V1.0.0 11/15/20
CVDP-LE created by Adam Phillips and Clara Deser, with support from John Fasullo, Isla Simpson and Dave Schneider (NCAR/CGD/CAS).

Before attempting to run the Climate Variability Diagnostics Package for Large Ensembles (CVDP-LE) please read the following document. For more information about the CVDP-LE project please see the CVDP-LE website: http://www.cesm.ucar.edu/working_groups/CVC/cvdp-le

Required software

- NCL v6.2.0 or newer (v6.6.2 recommended)
- Image Magick
- Python 2.7.X or newer (if running in parallel; see the run_parallel option in Step 3 below.)

General notes

The CVDP-LE is almost completely written in NCL, but knowledge of NCL is not required. The input files must be on local disk or on OPeNDAP servers. The CVDP-LE creates plots that are based on the following variables: TREFHT (tas), TS (ts), PRECT (pr), PSL (psl), SNOWDP (snd), MOC (msftmz/msftmyz/stfmmc) and aice (siconc/sic). Not all variables need to be present for a model simulation to be analyzed.

The CVDP-LE operates on monthly time series files covering a global domain (observational or model-based), such as those found in the CMIP6 archive or those distributed by NCAR that contain one variable and a number of timesteps. *The CVDP-LE does not work with CESM history files.* The CVDP-LE expects the file names of input model and observational data to end in a specific format of "YYYYMM-YYYYMM.nc" denoting the start year and month and end year and month of the dataset. "YYYY" MUST have 4 digits. This naming format is used for files in the CMIP6 archive along with NCAR's CESM post-processed files. If your model data file names do not end in this manner it is suggested that you either rename the files or use soft links. The CVDP-LE does not read the time variable out of input netCDF files; it relies on the times specified in the file names and checks that the number of timesteps indicated in the file name matches the number of timesteps in the file.

Multiple files containing the period of study are completely acceptable. The CVDP-LE does not expect overlapping time slices of data to be present in an identified data directory. For instance a directory that contained these tas files would be acceptable:
modelA.tas.190001-191912.nc modelA.tas.192001-193912.nc
A directory that contained these TS files would not be acceptable:
modelA.tas.190001-191912.nc modelA.tas.190001-193912.nc modelA.tas.192001-193912.nc

Non-monthly (ex. daily, 6-hourly) time series files should not be kept in the same directory as the monthly data. If they are, make sure the syntax used in namelist explicitly excludes non-monthly files. It is generally preferable for each model run to have its own directory, but it is not a necessity. (See the examples under *Step 1: set the namelist* below.)

One should avoid model file names that start with the variable name followed by a period. (ex. ts.mod.198001-201212.nc) One can use soft links or rename the files for use in the CVDP-LE.

Instructions for running the CVDP-LE

There are 2 files that must be set up to run the package (namelist, driver.ncl). A 3rd file can be set up if one wishes to include observations in the analysis (namelist_obs). These three files must be in the same directory. The CVDP-LE codebase (ncl_scripts/*) can be located anywhere and can be pointed to using the driver.ncl option zp.

Step 1: set the namelist

The namelist file contains information about which set of model data you would like to pass in to the CVDP-LE. You may enter as many simulations as you would like, but only specify one simulation per row.

Within the namelist file each row should follow the following format:
model name (arbitrary) | generic path to files | analysis start year | analysis end year | ensemble ID

" | " is used as a delimiter. Wildcard syntax (“{}” or "*" for example) is allowed. One can test that the specified path works to (only) identify the needed files by doing a "ls $path" at the command line.

The ensemble ID entry should be used to assign a model to a specific ensemble, and should be formatted as $Ensemble_Number-$Ensemble_Name. (ex. 1-CESM2, 2-MIROC6). The Ensemble Number should start with the number one, and rise sequentially by 1 for each ensemble. The Ensemble Name is arbitrary, and will be used by the CVDP-LE to title ensemble mean plots. See namelist examples below.

namelist Example 1:
CCSM4 Control 600-699 | /project/md/b40.1850.track1.1deg.006/ | 600 | 699 | 1-CCSM4 Control
CCSM4 Control 700-799 | /project/md/b40.1850.track1.1deg.006/ | 700 | 799 | 1-CCSM4 Control
CCSM4 Control 800-899 | /project/md/b40.1850.track1.1deg.006/ | 800 | 899 | 1-CCSM4 Control
CCSM4 Control 900-999 | /project/md/b40.1850.track1.1deg.006/ | 900 | 999 | 1-CCSM4 Control
CCSM4 Hist #1 | /p/cmip5/historical/*/CCSM4/r1i1p1/ | 1950 | 2005 | 2-CCSM4 Hist
CCSM4 Hist #2 | /p/cmip5/historical/*/CCSM4/r2i1p1/ | 1950 | 2005 | 2-CCSM4 Hist
CCSM4 Hist #3 | /p/cmip5/historical/*/CCSM4/r3i1p1/ | 1950 | 2005 | 2-CCSM4 Hist

The first four listed models are 100yr segments of the CCSM4 control where all the necessary files are in the stated directory. The latter three listed models are CMIP5 CCSM4 historical runs. /*/ syntax is used to span multiple directories.

namelist Example 2:
CESM1 LE #21 | /project/y/CESM1-LENS/b.e11.B*.f09_g16.021.* | 1979 | 2013 | 1-CESM1 LENS
CESM1 LE #22 | /project/y/CESM1-LENS/b.e11.B*.f09_g16.022.* | 1979 | 2013 | 1-CESM1 LENS
CESM1 LE #23 | /project/y/CESM1-LENS/b.e11.B*.f09_g16.023.* | 1979 | 2013 | 1-CESM1 LENS
GFDL-CM21 #1 | /p/cmip5/{Amon,Omon,OImon}/GFDL-CM21/r1i1p1/ | 1985 | 2005 | 2-GFDL
GFDL-CM21 #2 | /p/cmip5/{Amon,Omon,OImon}/GFDL-CM21/r2i1p1/ | 1985 | 2005 | 2-GFDL
GFDL-CM21 #3 | /p/cmip5/{Amon,Omon,OImon}/GFDL-CM21/r3i1p1/ | 1985 | 2005 | 2-GFDL

The first three simulations listed are for CESM1 Large Ensemble members #21-23. Note that a partial file name is provided to distinguish each member's data from other data in the specified directory. The latter three simulations specified are GFDL CMIP5 historical simulations that are identified by using "{ }" syntax where multiple directory names are needed.

General namelist notes:
- The paths specified in namelist should not end with the syntax "YYYYMM-YYYYMM.nc".
- Each member of an ensemble is required to span the exact same number of years. (The years however do not need to be the same.)
- If a directory path is specified it should end with a "/"
- The CVDP-LE can only analyze complete years. You can read in a simulation that starts or ends in an incomplete year, but those years cannot be set as being analyzed.
- Any instance of "/*/" syntax will be replaced with variable names when the CVDP-LE parses the input namelist for the creation of the variable namelists. (ex. For the creation of the namelist_byvar/namelist_psl "/*/" syntax will be replaced with "/{psl,slp}/".)
- *Atmospheric* data on curvilinear grids (such as the CESM spectral element grid) cannot be read into the CVDP-LE and should be regridded before CVDP-LE input. One can use NCL's ESMF regridding tools to accomplish this. (See https://www.ncl.ucar.edu/Applications/ESMF.shtml)

If after trying various syntax the CVDP-LE is still unable to correctly identify the desired files, one can always set up a new directory and create soft links within the directory pointing to the model files.

Step 2: set the namelist_obs (optional)

The namelist_obs file contains information about which observational datasets (if any) are to be used. This file is only used if the driver.ncl option obs is set to "True".

The namelist_obs file is formatted as follows:
variable | observation name (arbitrary) | path to file(s) | analysis start year | analysis end year

Note that " | " is used as a delimiter. The paths specified in namelist_obs (contrary to those in namelist) should be as specific as possible. One dataset should be specified per row, but multiple datasets can be specified for each variable. If an observational dataset is not specified for a particular variable that variable should be left off of the namelist_obs file.

namelist_obs Example 1:
TS | HadISST | /project/cas/DATA/hadisst.187001-201312.nc | 1920 | 2011
PSL | 20thC_ReanV2 | /project/cas/DATA/prmsl.mon.mean.187101-201112.nc | 1920 | 2011
TREFHT | MLOST | /project/cas/DATA/mlost.v3.5.2.188001-201212.nc | 1920 | 2011
PRECT | GPCC | /project/cas/DATA/full_data_v6_precip_10.190101-201010.nc | 1920 | 2009
SNOWDP | UDel | ~/Data/snow_depth.195001-201012.nc | 1960 | 2005
MOC | Obs_MOC1 | ~/Data/moc.observed.198101-201212.nc | 1981 | 2012
aice_nh | NASA B v2 | /project/cas/DATA/seaice.nsidc.nasa.bv2.197011-201412.nc | 1971 | 2014
aice_sh | NASA Team | /project/cas/DATA/seaice.nsidc.nasa.team.197011-201412.nc | 1979 | 2012

Example #1 shows how to set an observational dataset for each of the 8 CVDP-LE variables. Note that the analysis period can be different for each observed dataset.

namelist_obs Example 2:
PSL | 20thC_ReanV2 | /project/cas/DATA/prmsl.mon.mean.187101-201112.nc | 1920 | 2011
TREFHT | MLOST | /project/cas/DATA/mlost.v3.5.2.188001-201212.nc | 1920 | 2011
TS | HadISST | /project/cas/DATA/hadisst.187001-201312.nc | 1920 | 2011
TS | ERSSTv3b | /project/cas/DATA/ersstv3b.185401-201312.nc | 1920 | 2012
TREFHT | HadCRUT3v | /project/cas/DATA/hadcrut3v.temps.185001-201105.nc | 1920 | 2010
TREFHT | MLOST | /project/cas/DATA/mlost.v3.5.2.188001-201212.nc | 1950 | 2012

Example #2 shows how to specify multiple datasets per variable, while not setting observational datasets for PRECT, SNOWDP, MOC, aice_nh or aice_sh. The ordering of the rows is completely irrelevant, and one can specify a dataset twice but specify different periods of analysis (as was done for rows 2 and 6).

General namelist_obs notes:
- The CVDP-LE can only analyze complete years. You can read in an observational dataset that starts or ends in an incomplete year, but those years cannot be set as being analyzed. (See the PRECT row in namelist_obs Example 1 above for an example.)
- Each dataset can start and end at different years, and the period of analysis can be different for each dataset.
(continued on next page)

- In order to get ENSO SST/TS/PSL composites (and the metrics tables) the start and end years must match between the specified SST, TS and PSL datasets. In order to get ENSO PR composites the start and end years must match between the specified SST and PRECT datasets. Similar requirements are present for the Atmospheric Mode Regressions for TAS, TS and PR.
- The specified TS dataset(s) are used for sea-surface temperatures.
- aice_nh / aice_sh refers to sea ice concentration in the northern / southern hemispheres. If one has an observational file that spans both hemispheres that file can be specified twice in both aice_nh and aice_sh rows.
- Due to compositing requirements the CVDP-LE will set each variable namelist to be the same length in terms of the number of datasets listed. The length will be equal to the number of model simulations plus the maximum number of observational datasets per variable. Thus, if three PSL observational datasets are set and one observational dataset is set for every other variable, then each variable namelist will have three observational rows in addition to the specified model simulation. In this scenario the CVDP-LE will attempt to fill in the (non-PSL) namelists with the first specified observational dataset. If an observational dataset is not provided for a variable, the CVDP-LE would set each observational row to missing.

Step 3: modify and run driver.ncl

---

driver.ncl is the driving script of the CVDP-LE. There are user-adjustable options located at the top of driver.ncl. Each option has comments on the right explaining the various settings. Once driver.ncl is set, one can start the CVDP-LE by entering "ncl driver.ncl" in the terminal window. The command can also be put into background mode and the output sent to a file:
"ncl driver.ncl >&! a.out &"

Notable driver.ncl options:
namelists_only – Set to “True” when running with a new model or observational dataset in namelist or namelist_obs. This will allow you to examine the variable namelists that the CVDP-LE set up (based on your namelist and optional namelist_obs files). Within each file in namelist_byvar/ you will find a path for each dataset. You can execute a "ls $path" to see if the set path(s) are correct. If the path is listed as "missing" the CVDP-LE is not finding the file. Check your namelist/namelist_obs settings and verify that the specified path syntax is correct.

run_parallel – When set to “True”, the CVDP-LE will run in parallel mode and submit multiple CVDP-LE calculation scripts at once. This option can significantly reduce the CVDP-LE run time. To use this option python needs to be installed on your local machine along with the subprocess, sys, time, and os modules (all common). The number of scripts that can run concurrently is set via the *max_num_tasks* option. Note that terminal output may get intermixed when running in parallel mode. When set to “False” the CVDP-LE will submit calculation scripts serially.

machine_casesen – Set this option = “True” if your filesystem is case sensitive. If your filesystem is case insensitive (as most Macs are) you can set this option to “False”. In developing the namelists the CVDP-LE searches for files by using specific syntax that may cause case insensitive systems to see a file twice, thus causing the CVDP-LE to error out.

create_graphics – Set this option = “True” if you would like the CVDP-LE to create netCDF files and to create graphics. Set to “False” if you would only like netCDF files to be created.

Step 4: Examining CVDP-LE output

Final CVDP-LE output is written to the directory specified via the outdir option in driver.ncl. The output is displayed via HTML files. Open a browser and point to $outdir/index.hmtl to see CVDP-LE output. If you set the driver.ncl option tar_output = “True” you will have to untar the file prior to viewing the contents.

Known limitations of the CVDP-LE

- Does not run on spectral element or any other curvilinear grid for atmospheric data. (Curvilinear data uses two-dimensional latitudes and longitudes.)
- Only certain variable names (within the .nc files) are accepted:

```
 TS     = (/"TS","ts","sst",”t_surf”,”skt”/)
 PSL    = (/"PSL","psl","slp","SLP","prmsl",”msl”,”slp_dyn”/)
 TREFHT = (/"TREFHT","tas","temp","air","temperature_anomaly",\
                  ”temperature”,”t2m”,”t_ref”,”T2”,”tempanomaly”/)
 PRECT  = (/"PRECC","PRECL","PRECT","pr","PPT","ppt","p","P","precip",\
                "PRECIP",”tp”,”prcp”,”prate”/)
 MOC = (/”MOC”,”msftmyz”,”msftmz”,”stfmmc”/)
 aice_nh = (/”aice_nh”,”aice”,”sic”,”SIC”,”CN”,”ice”,”icec”,”siconc”/)
 aice_sh = (/”aice_sh”,”aice”,”sic”,”SIC”,”CN”,”ice”,”icec”,”siconc”/)
```

 If you wish to read in a different variable name you can alter lines 68-79, 360-362 and 587-592 of ncl_scripts/functions.ncl as necessary.

- Minimum length of simulation/observational data required: 5 years
- If your input data has any NaN’s in it, the CVDP-LE may hang and not complete. If this happens check your input data for instances of NaN’s and replace them with the appropriate value (_FillValue or otherwise). This issue is NCL-related and does not have anything to do with the CVDP-LE.