NEXTGEN SIMULATION DEVELOPMENT TOOLS: NGEN-DATASTREAM

LYNKER: JORDAN LASER, ZACH WILLS, MIKE JOHNSON, JUSTIN SINGH-MOHUDPUR, NELS FRAZIER, AUSTIN RANEY

ALABAMA WATER INSTITUTE: ARPITA PATEL, JAMES HALGREN, JOSH CUNNINGHAM, SHAHABUL ALAM, TRUPESH PATEL, HARI TEJA JAJULA, BENJAMIN LEE

SPRING 2024

OVERVIEW

- Motivation
- Development Roadmap
- Conceptual model
- Usage
- Research Datastream
- Workshop
- Future work

MOTIVATION

Uniform data pipeline that abstracts the laborious process of collecting input data and executing NextGen

→ ngen-datasteam will perform every step in executing NextGen. The user can compute steps separately and provide those files directly.
Batteries included while not dogmatic

Reproducibility

→ Enforces the NextGen In A Box standard run directory.

Metadata – all relevant information about user inputs, code versions, host architecture, etc. so that a NextGen execution can be understood and reproduced.

Need baseline NextGen dataset to evaluate new realizations against

Scaling from laptop to HPC

→ ngen-datastream can issue NextGen jobs to an AWS state machine that use Lambda functions to coordinate NextGen executions in the cloud. This infrastructure is behind the NextGen research datastream. In addition, allows users to customize their host to match their compute requirements.

NOAA CIROH THE UNIVERSITY OF ALABAMA Alabama Water Institute Lynker

DEVELOPMENT ROADMAP

ngen-datastream is already a powerful tool, but is still under development and has not been rigorously tested within the community

We encourage community feedback and questions. If you discover a bug, or would appreciate different functionality, let us know by submitting an issue to the repository

We are here!

- Identifying needs in community
- Software architecture decisions

- Writing the software
- Developer based testing

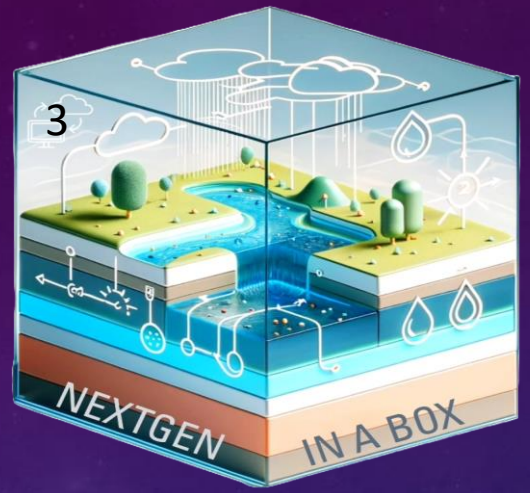
- Releasing software to community
- ngen-datastream version 1.0

- Continuous Integration Continuous Deployment (CI/CD)
- Feedback from community

- Add features

1 Prepare configuration and data inputs

2 Validation



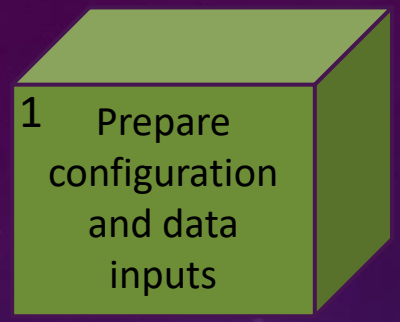
4 Versioning

NGEN-DATASTREAM

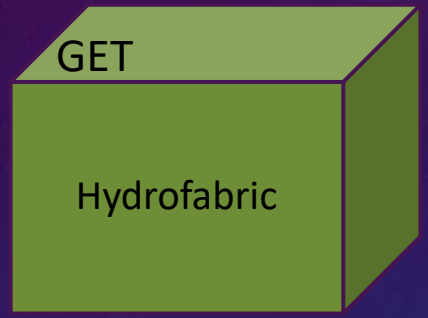
ngen-datastream refers to the software chain that builds and validates NextGen input packages (`ngen-run/`), executes NextGen through NextGen In A Box (NGIAB), and versions the entire run for reproducibility.

This enforces a standard folder `ngen-run/` which makes validation and versioning possible. A standard run folder also allows for other new tools to easily integrate with NGIAB (e.g., DataPreprocessor).

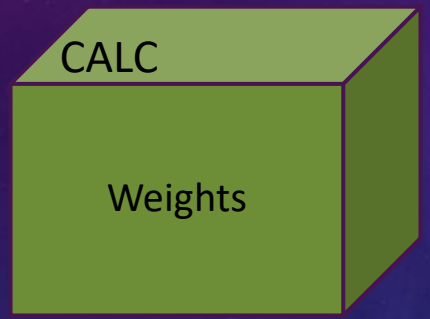
#	name	type	size
0	config	dir	288 B
1	forcings	dir	343.8 KiB
2	lakeout	dir	64 B
3	outputs	dir	592.3 KiB
4	restart	dir	64 B



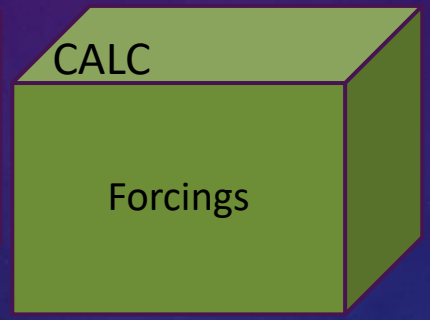
Required steps to build `ngen-run/config` and `ngen-run/forcings`



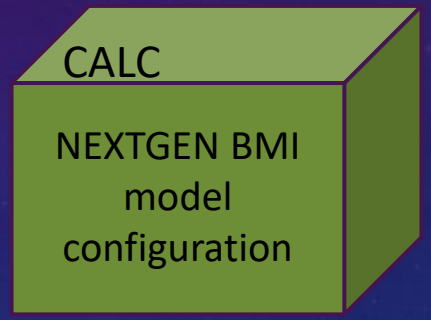
Defines spatial domain



Indices and coverage used to extract catchment averaged forcings



Performs conversion between National Water Model and NextGen forcings formats



Required files for NextGen BMI modules

#	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

https://www.lynker-spatial.com/#hydrofabric/v20.1/gpkg/

1 Prepare configuration and data inputs

GET

Hydrofabric

Defines spatial domain

Lynker

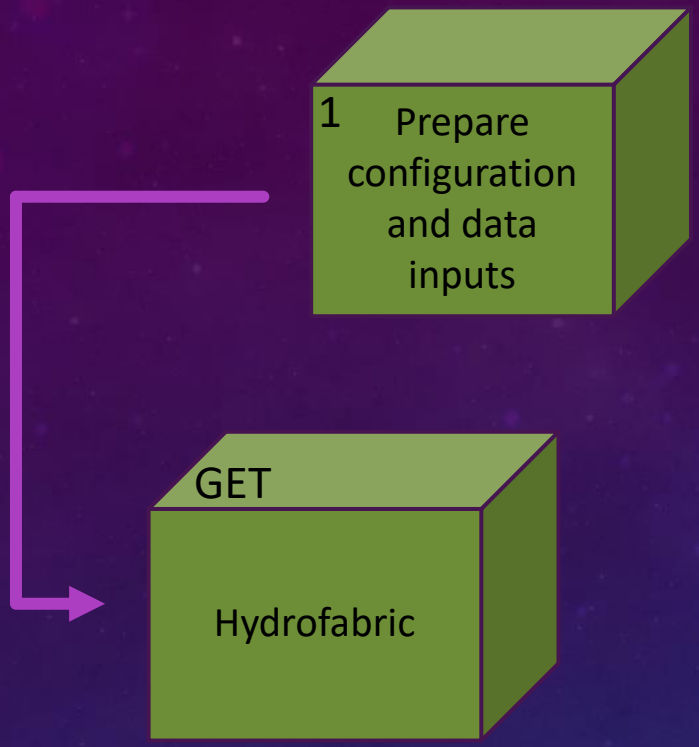
lynker-spatial / hydrofabric / v20.1 / gpkg

Hide folders? Folder Bucket 21

Show 50 entries Search:

Object	Last Modified	Timestamp	Size
nextgen_01.gpkg	20 days ago	2024-04-24 14:15:44	76 MB
nextgen_02.gpkg	20 days ago	2024-04-24 14:15:46	125 MB
nextgen_03N.gpkg	20 days ago	2024-04-24 14:15:49	117 MB
nextgen_03S.gpkg	20 days ago	2024-04-24 14:15:54	54 MB
nextgen_03W.gpkg	20 days ago	2024-04-24 14:15:56	126 MB

NOAA CIROH The University of Alabama Alabama Water Institute Lynker



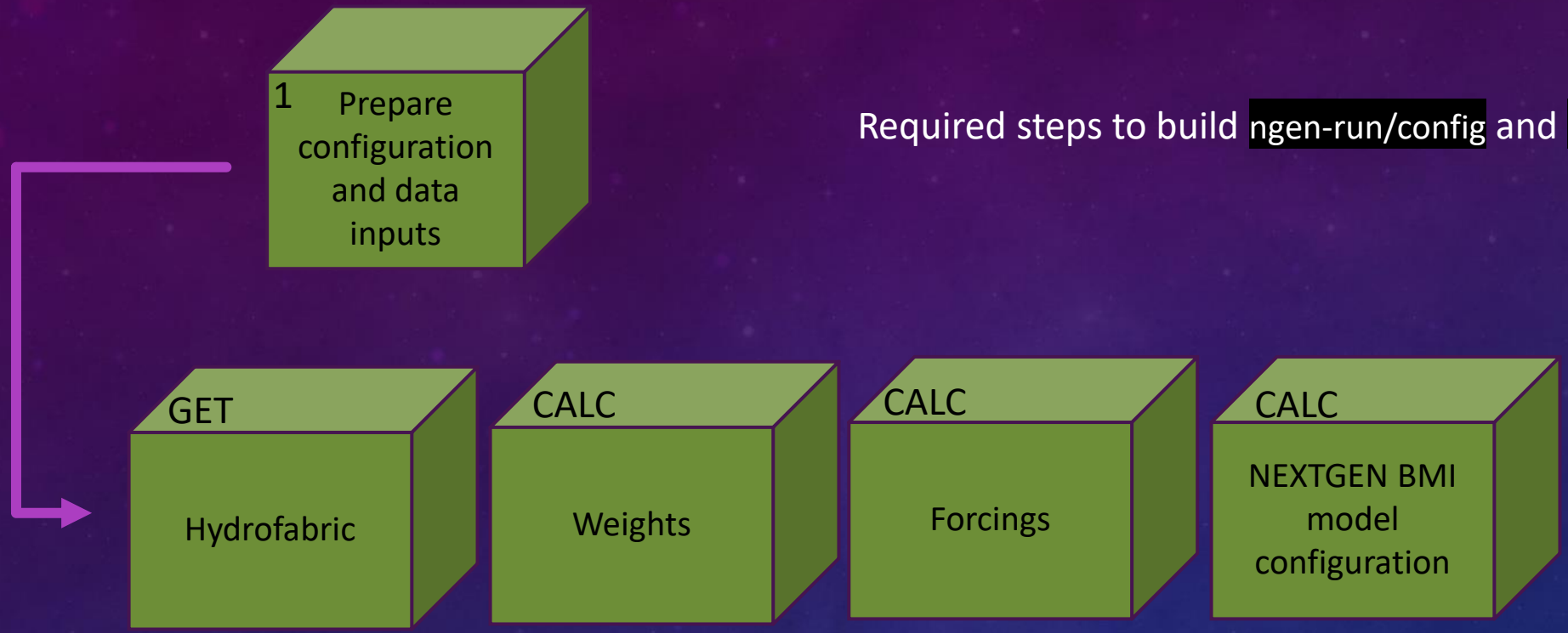
Defines spatial domain

"What if I want to define my own spatial domain?"

Do you know the catchment id you want to subset with?

Use the subsetting options!
hfsubset is integrated into ngen-datastream

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



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

#	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

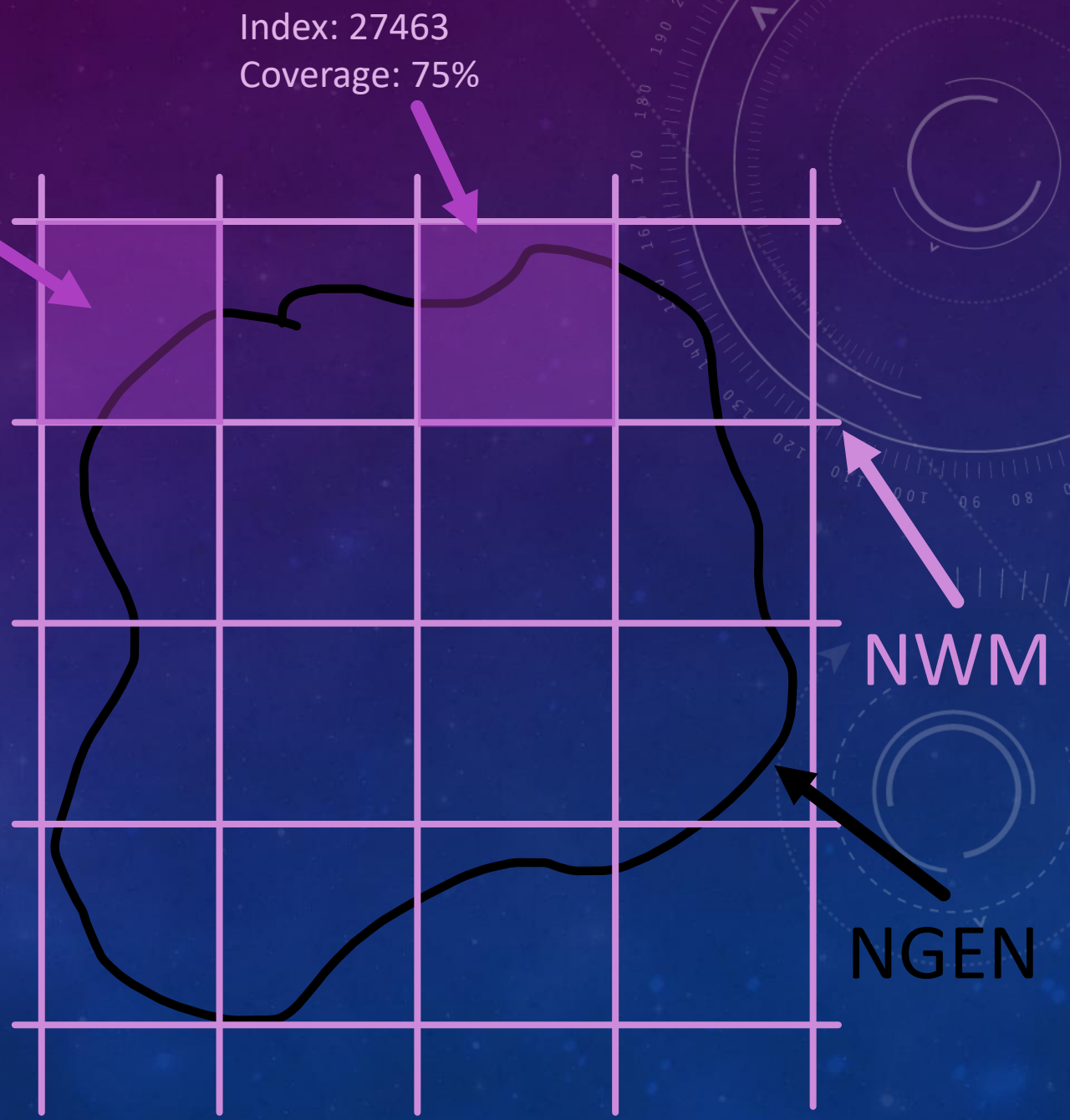
Hydrofabric

Defines spatial domain

CALC

Weights

Indices and coverage used to extract catchment averaged forcings



1 Prepare configuration and data inputs

GET
Hydrofabric

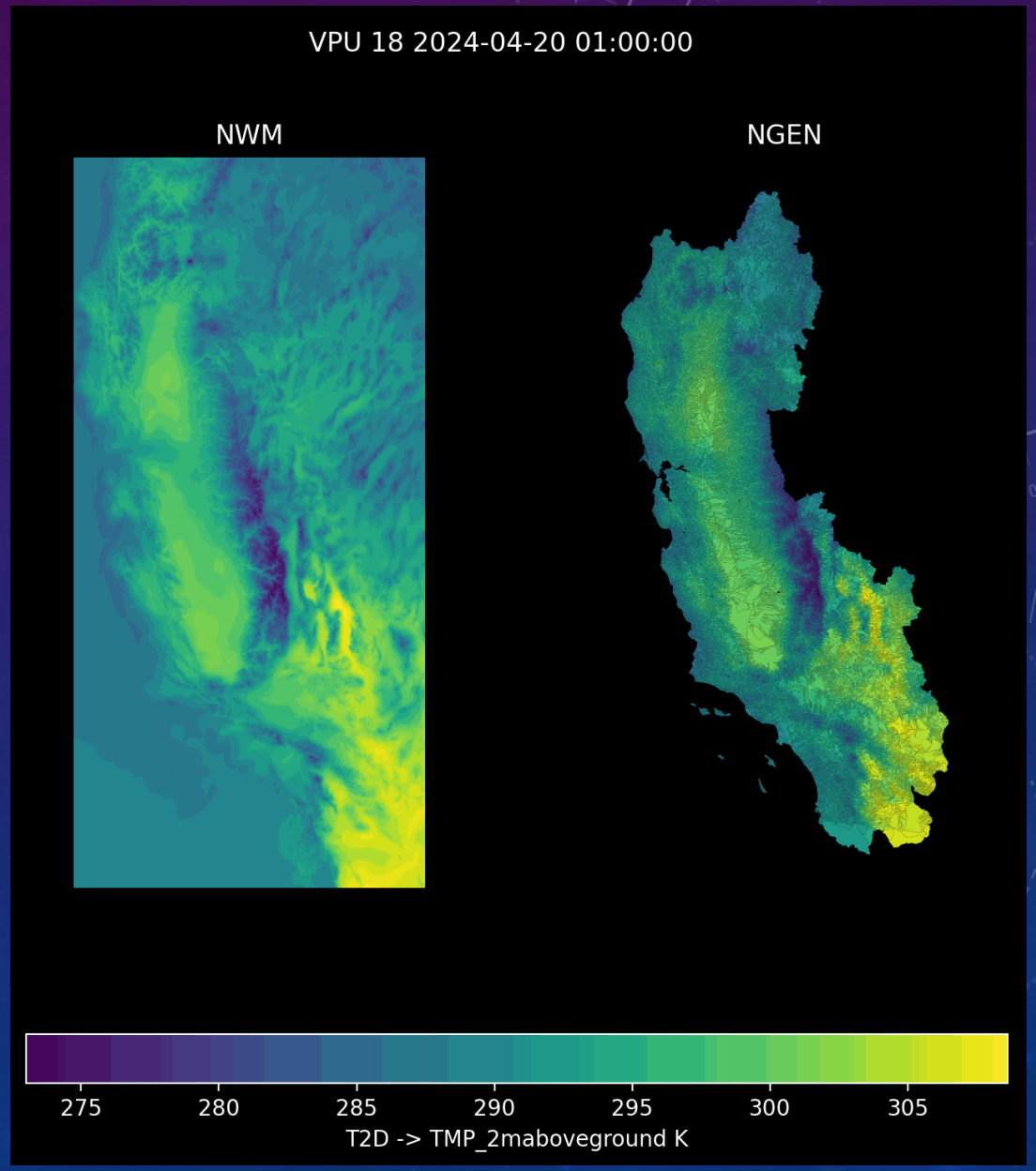
Defines spatial domain

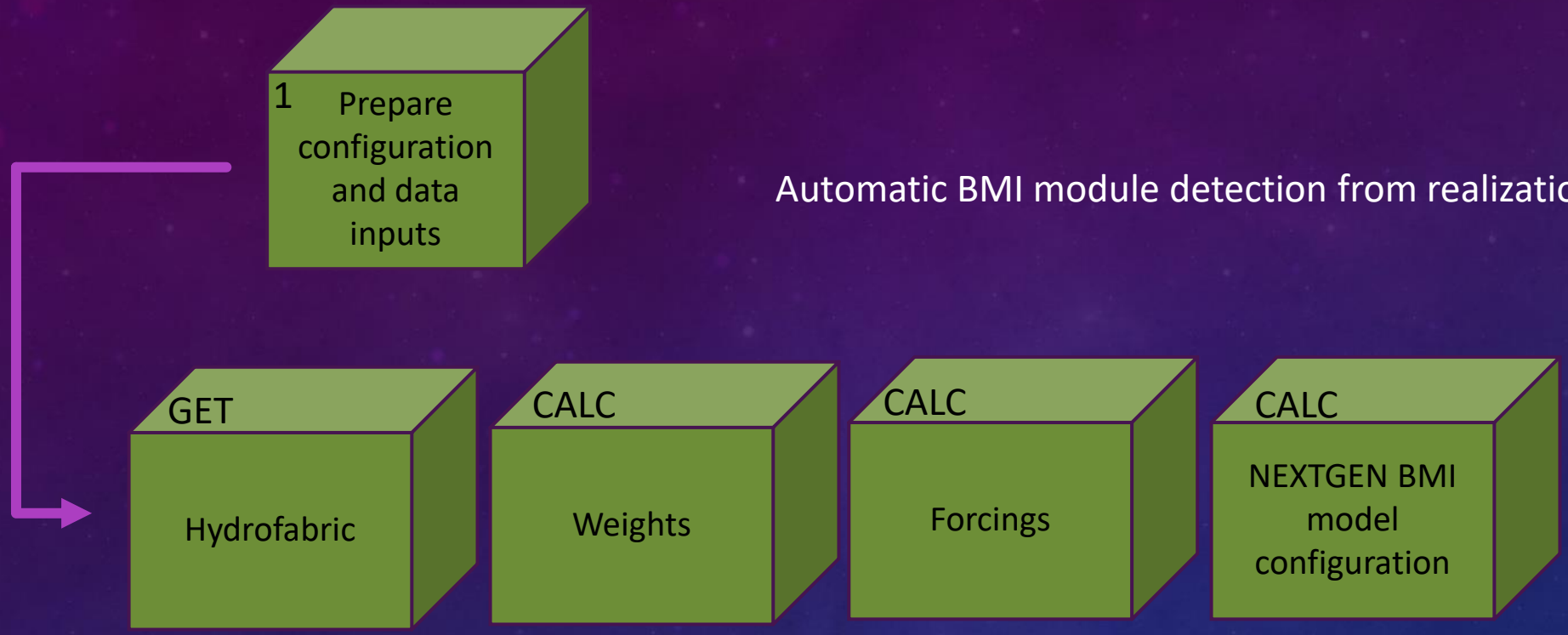
CALC
Weights

Indices and coverage used to extract catchment averaged forcings

CALC
Forcings

Performs conversion between National Water Model and NextGen forcings formats





Automatic BMI module detection from realization file

Defines spatial domain

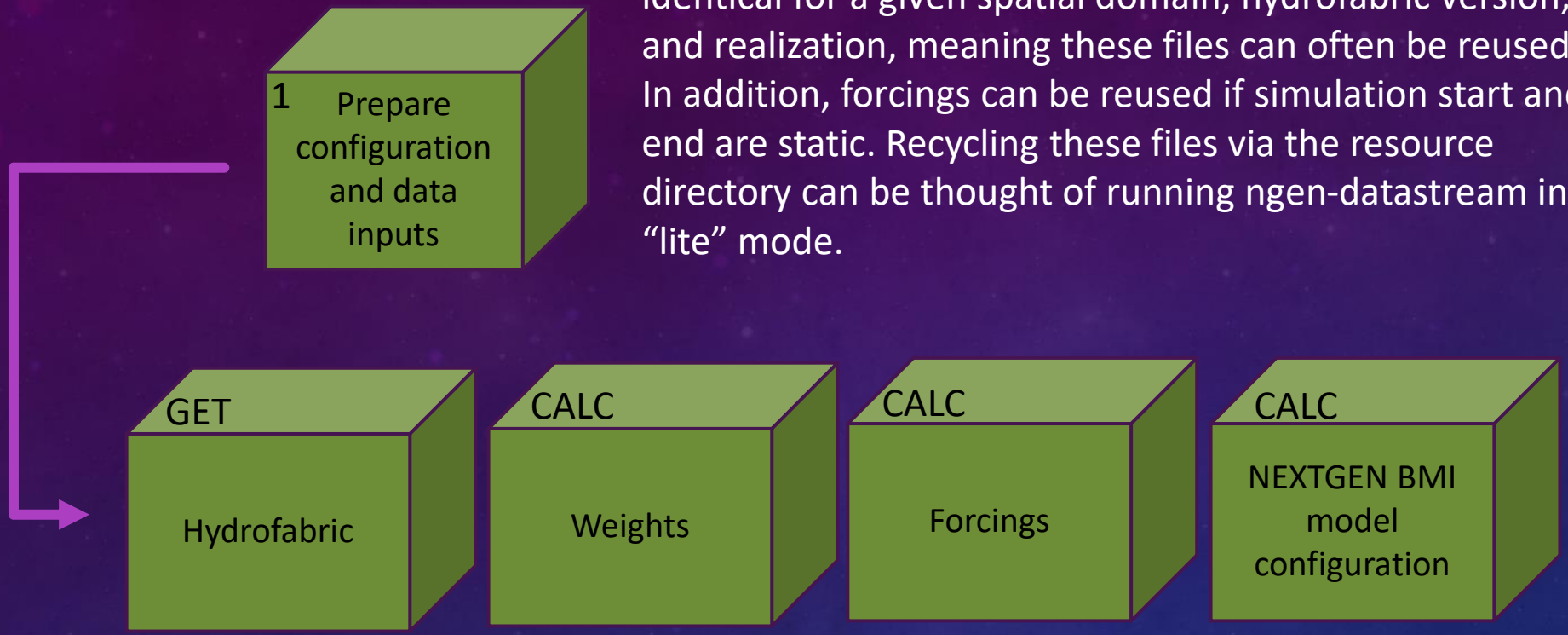
Indices and coverage used to extract catchment averaged forcings

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

Weights and BMI model configuration generation are identical for a given spatial domain, hydrofabric version, and realization, meaning these files can often be reused. In addition, forcings can be reused if simulation start and end are static. Recycling these files via the resource directory can be thought of running ngen-datastream in “lite” mode.



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

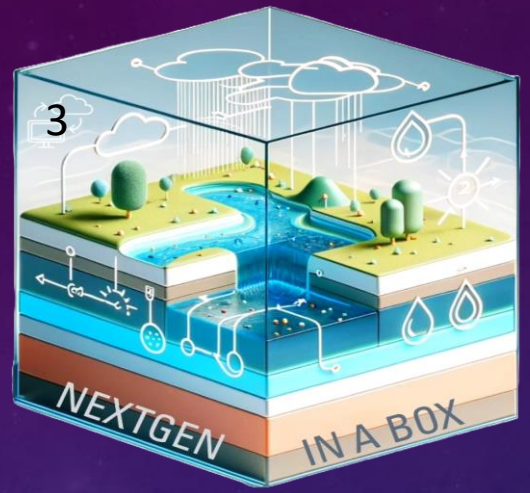
SAVES MONEY

```
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
|
```

https://github.com/CIROH-UA/ngen-datastream/tree/main?tab=readme-ov-file#resource_dir-datastream-resources

1 Prepare configuration and data inputs

2 Validation



4 Versioning

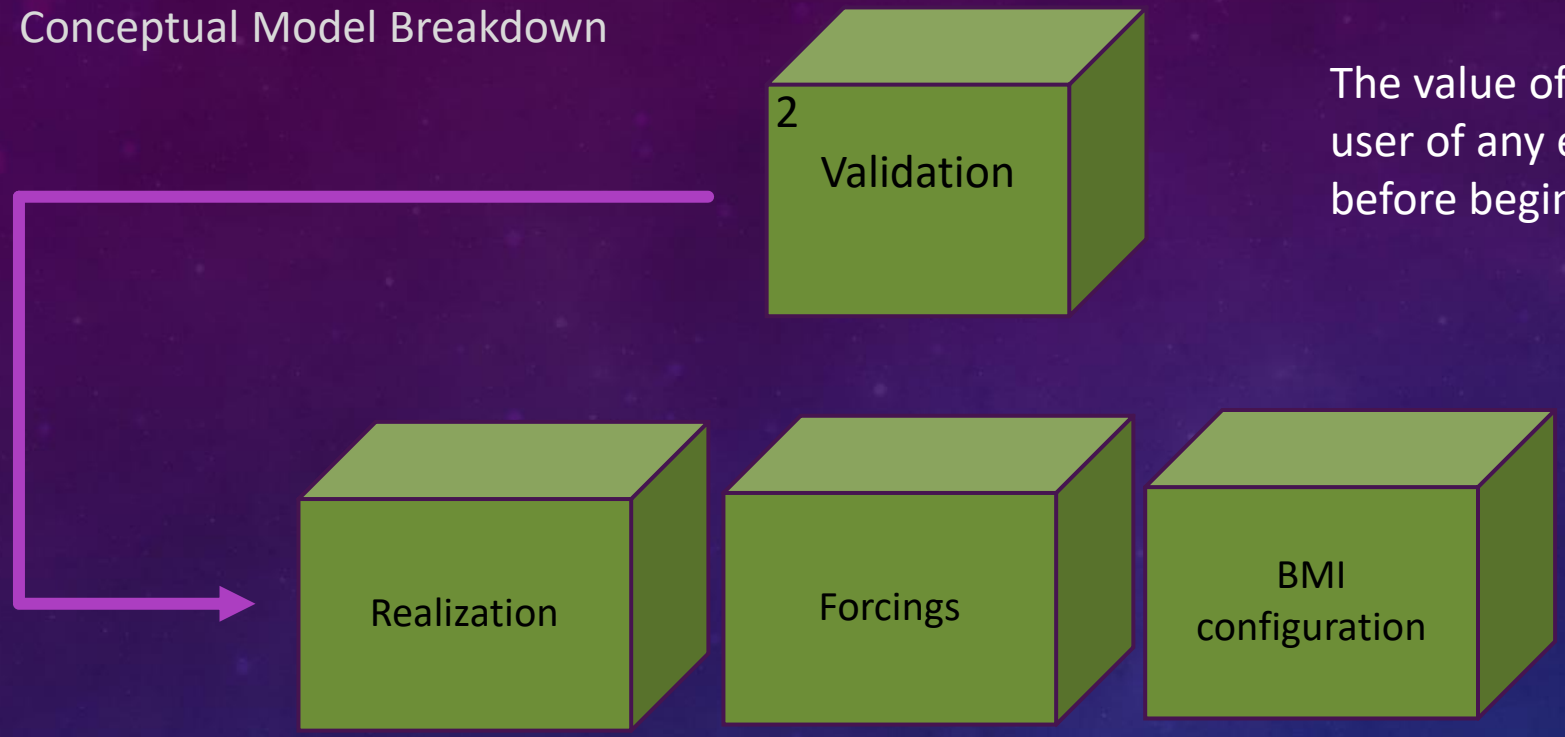
NGEN-DATASTREAM

ngen-datastream 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 makes validation and versioning possible. A standard run folder also allows for other new tools to easily integrate with NGIAB (e.g., DataPreprocessor).

#	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

NEXTGEN Water Modeling Framework Datastream Conceptual Model Breakdown



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

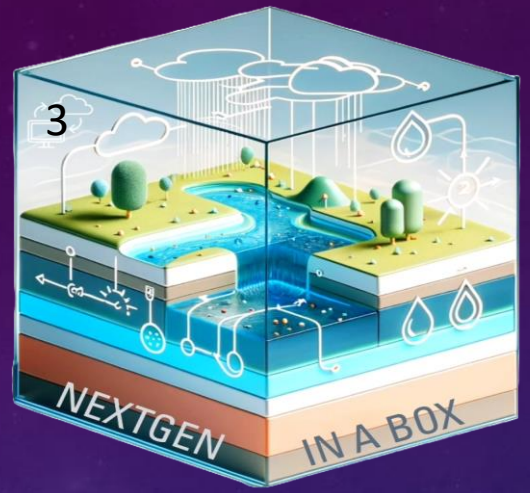
Ensures a forcing file exists for each catchment in the hydrofabric and for each time step specified in the realization

Ensures all BMI model configuration files exist.

#	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

2 Validation



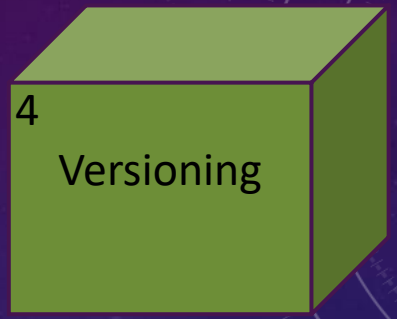
4 Versioning

NGEN-DATASTREAM

ngen-datastream 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 makes validation and versioning possible. A standard run folder also allows for other new tools to easily integrate with NGIAB (e.g., DataPreprocessor).

#	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

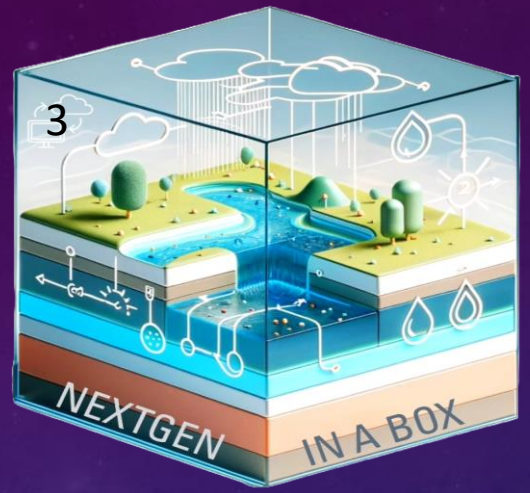


- Merkle Tree based hashing algorithm
  - “Root” hash allows for quickly identifying if two ngen-run directories are different.
  - Ability to query whether some file is a part of the tree represented by the root hash
  - Ability to compare files without opening them

```
[jlaser@LYNK-59WW6S3 ngen-datastream]$ docker run --rm -v $(pwd)/data/datast
ream_test_VPU09_0520_with_resources_new_realization:/mounted_dir zwills/merk
dir /merkdir/merkdir verify-file -t /mounted_dir/merkdir.file -n "ngen-run/c
onfig/realization.json"
OK: file is still verified by this Merkle tree
```

1 Prepare configuration and data inputs

2 Validation



4 Versioning

NGEN-DATASTREAM

ngen-datastream 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 makes validation and versioning possible. A standard run folder also allows for other new tools to easily integrate with NGIAB (e.g., DataPreprocessor).

#	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

CUSTOMIZE NEXTGEN SIMULATION RESOURCES IN AWS

ngen-datastream allows users to submit NextGen simulation jobs to cloud-based hosts in Amazon Web Services (AWS)

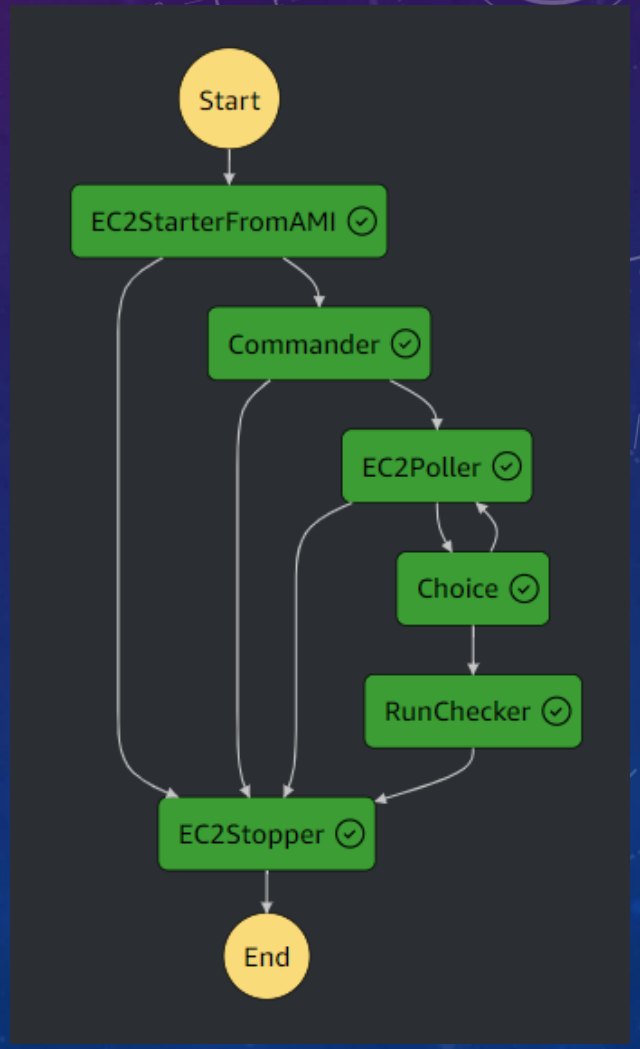
Users submit jobs via a customizable execution json to a generalizable AWS state machine that manages job execution

```
aws stepfunctions start-execution \
    --state-machine-arn $SM_ARN \
    --name $(env TZ=US/Eastern date +'%Y%m%d%H%M%S')\
    --input "file://"$EXEC_DIR""$file"" --region $REGION
```

- Terraform allows for quick building of complex AWS infrastructure
- Allows users to access HPC resources
- Highly configurable execution file allows users to finely tune resources to their ngen executions

Customizable parameters in the execution json include:

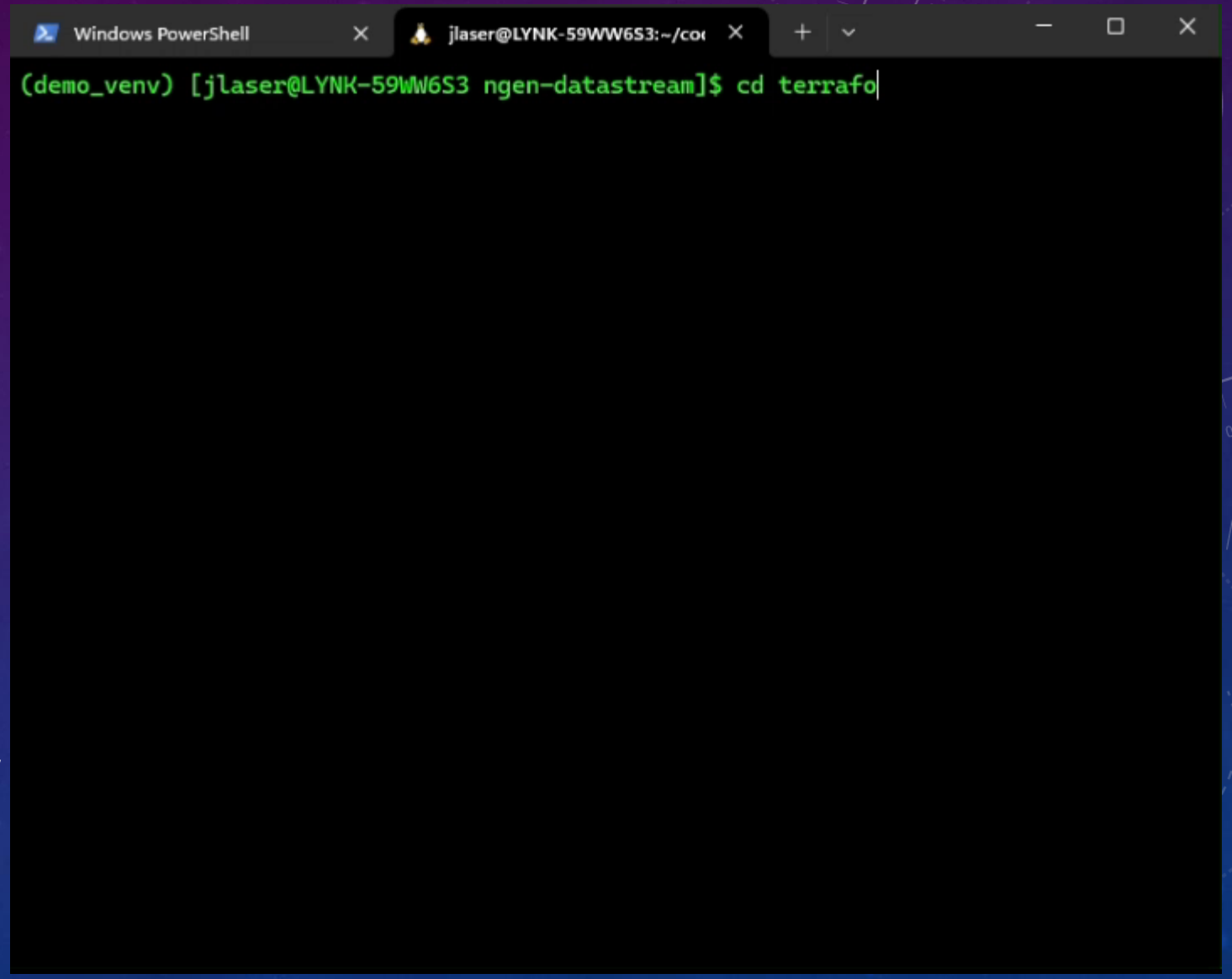
- Instance Type
- Image Id
- Number of instances
- Volume Size
- Region
- Security Groups / Instance profile (IAM)
- Commands



TERRAFORM

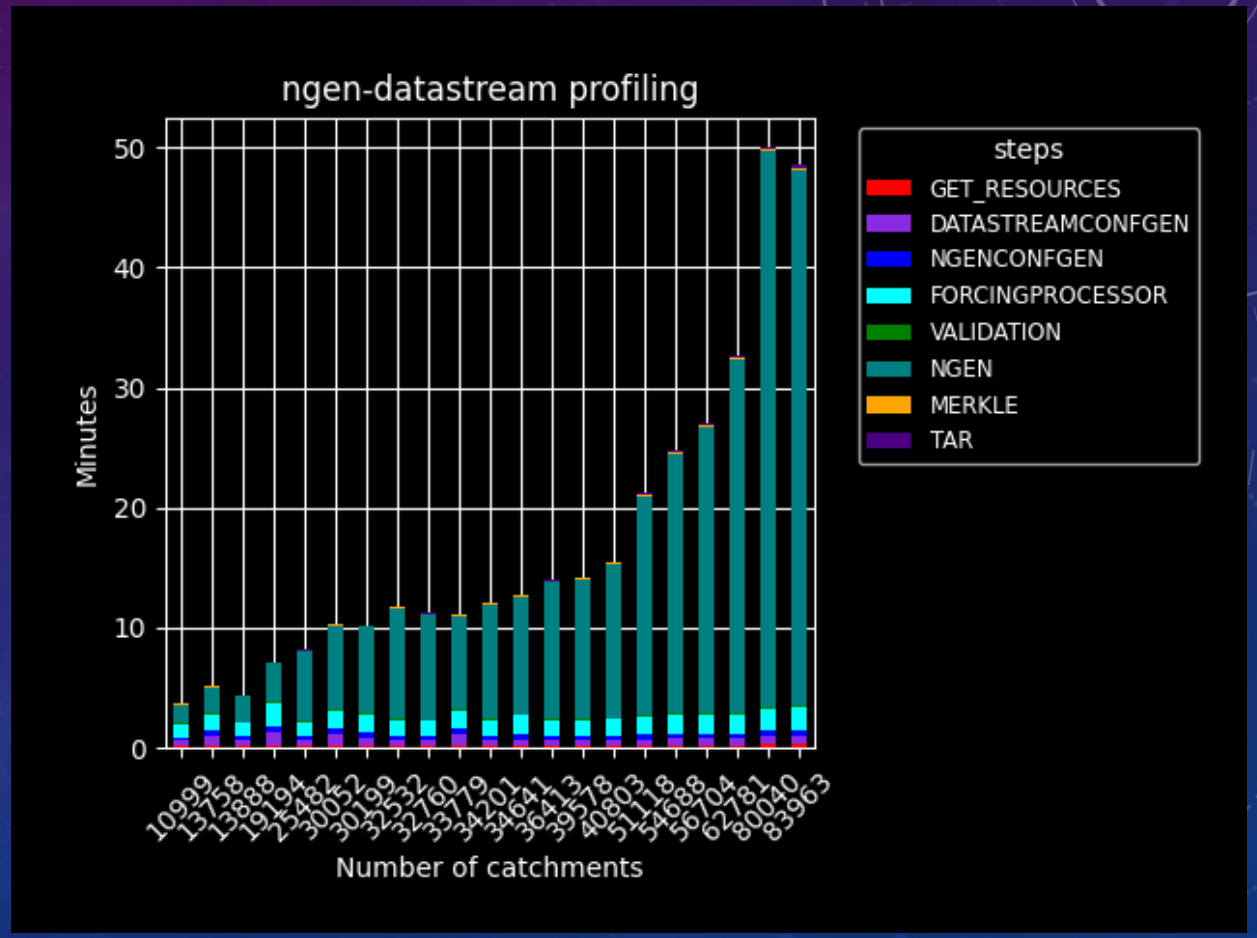
- Terraform (IaC) builds the AWS state machine
  - Build the cloud infrastructure with terraform init, plan, and apply
  - Edit the execution json to use an AWS Machine Image built from an ec2 with ngen-datastream installed. Set desired options
  - Submit job to AWS state machine via awscli

- Takeaways:
  - Terraform allows for quick building of complex AWS infrastructure
  - Allows users to access HPC resources ($$)
  - Highly configurable execution file allows users to finely tune resources to their ngen executions



RESEARCH DATASTREAM IN AWS CLOUD

- Split VPU daily run - An application of the ngen-datastream AWS state machine
  - 24-hour CONUS NextGen simulation scheduled each day with AWS EventBridge
  - 22 individual ec2 instances, 1 for forcingprocessor , 21 simultaneously processing for each VPU
  - CONUS runtime determined by the runtime of the largest VPU.



USAGE

- Be aware of your resources constrains (CPU, Memory, Network)
- In general, ngen-datastream memory footprint scales with
  - Number of catchments (size of the spatial domain)
  - Number of time steps (Simulation duration / output_interval)
- If a crash is experienced, either increase the host resources or decrease one or both above dimensions
  - Linux commands to monitor resources
    - Watch memory usage -> `free –h –s2`
    - Watch processes/cpu usage -> `top`
    - Check available processes -> `nprocs`
- https://github.com/CIROH-UA/ngen-datastream/blob/main/USAGE.md

OPTIONS

See the README.md in the repo for an explanation of each variable

```
> cd ngen-datastream && ./scripts/stream.sh --help

Usage: ./scripts/stream.sh [options]
Either provide a datastream configuration file
  -c, --CONF_FILE           <Path to datastream configuration file>
or run with cli args
  -s, --START_DATE          <YYYYMMDDHHMM or "DAILY">
  -e, --END_DATE            <YYYYMMDDHHMM>
  -D, --DOMAIN_NAME         <Name for spatial domain>
  -g, --GEOPACAKGE          <Path to geopackage file>
  -G, --GEOPACKAGE_ATTR     <Path to geopackage attributes file>
  -w, --HYDROFABRIC_WEIGHTS <Path to hydrofabric weights parquet>
  -I, --SUBSET_ID           <Hydrofabric id to subset>
  -i, --SUBSET_ID_TYPE      <Hydrofabric id type>
  -v, --HYDROFABRIC_VERSION <Hydrofabric version>
  -R, --REALIZATION         <Path to realization file>
  -d, --DATA_DIR            <Path to write to>
  -r, --RESOURCE_DIR        <Path to resource directory>
  -f, --NWM_FORCINGS_DIR    <Path to nwm forcings directory>
  -F, --NGEN_FORCINGS       <Path to ngen forcings tarball>
  -S, --S3_MOUNT            <Path to mount s3 bucket to>
  -o, --S3_PREFIX           <File prefix within s3 mount>
  -n, --NPROCS              <Process limit>
```

RUN options

simulation start and end

spatial domain

NextGen configuration

ngen-datasteam “lite”

AWS s3 mount

WORKSHOP

- https://github.com/CIROH-UA/ngen-datastream/blob/main/docs/CIROH_devcon_2024/workshop.md

TAKE-AWAYS

- ngen-datastream automates the process of collecting and formatting input data for NextGen, orchestrating the NextGen run through NextGen In a Box (NGIAB), and handling outputs. This software allows users to run NextGen in an efficient, relatively painless, and reproducible fashion.
- Flexibility in the design allows users to provide their own files if desired.
- Scalability via AWS state machine gives users access to HPC resources.
- Collaborative process! Feel free to contact me with questions

FUTURE WORK

- CI/CD
- Feedback/Collaboration with community
- Realization BMI module variable mapping validation
- Expand terraform to include more cloud providers
- Expand modules ngen-datastream is aware of
- Begin the science!
  - Find regional improvements to NextGen realization

TERMS

- Catchment – geographic area characterized by a single location, a nexus, where all precipitation in the area runs off through. A drainage basin.
- Nexus – the singular point where water flows into or out of a catchment. Often a point along a river.
- Subsetting – To reduce a large geopackage (many catchments) down to a smaller geopackage (fewer catchments) . In effect, this is choosing the domain over which ngen will run.
- Hashing – SHA256 algorithm applied to files to generate a unique id for a file. Useful for preserving and distinguishing unique inputs.
- Validation – Ensuring the ngen input directory data_dir has been constructed properly. Properly meaning that NextGen will not crash and will generate output data.

ACRONYMS

- NWM – National Water Model
- NGIAB – Next Generation National Water Model in a Box
- NGEN/NEXTGEN - Next Generation National Water Model
- IaC – Infrastructure as code
- VPU – Vector Processing Unit
- CFE – Conceptual Function Equivalent
- PET – Potential Evapotranspiration
- NOM – NOAA-OWP-Modular
- OWP – Office of water prediction
- BMI – Basic Model Interface

LINKS

- ngen-datastream
  - https://github.com/CIROH-UA/ngen-datastream/tree/main
- Forcingprocessor
  - https://github.com/CIROH-UA/ngen-datastream/tree/main/forcingprocessor
- Validation
  - https://github.com/CIROH-UA/ngen-datastream/blob/main/python/README.md
  - https://github.com/NOAA-OWP/ngen-cal
  - https://github.com/CIROH-UA/ngen-datastream/blob/main/python/src/datastream/run_validator.py
- NextGen BMI module config generation
  - https://github.com/CIROH-UA/ngen-datastream/blob/main/python/src/datastream/ngen_configs_gen.py
- Hydrofabric Subsetting
  - https://github.com/LynkerIntel/hfsubset
- Versioning
  - https://github.com/makew0rld/merkdir
- NGIAB
  - https://github.com/CIROH-UA/NGIAB-CloudInfra
- ngen-datastream AWS state machine
  - https://github.com/CIROH-UA/ngen-datastream/tree/main/terraform
- https://docs.ciroh.org/
- https://github.com/NOAA-OWP/ngen/wiki
- https://ciroh.ua.edu/