Implementing netCDF format output from the Unified Model

Jeff Cole
NCAS-CMS, Department of Meteorology, University of Reading

Computational Modelling Services
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

The Unified Model produces output in its own data format, most researchers prefer netCDF data which is a standard format, widely used by data processing and plotting software. Converting Unified Model data output into netCDF format as a post-processing task can be time consuming , especially with large resolution datasets. For example the Cascade project

- 80TB of PP data to be converted to netCDF
- NetCDF conversion rate - 10hr/TB on the HECToR NERC post processor
- 800hrs converting PP to netCDF
- Model data production 10 wall clock hrs/TB

Effectively the data conversion step doubles the time taken to produce the model results. Generating netCDF data directly from the Unified Model can therefore make a large improvement in the efficiency of a research project.

Another example is the SWAMMA (Sahara West African Monsoon Multiscale Analysis) project. Here a early version of the netCDF writing code was used and enabled a considerable time saving to be made. The various Unified Model runs for this project has produced 94 Terabytes of compressed netCDF data so far. Assuming a similar data conversion rate to that used above, forty additional days of computation time would have been needed if conversion to netCDF had been done as a post-processing step. For further details of the SWAMMA project see the poster by William McGinty.

A high resolution model with a large amount of STASH output was run on the Archer supercomputer using 768 processors (32 nodes). Both standard UM and netCDF output formats with varying levels of compression were written. The timings for the runs were split into three categories, the time to do STASH processing, the time to write the data and everything else. As the graph opposite shows the major factor in the total run time is the amount of time data writing takes, which in turn depends on the size of data written and the amount of time taken by the netCDF library to compress the data. The reasons for the large run time due to netCDF compression are, the high data resolution means a large amount of data needs to be compressed and due to the design of the Unified Model this data compression happens on only 1 processor and is hence a huge bottleneck. Later versions of the Unified Model (> 7.5) have implemented an IO server which totally decouples the task of writing STASH and start dump data from running the atmospheric model. Thus using asynchronous writing the run time can be brought down close to the run time without writing STASH output, at the cost of needing extra resources for the IO server, 1 extra node in this case. Therefore using a combination of UM netCDF output and an IO server, high resolution runs can have the advantage of not having to convert the data and apply appropriate levels of data compression. Thus significantly reducing post-processing time and disk storage requirements.

Objectives

- Directly write UM STASH output in netCDF format
- Implement the whole UMUI STASH panel
- Convert PP file metadata into CF metadata
- Support standard and mean PP files
- Support netCDF versions 3 and 4, including data compression
- Integrate into the UM code repository, so netCDF output will be available in all future UM releases

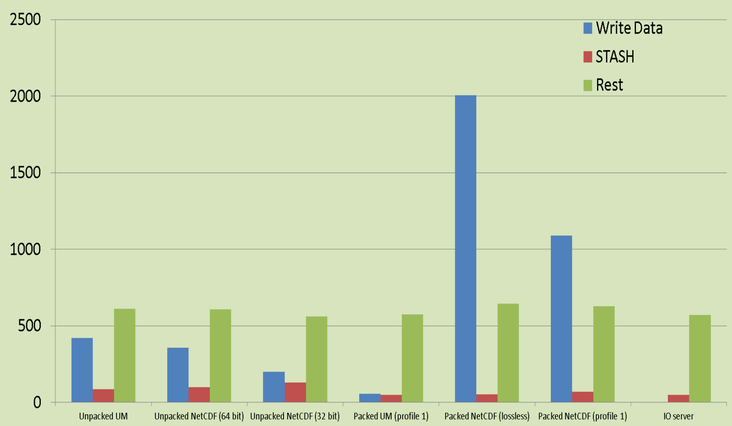
Coding decisions

- All netCDF library calls contained in one f90 module
- Minimize changes to existing UM code
- All model metadata is written when the netCDF files are first opened
- Use STASH domain and time profile names as vertical and time coordinate names
- Standard names and units are taken from file STASH_to_CF.txt, produced by David Hassell. Future plans are to replace this file with one generated using Metarelate (metarelate.net)

Work still to do

- Add support for STASH time-series output
- Add support for mean PP files
- Update code for latest UM release and incorporate into the UM code repository so all future UM releases include this feature

PS30 UKV, vn8.2, 744x928x70, 6 hour run (432 time steps), 200 diagnostics



Size of data in Gbytes,