CEMAC User Documentation: A-CURE

1 Python scripts

1.1 pp2nc_3hrly.py

1.1.1 Purpose

Extract and condense information from 3-hourly pp files belonging to the UKCA26AER perturbed parameter ensemble (PPE) set into netCDF format. Each netCDF file contains aerosol optical depth (AOD) at 550nm and column-integrated cloud droplet number concentration (CDNC) fields for all 235 PPE members for one day (8 timesteps per file) on a coarsened grid (N96 down to N48).

1.1.2 Usage

The script can be run from the command line as follows:

```
$ ./pp2nc_3hrly.py <ppRoot> <orogFile> <ncRef> <ncRoot> <startDate> <endDate>
```

where:

- `<ppRoot>` is the path (either relative to the current directory, or full) to the root directory containing the pp files. The expected file naming convention under this root directory is described further below.
- `<orogFile>` is the path (relative or full) to the ancilliary UM file containing the orgraphy data associated with the pp files (file typically called 'qrparm.orog')
- `<ncRef>` is the path (relative or full) to a reference netCDF file whose coordinate system is at the desired coarsened resolution (N48) onto which the original high-resolution (N96) data should be regridded.
- `<ncRoot>` is the path (relative or full) to the desired output directory.
- `<startDate>` is in the format YYYYMMDD and refers to the first day to be processed
- `<endDate>` is in the format YYYYMMDD and refers to the last day (inclusive) to be processed

It is also possible to see the above information in the terminal by typing:

```
$ ./pp2nc_3hrly.py --help
```

1.1.3 Dependencies

The script has been designed to run on the JASMIN analysis servers and/or LOTUS (which have access to the pp files that reside in the GASSP/UKCA group workspaces). Running the script will automatically choose the python2.7 interpreter, as it is this version of python that has access to all the scientific

packages (e.g. IRIS) installed on JASMIN.
For production runs, it is recommended to use LOTUS (batch computing) rather than run interactively on the JASMIN analysis servers, which can become slow if there are many users running interactively at the same time. To run on LOTUS, the following job submission script example can be modified as required (submit using "bsub < scriptName.sh"). Users should allow around 20-30 minutes of wall-clock time for each day of pp files to be processed.

```
#!/bin/bash
#BSUB -q short-seial
#BSUB -J pp2nc_3hrly
#BSUB -o pp2nc_3hrly.out
#BSUB -e pp2nc_3hrly.err
#BSUB -W 05:00
./pp2nc_3hrly.py /ppRoot/path /orogFile/path /ncRef/path /ncRoot/path ...
   ... 20080701 20080710
```

1.1.4 Output

Running the script will generate two netCDF files (one for AOD and one for CDN) per day of processed data (there are thus eight 3-hourly timesteps in each file). The files will be written to the directory specified by the user through the 'ncRoot' command-line argument. The naming convention of the files is '[aod550/cdn]_tebaa-tebiz_teafw_pbYYYYMMDD_N48.nc'. The first part refers to the main variable within the file, the next parts refer to the start, end and median job ids, the 'pb' part indicates 3-hourly model data, the date stamp refers to the date of the data within the file, and the 'N48' part references the grid resolution. A log file ('logfile.log') is also generated, which can be used to keep track of the script's progress during execution, as well as to check for any generated error/warning messages.

1.1.5 Further details

The script expects the input pp files to have the following directory structure/filename convention:

`/ppRoot/<jobid>/<jobid>a.pbYYYYDDMM.pp`

The main root directoy 'ppRoot' is provided as a command line argument. 'pb' indicates 3-hourly output files ('pm', 'pa' and 'pc' indicate monthly, daily and hourly output files).

The script expects 8 timesteps per day in each input pp file (3-hourly). However, it has also been designed to deal with the case where there are only 7 timesteps in the six AOD variables (modes). This is because it has been identified that some of the pp files for the first day of a calendar month are missing data for first AOD timestep (00:20). When this occurs, the 00:20 field is re-inserted into the output netCDF file and filled with missing values (NaN) so that the grid remains regular (as is necessary). As an example, there are 16 (out of 235) PPE members with a missing first timestep on 2008-07-01.

The 550nm AOD fields are calculated on the fine domain (N96) as the sum over the six 'modes', with the STASH codes: m01s02i500, m01s02i501, m01s02i502, m01s02i503, m01s02i504, m01s02i505. The column-integrated CDNC fields (i.e. CDN per m^2) are calculated on the fine domain by first multiplying

each CDNC value (STASH code m01s38i479) by its cell height and then summing over all cells within a given model column. The cell heights are currently obtained using IRIS's 'HybridHeightFactory' function, which takes orography data and the model sigma levels as input and gives the altitude bounds of each grid cell as output. However, it is planned to also implement Masaru's alternative approach of using the pressure and theta fields along with the hydrostatic equation; the method to use at execution could be chosen via an optional command line flag.

The fine domain data is then regridded onto the coarser (N48) domain using IRIS's 'regrid' function, where the coarse grid coordinates are extracted from the 'ncRef' file (supplied via a command-line argument). The 'Linear' regridding method is currently used. Discussion is ongoing as to whether this or the 'AreaWeighted' regrdding method is the most appropriate.

Efforts have been made to capture as many potential errors as possibe in a 'nice' manner (i.e. with descriptive erorr/warning messages written to the log file). These include:

- Checking that the paths given in the command-line arguments exist
- Checking that the date stamps given in the command-line arguments are in the expected format, with the start date preceding the end date.
- Checking that the data cubes can be loaded by IRIS (if this fails, it is most likely due to a missing/unexpected STASH codes)
- Checking that all the relevant data cube dimensions are as expected, i.e.:
    - The fine resolution (N96) fields have dimensions (time,lat,lon)=(8,145,192) (and 52 vertical levels in the case of CDNC)
    - The coarse resolution (N48) reference data file and orography file have dimensions (lat,lon)=(73,96)