

NextGen Research DataStream: Community Contributions Towards Improved Hydrologic Predictions

Lynker: Jordan J. Laser, Zach Wills, Nels Frazier
Alabama Water Institute: James Halgren, Arpita Patel



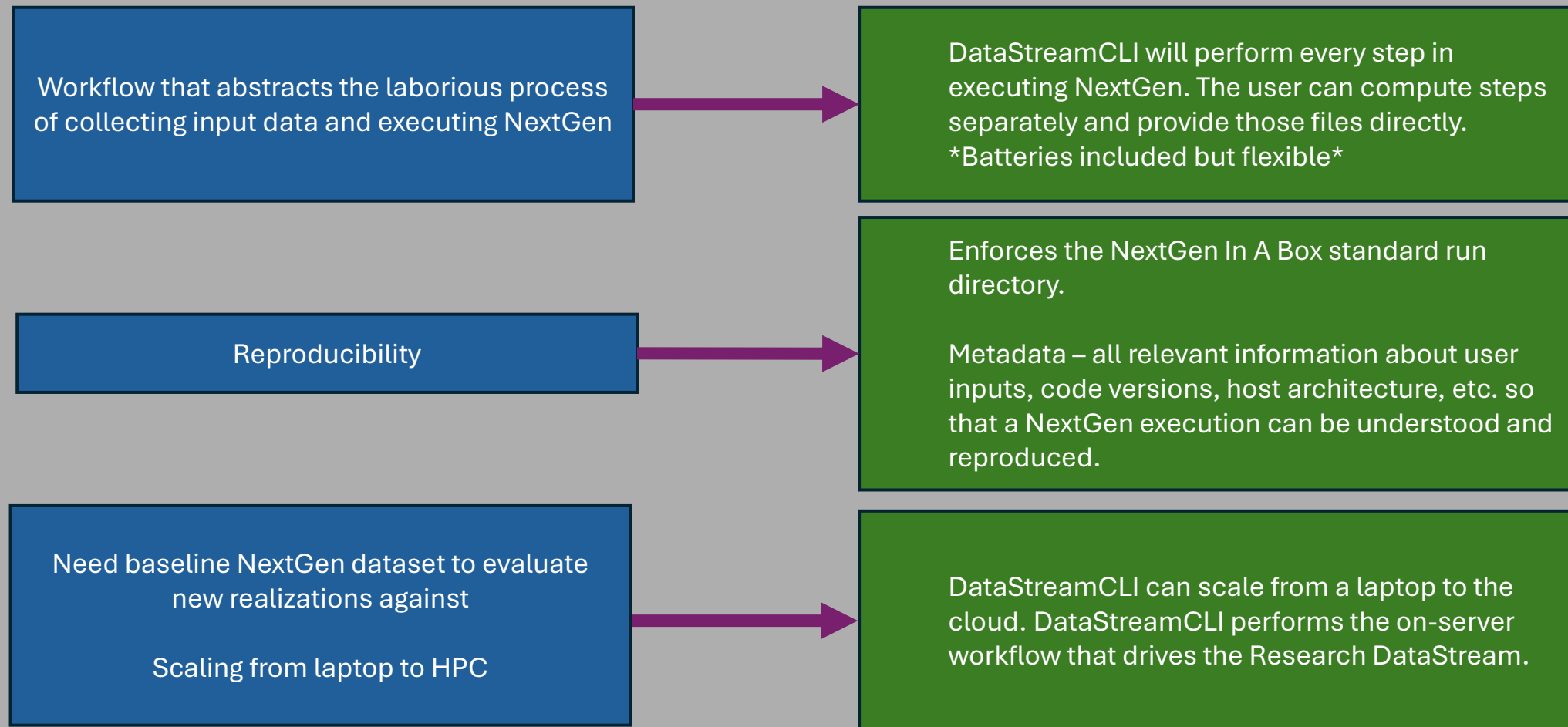
Road Map

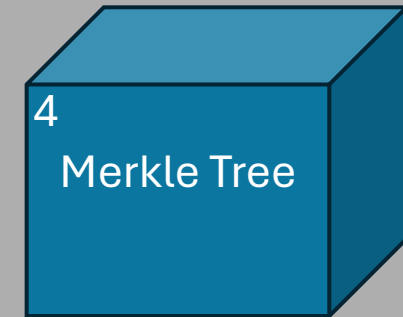
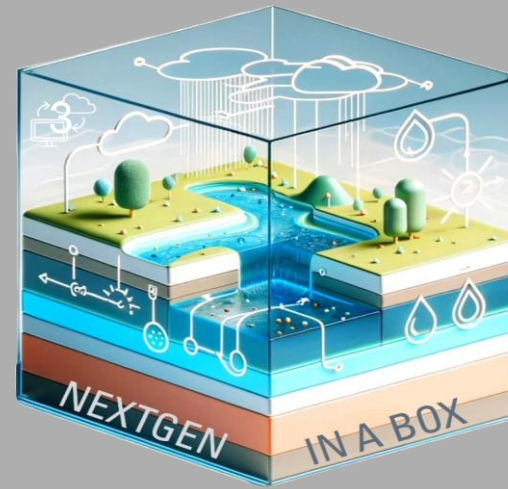
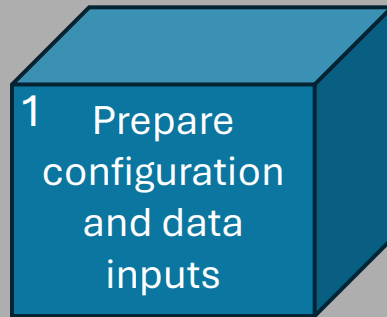
1. DataStreamCLI Motivation
2. DataStreamCLI Design
3. Research DataStream Motivation
4. Research DataStream Design
5. Research DataStream State
6. Hands-on workshop
 1. https://github.com/CIROH-UA/ngen-datastream/blob/main/docs/CIROH_devcon_2025/workshop.md

Please ask questions!
They are anonymous.



DataStreamCLI Motivation





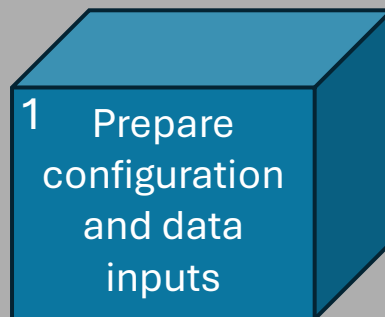
DataStreamCLI refers to the software chain that builds and validates NextGen input packages (ngen-run/), executes NextGen through NextGen In A Box (NGIAB), and versions the entire run for reproducibility.

This enforces a standard folder (ngen-run/), which enables interoperability and reproducibility.

ngen-run/

#	name	type	size
0	config	dir	288 B
1	forcings	dir	343.8 KiB
2	lakeout	dir	64 B
3	outputs	dir	592.3 KiB
4	restart	dir	64 B





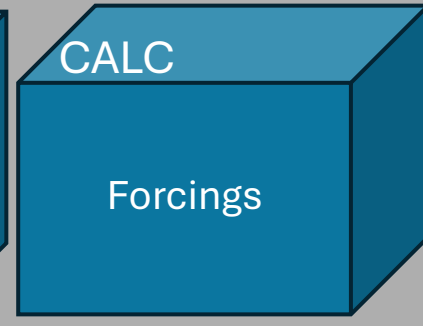
Required steps to build ngen-run/config and ngen-run/forcings



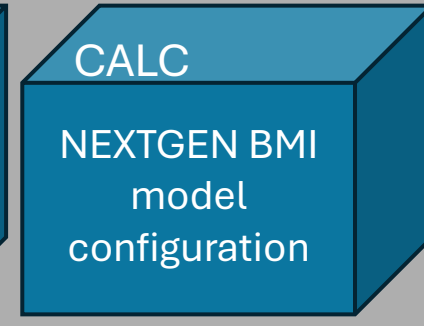
Defines spatial domain



Indices and coverage used to extract catchment averaged forcings. Calculated by exactextract.



Performs conversion between National Water Model and NextGen forcings formats



Required files for NextGen BMI modules

ngen-run/			
#	name	type	size
0	config	dir	288 B
1	forcings	dir	343.8 KiB
2	lakeout	dir	64 B
3	outputs	dir	592.3 KiB
4	restart	dir	64 B

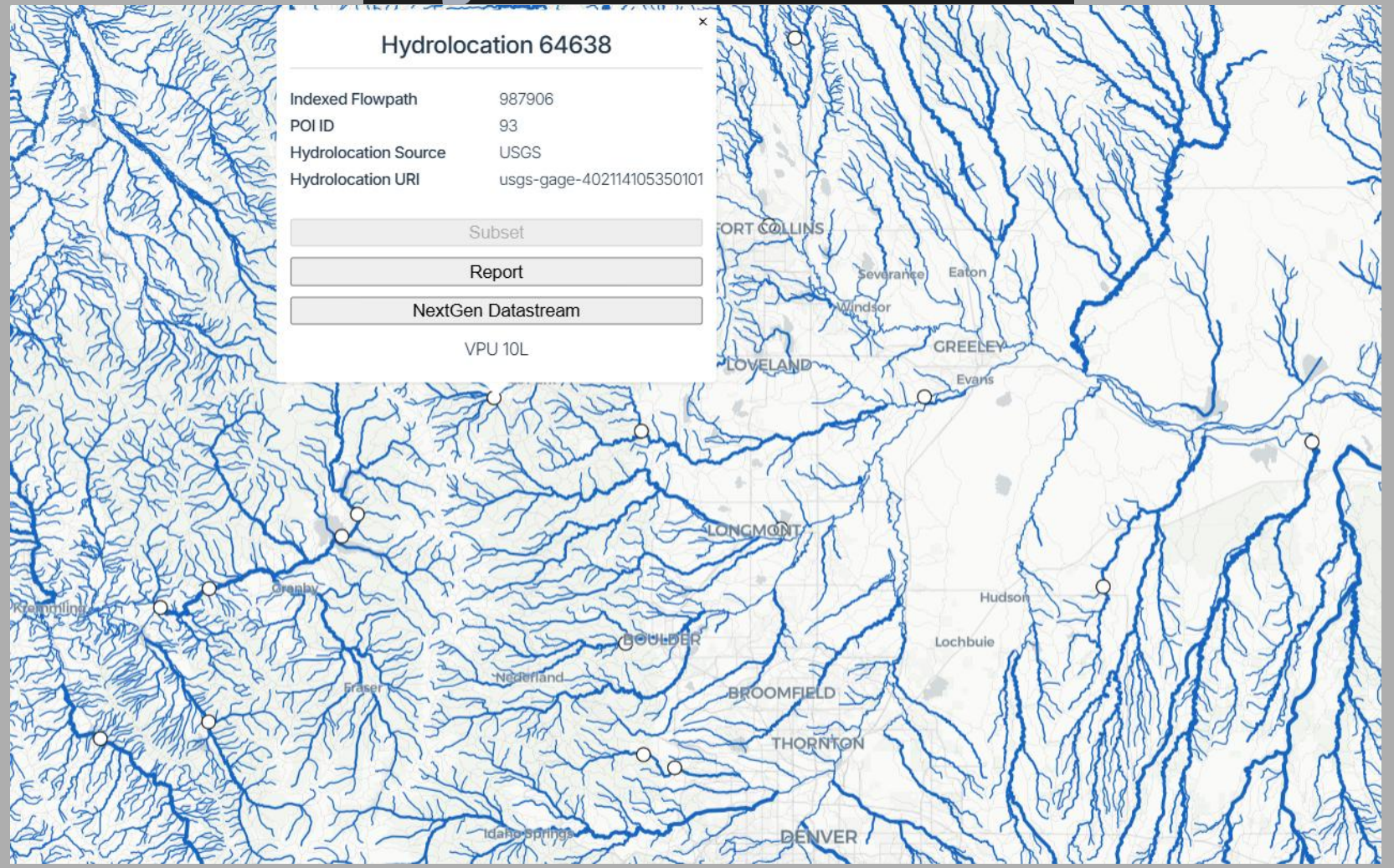


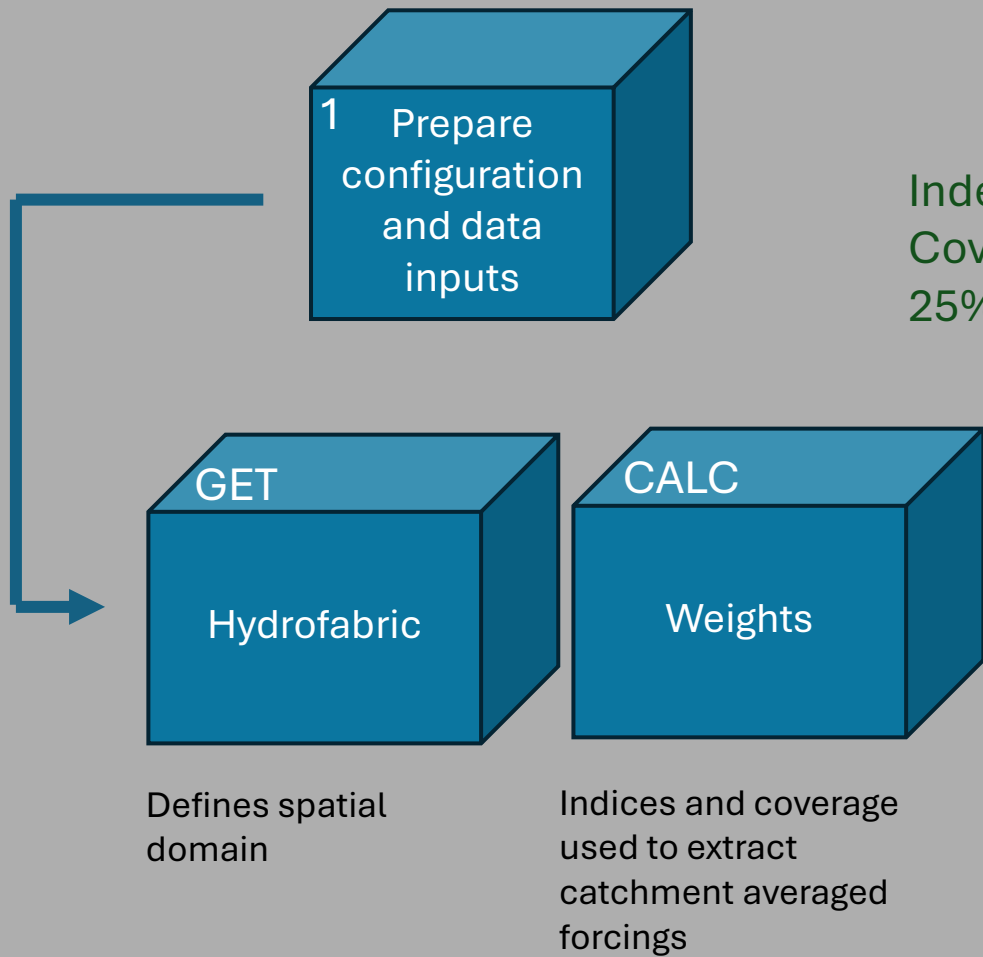
1 Prepare configuration and data inputs



GET
Lynker Spatial Hydrofabric

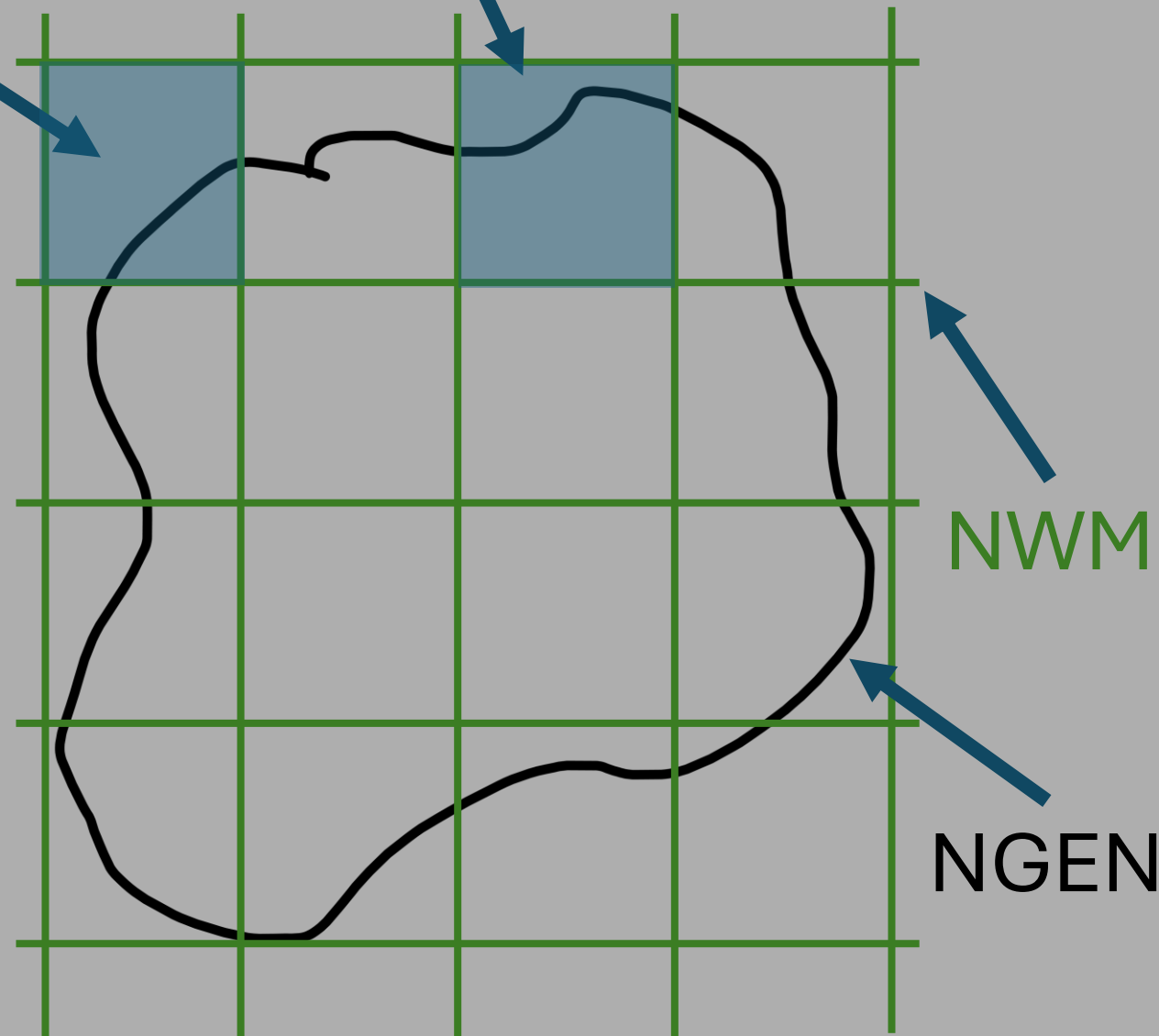
Defines spatial domain

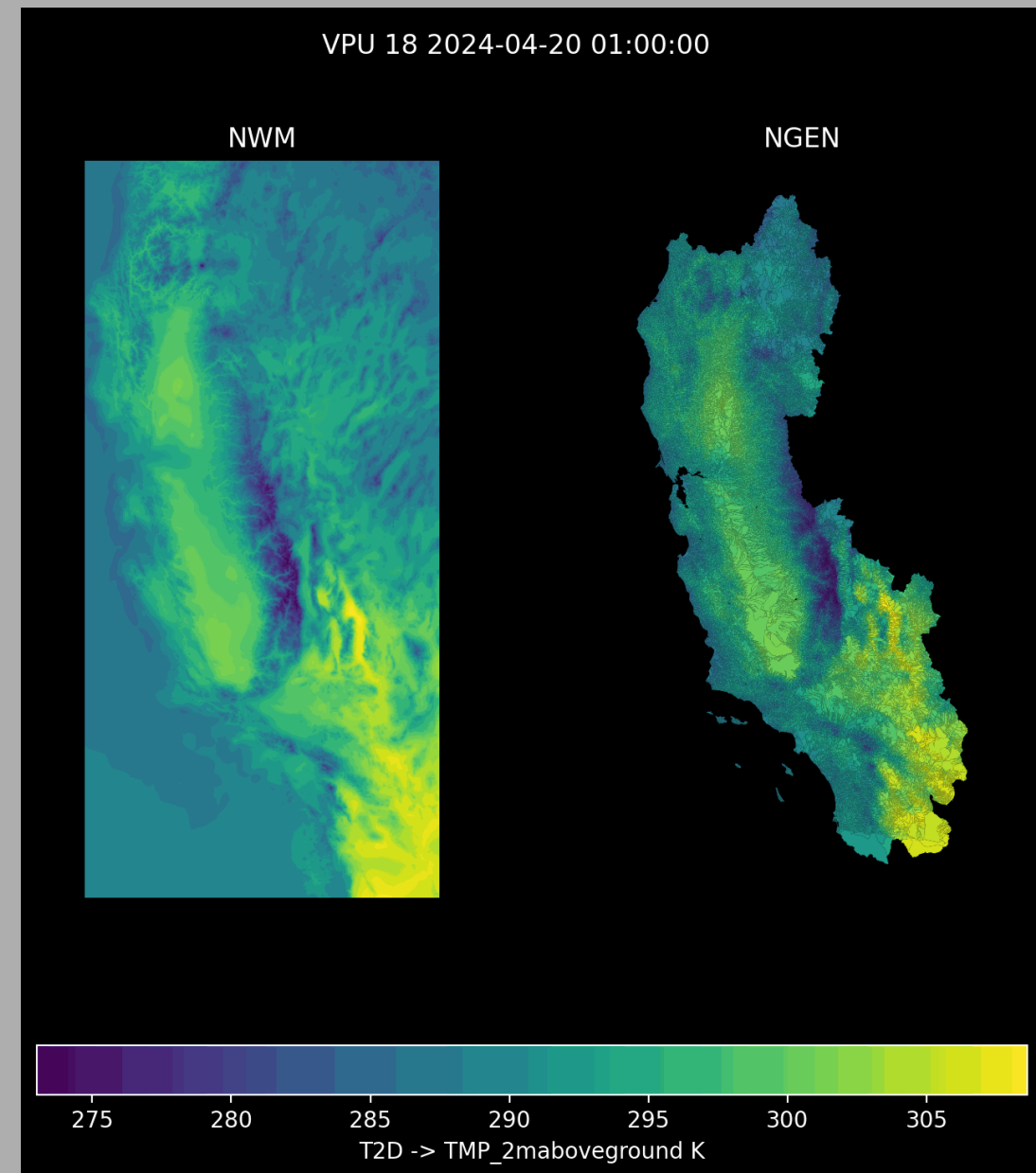
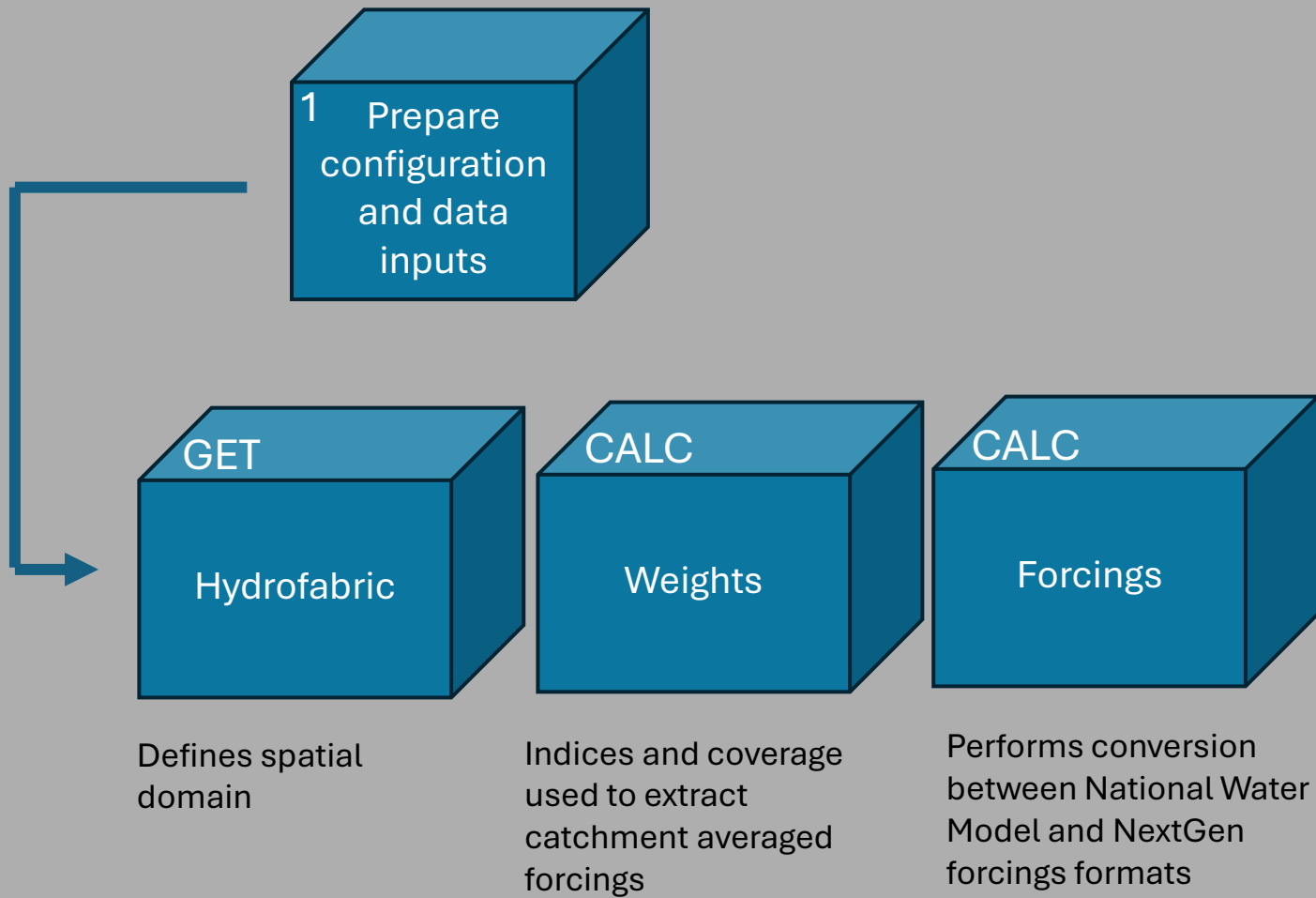


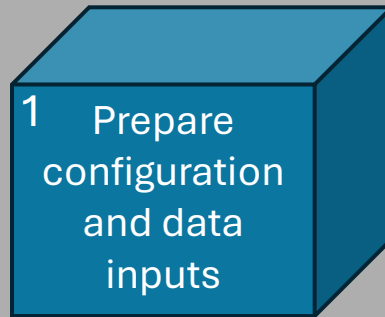


Index: 27461
Coverage:
25%

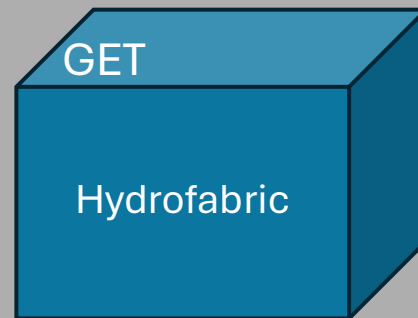
Index: 27463
Coverage:
75%



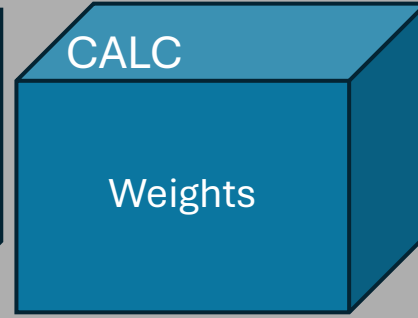




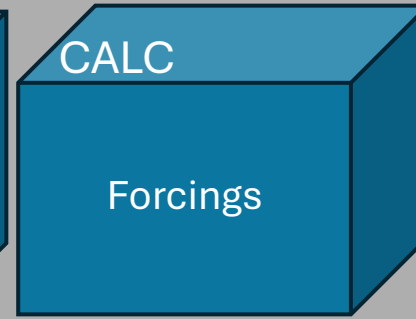
Automatic BMI module detection from realization file



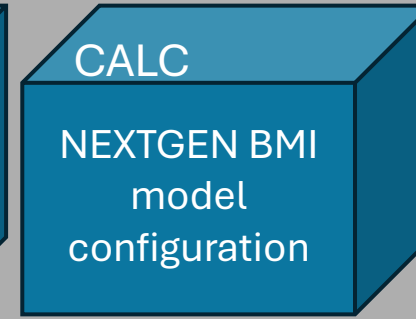
Defines spatial domain



Indices and coverage used to extract catchment averaged forcings

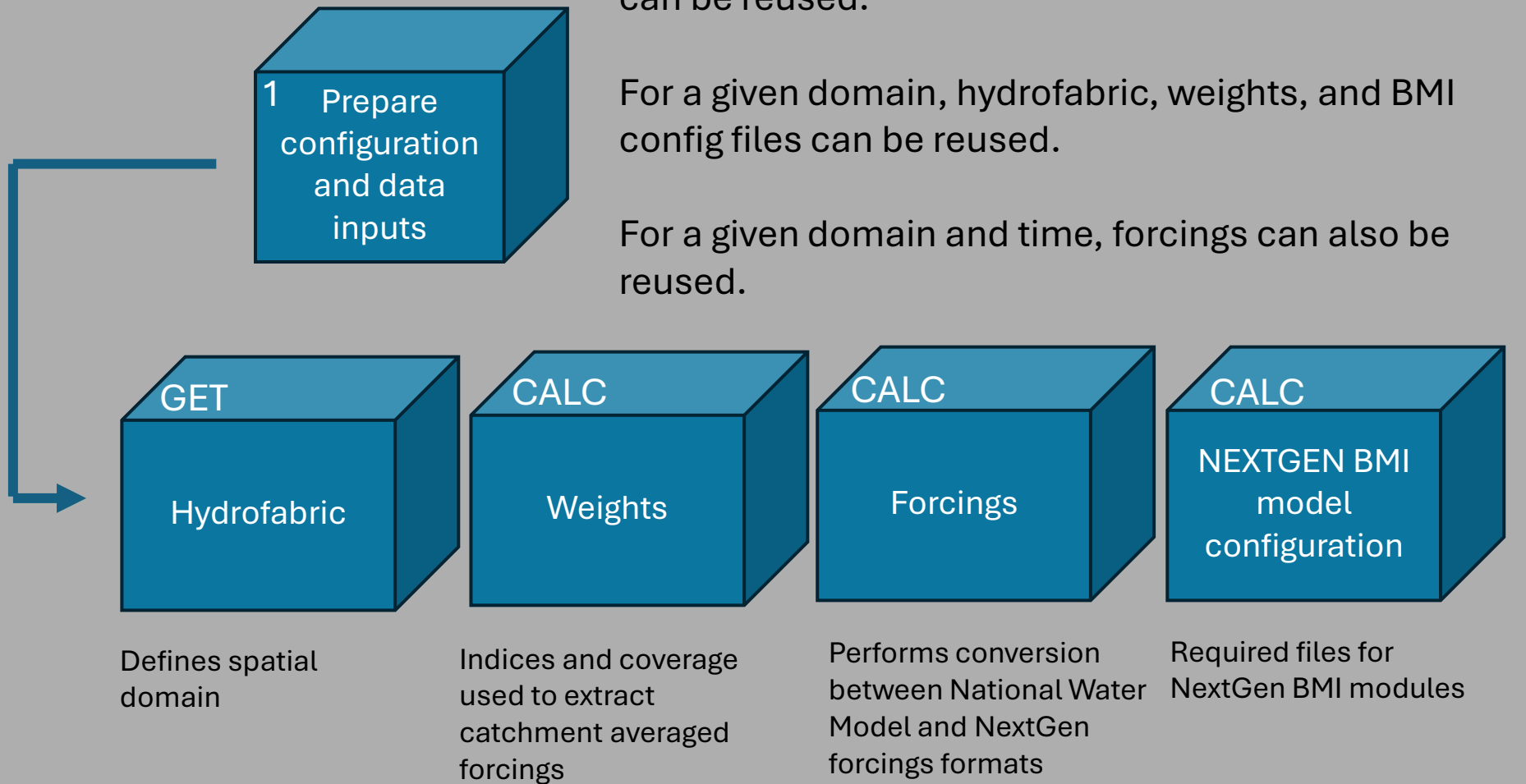


Performs conversion between National Water Model and NextGen forcings formats

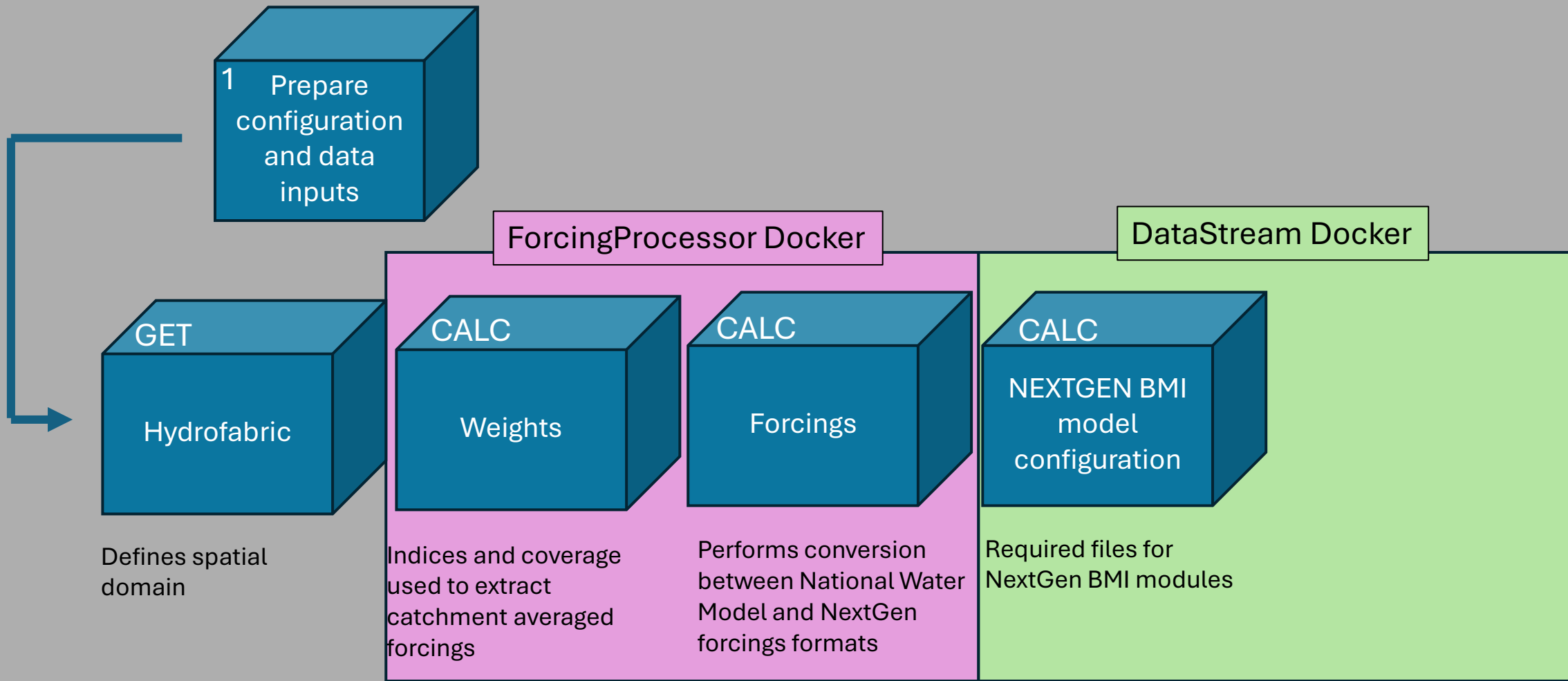


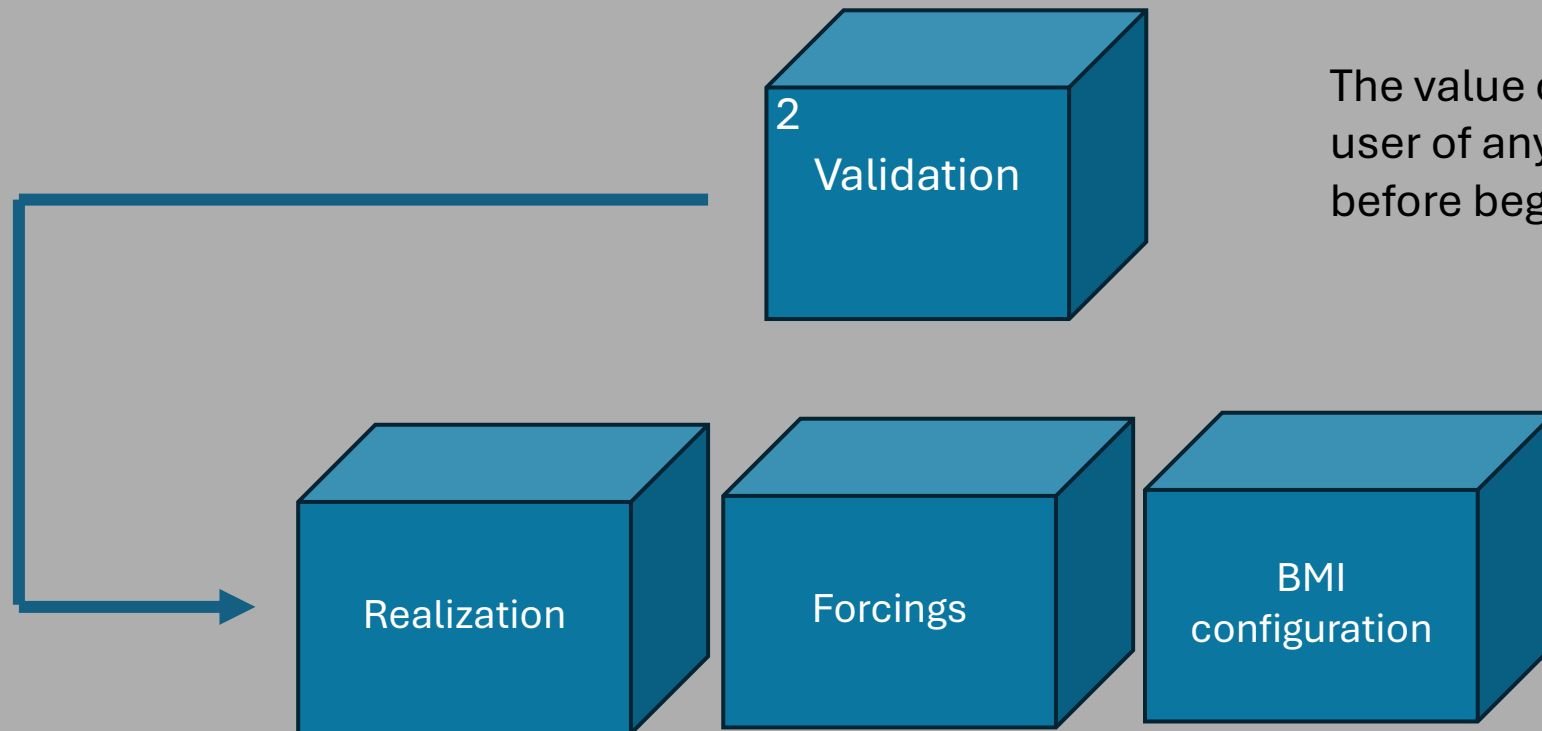
Required files for NextGen BMI modules

- Supported BMI module config generation
 - PET, CFE, Noah-OWP-Modular, t-route
- Coming soon
 - SoilFreezeThaw, TopModel, LSTM, others



```
RESOURCE_DIR/  
├── config/  
│   ├── ngen-bmi-configs.tar.gz  
│   └── realization.json  
├── datastream  
│   ├── partitions.json  
│   └── weights.json  
├── hydrofabric  
│   ├── nextgen_01.gpkg  
│   ├── nextgen_01.parquet  
│   └── weights.parquet  
├── nwm-forcings/  
│   ├── nwm.t00z.medium_range.forcing.f001.conus  
│   └── ...  
└── ngen-forcings/  
    └── forcings.tar.gz
```



The value of the validation step is to notify the user of any errors in the NEXTGEN run package before beginning the execution

Coming soon -> BMI module variable mapping validation

Required for

Ensures the user has supplied a valid realization file to configure NEXTGEN

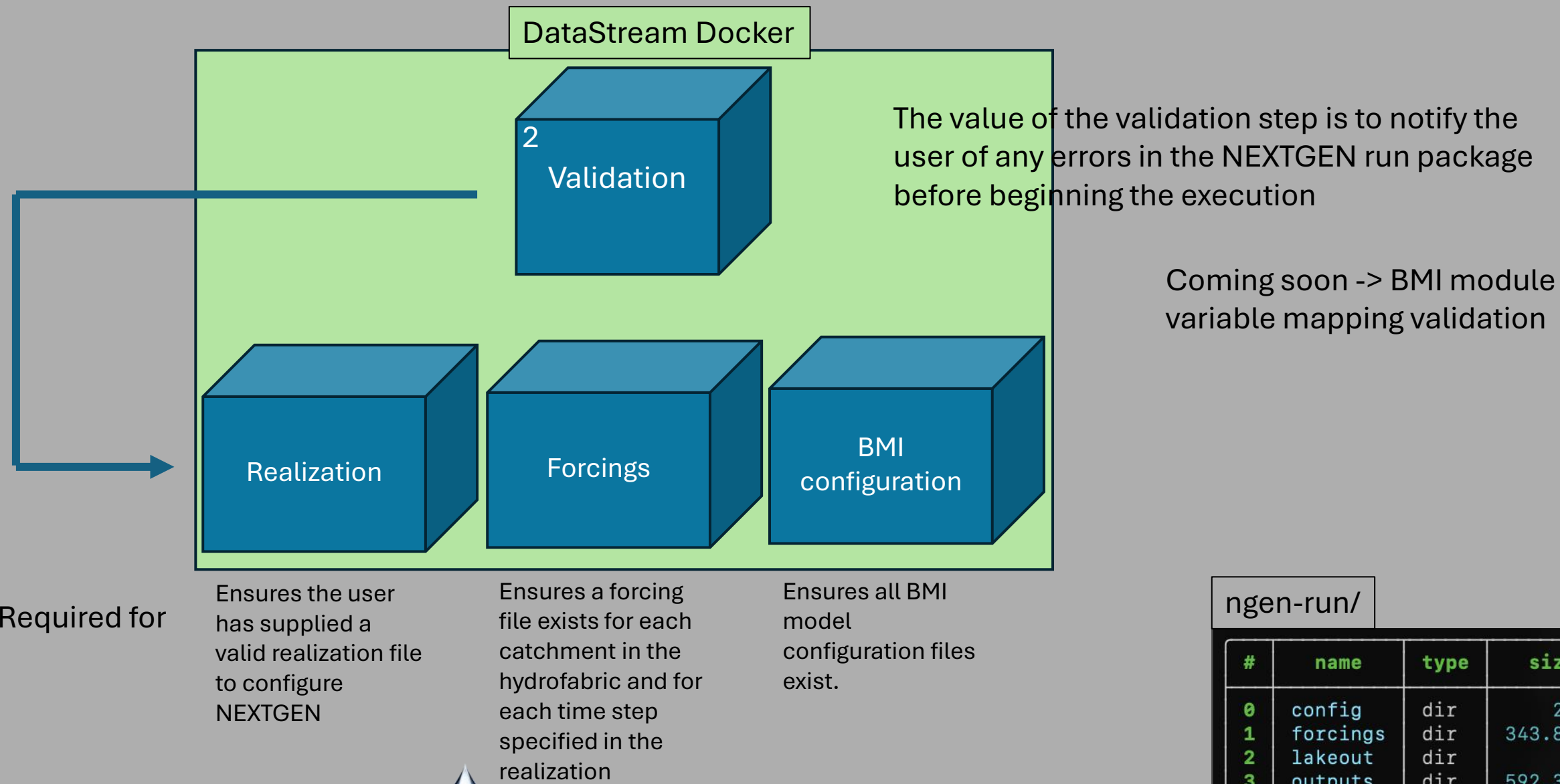
Ensures a forcing file exists for each catchment in the hydrofabric and for each time step specified in the realization

Ensures all BMI model configuration files exist.

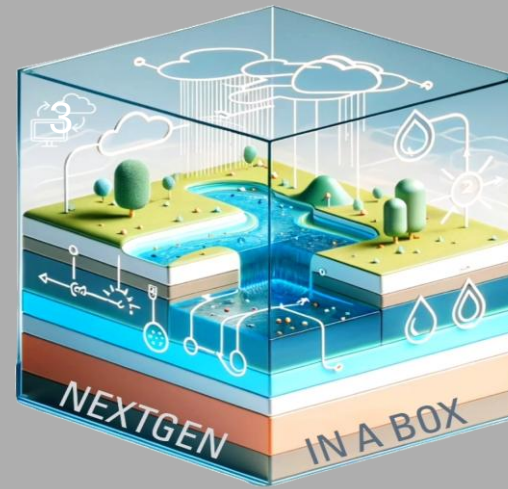
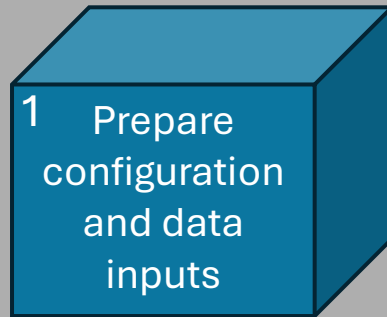


ngen-run/

#	name	type	size
0	config	dir	288 B
1	forcings	dir	343.8 KiB
2	lakeout	dir	64 B
3	outputs	dir	592.3 KiB
4	restart	dir	64 B



ngen-run/			
#	name	type	size
0	config	dir	288 B
1	forcings	dir	343.8 KiB
2	lakeout	dir	64 B
3	outputs	dir	592.3 KiB
4	restart	dir	64 B



DataStreamCLI refers to the software chain that builds and validates NextGen input packages (ngen-run/), executes NextGen through NextGen In A Box (NGIAB), and versions the entire run for reproducibility.

This enforces a standard folder (ngen-run/), which enables interoperability and reproducibility.

ngen-run/

#	name	type	size
0	config	dir	288 B
1	forcings	dir	343.8 KiB
2	lakeout	dir	64 B
3	outputs	dir	592.3 KiB
4	restart	dir	64 B

- Merkle Tree based hashing algorithm
 - “Root” hash allows for quickly identifying if two ngen-run directories are different.
 - Ability to query whether some file is a part of the tree represented by the root hash
 - Ability to compare files without opening them

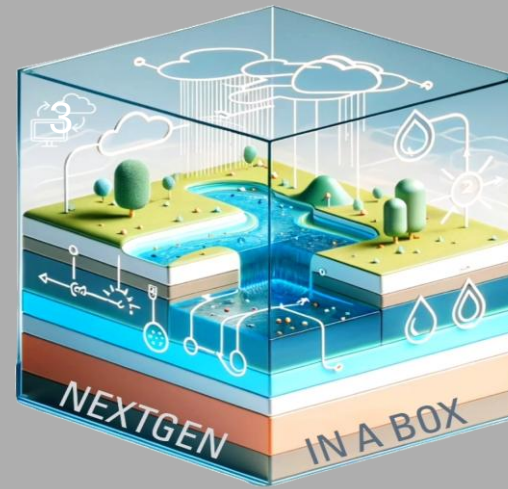
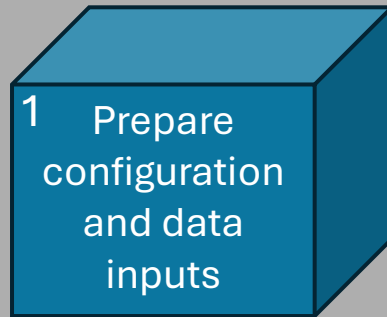


```
[jlaser@LYNK-59WW6S3 ngen-datastream]$ docker run --rm -v $(pwd)/data/datastream_test_VPU09_0520_with_resources_new_realization:/mounted_dir zwills/merkdir /merkdir/merkdir verify-file -t /mounted_dir/merkdir.file -n "ngen-run/config/realization.json"
OK: file is still verified by this Merkle tree
```

ngen-run/

#	name	type	size
0	config	dir	288 B
1	forcings	dir	343.8 KiB
2	lakeout	dir	64 B
3	outputs	dir	592.3 KiB
4	restart	dir	64 B





DataStreamCLI refers to the software chain that builds and validates NextGen input packages (ngen-run/), executes NextGen through NextGen In A Box (NGIAB), and versions the entire run for reproducibility.

This enforces a standard folder (ngen-run/), which enables interoperability and reproducibility.

ngen-run/

#	name	type	size
0	config	dir	288 B
1	forcings	dir	343.8 KiB
2	lakeout	dir	64 B
3	outputs	dir	592.3 KiB
4	restart	dir	64 B



3. Research DataStream Motivation

Need for regionalized parameterization and modeling

Within the NextGen Framework, models and parameters can be configured individually for each catchment

High on-premise cost of HPC

Cloud providers offer a cost-effective alternative to purchasing and maintaining expensive on-premise hardware.

The Research DataStream is written in Terraform and made publicly available.

Research 2 Operations

The Research DataStream is open to community contributions.

An evaluation workflow is under development to ensure continued improvement of the system.



4. Research DataStream Design

- CONUS wide
 - Distributed processing by Vector Processing Unit (VPU)
 - Regional hydrologic processes map to compute resources
 - Outputs are delineated by VPU
- Mimic NWM forecast cycles
 - Short range (18 hourly time steps) 24 times per day
 - Medium range (240 hourly time steps) 4 times per day
 - Analysis assim extend (28 hourly time steps) 1 per day
- Publicly available and editable NextGen configuration files
 - Automated evaluation drives improvement.
- AWS Step Functions state machine
 - Manages infrastructure workflow
- DataStreamCLI
 - Manages on-server workflow



4. Research DataStream State

- VPU available : 02, 03N, 03S, 03W, 04, 05, 06, 08, 09, 10L, 10U, 11, 12, 13, 14, 15, 16, 18
 - (05, 10L, 10U, 11 not available for medium range)
- Run Types –
 - short range (all initialization cycles),
 - medium range (all cycles, 1st member),
 - analysis assim extend
- Cold start
- NextGen configuration - NOAH-OWP, PET, CFE, and troute.
 - Dynamically read on each execution from publicly available realizations that now hold mutable community parameters.



Research DataStream: Workshop

- https://github.com/CIROH-UA/ngen-datastream/blob/main/docs/CIROH_devcon_2025/workshop.md
- Ask at least 1 anonymous question.



Future Work

- Implement community contribution workflow
 - Evaluation
 - Validation
- GUI for DataStreamCLI
- Stay up-to-date with hydrofabric
- Academic article

