High speed data access for spatial crop models running in the cloud

N. Herrmann [a] and **T. Erwin** [b]

[a] *CSIRO Agriculture and Food, Canberra, Australia*
[b] *CSIRO Information Management and Technology, Aspendale, Australia*
Email: neville.herrmann@csiro.au

Abstract: As biological models have improved, the acceptance of them for modelling large scale areas has meant that more computing resources have been allocated for their processing. Frequently these compute resources are found in cloud-based platforms that can host many CPU processes at minimal cost. Typically, many models require specific site information. As these systems scale up in size, it leads to a demand for a significant amount of input data. Distributing this data for use by each process instance can be a challenging task.

The use case that highlighted these issues was the generation of crop production data cubes for the Australian wheat belt using multiple future climate models. To run a crop model over a large area of Australia we used APSIM (Holzworth et al., 2018) hosted on an Azure compute cloud. This required climate data for around 49,000 sites for 45 climate scenarios. The total volume of climate data required was approximately 1TB. This data was made available through a web API that was optimized for fast access by many processes. During the execution period there were approximately 250 CPU's running concurrently each requesting their 500KB climate file.

The provision of climate data for many concurrent CPU processes where the source data is contained in NetCDF files, meant there needed to be a number of techniques to optimize the data retrieval. Several NetCDF data reorganizing or rechunking[1] methods were tested for data retrieval performance. An optimal one was found that improved the access for temporal spans of data. The climate model data found in the NetCDF files was then rechunked. The next stage was the extraction of the specific sites into APSIM climate format files. These files were then compressed into an archive. This process was replicated for all 45 climate scenarios using a high-performance multi-node computing infrastructure.

The web API was developed using a FastAPI (https://fastapi.tiangolo.com) server behind an Apache web server. The API allowed access to the compressed archive of climate files and also direct access to the NetCDF data. Requests for weather data returned files ready for use by APSIM. The improved performance gained by using the compressed data archive meant that the Azure cloud processes could get multiple returns per second, whereas the retrieval from the NetCDF files was significantly slower. Other alternatives such as retrieving data for a region from one API call were tested with some improvement in performance. One of the issues with data retrieval directly from the NetCDF storage was being restricted to a single threaded NetCDF library.

The success of this infrastructure has shown that there are viable alternatives to data provision for cloud processes other than purchasing cloud data storage. Cloud storage and transferring data usually involves costs. An alternative is to store the data locally and allow efficient remote access. This maintains flexibility for the data curator.

[1] https://www.unidata.ucar.edu/blogs/developer/entry/chunking_data_why_it_matters

REFERENCES

Holzworth, D.P., Huth, N.I., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N.I., Zheng, B.Y., Snow, V.O., 2018. APSIM Next Generation: overcoming challenges in modernising a farming systems model. Environ. Model. Softw. 103, 43–51.doi:10.1016/j.envsoft.2018.02.002

Keywords: *Climate change, APSIM, cloud computing*