Downscaling file descriptions, directory structure, and data access

Stefan Rahimi & Lei Huang (updated 18 January 2022)
Description website: https://dept.atmos.ucla.edu/alexhall/downscaling-cmip6
On AWS: https://registry.opendata.aws/wrf-cmip6, s3://wrf-cmip6-noversioning/

This document describes the directory structure inside the container hosting the downscaling data and the files therein on the storage medium. All files are in NetCDF-4 format.

1. Data tiers

Each of the three tiers of data contains output from all grids: the 45- 9-, and 3-km California and Wyoming grids. Users of the data may wish to use data of different tiers depending on their research interest or analysis objectives (see Figure below). Only a subset of GCMs have been downscaled to the highest resolution (3-km), and all GCMs are downscaled to 9-km.

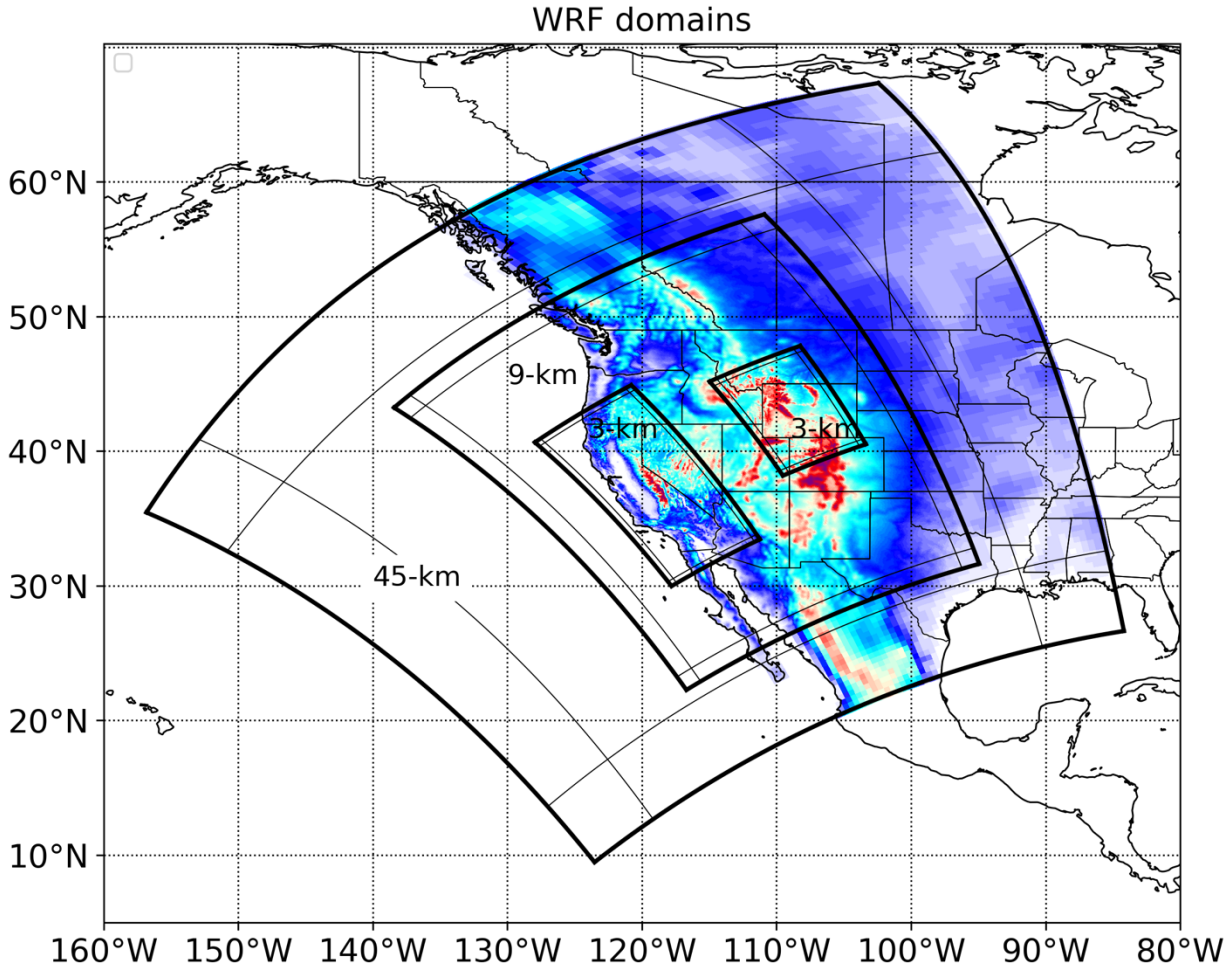

Figure 1: Domain setup of WRF simulations with a colorfill representing the complex topography across the region.

Tier 1
Tier 1 data files contain a datastream of 200+ WRF variables at 6-hour intervals. This includes variables on 39 vertical levels. Tier 1 files can be used to analyze, and process variables of interest not provided in Tier 2 and Tier 3 files. These files may also be used in conjunction with ndown.exe to dynamically downscale across other regions of western North America, bypassing the need to rerun the coarser resolution experiments or preprocess the GCM input. These files are date and time specific, with the following naming convention:

wrfout_d01_<yyyy>-<mm>-<dd>_<hh>:00:00,

where *yyyy*, *mm*, *dd*, and *hh*, are the 4-digit year, the 2-digit month, the 2-digit day, and 2-digit hour, respectively.

File sizes for each 6-hourly file are approximately 22, 140, 170, and 90 Mb for the 45-, 9-, 3-km CA, and 3-km WY domains, respectively.

Note that all variables in this tier are in their WRF-native form.

Tier 2
Tier 2 data files contain an auxiliary WRF datastream which contains 21 variables at 1-hour intervals. These variables are:

Name	**Units**
1. 2-m temperature	[K]
2. 2-m specific humidity	[kg kg^{-1}]
3. Surface pressure	[Pa]
4. 10-m u-component of the wind (grid relative)	[m s^{-1}]
5. 10-m v-component of the wind (grid relative)	[m s^{-1}]
6. Snow water equivalent	[mm]
7. Skin temperature	[K]
8. Non-convective precipitation (cumulative)	[mm]
9. Convective precipitation (cumulative)	[mm]
10. Cumulative snowfall equivalent	[mm]
11. Diffuse downwelled solar radiation	[W m^{-2}]
12. Surface upwelled solar radiation (all sky)	[W m^{-2}]
13. Surface upwelled solar radiation (clear sky)	[W m^{-2}]
14. Surface downwelled solar radiation (all sky)	[W m^{-2}]
15. Surface downwelled solar radiation (clear sky)	[W m^{-2}]
16. Surface upwelled longwave radiation (all sky)	[W m^{-2}]
17. Surface upwelled longwave radiation (clear sky)	[W m^{-2}]
18. Surface downwelled longwave radiation (all sky)	[W m^{-2}]
19. Surface downwelled longwave radiation (clear sky)	[W m^{-2}]
20. Surface runoff	[mm s^{-1}]

21. Sub-surface runoff [mm s^{-1}]

As for Tier 1 data, these files are date and time specific, with the following naming convention:

auxhist_d01_<yyyy>-<mm>-<dd>_<hh>:00:00,

where *yyyy*, *mm*, *dd*, and *hh*, are the 4-digit year, the 2-digit month, the 2-digit day, and 2-digit hour, respectively. Note that all variables in this tier are in their WRF-native form.

Tier 3

Tier 3 data files contain either daily averaged, daily maximum, or daily minimum quantities for 37 variables. These variables are:

Name	Units	Label
1. 2-m average temperature	[K]	't2'
2. 2-m minimum temperature	[K]	't2min
3. 2-m maximum temperature	[K]	't2max'
4. Maximum hourly precipitation	[mm h^{-1}]	'prec_max'
5. 2-m specific humidity	[kg kg^{-1}]	'q2'
6. Maximum 10-m wind speed	[m s^{-1}]	'wspd10max'
7. Snow water equivalent	[mm	'snow'
8. Precipitation rate	[mm d^{-1}]	'prec'
9. Snow precipitation rate	[mm d^{-1}]	'prec_snow'
10. Relative humidity	[0-100]	'rh'
11. Integrated vapor transport (zonal and meridional components; earth relative)	[kg s^{-1} m^{-1}]	'ivt'
12. Ice water path	[kg m^{-2}]	'iwp'
13. Liquid water path	[kg m^{-2}]	'lwp'
14. Soil moisture	[m^3 m^{-3}]	'soil_m'
15. Soil temperature	[K]	'soil_t'
16. Skin temperature	[K]	'tskin'
17. Surface pressure	[Pa]	'psfc
18. Surface runoff	[mm d^{-1}]	'sfc_runoff'
19. Sub-surface runoff	[mm d^{-1}]	'subsfc_runoff'
20. Evaporation	[mm d^{-1}]	'evap_sfc'
21. Evapotranspiration	[mm d^{-1}]	'etrans_sfc'
22. Downwelled SW at surface (> 0 into sfc)	[W m^{-2}]	'sw_dwn'
23. Downwelled LW at surface (> 0 into sfc)	[W m^{-2}]	'lw_dwn'
24. Net SW flux at the surface (> 0 into sfc)	[W m^{-2}]	'sfc_sfc'
25. Net LW flux at surface (> 0 into atm)	[W m^{-2}]	'lw_sfc'
26. Sensible heat flux at surface (> 0 into atn)	[W m^{-2}]	'sh_sfc'
27. Latent heat flux at surface (> 0 into atm)	[W m^{-2}]	'lh_sfc'
28. Ground heat flux at surface (> into atm)	[W m^{-2}]	'gh_sfc'
29. 3-D q	[kg kg^{-1}]	'q_3d

30. 3-D w	[m s^{-1}]	'w_3d
31. 10-m u, v (earth relative)	[m s^{-1}]	'uv10'
32. 3-D u (earth relative)	[m s^{-1}]	'u_3d'
33. 3-D v (earth relative)	[m s^{-1}]	'v_3d'
34. 3-D geopotential height	[m^2 s^{-2}]	'phi_3d'
35. 3-D temperature	[K]	't_3d'
36. Convective precipitation*	[mm d^{-1}]	'prec_c'
37. Mean 10-m wind speed	[m s^{-1}]	'wspd10mean'

(*) denotes that the variable is available only for d01 and d02.

Tier 3 files are labeled by variable, GCM, experimental ID (e.g. 'hist' for historical or 'ssp370' for SSP3-7.0), variant, domain, and year in the file name:

<variable>.daily.<gcm>_<exp_id>_<variant>_<domain>_<year>.nc,

For example, wspd10max.daily.mpi-esm1-2-lr_ssp370_r7i1p1f1_d02_2092.nc.

The domain naming convention (d0*) is as follows:

- Domain 1 – d01: The 45-km domain
- Domain 2 – d02: The 9-km domain
- Domain 3 – d03: The 3-km California domain
- Domain 4 – d04: The 3-km Wyoming domain

An additional string may appear in the Tier 3 data file names, 'BIAS_CORRECT'. This designates GCMs that have been bias corrected to ERA5 prior to dynamical downscaling.

Tier 3 files are much smaller than Tier 1 and Tier 2 files, and thus are great for analyses.

The 3-D variables are interpolated to the 1000, 925, 850, 800, 700, 600, 500, 400, 300, and 250 hPa isobaric surfaces.

The soil fields are defined on the 4 Noah-MP soil levels of 5, 25, 70, and 150 centimeters.

Metadata
Invariant fields are provided in grid-specific *wrfinput_<domain>* files as well as the Tier 1 files. These files are located in:
/wrf-cmip6-noversioning/downscaled_products/wrf_coordinates/

2. Directory structure

Each downscaled dataset follows an identical directory structure, which begins at the top directory whose name is that of the dataset being dynamically downscaled. For

instance, the ERA5 downscaled dataset's parent directory name is "era5" in /reanalysis subfolder of /downscaled_products. The datasets are still being generated, so these directory names are not listed in this document at the current time. Please see https://dept.atmos.ucla.edu/alexhall/downscaling-cmip6
for a list of ongoing and completed dynamically downscaled products.

Within each dataset's directory lives 4 subdirectories: 6hourly, hourly, postprocessed, and spin_up that contain the Tier 1, Tier 2, Tier 3, and "spin-up" files, respectively. Spin-up directories contain the full months' spin-up data used for each year-long simulation. Within the 6hourly, hourly, spin_up directories, each year is a subdirectory, inside which lives domain directories containing the files therein. The overall directory structure is given as:

/<data_name>/<file_type>/<year>/<domain>,

Where <data_name> is the label of the product being downscaled, <file_type> is either 6hourly, hourly, or spin_up, <year> is the 4-digit year, and <domain> is d01, d02, d03, or d04, which are the domain number labels. Recall,

- Domain 1 – d01: The 45-km domain
- Domain 2 – d02: The 9-km domain
- Domain 3 – d03: The 3-km California domain
- Domain 4 – d04: The 3-km Wyoming domain

The postprocessed files are all contained in a single directory within /<data_name>/postprocessed/<domain> given their relatively small size.

Of note, each simulated year is conducted independently of other years. Files contained in the 6hourly and hourly directories span the time range from 1 September through 1 September of the following year. Consider a non-leap year in an hourly directory: 8,761 files are stored within. For fiscal year 1988 (directory /1988/<domain>), the first auxhist file has a time stamp of 1988-09-01_00, while the last file has a time stamp of 1989-09-01_00. However, we note that the directory for the next fiscal year 1989 also has a file time stamped 1989-09-01_00. The question is begged, "Which file should we use?" The short answer is, "it depends".

Suppose you are interested in computing hourly precipitation rates in [mm/h]. Since RAINNC and RAINC are output cumulatively, you would need to apply a running difference in time at each grid point to get hourly precipitation increments. Since we have 8,761 files in /1988, you can compute the precipitation rate for 8,760 time slices in fiscal year 1988 within a single directory. However, if you are interested in hourly precipitation events over a 5-year period, say, fiscal years 1988-1993, you would compute the hourly rates within each yearly directory separately from one another, and concatenate the results of multiple years. This is an artifact of how we output precipitation in WRF (and a few other variables), for which we just let precipitation accumulate in the output. To conclude, if you compare the 1989-09-01_00 file in /1988

with the same file in /1989, you will see a much lower precipitation value in the /1989 file because it has only run for 1 month as opposed to 13 months.

Besides the cumulative variables (RAINNC, RAINC, SNOWNC, etc.), feel free to disregard the final 09-01_00 file in each yearly subdirectory in analysis.

3. Data access

There are two ways to access the data. Because it is open source and, in our experience, the fastest method, we only discuss the Amazon Web Services Command Line Interface (AWS CLI) method here.

The data are located in a publicly accessible bucket on AWS' Simple Storage Service (S3). Buckets are simply data storage centers that can be accessed to either download data or analyze the data using Python via Jupyter using Amazon Sagemaker.

Focusing on downloading using AWS CLI, make sure that your system has the appropriate prerequisites to properly download, unpack, install, and use AWS CLI (see official documentation here). Begin the process by downloading AWS CLI version II for Linux and creating a directory in your home space, we call it /Awscli:

```
Mkdir Awscli

curl "https://awscli.amazonaws.com/awscli-exe-linux-x86_64.zip"
-o "awscliv2.zip"

unzip awscliv2.zip

mkdir bin
mkdir aws-cli
```

You will see both the bin and aws-cli directories within your Awscli directory.

Now, it is time to install the software, which will give you access to the "aws" command that allows you to transfer, peruse, inventory, and manage data on S3. Within Awscli, install aws by typing:

```
 ./aws/install -i <absolute path to aws-cli>/aws-cli -b
<absolute path to bin>/bin
```

Within Awscli, you will now be able to run:

```
bin/aws –version
     aws-cli/2.1.6 Python/3.7.3 Linux/4.12.14-95.51-default
     exe/x86_64.sles.12 prompt/off
```

Now it is time to configure your AWS CLI. In your home directory, you should see the hidden directory, /.aws. Add the following lines to the end of the file:

```
s3 =
 max_concurrent_requests = 4
 max_queue_size = 1000
```

The max_concurrent_requests is set to 4 (instead of the default 10) for our purposes because we are interested in transferring <domain> directories individually. We recommend transferring no more than 40 directories' worth of data at a given time. AWS S3 transfers are multithreaded; that is, file transfers are parallelized. This parallelization can quickly overwhelm the system if too many transfers are initiated.

Transfers

Use the aws sync and cp commands to transfer data from the wrf-cmip6-noversioning bucket to your local machine. For example, say I want to transfer the /mpi-esm1-2-lr_r7i1p1f1_ssp370/hourly/2059/d03 directory to my local machine. I would type:

```
/glade/u/home/srahimi/Awscli/bin/aws s3 cp s3://wrf-cmip6-noversioning/downscaled_products/gcm/mpi-esm1-2-lr_r7i1p1f1_ssp370/hourly/2059/d03/ d03 --recursive --no-sign-request
```

Note that I have used the absolute path to invoke my own aws command. The "--recursive" flag recursively copies all of the directory's contents into a newly created directory d03. Note that you do not need to create d03 a priori.