OUTPUT DOWNSCALED Data Directory Structure and File Name Conventions for FUDGE darkchocolate and Beyond

For FUDGE darkchocolate versions and beyond, the directory structure and file name conventions to be applied to downscaled output will remain quite similar to what has been used in FUDGE cinnamon. Differences include:

- Elimination of the "NOAA-GFDL" and the ending downscaling version (formerly referred to as $dsVersion) sub-directories.
  We perceive no compelling reason to continue to include these sub-directories in our standard output directory structure. Eliminating it reduces path lengths by 19 characters. Essentially a vestigial carryover from the data publication of early NCPP work, it provides negligible value to our in-house day-to-day research efforts.
- A $maskedRegion sub-directory and inclusion of $maskedRegion in the file name are necessary.
  We do not need a $gridDomain sub-directory in output directory structure, because, according to our in-house conventions, all experiments for a given $SUBPROJECT will always use the same $gridDomain, but will not necessarily use the same spatial mask within that full $gridDomain.
- Use of a single XML element to specify a fairly long character string (a set of contiguous sub-directories) defining a segment of the downscaled output file path. Described in more detail below as
  "$dataSource/$epoch/$freq/$realm/$misc/$rip/$dataVersion"
  For FUDGE cinnamon, these were taken from the single future predictor variable file pathname. In darkchocolate and beyond, we may run experiments where multiple predictors are used; hence, FUDGE needs to know the location of single output directory, and an unambiguous way of accomplishing that is to pass this long string in from the XML.

==========

DOWNSCALED OUTPUT DIRECTORY FORMAT =
$outRoot/$SUBPROJECT/$dataType/$dataSource/$epoch/$freq/$realm/$misc/$rip/$dataVersion/$experimentName/$target/$maskedRegion/$dim/

$outRoot = root (or parent) directory path, below which downscaled output files generated by FUDGE are stored.
Will often be /archive/esd/PROJECTS/DOWNSCALING

$SUBPROJECT = the short name for the downscaling (sub)project of interest. Typically, an individual subproject will be associated with a focused research effort linked to a particular grant or manuscript -or- be associated with a particular grouping of test case results.
The $SUBPROJECT character string is to be set in the XML (e.g., "RR" for Red River subproject, "PM" of Perfect Model subproject, "S1" for the first synthetic data test case, etc.)

$dataType = will always have the value "downscaled" for downscaled output generated by FUDGE.
(A wider range of $dataType values are used for input data)

$dataSource/$epoch/$freq/$realm/$misc/$rip/$dataVersion = A contiguous set of character strings treated as a single unit in experGen and postProc processing. The meanings of the FUDGE downscaled output sub-directories are analogous to their meanings for FUDGE input files.

Currently, within expergen, this contiguous string is taken from the future predictor input file pathname … an assumption that works for cases in which the single input climate variable is the same as the downscaled output climate variable. However, this assumption will not hold for multivariable downscaling. For this reason, and inline with the general theme of moving the need for logic, table look-ups etc. from expergen and postProc up to XMLgen, this long character string will be included in the XML (perhaps with a tag such as "**$outSubdirs**").

The following info is reproduced the INPUT_FILE_PATHNAMES_darkchocolate documentation, and is relevant for interpreting this contiguous set of subdirectory character strings:

$dataSource = unique identifiers for the source of the data, re-using names from original sources when applicable.
For example, re-use CMIP5 model names such as MPI-ESM-LR, GFDL-HIRAM-C360, MIROC5 per conventions at http://cmip-pcmdi.llnl.gov/cmip5/availability.html. However, at times we may want to modify names to reflect GFDL in-house needs. And when reasonable established names do not exist, we will develop descriptive names, such as we did with "livneh", "prism" and "daymet" under OBS_DATA/GRIDDED_OBS.

$epoch = This sub-directory level provides a general description of the time period or epoch covered by the data sets below. Usually, different dataSources grouped under a single $dataType will have the same set of $epoch sub-directories.
Current examples include "amip", "historical", "future", "rcp26", "rcp45", "rcp85", & "sst2090". The sub-directory level can be re-purposed SYN_DATA cases for which the concept of a temporal epoch does not apply.

$freq = Patterned after GFDL Fre and CMIP styles, this identifies the time interval between successive data points in the time series files (aka sampling frequency).
Through cinnamon, we have only used "day". During 2015 we may have need for others, such as "mon", "1yr", and "ssn". Synthetic data designed to mimic certain temporal sampling frequencies could adopt the same directory names, while more generic synthetic data could re-purpose this level and adopt another name.

$realm = Patterned after GFDL Fre and CMIP conventions, this is generally intended to identify the component of the physical climate system represented by the data sets below.
Through cinnamon, we have only used "atmos". The most common 4 examples are: "atmos", "land", "ocn", "ice". It's possible that some other designation might become necessary, for example, in a case where the data is derived from two or more variables that are from different realms.

$misc = This sub-directory level will serve somewhat different purposes, affording us another degree of freedom to categorize our data.
In some cases, using a MIP table name could be reasonable. But often the data will not be directly associated with a MIP table. In those cases, $misc will be re-purposed. For example, data that has been processed into anomaly fields via use of a 15 day boxcar smoother could be given the value of "anom15SBX". Some indices computed from input data sets (e.g., GCMs or observations) perhaps could be grouped under "climdex" , etc.

$rip = A string that identifies ensemble members for different ways of generating multi-member ensembles. Generally in form "rKiMpN" where K,M,N are integers. The general convention is r=realization (varies for climate model ensemble members started from different initial conditions); i = initialization_method (we may encounter variations in this if we downscale seasonal forecasts); p = physics_version (varies in perturbed physics ensembles).
For CMIP data sets, the $rip identifier will be identical to that found in the CMIP archive. For non-CMIP data sets, we will adapt this to provide info meaningful to our applications. (e.g. We have use "r0i0p0" and "r0i0p1" as in-house conventions for *observation* data sets as a simple means to differentiate between time series which used Julian calendar (r0i0p0) vs NOLEAP calendar (r0i0p1).)

$dataVersion = version of the data.

$experimentName = A character string intended to provide a unique identifier for each FUDGE downscaling experiment, constructed in a manner that attempts to balance conciseness with info

about the configuration of the downscaling experiment. Adheres to experiment naming conventions described in http://cobweb.gfdl.noaa.gov/~esd/PDFs/FUDGE_exper_names.pdf
To date, $experimentName has been constructed within expergen using our in-house experiment naming conventions. Also, $experimentName is constructed in XMLgen. **We suggest we continue the approach of having the very important $experimentName string be constructed in both XMLgen and expergen, and that a check be included in expergen to confirm that the two match.**
(We also note that the $platform_ID within the experiment name has thus far been "p1" for all production jobs. However, this could vary at a later date – perhaps in 2015, though that is unclear. Currently, the platform_ID = "p1" is *not* passed as an individual element in the XML. Rather, it is derived from logic in expergen. whereas XMLgen ASSUMES "p1". For darkchocolate the current assumptions may suiffice. After darkchocolate, we will want to return to consider uses of $platform_IF in more detail.)

$target = long name of the downscaled output climate variable, and by convention always equal to the ***historical target*** variable for FUDGE downscaling methods likely to be considered in 2015.

$maskedRegion = character string used to name the subset of I,J grid points within the full geographic domain ("$gridDomain" in the input file directory specs) to be considered for downscaling. The specific points for which the R-code generates downscaled output is controlled via the setting of values (1.0 or missing_value) in the spatial masks input file.
The proposed name for XML element for the $maskedRegion character string = "**spat_mask_ID**". This will make obsolete the current practice in which $maskedRegion is based on the "project" element value in the XML and look-up tables within expergen. Examples of current and potential future $maskedRegion values include RR, US48, OKL, TEX, SWUS, etc.

$dim = the same $dim as for the ***historical target***. It identifies the sub-directory in which the input files are stored at the lowest level. **$dim** refers to the spatial dimension.

DOWNSCALED MINIFILE NAME IS DEFINED AS:

$target_$freq_$experimentName _$epoch_$rip_$maskedRegion _$trange.I$Islice_${file_j_range}.nc

---

where

$target = long name of the ***historical target*** variable

$experimentName, **$maskedRegion** (see above.)

$freq, **$epoch**, **$rip**, **& $trange, $Islice, ${file_j_range}** come from the ***future predictor*** input file values

e.g. /archive/esd/PROJECTS/DOWNSCALING/RR/downscaled/MPI-ESM-LR/rcp85/day/atmos/day/r1i1p1/v20111014/RRtxp1-CDFt-B38atL01K00/tasmax/RR/OneD/tasmax_day_RRtxp1-CDFt-B38atL01K00_rcp85_r1i1p1_RR_20060101-20991231.I240_J31-170.nc

---

DOWNSCALED LAT-LON (concatenated) FILE NAME IS DEFINED AS:

$target_$freq_$experimentName _$epoch_$rip_$maskedRegion _$trange.nc

---

and should be stored up one directory level up from the minifiles (directly under the $maskedRegion sub-directory.)

e.g. /archive/esd/PROJECTS/DOWNSCALING/RR/downscaled/MPI-ESM-LR/rcp85/day/atmos/day/r1i1p1/v20111014/RRtxp1-CDFt-B38atL01K00/tasmax/RR/tasmax_day_RRtxp1-CDFt-B38atL01K00_rcp85_r1i1p1_RR_20060101-20991231.nc