Contribution ID: **12** Type: **Oral presentation**

Automated production of high value air quality forecasts with Pangeo, Papermill and Krontab

Tuesday, 15 October 2019 09:50 (20 minutes)

In many ways, a Jupyter notebook describes a data processing pipeline: you select some data at the top of the notebook, define reduction and analysis algorithms as the core of the notebook's content, and generate value –often in the form of plots or new insight –at the end of the notebook by applying algorithm to data. Value can be added to analysis and insight by including textual metadata throughout the notebook that describes the analysis applied and interpretation of the insight generated in the notebook.

It is a common requirement to want to apply the same processing pipeline, described by a Jupyter notebook, to multiple datasets. In the case of air quality forecasts, this might mean executing the same processing pipeline on all chemical species implicated in a particular air quality study.

In this talk we will present Pangeo as an open-source, highly customisable, scalable, cloud-first data processing platform. We will demonstrate using Pangeo to run a defined data processing pipeline in a Jupyter notebook, and move on to explore running this notebook multiple times on a range of input datasets using papermill. Finally we will demonstrate running the processing pipeline automatically to a schedule defined with krontab, a crontab-like job scheduling system for kubernetes.

Primary authors: KILLICK, Peter (Met Office Informatics Lab); ROBINSON, Niall (Met Office Informatics Lab); DONKERS, Kevin (Met Office)

Presenter: KILLICK, Peter (Met Office Informatics Lab)

Track Classification: Workshop: Building reproducible workflows for earth sciences