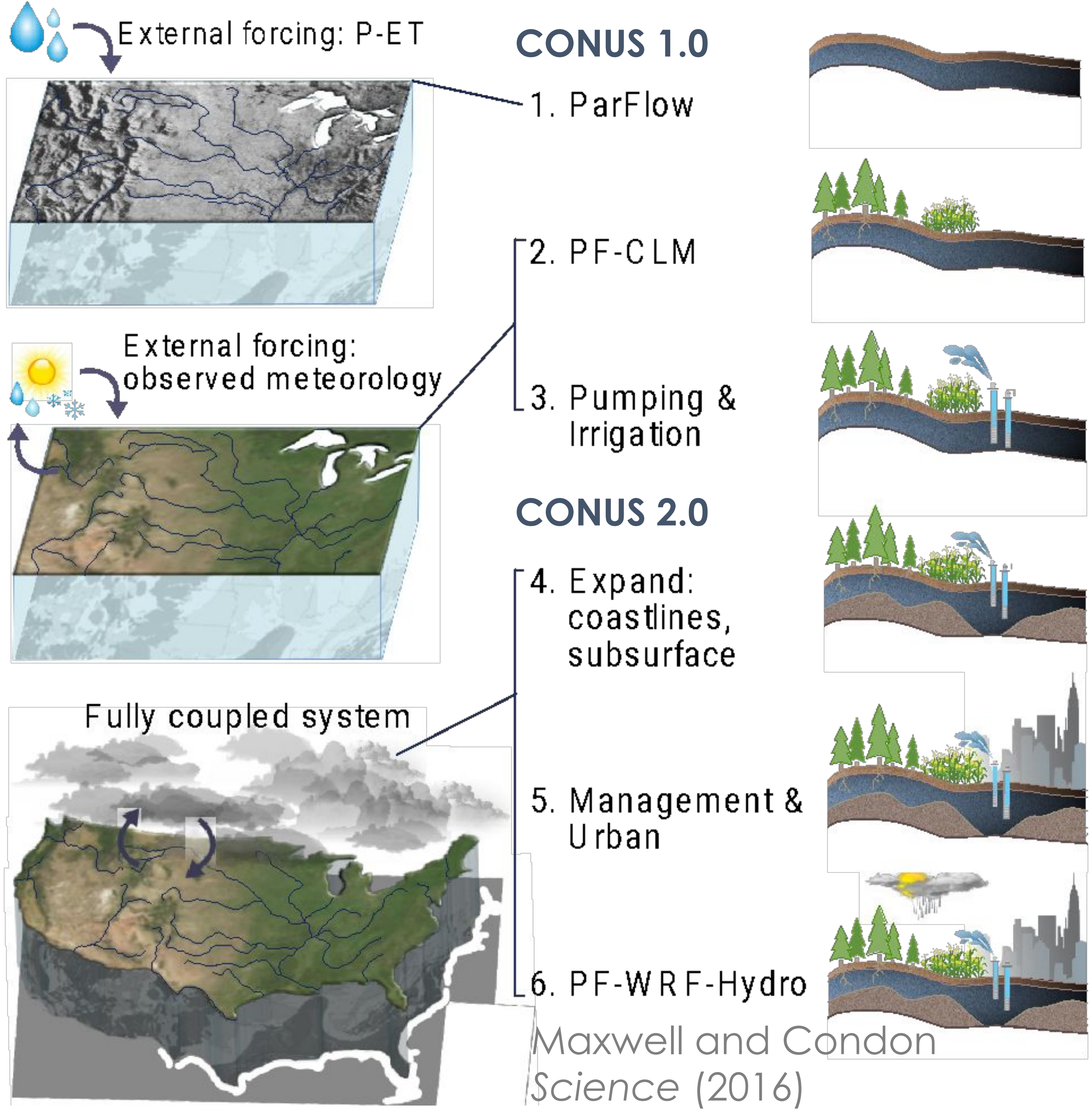
Tijerina et al 2021

O'Neill et al 2021

Zhang et al (2022)

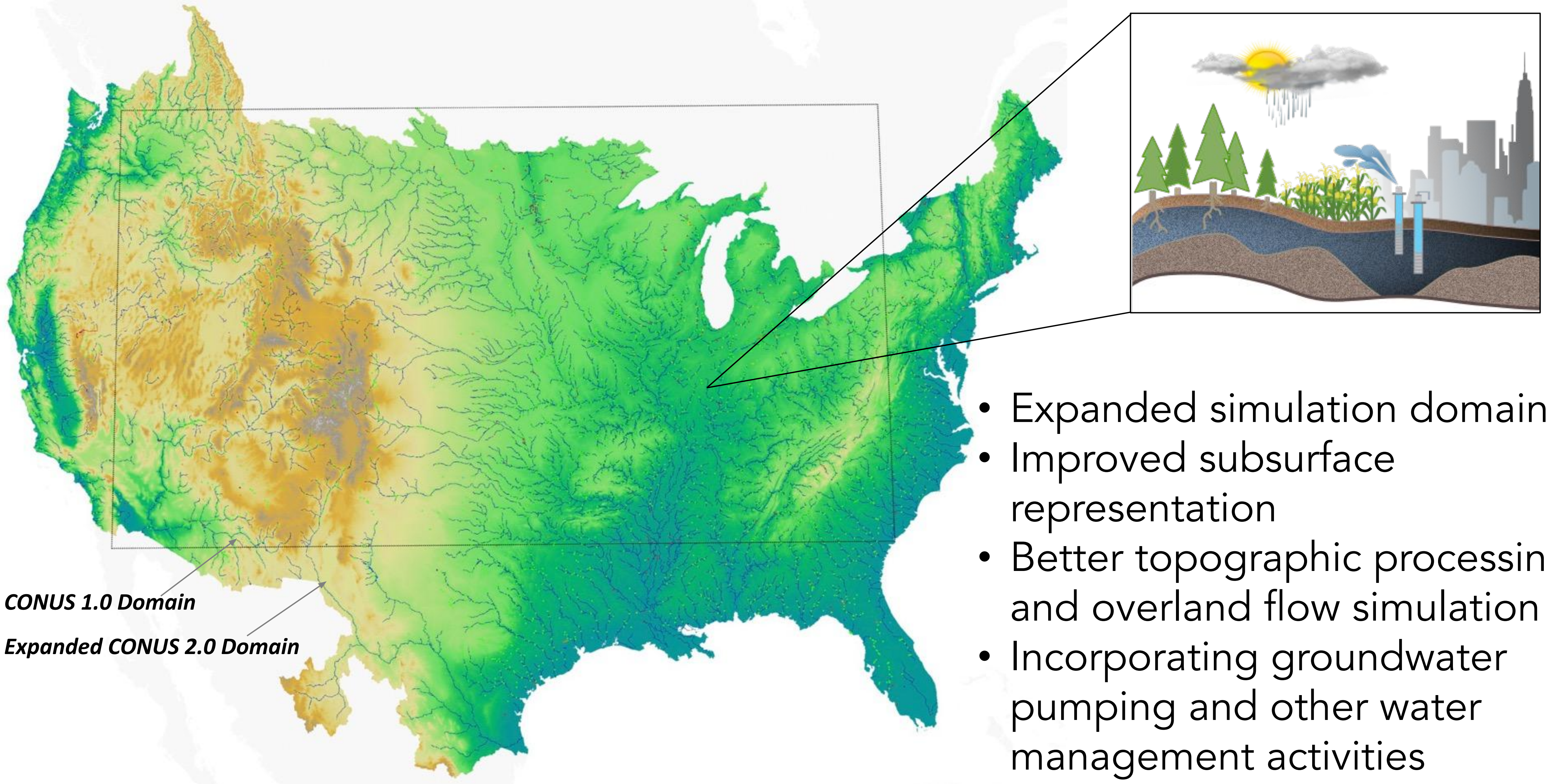
Tijerina et al (2023)

Yang et al (2023)



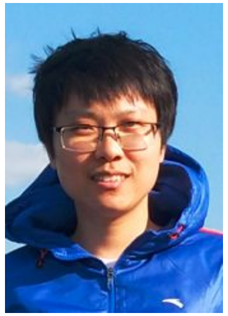
Maxwell and Condon *Science* (2016)

CONUS 2.0 Model development



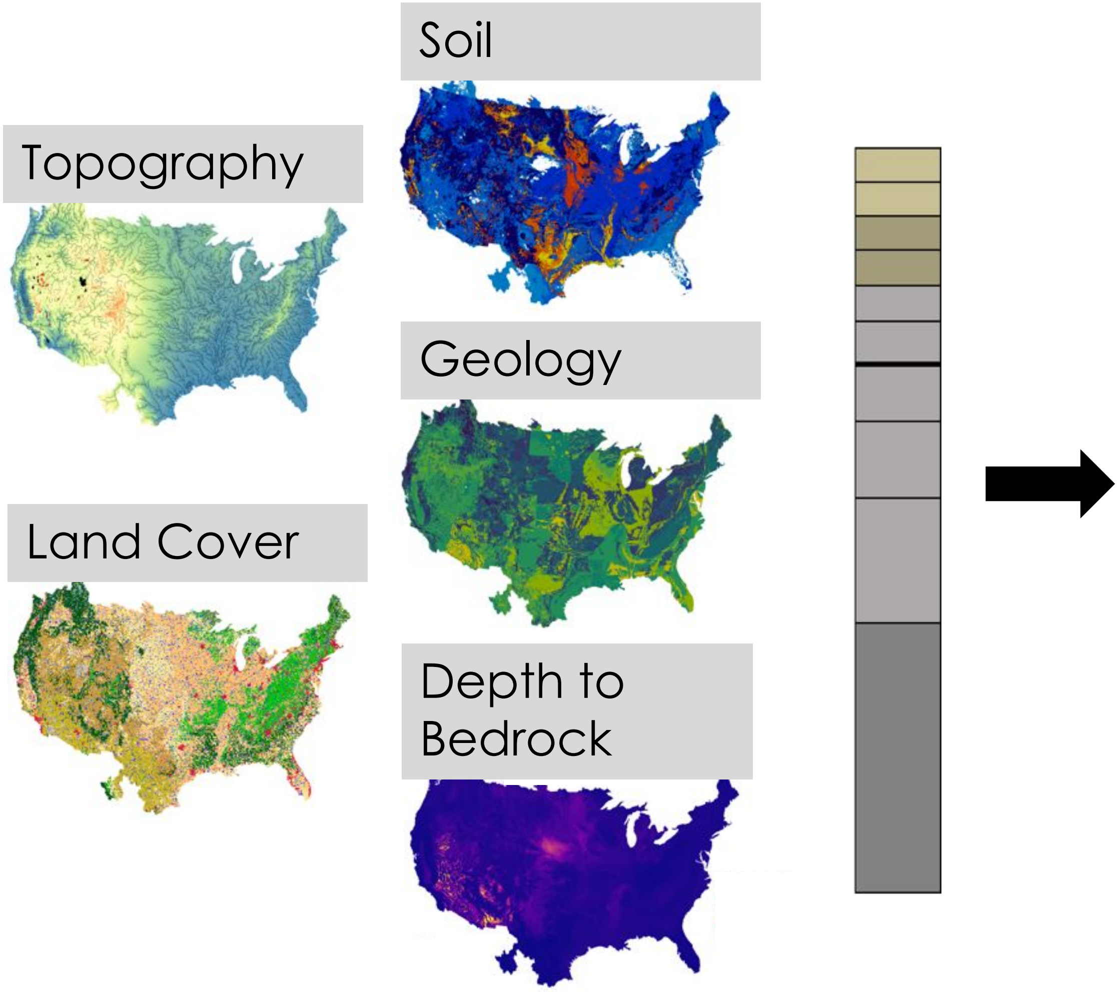
- Expanded simulation domain
- Improved subsurface representation
- Better topographic processing and overland flow simulation
- Incorporating groundwater pumping and other water management activities

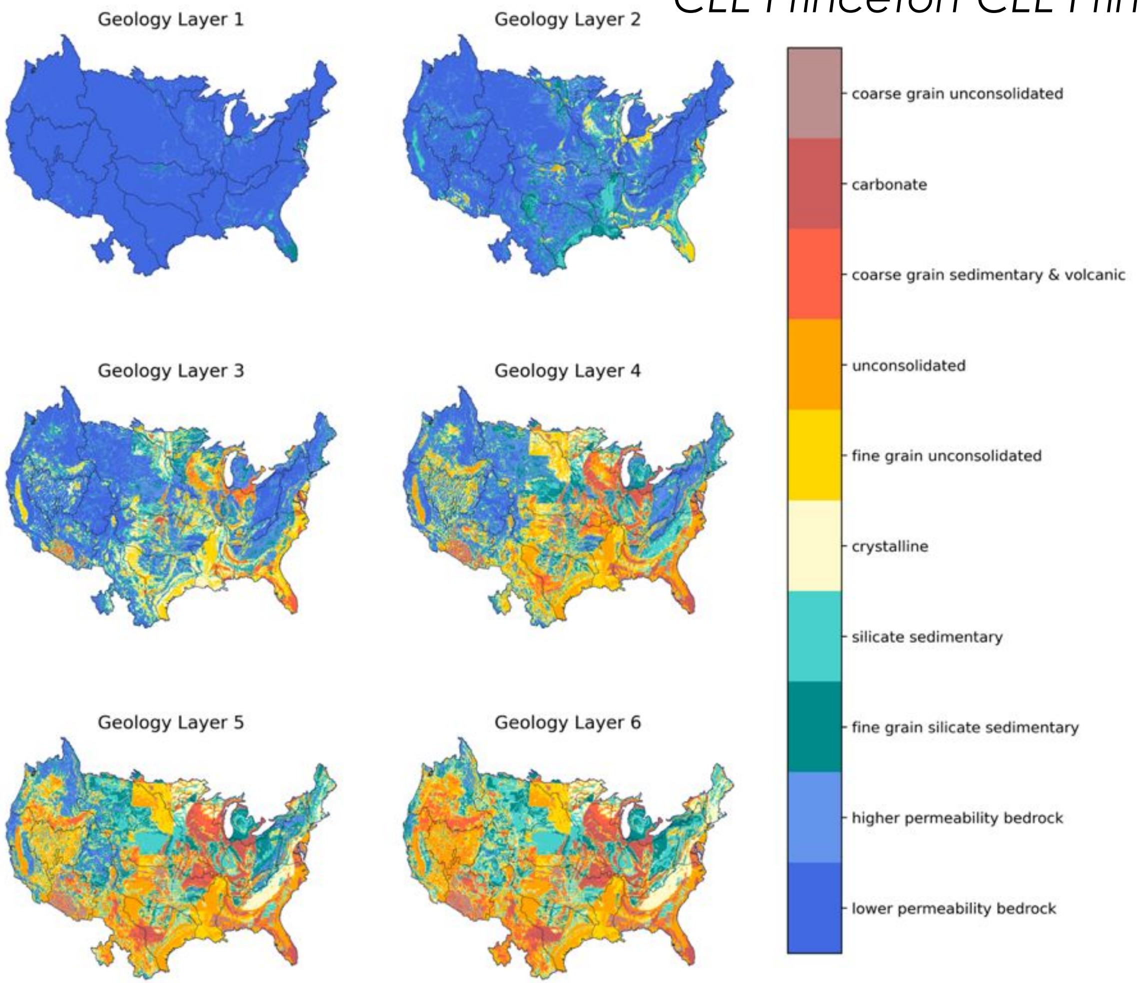
National **model development** and subdomain testing is a **huge endeavor**



Chen Yang
CEE Princeton

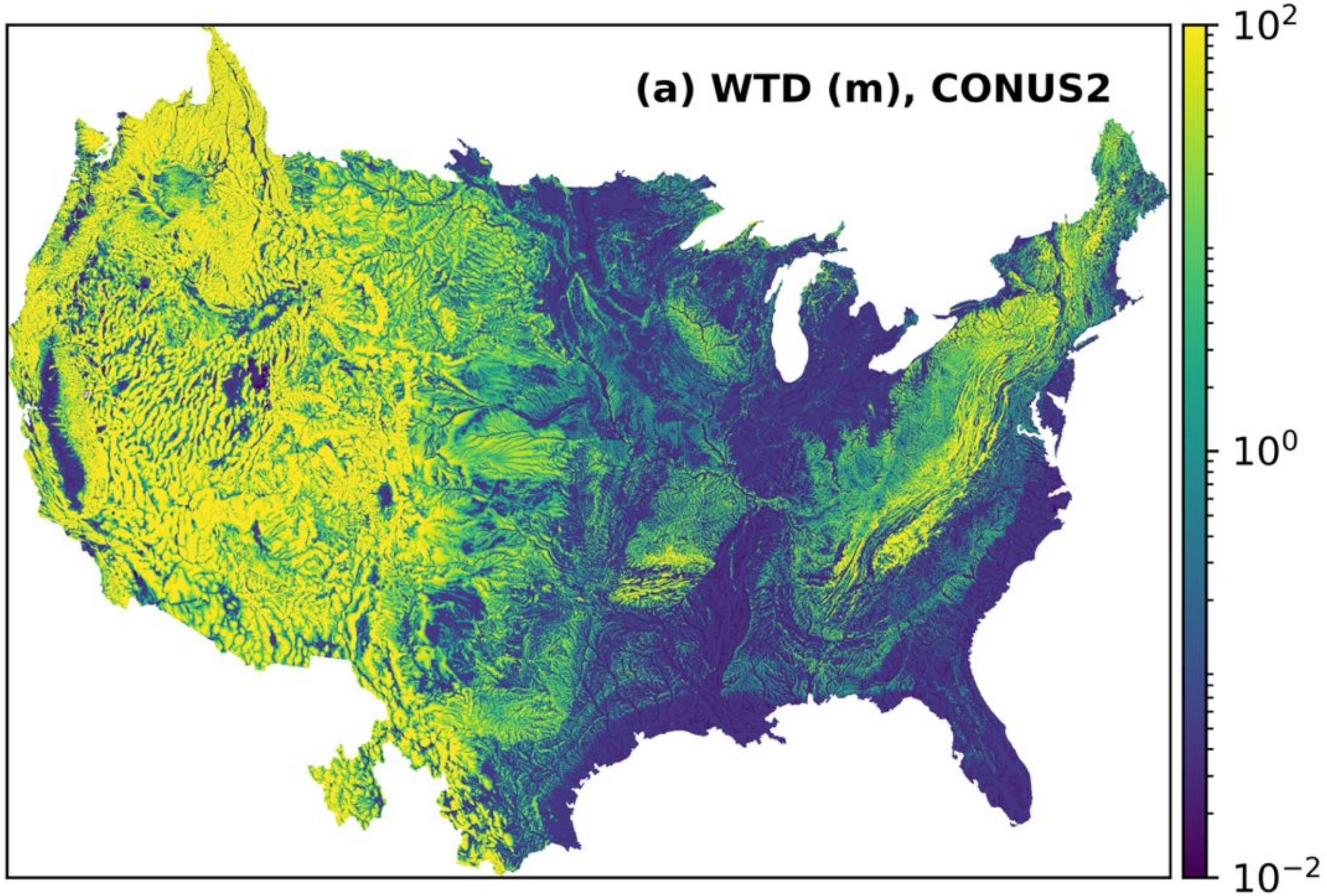
Danielle Tijerina
CEE Princeton





Tijerina-Kreuzer et al *Groundwater* (2023)

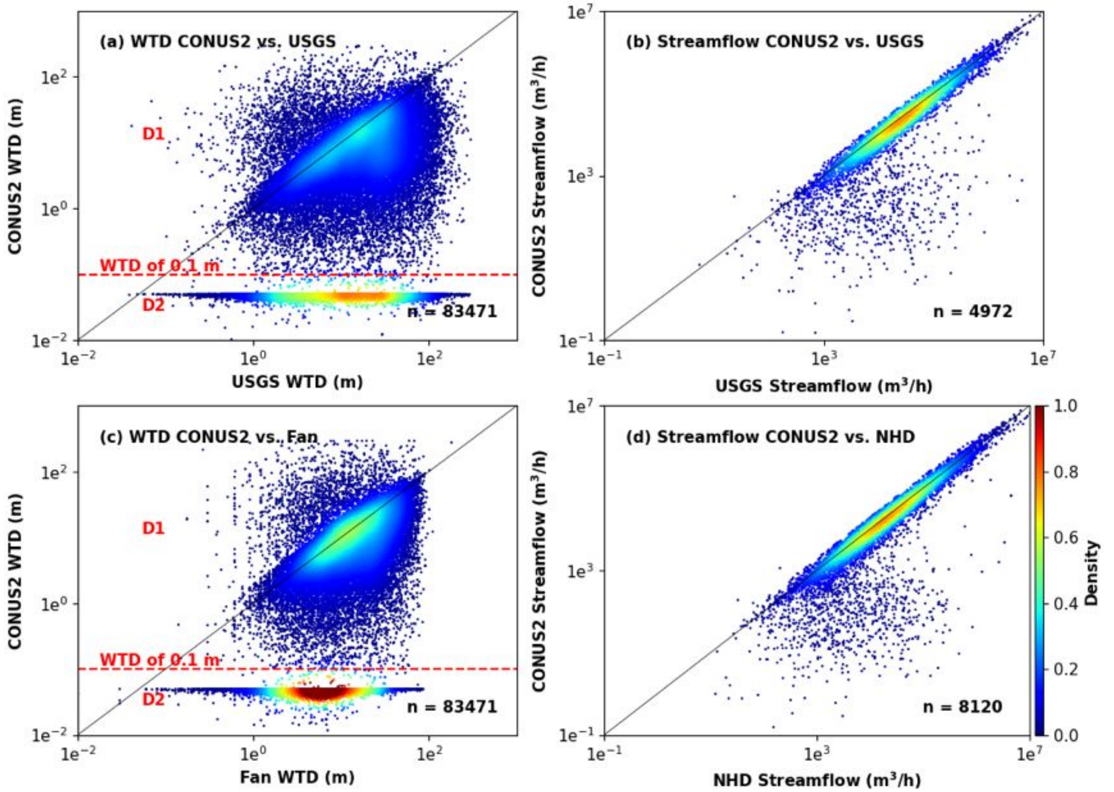
Large scale **models** (e.g. PFCONUS2) show **promising performance**



Yang, Tijerina-Kreuzer et al *J of Hydrology* (2023)

Chen Yang
Postdoc
CEE Princeton

Danielle Tijerina
PhD Candidate
CEE Princeton

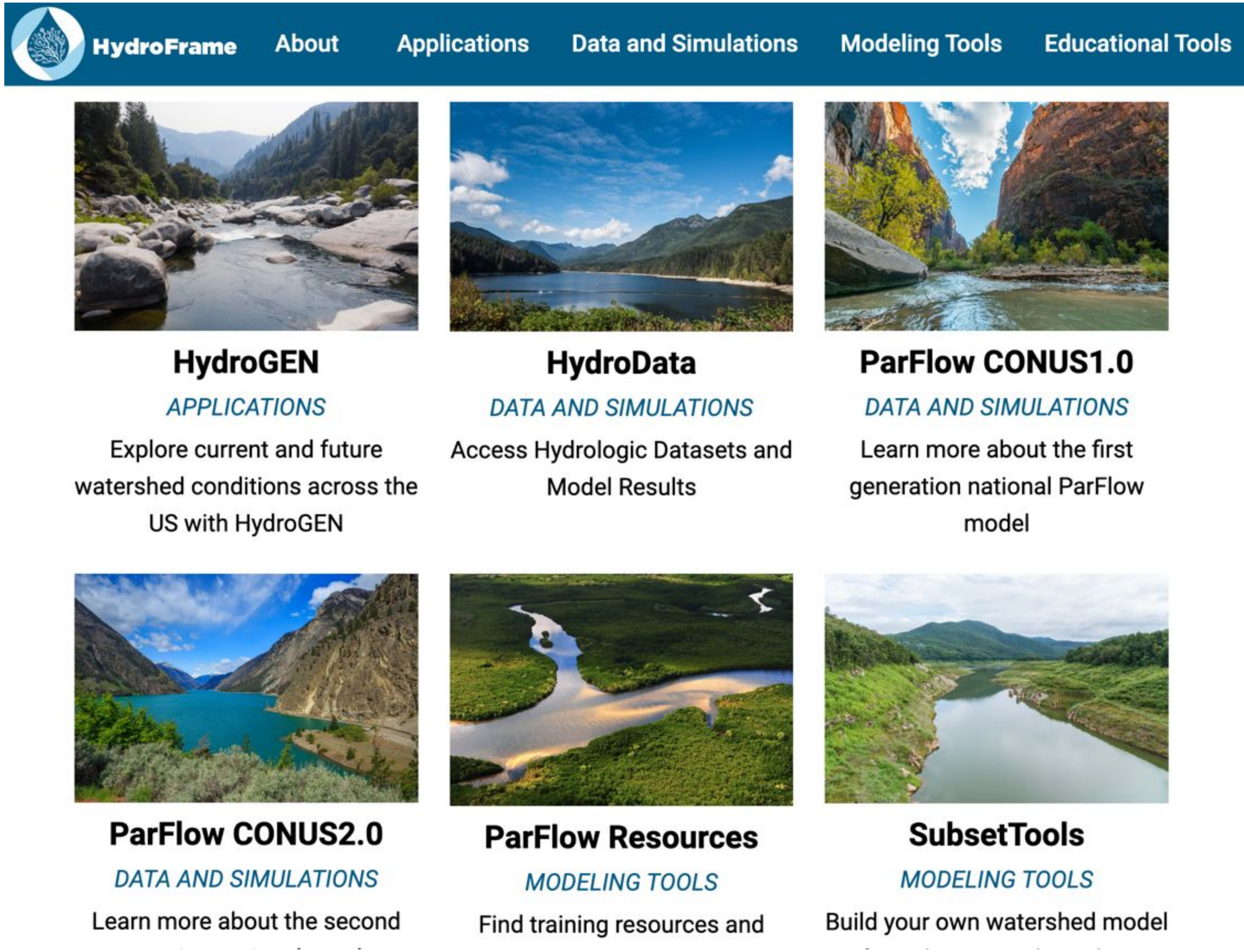


Agenda

1. Overview of the ParFlow CONUS2.0 model
2. **Overview of HydroFrame and HydroGEN toolset**
3. SubsetTools Package Overview and Activity
4. HydroData Package Overview and Activity

HydroFrame

- A framework for modeling and data tools related to the national modeling platform
- Our goal is to make national hydrologic simulations and datasets more accessible.



HydroData: Data Catalog

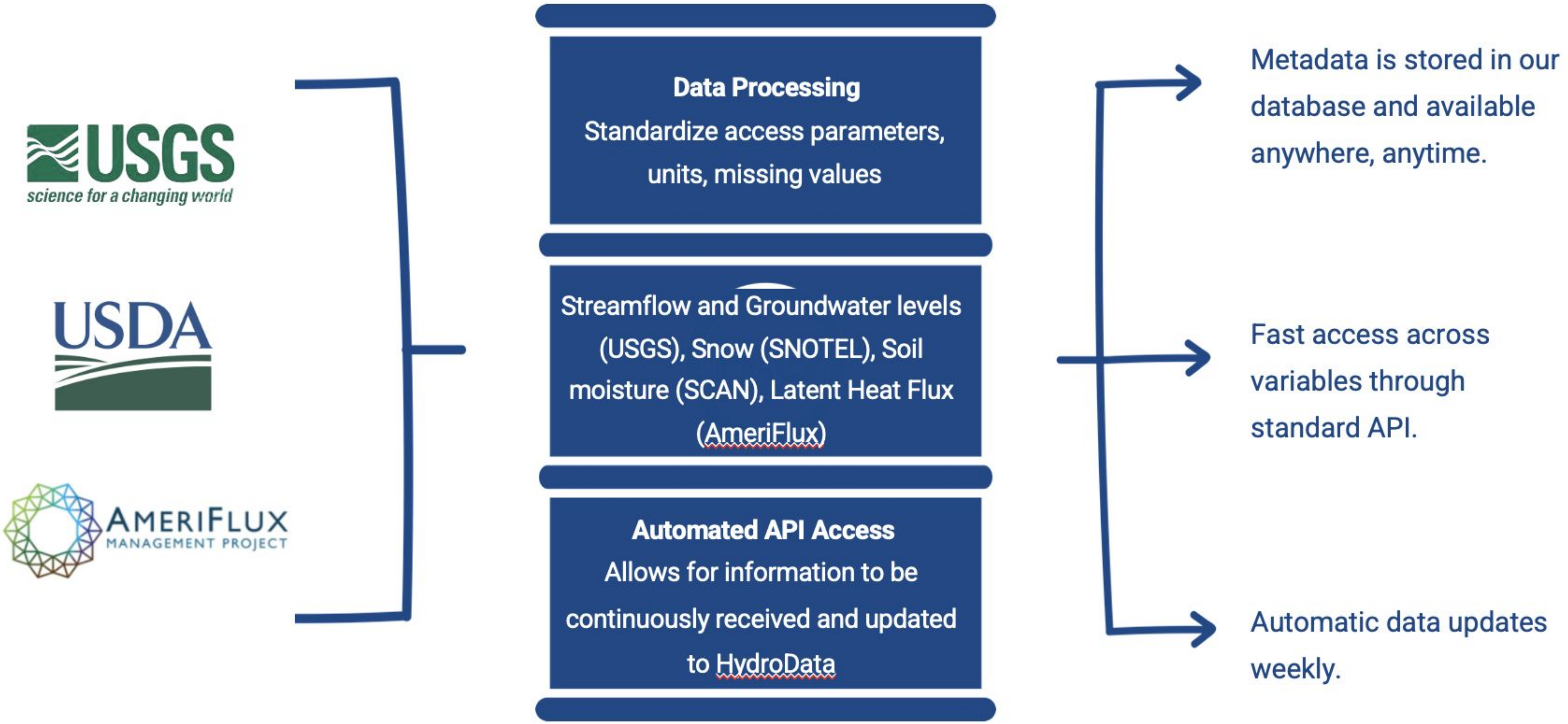
A database of continuously updated hydrologic simulations and observations that is housed at the Princeton HydroData center

1. **Point observations**: Point observations of hydrologic variables (e.g. streamflow and water table depth) that have been assembled and processed from a variety of external sources.
2. **ParFlow simulation outputs**: Gridded outputs from ParFlow simulations of the contiguous US.
3. **Meteorological forcings**: Gridded datasets that contain meteorological forcing variables (e.g. precipitation, temperature, and wind speed) needed to drive hydrologic models.
4. **Static model inputs and domain files**: Gridded datasets that contain static model inputs that are used directly in model definition (e.g. soil properties, land cover, and hydraulic conductivity) as well as additional domain files that are used in pre and post processing (e.g. distance to streams and watershed boundaries).
5. **Current conditions**: Current hydrologic conditions generated by the [HydroGEN project]
6. **Remote sensing**: Gridded observation products that are generated from remote sensing.

HydroData: Point Observations

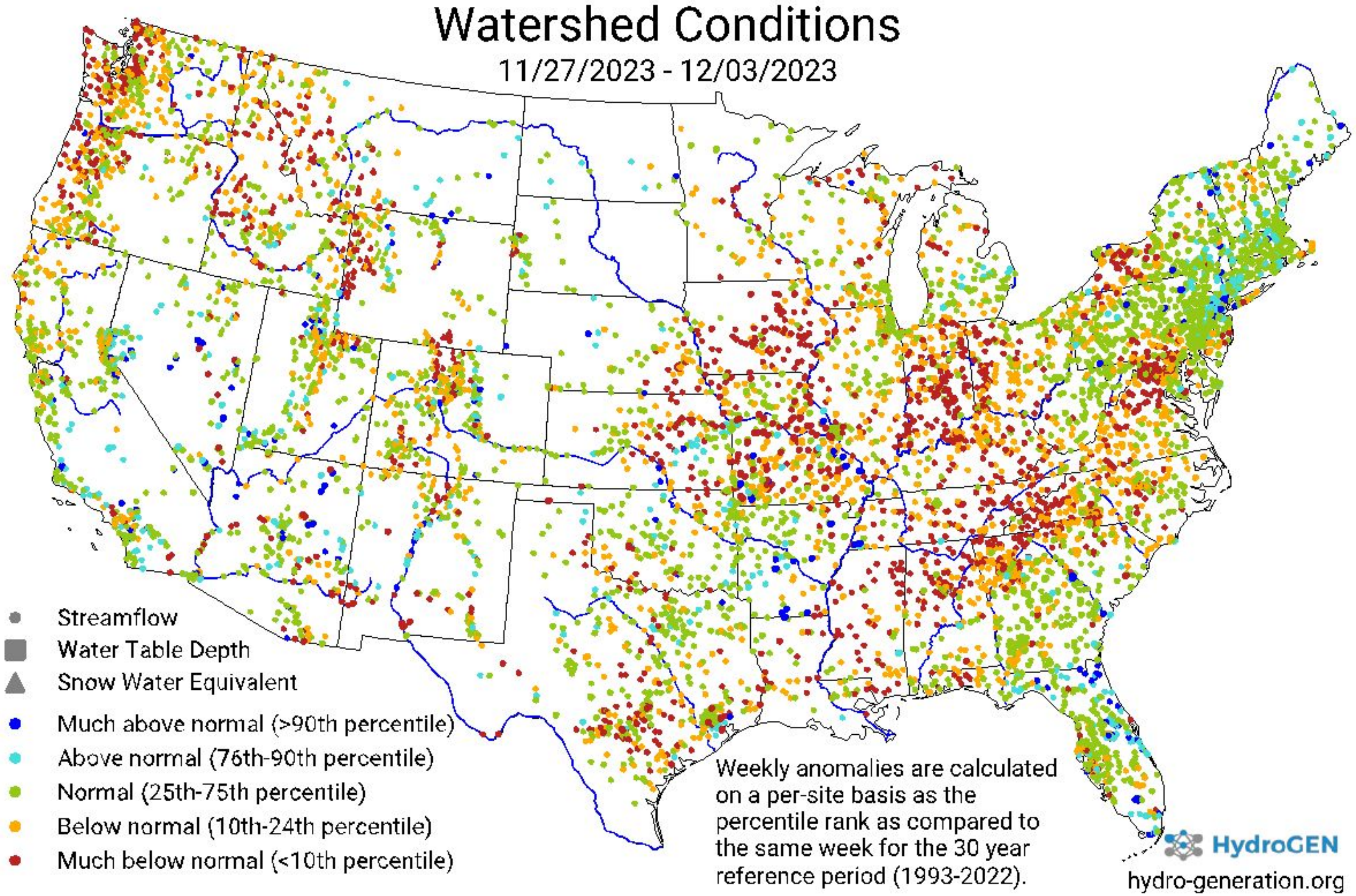
Point observations will be the focus of this Short Course

Continuously updated repository of national hydrologic observations



HydroData: Point Observations

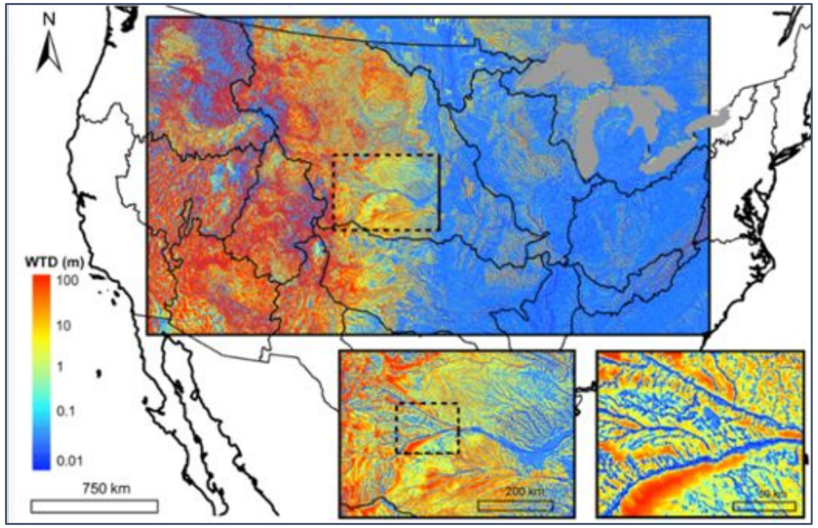
By combining data from multiple observation networks we can explore up-to-date watershed conditions across the US considering groundwater, surface water, and snowpack together.



HydroData: Model Datasets

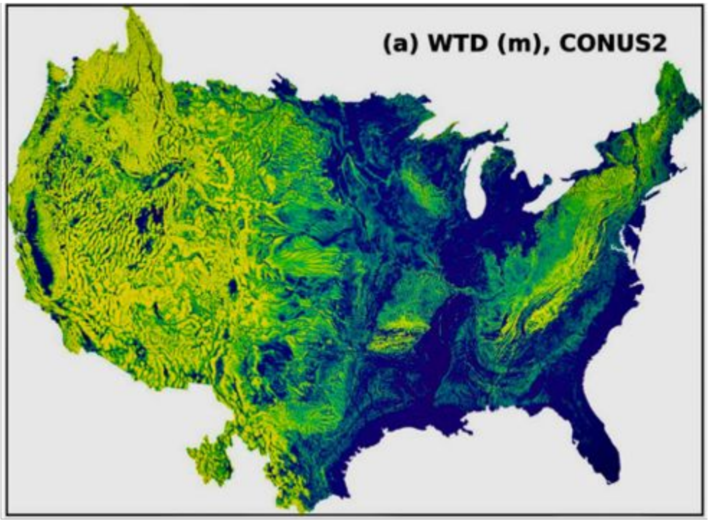
See our documentation for more information on these datasets

CONUS1 Datasets



- Complete model geofabric
- Steady state water table depth
- Multi-year transient simulation outputs
- NLDAS2 forcing used to run the CONUS1 simulations
- ParFlow and CLM driver files

CONUS2 Datasets



- Updated geofabric for expanded CONUS2 domain
- Steady state water table depth
- CW3E forcing used to run the CONUS2 transient simulations
- ParFlow and CLM driver files

HydroData: Tools

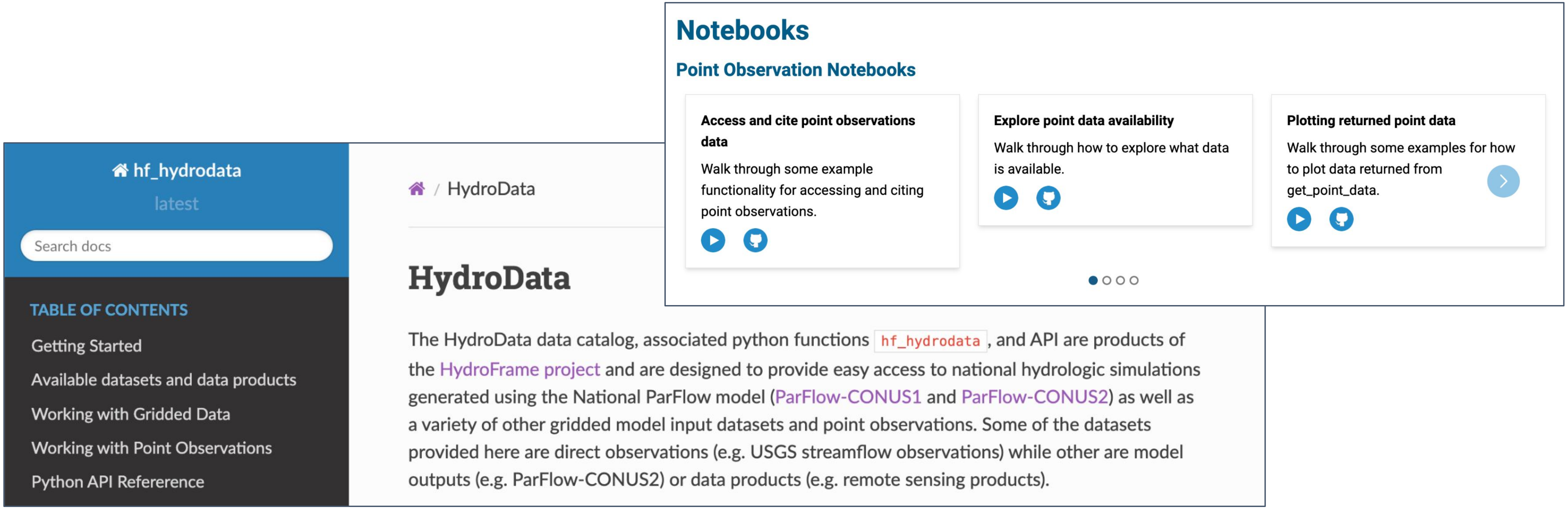
- hf_hydrodata is a Python package that provides a streamlined syntax for accessing data from **HydroData**
- The hf_hydrodata package includes functions to access and subset data, return dataset-level or site-level metadata, and relevant dataset citations

hf_hydrodata

The `hf_hydrodata` Python package is a product of the HydroFrame project and is designed to provide easy access to national hydrologic simulations generated using the National ParFlow model (ParFlow-CONUS1 and ParFlow-CONUS2) as well as a variety of other gridded model input datasets and point observations. Some of the datasets provided here are direct observations (e.g. USGS streamflow observations) while other are model outputs (e.g. ParFlow-CONUS2) or data products (e.g. remote sensing products).

HydroData: Tools

- Our documentation on [ReadtheDocs] includes information on the datasets that are available as well as full examples for how to access and work with the data



SubsetTools

- HydroData provides access to CONUS1 and CONUS2 model configurations and outputs including subsetting functionality.
- However, assembling a ParFlow model is very complex and can still pose a barrier to entry for many users
- **Many users want an easy way to accelerate the domain creation process including input data**

```
model.Geom.Perm.Names = "domain s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 g1 g2 g3

model.Geom.domain.Perm.Type = "Constant"
model.Geom.domain.Perm.Value = 0.02

model.Geom.s1.Perm.Type = "Constant"
model.Geom.s1.Perm.Value = 0.269022595

model.Geom.s2.Perm.Type = "Constant"
model.Geom.s2.Perm.Value = 0.043630356

model.Geom.s3.Perm.Type = "Constant"
model.Geom.s3.Perm.Value = 0.015841225

model.Geom.s4.Perm.Type = "Constant"
model.Geom.s4.Perm.Value = 0.007582087

model.Geom.s5.Perm.Type = "Constant"
model.Geom.s5.Perm.Value = 0.01818816

model.Geom.s6.Perm.Type = "Constant"
model.Geom.s6.Perm.Value = 0.005009435

model.Geom.s7.Perm.Type = "Constant"
model.Geom.s7.Perm.Value = 0.005492736

model.Geom.s8.Perm.Type = "Constant"
model.Geom.s8.Perm.Value = 0.004675077

model.Geom.s9.Perm.Type = "Constant"
model.Geom.s9.Perm.Value = 0.003386794

model.Geom.s10.Perm.Type = "Constant"
model.Geom.s10.Perm.Value = 0.004783973
```

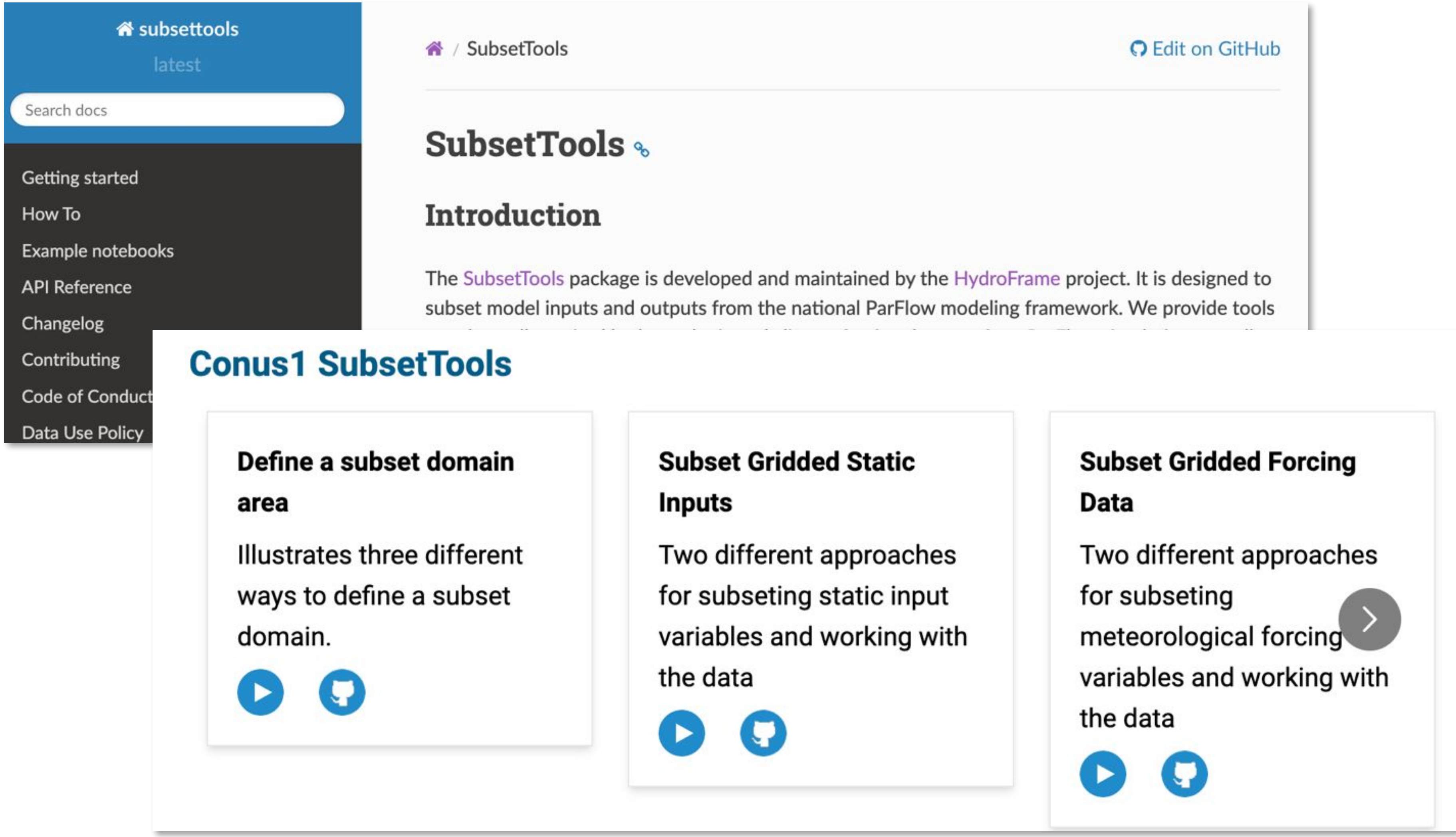
SubsetTools

Help users build watershed models from the national ParFlow modeling framework

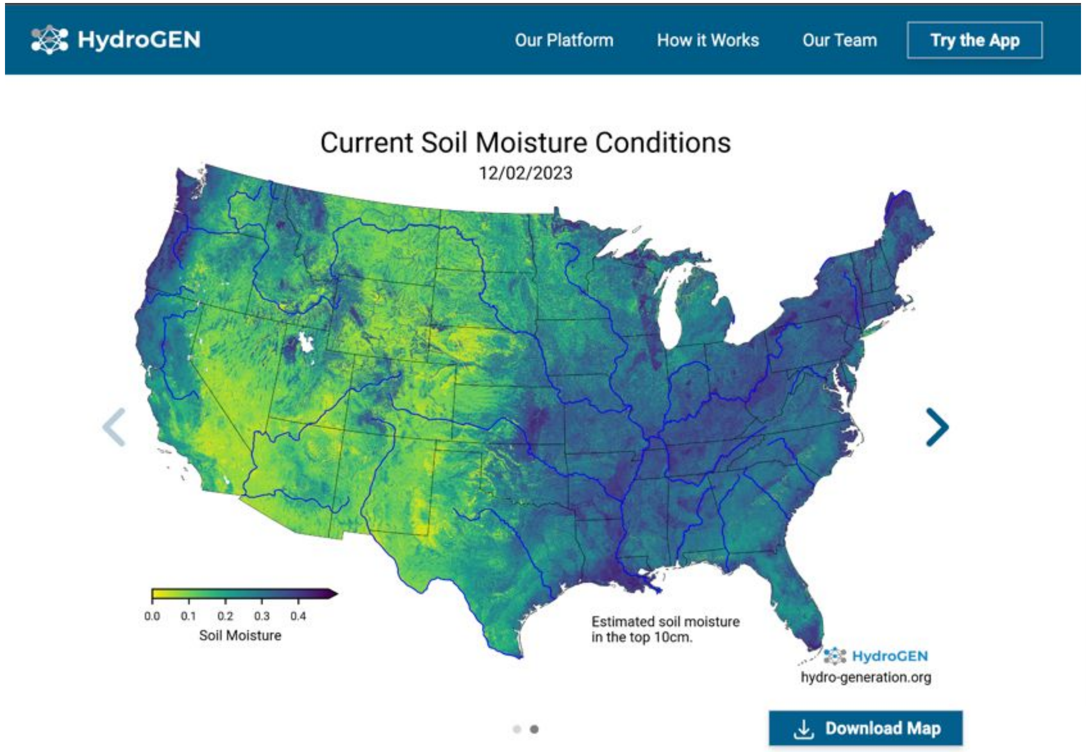
- Functions that simplify the process of setting up a modeling domain and running ParFlow
- Allows the user to subset all required hydrogeologic and climate forcing datasets from the HydroData filesystem
- Template runscripts for common use cases and examples for how to use them
- Helper functions to customize and run your subset model

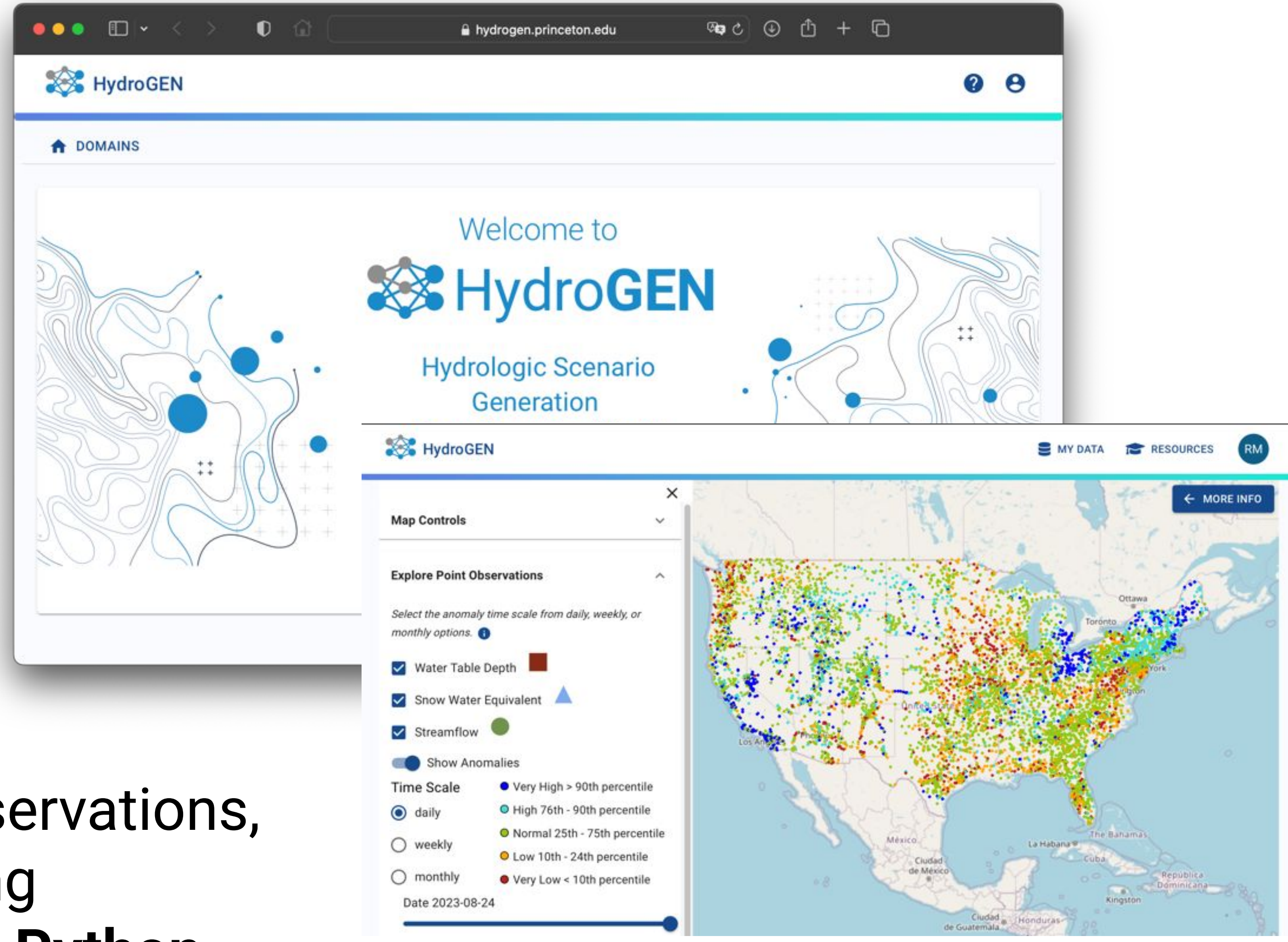
SubsetTools

Workbooks and tutorials provide end-to-end workflows to build and run steady state and transient ParFlow models



HydroGEN





A user friendly app to explore point observations, current conditions and machine learning emulators for ParFlow. **No modeling or Python experience required.**

Hydro-Generation.org

Agenda

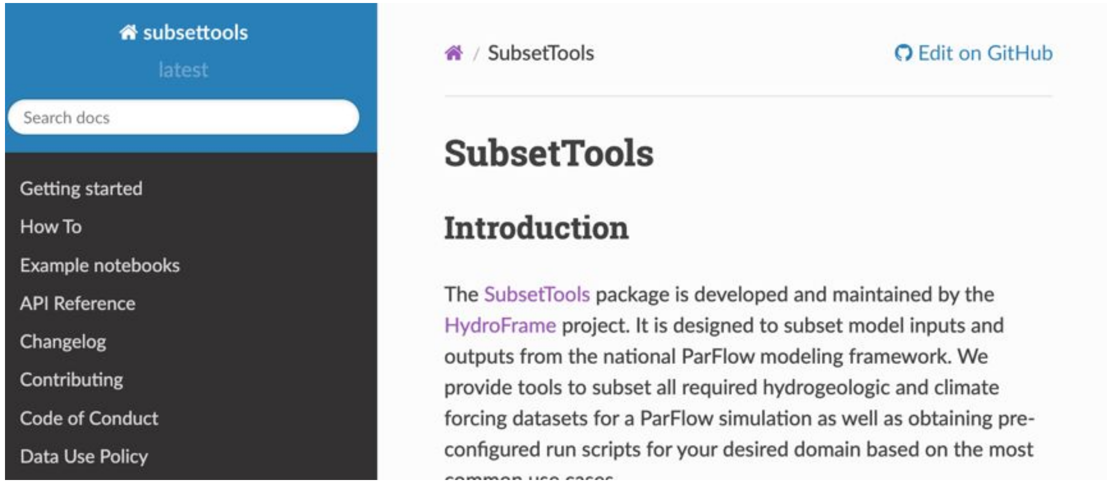
1. Overview of the ParFlow CONUS2.0 model
2. Overview of HydroFrame and HydroGEN toolset
3. **SubsetTools Package Overview and Activity**
4. HydroData Package Overview and Activity

SubsetTools: Usage and Installation

Getting access to the HydroData Catalog and SubsetTools is easy!

1. Sign up for a HydroFrame account
2. Get a PIN for API access
3. Install the subset tools package

```
$ pip install subsettools
```



Everything is documented on our Read the Docs page
(you can find the link to this on hydroframe.org)

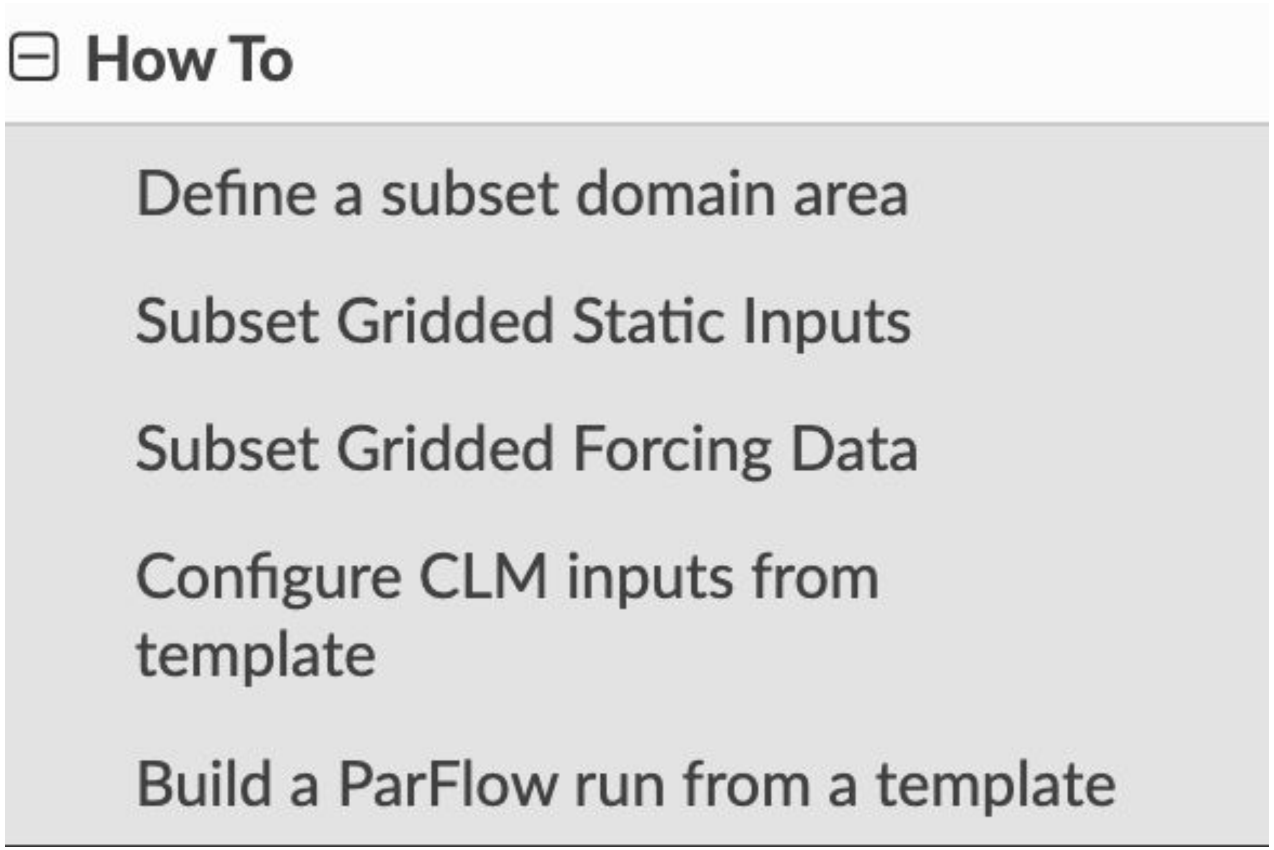
Ways to Interact with SubsetTools Resources

- Binder
  - Easiest: ParFlow and subsettools already installed, tutorials included and runnable
  - Limited resources (max 6 hours session, 1 CPU hour total)
  - Basic ParFlow version (no support for parallel runs, no optional packages)
- Docker:
  - Container with subsettools and ParFlow installed.
  - Linux version supported for both arm64 and amd64 architectures
  - Installing Docker is much easier than installing ParFlow
- Pip install in a virtual environment
  - Needs some knowledge about virtual environments
  - If a user wants to do a simulation as well, needs ParFlow installed and some environment variables configured

SubsetTools: Tutorials

We have detailed tutorial walking through examples of common workflow needs for

- Define a domain of interest using a bounding box or a watershed boundary
- Accessing model input files to configure a domain
- Working with atmospheric forcing data
- Accessing ParFlow and CLM template run scripts and configuring them for a custom domain/simulation



SubsetTools: Example Workflows

We currently have three end-to-end workflows for subsetting and running ParFlow simulations including steady state and transient examples for CONUS1 and CONUS2

(More examples will be coming soon)

1. **Performing a model initialization (spin up) with ParFlow on CONUS1**: This notebook will subset a HUC8 from the CONUS1 domain and will create a ParFlow run script to run a spinup simulation uses a constant recharge forcing to achieve a steady state water table depth (*NOTE*: This run does not include land surface simulation).

2. **Transient simulation with ParFlow-CLM on CONUS1**: This notebook walks through an example of subsetting a HUC8 from the CONUS1 domain. This example will subset everything needed to do a transient run with ParFlow-CLM and create a run script. This simulation is designed to start from a steady state groundwater condition. Here we subset the initial conditions from the pre-generated national steady state water table depth simulations, but the starting point could also be generated from the spin up workbook.

3. **Transient simulation with ParFlow-CLM on CONUS2**: This notebook walks through an example of subsetting a HUC8 from the CONUS2 domain. This example will subset everything needed to do a transient run with ParFlow-CLM and create a run script. This simulation is designed to start from a steady state groundwater condition. Here we subset the initial conditions from the pre-generated national steady state water table depth simulations, but the starting point could also be generated from the spin up workbook.

Hands-on activity 1

We will run a couple of examples to get familiar with common workflow steps:

- Defining a subset domain
- Getting forcing data from HydroData

- Launch Docker

- Docker paths:
  - Subsetting_ShortCourse/subsettools/definte_subset.ipynb
  - Subsetting_ShortCourse/subsettools/subset_forcing_data.ipynb

- First try running the notebooks as they are. Then, try modifying them to define your own subset domain and/or start and end dates for your forcing data.

Walk through of a subsetting workflow

/ Subset CONUS and run ParFlow-CLM

Edit on GitHub

Subset CONUS and run ParFlow-CLM

To launch this notebook interactively in a Jupyter notebook-like browser interface, please click the “Launch Binder” button below. Note that Binder may take several minutes to launch.

launch binder

This notebook walks through an example of subsetting a HUC8 from the CONUS2 domain. This example will subset everything needed to do a transient run with ParFlow-CLM. This is includes all hydrogeologic datasets as well as climate forcing data from CW3E. All of the data is written to a folder for the specified days to run. This example uses the template runscript conus2_pfclm_transient_solid.yaml and edits it to correspond with the domain subset. It also sets-up and performs the designed simulation. The result will be model output pressure and saturation pfbs according to the days specified.

This notebook has two principal sections:

1. Subset all static inputs and climate forcings from a CONUS run stored in Hydrodata
2. Load and alter a reference run to set up and perform your ParFlow-CLM subset.

Import the required libraries

```
import matplotlib.pyplot as plt
import numpy as np
import os
from parflow import Run
```

Hands-on activity 2

Now you will try running the subsetter workflow.

- Launch Docker

- Docker path:
  Subsetting_ShortCourse/subsettools/conus2_subsetting_transient.ipynb

- First you can run it as is then modify to select a different HUC and/or start and end dates.

Agenda

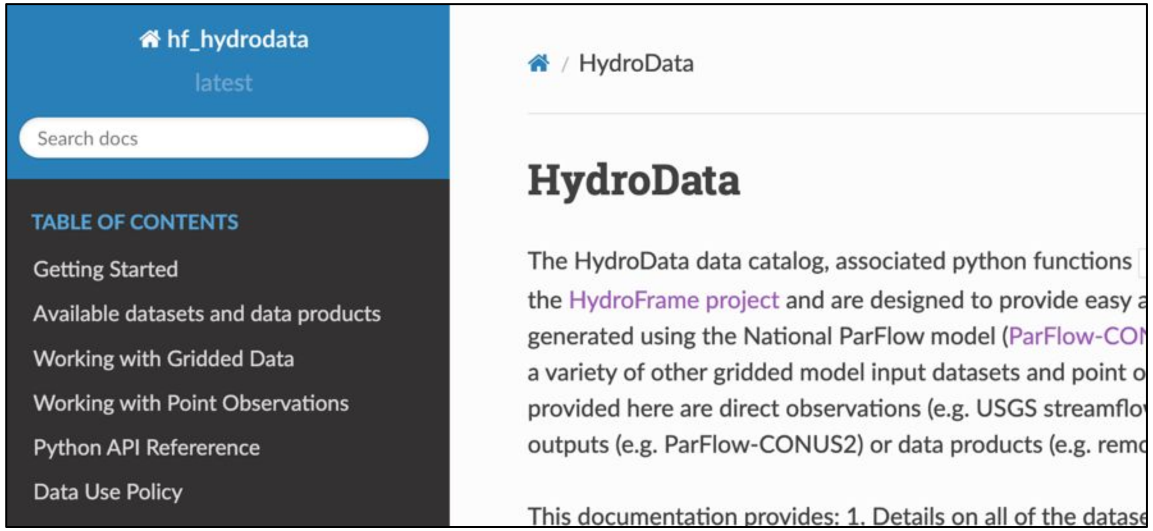
1. Overview of the ParFlow CONUS2.0 model
2. Overview of HydroFrame and HydroGEN toolset
3. SubsetTools Package Overview and Activity
4. **HydroData Package Overview and Activity**

hf_hydrodata: Usage and Installation

Similar to the Subsettools package:

1. Sign up for a HydroFrame account
2. Get a PIN for API access
3. Install the hf_hydrodata package

```
$ pip install hf_hydrodata
```



Everything is documented on our ReadtheDocs page

HydroData: Exploring the Catalog

All available datasets as well as variable definitions, grid coordinates, and metadata definitions are defined on the HydroData ReadtheDocs

TABLE OF CONTENTS

Available datasets and data products

The HydroData catalog contains a variety of datasets and data products to support national hydrologic modeling and analysis. We have the following categories of datasets:

1. **Point observations**: Point observations of hydrologic variables (e.g. streamflow and water table depth) that have been assembled and processed from a variety of external sources.
2. **ParFlow simulation outputs**: Gridded outputs from ParFlow simulations of the contiguous US.
3. **Meteorological forcings**: Gridded datasets that contain meteorological forcing variables (e.g. precipitation, temperature, and wind speed) needed to drive hydrologic models.
4. **Static model inputs and domain files**: Gridded datasets that contain static model inputs that are used directly in model definition (e.g. soil properties, land cover, and hydraulic conductivity) as well as additional domain files that are used in pre and post processing (e.g. distance to streams and watershed boundaries).
5. **Current conditions**: Current hydrologic conditions generated by the HydroGEN project
6. **Remote sensing**: Gridded observation products that are generated from remote sensing.

HydroData: Exploring the Catalog

Every dataset has its own documentation page providing a detailed description of how it was generated and all available variables

Modern CONUS1 simulations WY 2003-2006

A simulation of water years 2003-2006 completed with the ParFlow CONUS1 model. This is the most recent transient simulation that was completed with the ParFlow CONUS1 model (i.e. after the conus_baseline_85 simulation). The simulation was completed using the NLDAS2 forcing dataset at 1km resolution and a 1 hour time step. This dataset contains all of the outputs from the simulation as well as the ParFlow and CLM run scripts. The model inputs can be found in the conus1_domain dataset. Refer to the citations below for additional information on the model development and findings.

Dataset Name: conus1_baseline_mod

Data Source: hydroframe

Citations:

Please refer to the following citations for more information on this dataset and cite them if you use the data

- https://doi.org/10.5194/gmd-14-7223-2021

Extent and Resolution:

- Available Date Range: 10/1/02 to 9/30/06
- Grid: conus1
  - Spacial Resolution: 1000 meters

variable	description	temporal_resolut
evapotranspiration	Evapotranspiration	daily, hourly
latent_heat	Latent heat flux from canopy height to atmosphere	hourly
outward_longwave_radiation	Outgoing long-wave radiation from ground+canopy	hourly
sensible_heat	Sensible heat from canopy height to atmosphere	hourly
ground_heat	Ground heat flux	hourly
ground_evap	Ground surface evaporation rate	hourly
ground_evap_heat	Evaporation heat flux from ground	hourly
transpiration_leaves	Evaporation+transpiration from leaves	hourly
transpiration	Transpiration rate	hourly

Run File Variables

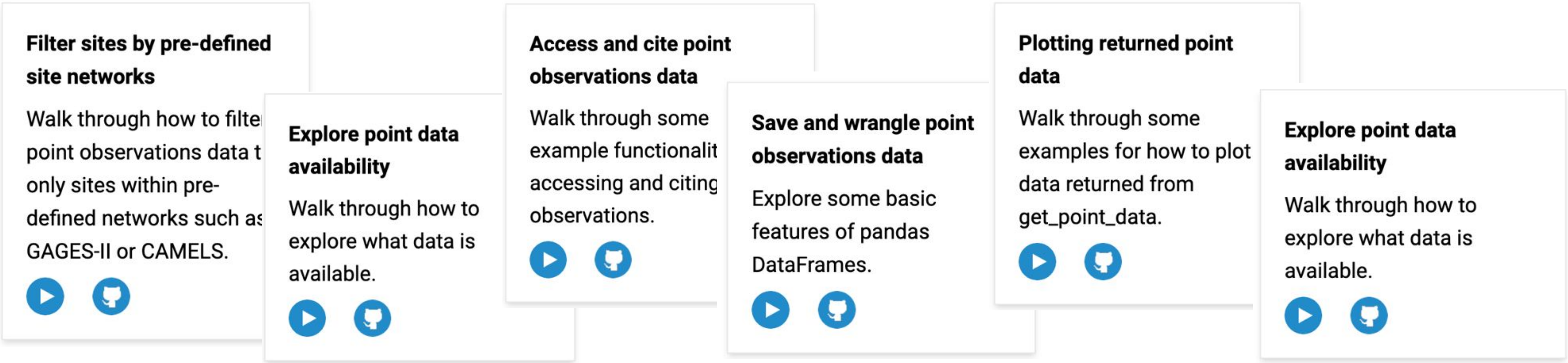
variable	description	temporal_resolution	units
parflow_run	Run scripts and other model configuration files	static	unitless
clm_run	No description	static	unitless

Subsurface Variables

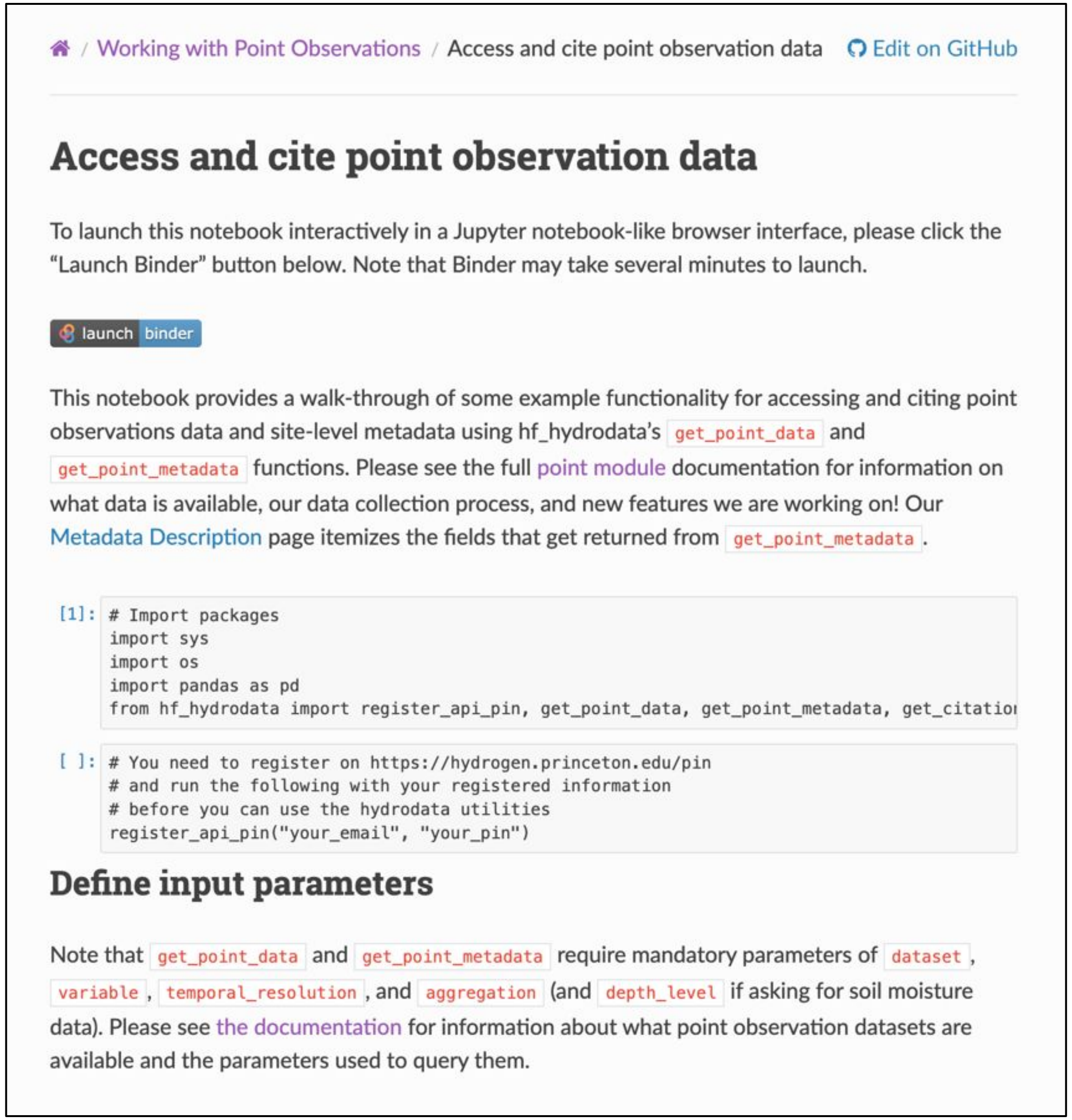
variable	description	temporal_resolution	units
soil_moisture	Soil moisture	daily	kg/m2
water_table_depth	Water table depth	daily	m
pressure_head	Map 3d grid point to pressure head	daily, hourly	m
pressure_next	Pressure head at the following timestep	daily	m
saturation	Map 3d grid point to saturation value	hourly	unitless

HydroData: Point Data Tools

We have a series of example notebooks illustrating how to subset, analyze and cite point observations across the HydroData Catalog



Walk through of an hf_hydrodata workflow



🏠 / Working with Point Observations / Access and cite point observation data   Edit on GitHub

Access and cite point observation data

To launch this notebook interactively in a Jupyter notebook-like browser interface, please click the "Launch Binder" button below. Note that Binder may take several minutes to launch.

launch binder

This notebook provides a walk-through of some example functionality for accessing and citing point observations data and site-level metadata using hf_hydrodata's `get_point_data` and `get_point_metadata` functions. Please see the full point module documentation for information on what data is available, our data collection process, and new features we are working on! Our Metadata Description page itemizes the fields that get returned from `get_point_metadata`.

[1]:
```
# Import packages
import sys
import os
import pandas as pd
from hf_hydrodata import register_api_pin, get_point_data, get_point_metadata, get_citatio
```

[ ]:
```
# You need to register on https://hydrogen.princeton.edu/pin
# and run the following with your registered information
# before you can use the hydrodata utilities
register_api_pin("your_email", "your_pin")
```

Define input parameters

Note that `get_point_data` and `get_point_metadata` require mandatory parameters of `dataset`, `variable`, `temporal_resolution`, and `aggregation` (and `depth_level` if asking for soil moisture data). Please see the documentation for information about what point observation datasets are available and the parameters used to query them.

Hands-on activity 1

Now you will try running a point observations workflow.

- Launch Docker

- Docker path:
  Subsetting_ShortCourse/hf_hydrodata/point/**example_get_data.ipynb**

- First you can run it as-is.
  - Then modify Example 1 to query "water_table_depth" from the USGS instead of streamflow. Adjust the start and end dates of your query.

Hands-on activity 2

<u>Option 1</u>

Want to explore how to visualize outputs from the point observations module?

- Docker path:
  Subsetting_ShortCourse/hf_hydrodata/point/**example_plot_data.ipynb**

<u>Option 2</u>

Want to get more familiar with how to work with pandas DataFrames?

- Docker path:
  Subsetting_ShortCourse/hf_hydrodata/point/**example_pandas.ipynb**


THANK YOU