E3SM Processflow

AUTOMATED POST PROCESSING AND DIAGNOSTICS FOR GCM MODEL DATA

Sterling Baldwin
E3SM workflow group
Lawrence Livermore National Lab
https://github.com/ACME-Climate/acme_processflow

Too much data

- Global Climate Models produce a lot of data. E3SM plans to produce ~2000 years of simulated data in the next 6 months.

Too much data

- Global Climate Models produce a lot of data. E3SM plans to produce ~2000 years of simulated data in the next 6 months.
- The data’s not the product, scientific analysis is the product.

Too much data

- Global Climate Models produce a lot of data. E3SM plans to produce ~2000 years of simulated data in the next 6 months.
- The data’s not the product, scientific analysis is the product.
- Manually post processing is extremely time consuming

Post Processing Requirements

Post Processing Requirements

Post processing steps

1) Data transport
    - 28.6TB from HPC facility to post processing and data storage

Post Processing Requirements

Post processing steps

1) Data transport
   - 28.6TB from HPC facility to post processing and data storage

2) Regrid
   - From ne30 grid to fv129x256 grid
   - Extract time series variables

Post Processing Requirements

Post processing steps

1) Data transport
    - 28.6TB from HPC facility to post processing and data storage

2) Regrid
    - From ne30 grid to fv129x256 grid
    - Extract time series variables

3) Diagnostics
    - AMWG Diagnostics (atm only)
    - E3SM Diagnostics (atm only)
    - A-Primary Diagnostics (atm, ocn, ice)

Post Processing Example

E3SM OUTPUT

500 Years:

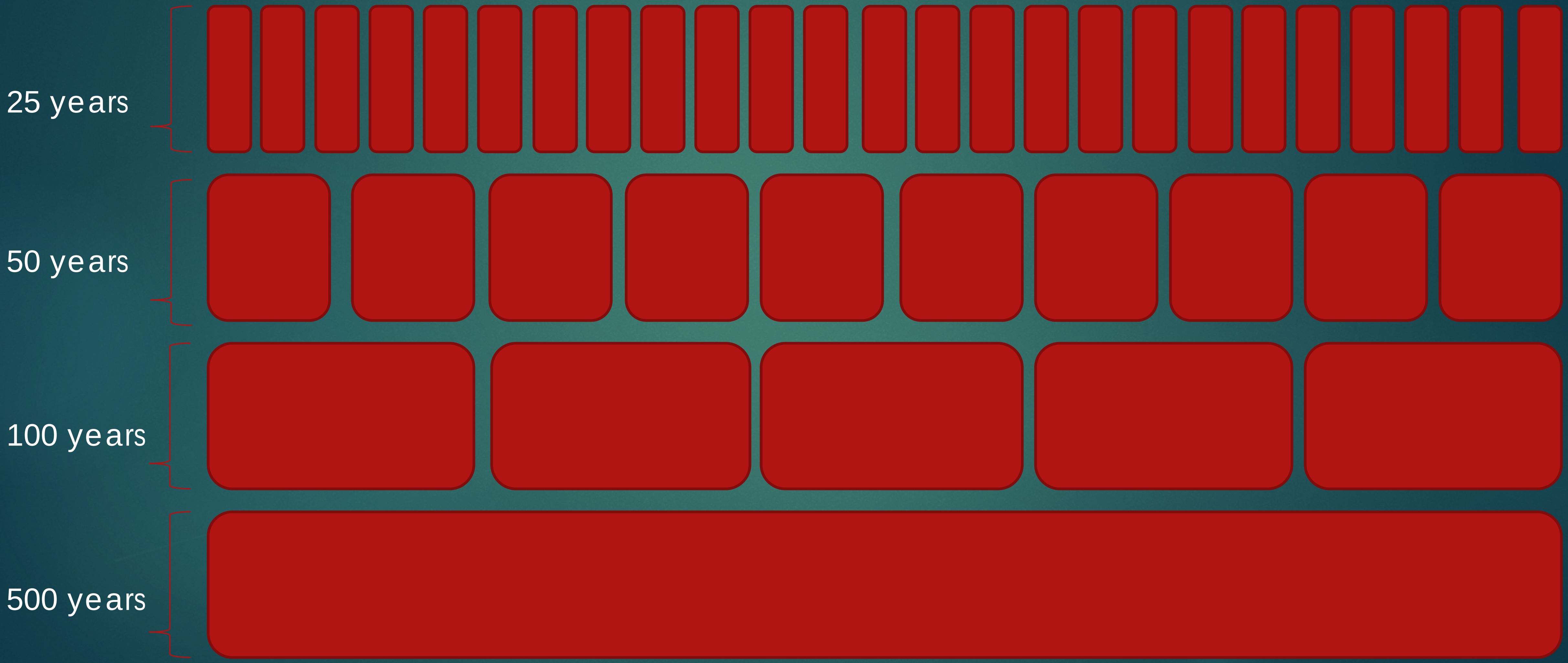
- Atmosphere
- Ocean
- Land
- Ice

Post Processing Example

E3SM OUTPUT

- 25 years: regrid + E3SM diags
- 50 years: regrid + AMWG diags
- 100 years: regrid + AMWG + Aprime
- 500 years: regrid + AMWG + Aprime + Time series

Post Processing Example



25 years

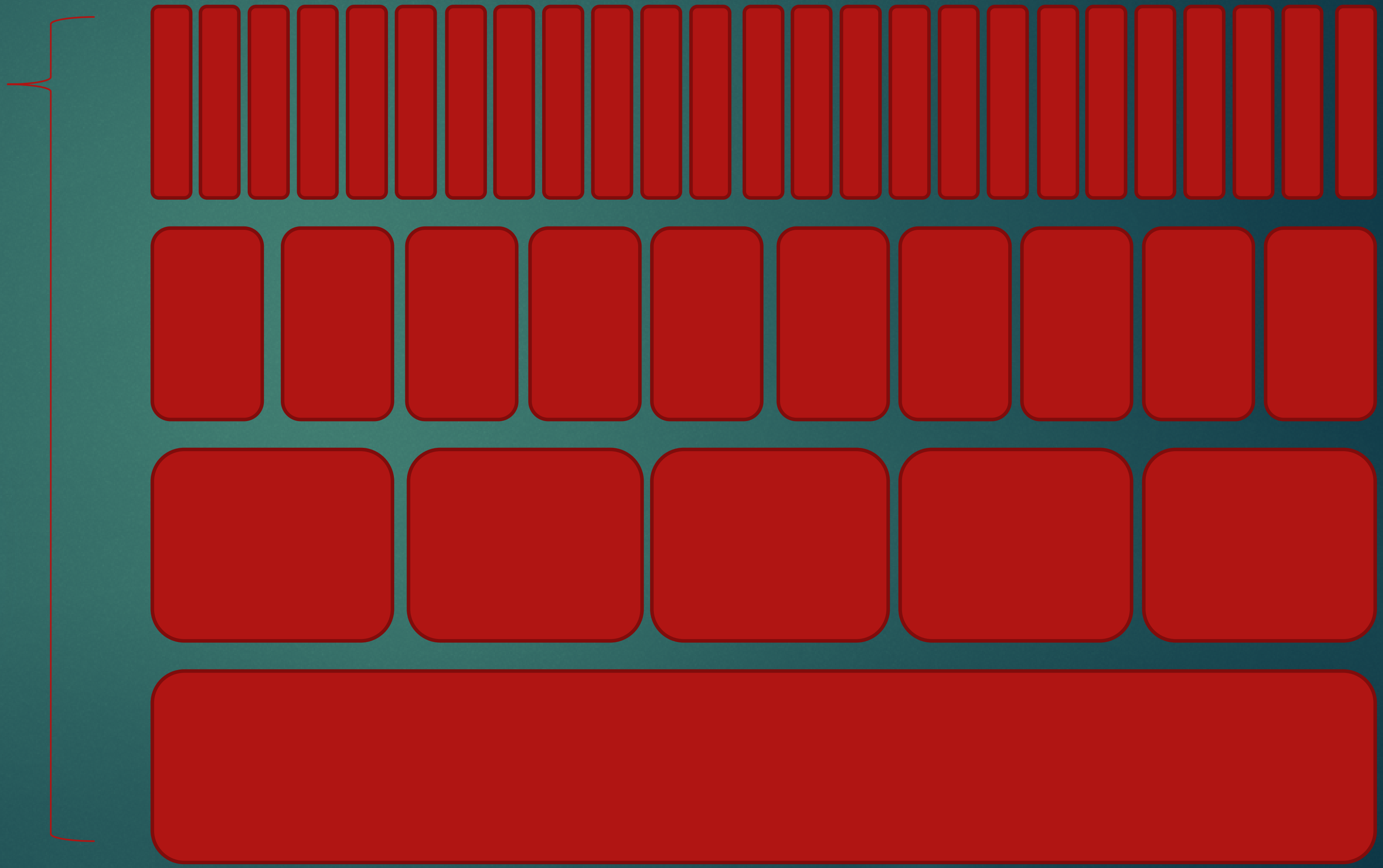
50 years

100 years

500 years

Post Processing Example

89 Jobs + Data Transfer!



Processflow: Automate Everything

Processflow: Automate Everything

- One Job to manage all your jobs

Processflow: Automate Everything

- One Job to manage all your jobs
- One config to setup and run everything

Processflow: Automate Everything

- One Job to manage all your jobs
- One config to setup and run everything
- Move your data, host your diagnostics

Processflow: Automate Everything

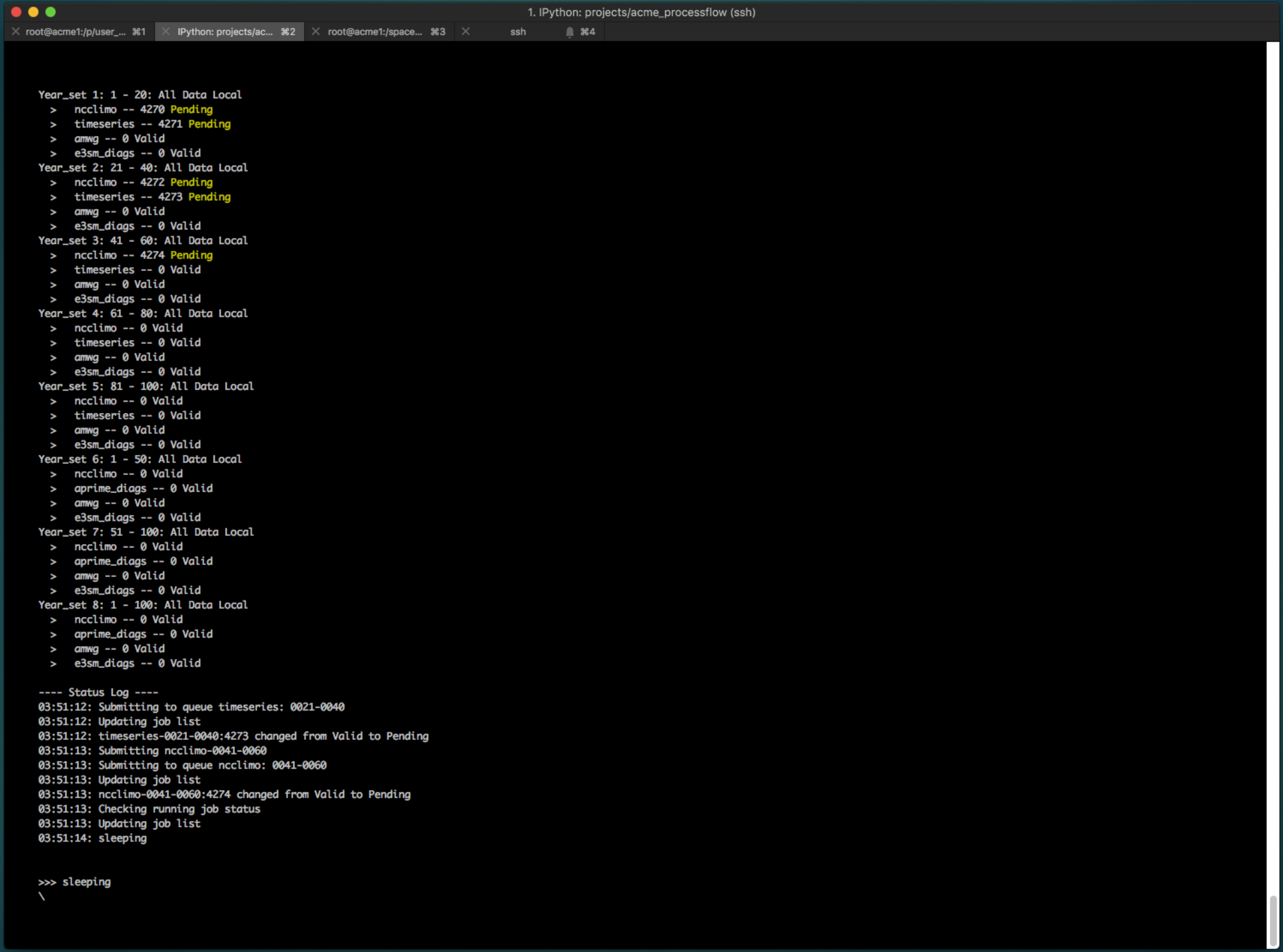
- One Job to manage all your jobs
- One config to setup and run everything
- Move your data, host your diagnostics
- Portable, extensible, robust, parallel

Example Configuration

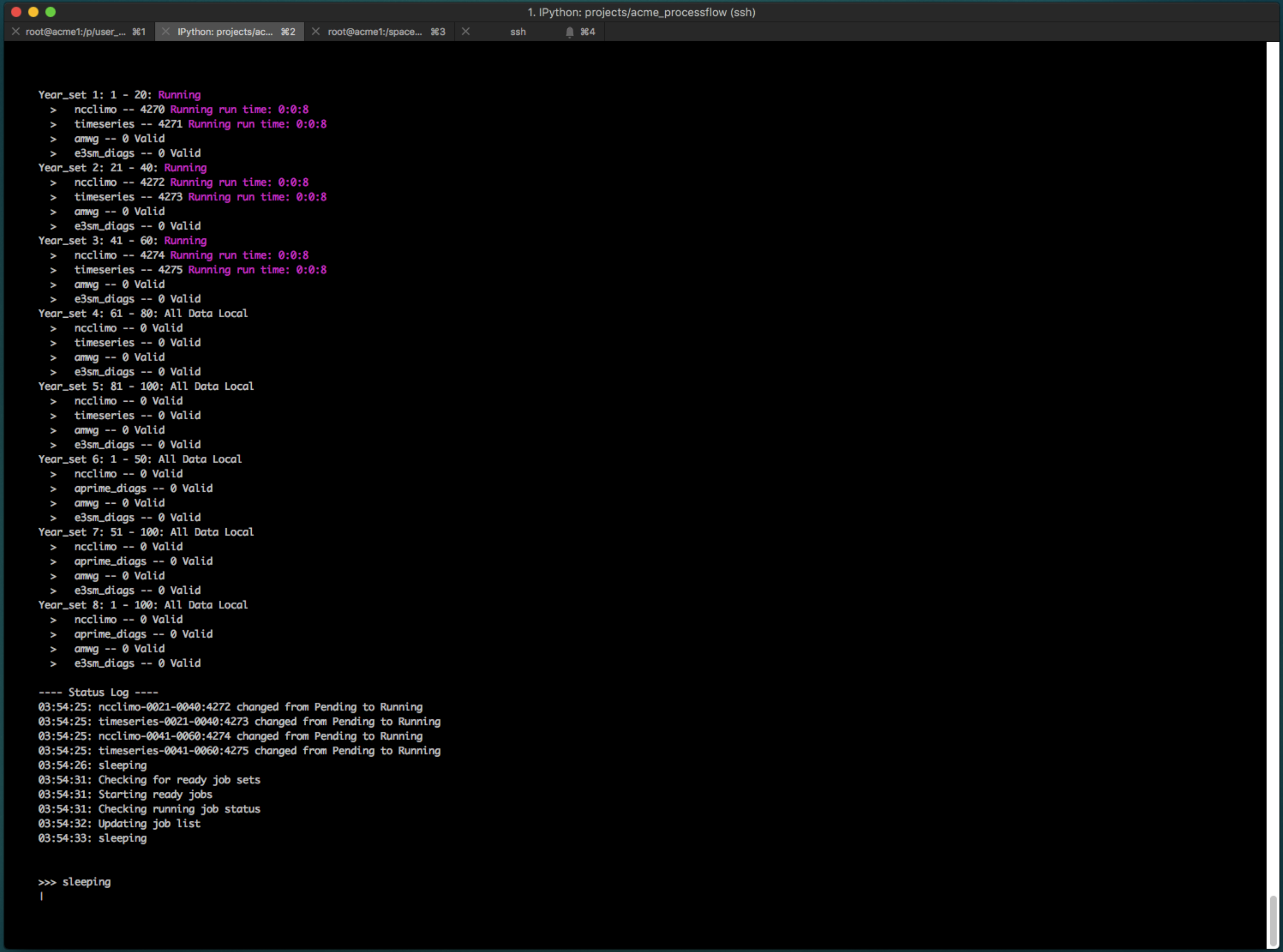
- project_path = /p/cscratch/baldwin32/DECKv1b_piControl
- source_path = /global/cscratch1/…/20180129.DECKv1b_piControl.ne30_oEC.Edison
- experiment = 20180129.DECKv1b_piControl.ne30_oEC.edison
- short_name = DECKv1b
- simulation_start_year = 1
- simulation_end_year = 100
- set_frequency = 20, 50, 100
- host_directory = /var/www/acme/acme-diags/

Example Configuration

- [[set_jobs]]
  - ncclimo = 20, 50, 100
  - timeseries = 20
  - e3sm_diags = 20, 50, 100
  - amwg = 20, 50, 100
  - aprime_diags = 50, 100

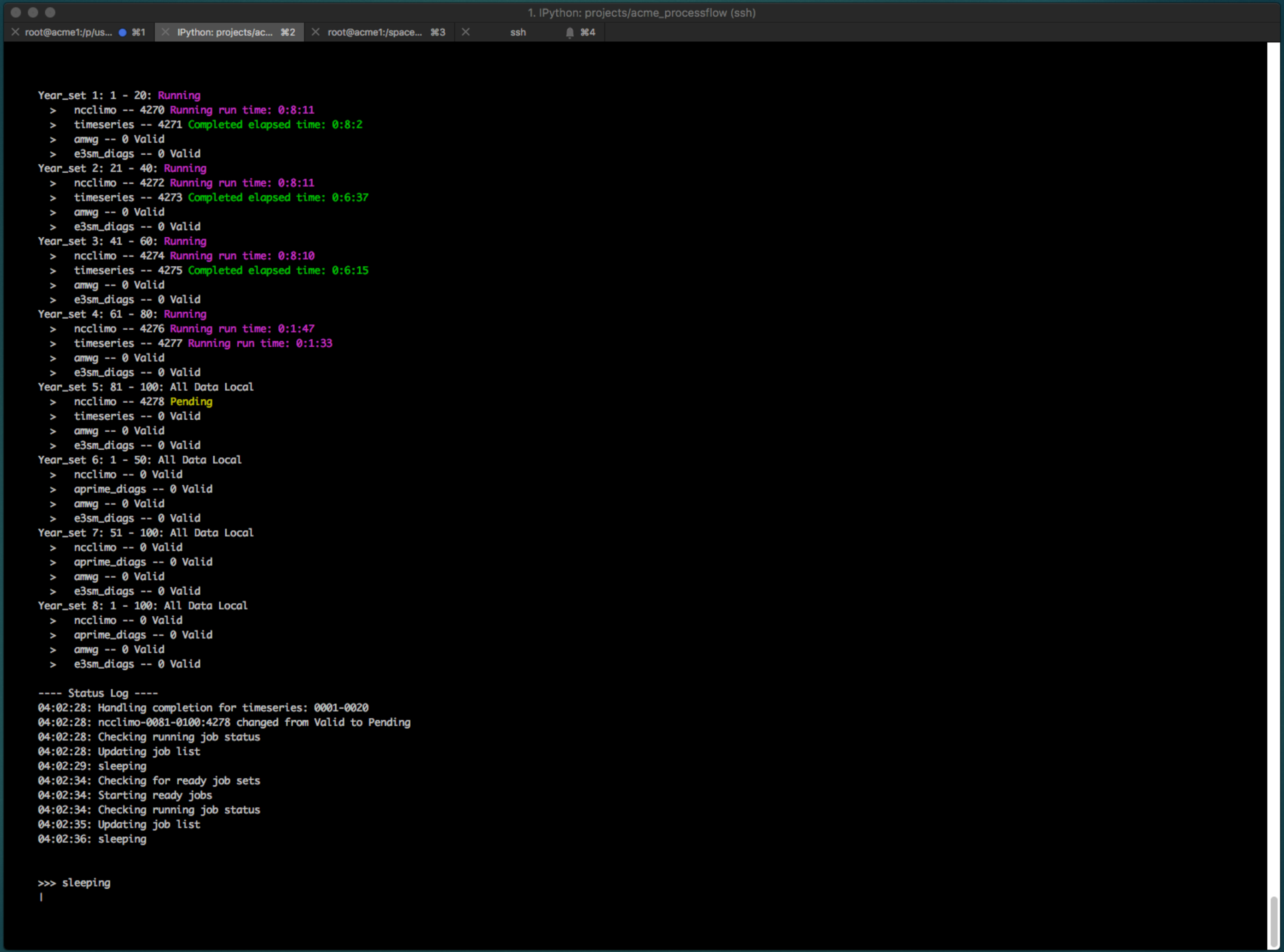

1. IPython: projects/acme_processflow (ssh)
root@acme1:/p/user_... ⌘1
IPython: projects/ac... ⌘2
root@acme1:/space... ⌘3
ssh
⌘4

```
Year_set 1: 1 - 20: All Data Local
  >  ncclimo -- 4270 Pending
  >  timeseries -- 4271 Pending
  >  amwg -- 0 Valid
  >  e3sm_diags -- 0 Valid
Year_set 2: 21 - 40: All Data Local
  >  ncclimo -- 4272 Pending
  >  timeseries -- 4273 Pending
  >  amwg -- 0 Valid
  >  e3sm_diags -- 0 Valid
Year_set 3: 41 - 60: All Data Local
  >  ncclimo -- 4274 Pending
  >  timeseries -- 0 Valid
  >  amwg -- 0 Valid
  >  e3sm_diags -- 0 Valid
Year_set 4: 61 - 80: All Data Local
  >  ncclimo -- 0 Valid
  >  timeseries -- 0 Valid
  >  amwg -- 0 Valid
  >  e3sm_diags -- 0 Valid
Year_set 5: 81 - 100: All Data Local
  >  ncclimo -- 0 Valid
  >  timeseries -- 0 Valid
  >  amwg -- 0 Valid
  >  e3sm_diags -- 0 Valid
Year_set 6: 1 - 50: All Data Local
  >  ncclimo -- 0 Valid
  >  aprime_diags -- 0 Valid
  >  amwg -- 0 Valid
  >  e3sm_diags -- 0 Valid
Year_set 7: 51 - 100: All Data Local
  >  ncclimo -- 0 Valid
  >  aprime_diags -- 0 Valid
  >  amwg -- 0 Valid
  >  e3sm_diags -- 0 Valid
Year_set 8: 1 - 100: All Data Local
  >  ncclimo -- 0 Valid
  >  aprime_diags -- 0 Valid
  >  amwg -- 0 Valid
  >  e3sm_diags -- 0 Valid

---- Status Log ----
03:51:12: Submitting to queue timeseries: 0021-0040
03:51:12: Updating job list
03:51:12: timeseries-0021-0040:4273 changed from Valid to Pending
03:51:13: Submitting ncclimo-0041-0060
03:51:13: Submitting to queue ncclimo: 0041-0060
03:51:13: Updating job list
03:51:13: ncclimo-0041-0060:4274 changed from Valid to Pending
03:51:13: Checking running job status
03:51:13: Updating job list
03:51:14: sleeping


>>> sleeping
\
```



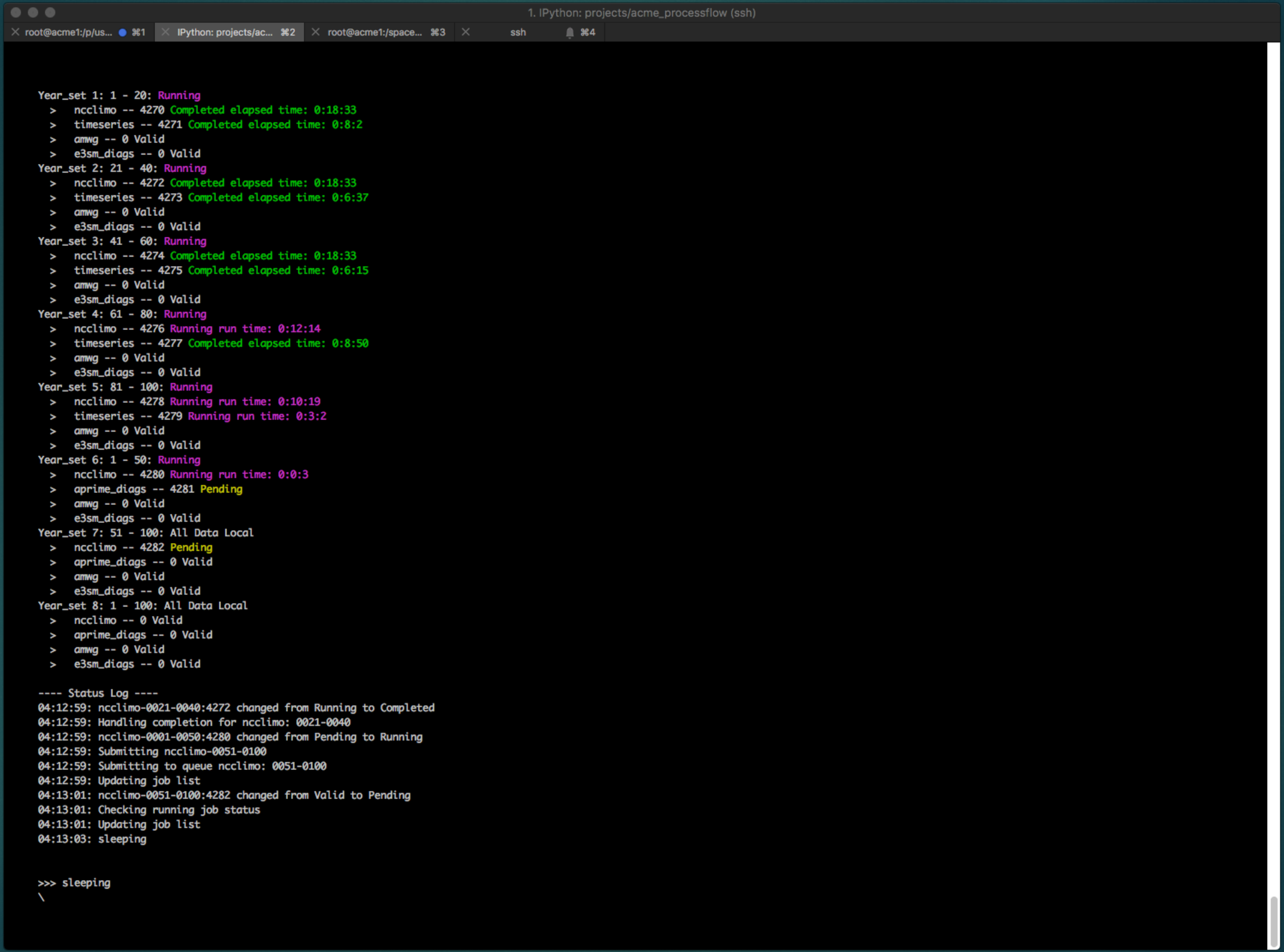
1. IPython: projects/acme_processflow (ssh)
root@acme1:/p/user_... ⌘1
IPython: projects/ac... ⌘2
root@acme1:/space... ⌘3
ssh
⌘4

```
Year_set 1: 1 - 20: Running
  >   ncclimo -- 4270 Running run time: 0:0:8
  >   timeseries -- 4271 Running run time: 0:0:8
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 2: 21 - 40: Running
  >   ncclimo -- 4272 Running run time: 0:0:8
  >   timeseries -- 4273 Running run time: 0:0:8
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 3: 41 - 60: Running
  >   ncclimo -- 4274 Running run time: 0:0:8
  >   timeseries -- 4275 Running run time: 0:0:8
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 4: 61 - 80: All Data Local
  >   ncclimo -- 0 Valid
  >   timeseries -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 5: 81 - 100: All Data Local
  >   ncclimo -- 0 Valid
  >   timeseries -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 6: 1 - 50: All Data Local
  >   ncclimo -- 0 Valid
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 7: 51 - 100: All Data Local
  >   ncclimo -- 0 Valid
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 8: 1 - 100: All Data Local
  >   ncclimo -- 0 Valid
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid

---- Status Log ----
03:54:25: ncclimo-0021-0040:4272 changed from Pending to Running
03:54:25: timeseries-0021-0040:4273 changed from Pending to Running
03:54:25: ncclimo-0041-0060:4274 changed from Pending to Running
03:54:25: timeseries-0041-0060:4275 changed from Pending to Running
03:54:26: sleeping
03:54:31: Checking for ready job sets
03:54:31: Starting ready jobs
03:54:31: Checking running job status
03:54:32: Updating job list
03:54:33: sleeping


>>> sleeping
|
```

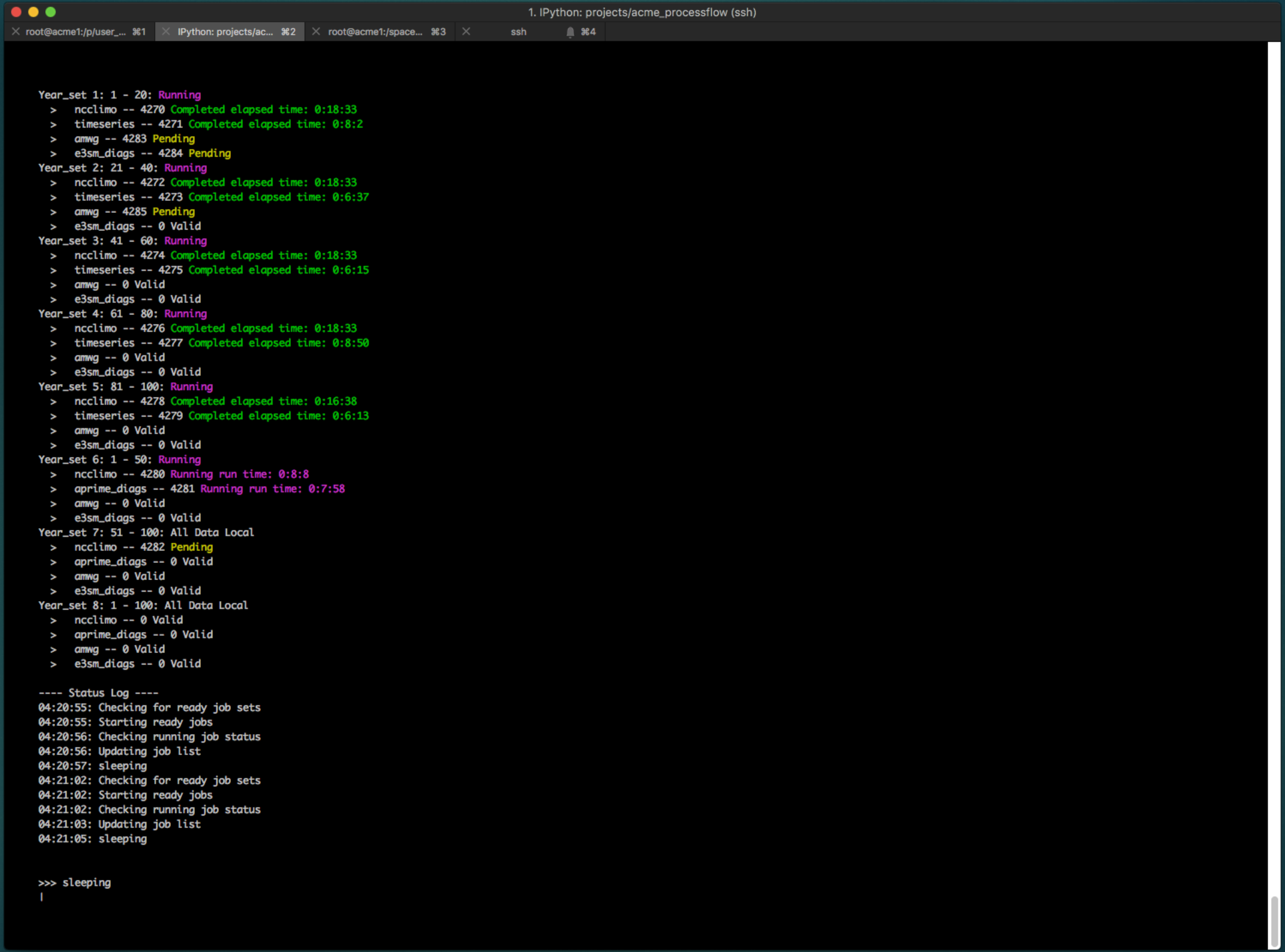


1. IPython: projects/acme_processflow (ssh)
root@acme1:/p/us... ⌘1
IPython: projects/ac... ⌘2
root@acme1:/space... ⌘3
ssh ⌘4

```
Year_set 1: 1 - 20: Running
  >   ncclimo -- 4270 Running run time: 0:8:11
  >   timeseries -- 4271 Completed elapsed time: 0:8:2
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 2: 21 - 40: Running
  >   ncclimo -- 4272 Running run time: 0:8:11
  >   timeseries -- 4273 Completed elapsed time: 0:6:37
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 3: 41 - 60: Running
  >   ncclimo -- 4274 Running run time: 0:8:10
  >   timeseries -- 4275 Completed elapsed time: 0:6:15
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 4: 61 - 80: Running
  >   ncclimo -- 4276 Running run time: 0:1:47
  >   timeseries -- 4277 Running run time: 0:1:33
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 5: 81 - 100: All Data Local
  >   ncclimo -- 4278 Pending
  >   timeseries -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 6: 1 - 50: All Data Local
  >   ncclimo -- 0 Valid
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 7: 51 - 100: All Data Local
  >   ncclimo -- 0 Valid
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 8: 1 - 100: All Data Local
  >   ncclimo -- 0 Valid
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid

---- Status Log ----
04:02:28: Handling completion for timeseries: 0001-0020
04:02:28: ncclimo-0081-0100:4278 changed from Valid to Pending
04:02:28: Checking running job status
04:02:28: Updating job list
04:02:29: sleeping
04:02:34: Checking for ready job sets
04:02:34: Starting ready jobs
04:02:34: Checking running job status
04:02:35: Updating job list
04:02:36: sleeping


>>> sleeping
```


1. IPython: projects/acme_processflow (ssh)
root@acme1:/p/us... ⌘1
IPython: projects/ac... ⌘2
root@acme1:/space... ⌘3
ssh
⌘4

```
Year_set 1: 1 - 20: Running
  >   ncclimo -- 4270 Completed elapsed time: 0:18:33
  >   timeseries -- 4271 Completed elapsed time: 0:8:2
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 2: 21 - 40: Running
  >   ncclimo -- 4272 Completed elapsed time: 0:18:33
  >   timeseries -- 4273 Completed elapsed time: 0:6:37
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 3: 41 - 60: Running
  >   ncclimo -- 4274 Completed elapsed time: 0:18:33
  >   timeseries -- 4275 Completed elapsed time: 0:6:15
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 4: 61 - 80: Running
  >   ncclimo -- 4276 Running run time: 0:12:14
  >   timeseries -- 4277 Completed elapsed time: 0:8:50
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 5: 81 - 100: Running
  >   ncclimo -- 4278 Running run time: 0:10:19
  >   timeseries -- 4279 Running run time: 0:3:2
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 6: 1 - 50: Running
  >   ncclimo -- 4280 Running run time: 0:0:3
  >   aprime_diags -- 4281 Pending
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 7: 51 - 100: All Data Local
  >   ncclimo -- 4282 Pending
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 8: 1 - 100: All Data Local
  >   ncclimo -- 0 Valid
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid

---- Status Log ----
04:12:59: ncclimo-0021-0040:4272 changed from Running to Completed
04:12:59: Handling completion for ncclimo: 0021-0040
04:12:59: ncclimo-0001-0050:4280 changed from Pending to Running
04:12:59: Submitting ncclimo-0051-0100
04:12:59: Submitting to queue ncclimo: 0051-0100
04:12:59: Updating job list
04:13:01: ncclimo-0051-0100:4282 changed from Valid to Pending
04:13:01: Checking running job status
04:13:01: Updating job list
04:13:03: sleeping


>>> sleeping
\
```


1. IPython: projects/acme_processflow (ssh)
root@acme1:/p/user_... ⌘1
IPython: projects/ac... ⌘2
root@acme1:/space... ⌘3
ssh
⌘4

```
Year_set 1: 1 - 20: Running
  >   ncclimo -- 4270 Completed elapsed time: 0:18:33
  >   timeseries -- 4271 Completed elapsed time: 0:8:2
  >   amwg -- 4283 Pending
  >   e3sm_diags -- 4284 Pending
Year_set 2: 21 - 40: Running
  >   ncclimo -- 4272 Completed elapsed time: 0:18:33
  >   timeseries -- 4273 Completed elapsed time: 0:6:37
  >   amwg -- 4285 Pending
  >   e3sm_diags -- 0 Valid
Year_set 3: 41 - 60: Running
  >   ncclimo -- 4274 Completed elapsed time: 0:18:33
  >   timeseries -- 4275 Completed elapsed time: 0:6:15
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 4: 61 - 80: Running
  >   ncclimo -- 4276 Completed elapsed time: 0:18:33
  >   timeseries -- 4277 Completed elapsed time: 0:8:50
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 5: 81 - 100: Running
  >   ncclimo -- 4278 Completed elapsed time: 0:16:38
  >   timeseries -- 4279 Completed elapsed time: 0:6:13
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 6: 1 - 50: Running
  >   ncclimo -- 4280 Running run time: 0:8:8
  >   aprime_diags -- 4281 Running run time: 0:7:58
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 7: 51 - 100: All Data Local
  >   ncclimo -- 4282 Pending
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid
Year_set 8: 1 - 100: All Data Local
  >   ncclimo -- 0 Valid
  >   aprime_diags -- 0 Valid
  >   amwg -- 0 Valid
  >   e3sm_diags -- 0 Valid

---- Status Log ----
04:20:55: Checking for ready job sets
04:20:55: Starting ready jobs
04:20:56: Checking running job status
04:20:56: Updating job list
04:20:57: sleeping
04:21:02: Checking for ready job sets
04:21:02: Starting ready jobs
04:21:02: Checking running job status
04:21:03: Updating job list
04:21:05: sleeping


>>> sleeping
|
```

Email Notification

```
YearSet 1-50: SetStatus.COMPLETED
   > ncclimo - COMPLETED :: /p/user_pub/e3sm/baldwin32/ACME_simulations/20180129.DECKv1b_piControl.ne30_oEC.edison_2/output/pp/fv129x256/climo/50yr
   > aprime_diags - COMPLETED :: https://acme-viewer.llnl.gov/baldwin32/20180129.DECKv1b_piControl.ne30_oEC.edison/a-prime/0001-0050
   > amwg - COMPLETED :: https://acme-viewer.llnl.gov/baldwin32/20180129.DECKv1b_piControl.ne30_oEC.edison/amwg/0001-0050
   > e3sm_diags - COMPLETED :: https://acme-viewer.llnl.gov/baldwin32/20180129.DECKv1b_piControl.ne30_oEC.edison/e3sm-diags/0001-0050


YearSet 51-100: SetStatus.COMPLETED
   > ncclimo - COMPLETED :: /p/user_pub/e3sm/baldwin32/ACME_simulations/20180129.DECKv1b_piControl.ne30_oEC.edison_2/output/pp/fv129x256/climo/50yr
   > aprime_diags - COMPLETED :: https://acme-viewer.llnl.gov/baldwin32/20180129.DECKv1b_piControl.ne30_oEC.edison/a-prime/0051-0100
   > amwg - COMPLETED :: https://acme-viewer.llnl.gov/baldwin32/20180129.DECKv1b_piControl.ne30_oEC.edison/amwg/0051-0100
   > e3sm_diags - COMPLETED :: https://acme-viewer.llnl.gov/baldwin32/20180129.DECKv1b_piControl.ne30_oEC.edison/e3sm-diags/0051-0100


YearSet 1-100: SetStatus.COMPLETED
   > ncclimo - COMPLETED :: /p/user_pub/e3sm/baldwin32/ACME_simulations/20180129.DECKv1b_piControl.ne30_oEC.edison_2/output/pp/fv129x256/climo/100yr
   > aprime_diags - COMPLETED :: https://acme-viewer.llnl.gov/baldwin32/20180129.DECKv1b_piControl.ne30_oEC.edison/a-prime/0001-0100
   > amwg - COMPLETED :: https://acme-viewer.llnl.gov/baldwin32/20180129.DECKv1b_piControl.ne30_oEC.edison/amwg/0001-0100
   > e3sm_diags - COMPLETED :: https://acme-viewer.llnl.gov/baldwin32/20180129.DECKv1b_piControl.ne30_oEC.edison/e3sm-diags/0001-0100
```

Key Features

- Run while the model is running
  - Start the processflow while the model is running. Data will be moved, and jobs will start when they’re ready
- Diagnostic output hosted when jobs finish
- Restarting the processflow wont rerun completed jobs
- Don’t need to learn ins and out of every sub-tool

Future Work

- Additional diagnostic suits (LMWG, ilamb)
- More configurability (file types, job config)
- Long term archive
- CMORization and ESGF publication

Questions?