Housekeeping and Presentations, September 12, 2022, 9:30 AM - 11:00 AM

FAIR Research Software

Dr. Aleksandra Pawlik
Manaaki Whenua Landcare Research
PawlikA@landcareresearch.co.nz

ABSTRACT / INTRODUCTION (Up to 200 words)

Research outputs span beyond publications and reports. Research data and software are equally significant and ubiquitous outcomes of scientific work. In my talk I will look at FAIR principles for research software. FAIR is an acronym for "findability, accessibility, interoperability, and reuse of research". Whilst seemingly self-explanatory, these principles need to be well understood, underpinned by relevant infrastructure and best practices. I will talk about international efforts around defining and implementing FAIR for research software. I will cover selected examples of practical approaches towards making scientific software findable, accessible, interoperable and reusable.

ABOUT THE AUTHOR(S)

Aleksandra Pawlik
Aleksandra Pawlik is an eResearch Capability Specialist at Manaaki Whenua Landcare Research. For over 10 years she has been working with researchers helping them with the computational side of their research, from data management and analysis to improving their programming skills. She has helped develop Software and Data Carpentry training across several countries, and taught 60+ workshops which has been a very rewarding experience for her.

Interactive visualisation of modelled biological cells using a game engine

John Rugis[1], James Sneyd[2], David Yule[3]
[1,2]University of Auckland, [3]University of Rochester
[1]j.rugis@auckland.ac.nz

ABSTRACT / INTRODUCTION

Our research group is an interdisciplinary collaboration that combines real-world physical measurements with synthetic reconstructions, numerical modelling and 3D visualisation. The object under study in this demonstration is the collection of cells within mammalian parotid glands that are responsible for the initial creation of saliva. This process involves a complex chain of ion reactions and diffusion as well as time-varying fluid flow.

We have performed numerical simulations of this process within what we refer to as a "mini-gland" which is a collection of over 150 synthetic cells that closely match the measured structure in real glands. The simulation output data includes both the ion concentrations and fluid flow.

Our data visualisation goal was to clearly display the simulation results in 3D using the synthetic cell meshes. The visual density of the data precluded any predetermined or scripted point-of-view display. Our solution was to use a state-of-the-art game engine to provide dynamic user-driven interaction with the simulation results.

In this demonstration we will present both the visualisation development tools and the final stand-alone executable.

ABOUT THE AUTHORS

- John Rugis holds a PhD in Computer Science from the University of Auckland and currently works in the Department of Mathematics at the University of Auckland. John specialises in scientific data visualisation.
- James Sneyd is Professor of Mathematics at the University of Auckland.
- David Yule is a Professor at the University of Rochester Medical Center.

Category: Presentations

Using {golem} to create sustainable and reproducible apps in R

Richard Dean[1], Lillian Lu[1]
[1] Institute of Environmental Science and Research Limited (ESR)
richard.dean@esr.cri.nz , lillian.lu@esr.cri.nz

ABSTRACT / INTRODUCTION

Getting started with Shiny in Rstudio is easy – within four clicks from "new project" you can be running a wireframe application. However – from that point on, building and then maintaining a production-grade product can be daunting, with many different routes to achieve the same goal. This can make large scale R Shiny apps unsustainable and difficult to maintain. The {golem} framework is a proposed solution to provide structure and create a reproducible development workflow.
In this presentation, Richard and Lillian will explain the main principles behind the {golem} framework, and how this allows data scientists and non-web developers to develop and maintain robust, sustainable and reproducible R shiny applications.
While the presentation will use the example of ESR's public-facing SARS-CoV-2 wastewater dashboard, the basic principles should be widely applicable to other data scientists and research software engineers wishing to adopt a standard production process.

ABOUT THE AUTHOR(S)

Richard Dean is a senior data scientist at ESR. He works across the organisation on projects that gain insight from big data sets to tackle nitty gritty real-world problems affecting human communities in Aotearoa, covering everything from forensic science to human health, and the environment. He has a BSc in Information Systems Management from Durham University and wrote an MSc thesis on public health data interoperability standards while working at the Wolfson Research Institute for Health and Wellbeing in Durham.

Richard was the first member of staff at Public Health England to graduate from the UK government digital service 'data science accelerator' programme. In 2019, he brought the scheme to New Zealand through an internal accelerator programme within ESR which has since trained three cohorts of data scientists.
Lillian Lu is a data scientist at ESR, with a background in digital marketing and business development. She returned to university in 2019 and completed her master's degree in Analytics. While studying, Lillian worked for local and US companies as a data consultant and decided to kick off her data science journey at ESR after she graduated. Lillian's current work mainly focuses on public health and toxicology.

Tips and Tools for RSE's, September 12, 2022, 11:20 AM - 12:50 PM

What Flexi-HPC could mean for RSEs and DevOps

Jun Huh, Georgina Rae, Nick Jones, Blair Bethwaite
NeSI
jun.huh@nesi.org.nz, georgina.rae@nesi.org.nz, nick.jones@nesi.org.nz, blair.bethwaite@nesi.org.nz

ABSTRACT / INTRODUCTION (Up to 200 words)

NeSI's new Flexible HPC Platform provides programmable infrastructure for research software and data collaboration. From a platform capability perspective, Flexi HPC enables a cloud-native eResearch service development and hosting platform. It supports access to and hosting of emerging technologies, and a more inclusive and equitable approach to eResearch and HPC infrastructure for researchers across Aotearoa.

Early partners on the platform include an all-in-one eResearch environment with AgResearch, bare-metal hosting of advanced GPU capabilities for the Strong AI Lab out of University of Auckland, hosting of the Aotearoa Genomic Data Repository and eventually the Rakeiora Pathfinder interactive data analysis environment both of which have Māori Data Sovereignty requirements at their core.

In this lightning talk we want to hear your views on the service and how it might meet your needs.

ABOUT THE AUTHOR(S)

- Jun Huh is a product manager at NeSI and has focused on the data repository project over the past 2 years to bring new data management capabilities for Genomics Aotearoa and NeSI.
- Georgina is the Science Engagement Manager at NeSI where she ensures that NeSI is building strong relationships with the research sector. Prior to NeSI she has worked in molecular biology and intellectual property. She is passionate about enabling research and is interested in the fundamental shifts required to level up scientific research.
- Nick Jones is NeSI's founding Director, having established and led NeSI alongside a team of colleagues and peers since inception in mid-2011. Nick is responsible for NeSI's strategic directions and performance overall, bringing together a talented and diverse array of people, and their institutions and interests. Nick has over 20 years' experience in innovating in advanced information/computing technology in sectors including education, science and research. Nick established the eResearch NZ conference series in 2010 to support the sector coming together in the spirit of community to share experiences and explore directions in an area so critical to our future prosperity as a nation.
- Blair Bethwaite
- Bio

Tips and Tools for RSE's, September 12, 2022, 11:20 AM - 12:50 PM

The Regional Models Evaluation and Development toolbox is a python package for verification and diagnostics of Limited Area Numerical Weather Prediction models (NWP-LAMS). It is used by research teams across the Unified Model Partnership to assess changes to those models and to understand their biases and systematic errors, comparing model output and a variety of observation sources.
The original remit of the toolbox was to facilitate validation work within a single team at the Met Office (UK), then the scope was expanded to service several teams within the same organisation, and later to be used by teams across several organisations within an international consortium of weather and climate research and forecasting institutions, the UM Partnership.
The installation of the package at different organisations was fraught with many portability problems; furthermore, some time after, as the HPC systems in those organisations were updated, portability was lost again, creating a barrier to deliver collaborative research.
In the last year NIWA has been leading a concerted effort accross the UM Partnership to regain portability of this toolbox and put in place measures to preserve it. In this presentation I will use this work to discuss ideas on portability in an international collaborative development context.

Workflows, September 12, 2022, 2:00 PM - 3:00 PM

Cylc 8 Workflow Orchestration on NeSI HPC Platforms (or Your Laptop)

Hilary Oliver

NIWA

hilary.oliver@niwa.co.nz

ABSTRACT / INTRODUCTION

Cylc is an Open Source workflow management system, used around the world, that originates at NIWA in New Zealand. The recent release of Cylc 8.0 marks fruition of a ~4-year programme to re-engineer Cylc for Python 3, modern web technologies, a new scheduling algorithm, and more. Cylc 8 is more powerful and easier to use than earlier versions. Despite its origins it is not specialized to weather forecasting, and it is available on NeSI HPC Platforms. Hilary Oliver will give a brief teaser of Cylc 8 capabilities and demo ease of use with a small workflow on the Mahuika HPC.

ABOUT THE AUTHOR

- Hilary Oliver is a Principal Scientist at NIWA where he works on software infrastructure for weather, climate, and environmental forecasting systems on HPC. He leads the Open Source Cylc Workflow Engine project and helps chair the Technical Advisory Group of the Unified Model Partnership.

RStudio Workflow in an HPC Environment

Matt Bixley
NeSI/University of Otago
matt.bixley@nesi.org.nz

ABSTRACT

RStudio is the predominant IDE for users of R, who make up a significant portion of the NeSI code base. As and IDE there is excellent integration with other languages, particularly Python, BASH, Julia and other command line tools which allows for multilingual development, workflows and collaboration. RStudios embedded RMarkdown and now QUARTO provide additional tools for reports, papers, sharing your research and reproducibility. BUT!!! – RStudio does not integrate well with a shared resource infrastructure. R/RStudio is good for the cleaning, reporting and developing your workflow. Don't be afraid of the command line.

Here we will take a look at how to make use of RStudio, it quirks, some tips and tricks to get the best out of the NeSI resources.

ABOUT THE AUTHOR

- Matt Bixley
- Matt works in the Applications Support Team at NeSI (Otago) and comes from a background of Bioinformatics and Quantitative Genetics. He first learnt R sometime in the early 2000s and is still trying to work it out.

Building High Performance Computing Workflows (feat. Bash, Slurm)

Callum Walley
New Zealand eScience Infrastructure
callum.walley@nesi.org.nz

ABSTRACT / INTRODUCTION

Running many jobs, or complex multi-stage jobs can become tedious or even impossible without the use of some workflow automation. Using a dedicated workflow engine is one approach, however this can also add an unnecessary layer of complexity or be frustratingly inflexible.

We will be exploring when a pre-made solution isn't right for you and how to build your own workflow with readily available tools. This will mostly cover what can be done with bash shell and the Slurm scheduler as used on the NeSI clusters.

Note: We will not be covering a particular *workflow engine* (link to Alex workshop), only principles and general tools you can use to build your own workflows.

ABOUT THE AUTHOR(S)

- Callum is a mechanical engineer and applications support member at NeSI.

Are workflows the next big thing in scientific computing?

Alexander Pletzer[1], Chris Scott[2], Maxime Rio[1] and Dinindu Senanayake[2]
NIWA/NeSI[1] and University of Auckland/NeSI[2]
alexander.pletzer@nesi.org.nz

ABSTRACT / INTRODUCTION

As the complexity of research computing increases, one reaches a point where single, top-down execution streams no longer apply. Examples are postprocessing tasks that require multiple input files, each generated by a different executable. While it is possible to write custom shell scripts to orchestrate the execution of such tasks, using a workflow engine such as Snakemake or Cylc can have many advantages, including portability, reproducibility, scalability and improved maintenance. Perhaps one of the most common workflow patterns involves executing many embarrassingly parallel jobs and combining the results to produce visualisation and/or summary statistics. We will show how to implement such a pattern with Snakemake and Cylc and how this can increase your research productivity.

ABOUT THE AUTHOR(S)

- Alexander Pletzer
    - Alex is High Performance Research Software Engineer at NIWA for NeSI, helping scientists run better and faster. On his spare time, Alex enjoys windsurfing and cycling.
- Chris Scott
- Chris is Computational Science Team Lead at NeSI
- Maxime Rio
- Maxime is Data Science Engineer for NeSI at NIWA
- Dinindu Senanayake
- Dini is Bioinformatics/Genomics Application Support Specialist at NeSI

Virtual Reality – More Than Just Science Engagement

Ben Jolly
Manaaki Whenua – Landcare Research
jollyb@landcareresearch.co.nz

ABSTRACT / INTRODUCTION **(Up to 200 words)**

Individual tree structure metrics (height, canopy diameter, etc) provide important information for scientists studying forest dynamics, carbon sequestration, and species distributions. Point clouds from aerial and terrestrial lidar scans are used to calculate these metrics, but first each tree must be identified and the points tagged. This is difficult to achieve in natural forests that are structurally complex which stretches the capabilities of most automated approaches. Manual techniques also struggle with visualisation and interaction of these data on inherently two-dimensional equipment (screen and mouse). Virtual Reality (VR) is very effective at visualising three-dimensional data, however most applications use pre-built environments rendered with game engines and cannot ingest science data. The open-source 'LidarViewer' software is able to deal with standard lidar formats and provides a very powerful environment useful for viewing, labelling, and analysing complex data in VR. However, it is notoriously difficult to install and launch. In this presentation I address the steps taken to get to a working LidarViewer environment and present an example of the use of VR to solve a science problem: segmentation of trees in lidar data to calculate structural metrics and provide a training dataset for future segmentation efforts using artificial intelligence (AI).

ABOUT THE AUTHOR(S)

- Ben Jolly
- Ben is a Remote Sensing Researcher at Manaaki Whenua – Landcare Research with an Electronics and Computer Systems Engineering degree from Massey University and a PhD in Atmospheric Physics from the University of Canterbury. Current research interests range from sea-ice to vegetation mapping utilising both desktop and HPC resources.

NIWA is ideally positioned to leverage data science due to the vast amounts of data it produces, the domain expertise to analyse them and access to HPC facilities. However, due to NIWA’s broad range of scientific disciplines, the experiences, knowlegde and tools have not been optimally shared in the past. This has been addressed by a solid data science strategy, the creation of a new data science team that cuts across domains, and a community of practice, NIWA’s data science journal club.

The Data Science Journal Club is an initiative to bring NIWA’s data scientists and researchers together. It is a platform to exchange ideas and learn from each other. This talk reflects on the two years since the Data Science Journal Club's inception, shares tips on how to bring communities together and discusses the challenge of how to keep the momentum going.

Category: Presentations

Mobile first – using shinyMobile to improve engagement of science in communities

Lillian Lu[1], Richard Dean[1]
[1] Institute of Environmental Science and Research Limited (ESR)
lillian.lu@esr.cri.nz, Richard.dean@esr.cri.nz

ABSTRACT / INTRODUCTION

It is important for data scientists to communicate research results and findings to a wide audience. However, standard dashboarding tools favour desktop-based solutions and there are significant barriers which make building a mobile-friendly app time and cost-consuming.
R is a language preferred in fields such as statistics and data science; its possibility to produce a mobile-app-like dashboard has often been overlooked. When our team began the development of a public facing Wastewater Surveillance Tool, the ability to use the dashboard on a mobile device was important.
This talk will share how SARS-CoV-2 wastewater data was communicated with the public using a lightweight mobile-friendly dashboard. One of the main aims of this project was to aid communities in personal decision-making in addition to workers in the health sector. The main R package used to make the dashboard was ShinyMobile. The challenges faced in implementing this along with modularising the code added many lessons throughout the pipeline to production. The use of CSS and JavaScript were essential in improving the dashboard UI and functionality. This experience illustrates opportunities for research organisations across Aotearoa to innovate and improve public engagement of science with limited resources.

ABOUT THE AUTHOR(S)

Lillian Lu is a data scientist at ESR, with a background in digital marketing and business development. She returned to university in 2019 and completed her master's degree in Analytics. While studying, Lillian worked for local and US companies as a data consultant and decided to kick off her data science journey at ESR after she graduated. Lillian's current work mainly focuses on public health and toxicology.
Richard Dean is a senior data scientist at ESR. He works across the organisation on projects that gain insight from big data sets to tackle nitty gritty real-world problems affecting human communities in Aotearoa, covering everything from forensic science to human health, and the environment. He has a BSc in Information Systems Management from Durham University and wrote an MSc thesis on public health data interoperability standards while working at the Wolfson Research Institute for Health and Wellbeing in Durham.

Running a Carpentries Workshop

Murray Cadzow
University of Otago
murray.cadzow@otago.ac.nz

- ABSTRACT / INTRODUCTION **(Up to 200 words)**

This Birds of a Feather session is for people who are interested in running a Carpentries workshop and would like to know more about what is involved. Find out what resources are available from The Carpentries, how to create a workshop website, and learn from the experiences of previous workshop events.

ABOUT THE AUTHOR(S)

- Dr Murray Cadzow
- Murray is a Scientific Programmer on the Research Teaching IT Support team at the University of Otago. From 2010, he has worked on research into the genetic basis of gout. He received his PhD (Biochemistry) in 2018, and has been heavily involved in computational literacy and bioinformatic training at the University of Otago - organising ResBaz Dunedin and the Otago Bioinformatics Spring School. He is both a Carpentries instructor and instructor trainer. His teaching has focused on delivering digital literacy training to researchers, and the development and support of the local Carpentries community at Otago. Murray is a Genomics Aotearoa training associate, and a member of the NeSI Research Reference Group.

Sustaining Research Software: Issues and Actions

Dr Daniel Katz
National Centre for Supercomputing Applications (NCSA)
dskatz@illinois.edu

ABSTRACT / INTRODUCTION (Up to 200 words)

Research software is pervasive in research today; most research is dependent on software. Much of this research software is built and maintained by communities, typically comprising research software engineers and researchers. And most of the communities have or develop a strong voluntary aspect, at least over time, particularly because any one funding activity for research software is usually shorter term that the life span of the software itself. Additionally, software that is not maintained will eventually stop working or stop being useful, because of changes in the underlying hardware and software environment and changes in user needs. Therefore, sustaining the software requires ongoing human effort, and can be seen as balancing that human effort with the specific needs of the software project. This talk will discuss these issues and actions that effect and affect this balance that be taken by individual projects, communities, funders, and institutions.

ABOUT THE AUTHOR(S)

Daniel Katz
Daniel S. Katz is Chief Scientist at the National Center for Supercomputing Applications (NCSA), Research Associate Professor in Computer Science (CS), Research Associate Professor in Electrical and Computer Engineering (ECE), Research Associate Professor in the School of Information Sciences (iSchool), and Faculty Affiliate in Computational Science and Engineering (CSE) at the University of Illinois Urbana-Champaign. Dan's interest is in the development and use of advanced cyberinfrastructure to solve challenging problems at multiple scales. His technical research interests are in applications, algorithms, fault tolerance, and programming in parallel and distributed computing, including HPC, Grid, Cloud, etc. He is also interested in policy issues, including citation and credit mechanisms and practices associated with software and data, organisation and community practices for collaboration, and career paths for computing researchers.

Category: Lightning talks

Defining success of a dashboard product with user analytics

Lillian Lu
Institute of Environmental Science and Research Limited (ESR)
lillian.lu@esr.cri.nz

ABSTRACT / INTRODUCTION

When we develop a dashboard to communicate our science with the public and fellow researchers, how do we know what we share is what they are interested in? How do people interact with the UI and functionality of a dashboard and how can we improve them?
Web analytics tools are commonly used by different organisations to understand user behaviours on their website. Since R Shiny dashboards are primarily web-based, we can easily implement tools such as Google Analytics with a few lines of JavaScript codes and start to gain insights into our audience, even in real time. User analytics provides tremendous value in improving the efficiency of science communications, defining success of a dashboard product and equivalently important - knowing when we should let something go.
In this talk, Lillian will walk you through the what, why and how and prep you to implement your own analytics within 5 minutes.

ABOUT THE AUTHOR(S)

Lillian Lu is a data scientist at ESR, with a background in digital marketing and business development. She returned to university in 2019 and completed her master's degree in Analytics. While studying, Lillian worked for local and US companies as a data consultant and decided to kick off her data science journey at ESR after she graduated. Lillian's current work mainly focuses on public health and toxicology.

Category: Demos

Open source down the drain – using R, Leaflet and QGIS to track COVID in wastewater

Helen Morris[1]
[1] Institute of Environmental Science and Research Limited (ESR)
helen.morris@esr.cri.nz

ABSTRACT / INTRODUCTION

Aotearoa's response to the COVID-19 pandemic generated large amounts of geospatial data. In addition to traditional datasets such as reported cases, the emergence of wastewater-based epidemiology (WBE) provides an independent metric and has been used in key decision making to determine traffic light settings.
ESR relied heavily upon open-source geospatial tools such as QGIS and R's spatial manipulation (sf) library to interpret wastewater results. In addition, relying on support from wastewater treatment plants and local councils was essential for collecting related wastewater network data.
In this talk, Helen will walk you through the pipeline used to interpret geospatial data and track cases in wastewater – from collecting maps of wastewater pipes to creating catchment maps, assigning cases to catchments and using the maps to interpret wastewater results from over 200 sites across Aotearoa. The talk will conclude by considering lessons learned from COVID can be applied to other research domains – such as monitoring environmental health and Aotearoa's response to climate change.

ABOUT THE AUTHOR(S)

Helen Morris is a graduate data scientist working at The Institute for Environmental Science and Research (ESR). Her background includes a BSc in Biological Sciences with an endorsement in Biotechnology and a PGDipSci in Molecular and cellular biology. Her work currently focuses on public health and metagenomics.

Tools and Applications, September 13, 2022, 10:50 AM - 12:30 PM

A look at how NeSI's new processors can lift researchers' productivity

Alexander Pletzer[1], Gene Soudlenkov[2] and Chris Scott[2]
NIWA/NeSI[1] and University of Auckland/NeSI[2]
alexander.pletzer@nesi.org.nz

ABSTRACT / INTRODUCTION **(Up to 200 words)**

The New Zealand eScience Infrastructure (NeSI) has recently purchased and installed 8,448 processors (physical cores) to alleviate some of the current computational load experienced by users on the Mahuika platform. Some of the most salient features of these new AMD EPYC Milan CPUs are: large cache sizes and 128 cores per node. Here, we show how you can leverage this new computational resource and provide a glimpse of which types of code might benefit from running on the Milan processors.

ABOUT THE AUTHOR(S)

- Alexander Pletzer
    - Alex is High Performance Research Software Engineer at NIWA for NeSI, helping scientists run better and faster. On his spare time, Alex enjoys windsurfing and cycling.
- Gene Soudlenkov
- Gene is Research Support Team Lead at NeSI.
- Chris Scott
- Chris is Computational Science Team Lead at NeSI

A new decentralised data management model to fit the challenges in multi-agency collaboration in flood management in NZ

Phil Mourot[1,2], Guilherme Weigert Cassales[1], Albert Bifet[1], Nick Lim[1], Justin Liu[1]
[1] Te Ipu o te Mahara, Artificial Intelligence Institute, University of Waikato
[2] Waikato Regional Council
phil.mourot@waikato.ac.nz

ABSTRACT / INTRODUCTION (Up to 200 words)

When we started to work on flood prediction using a data-driven approach, we focused on building and tuning a machine learning model. This is mainly what you do when you want to build a predictive model using artificial intelligence. But the journey has just begun.
Scalability is a real issue for many AI projects. How many developed models are fully deployed in production? The infrastructure will slow down the deployment as your models will take on larger volumes of data and compute resources. We also need to ensure that data is consistent and trustworthy. It could be easily solved if we used datasets from only one source, one organisation.
To address the challenge of building resilience against flooding, we must change our approach from reactive to proactive and from single agency to partnerships. Effective collaboration among agencies is the key to a sustainable solution toward risk reduction (Sendai Framework priority 2 on strengthening disaster risk governance).
The TAIAO team from the AI Institute has developed a new data mesh-based solution to face the multi-source of data issue. This presentation presents a new data-driven modelling framework for a real-time flood impact prediction application.

ABOUT THE AUTHOR(S)

- Dr Phil Mourot
- Dr Phil Mourot has been working in Natural Hazards for more than 20 years, across different disciplines such as geoscience, computer science, electronics, geophysics and seismology, and completed a PhD in applied seismology in France to predict rockfalls and landslides (Grenoble). Phil is involved in the TAIAO project. With a foot in local government (Waikato Regional Council) and another foot in research (AI Institute), he can undertake research projects that align with local government issues and focus on helping New Zealand be a more resilient nation. Phil's current research focuses on predicting the impact of floods using machine learning.

Tools and Applications, September 13, 2022, 10:50 AM - 12:30 PM

Developing a package to fit piecewise nonlinear curves

Gemma Mason
NIWA
gemma.mason@niwa.co.nz

ABSTRACT / INTRODUCTION **(Up to 200 words)**

When do you use an existing tool, and when do you develop a new one? This talk will discuss my experiences as a data scientist, helping the Air Quality team at NIWA to analyse a large data set of classroom CO2 time series measurements. Automatically splitting the data into piecewise segments with nonlinear fitted curves was a challenge that had been somewhat addressed in the literature, on a theoretical level, but the existing available software packages had not been designed for it directly.
This talk will discuss how I built a Python package to implement a method that had been discussed but not directly tried in the existing literature. I will then compare the performance of my purpose-built package with an alternate strategy of customizing an existing package, and discuss the risks and benefits of each approach.

ABOUT THE AUTHOR(S)

Gemma Mason works as a Data Scientist at NIWA, where she specializes in machine learning for scientific simulations. She is originally from Christchurch but now lives in Auckland.

Tools and Applications, September 13, 2022, 10:50 AM - 12:30 PM

Biography:
Rose is currently a remote sensing scientist and a visiting researcher at the Geospatial research institute. Her work primarily focuses on combining geospatial data, primarily LiDAR point clouds, to produce hydrologically conditioned DEMs and roughness maps for use in river flood modelling. Her research interests centre on surface generation and attribute mapping from a wide array of spatial and geospatial datasets.

Title **(Up to 15 words) Automated generation of hydrologically conditioned Digital Elevation Models for national-scale hydrodynamic modelling.**

Authors name(s) Rose Pearson, Emily Lane, ??
Organisation(s) National Institute of Water and Atmosphere
Authors Email(s) rose.pearson@niwa.co.nz
Theme: Sustainability: Communities, Digital Practices and Tools. (Full guidelines)

- Lessons learned from experience
- Roadblocks and pitfalls of some approaches
- Upskilling requirements
- Data ethics and privacy considerations in sustainable research
- ***The role of reproducibility in sustainable research***
- ***New tools and approaches***
- ***How to leverage tools and workflows to facilitate sustainable research***

ABSTRACT / INTRODUCTION **(Up to 200 words)**

GeoFabrics is a tool for improving Digital Elevation Model (DEM) for flood modelling as part of the Mā te haumaru ō te wai: Flood resilience Aotearoa. GeoFabrics is an open-source Python package that combines LiDAR point clouds with ocean bathymetry surveys and information extracted from Open Street Maps (OSM) and river networks to improve a DEMs hydrological conditioning around rivers, drains, culverts and bridges. This is important as unconditioned DEMs give inaccurate inundation hazard maps. The described methodology has been configured into an automated pipeline on New Zealand eScience Infrastructure (NeSI) using the task-based scheduler Cylc. This allows GeoFabrics to be efficiently deployed on NeSI and applied across Aotearoa. GeoFabrics is only fifteen months old. Its rapid advancement would not be possible without open frameworks like Dask and Cylc. This is because these frameworks are mature and purpose built to solve a specific set of problems to a high level of excellence. These frameworks also reduce the maintenance load of GeoFabrics making it more sustainable for a small team. In turn, we have published our tools on PyPI and conda-forge making them more accessible to others, and further improving the reproducibility of the approach.

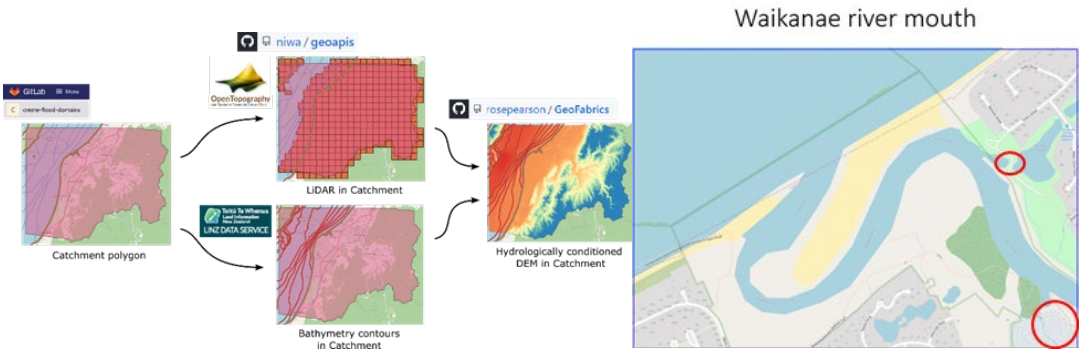

Figure 1: Process of including ocean, river, drain and culvert information into a DEM.

ABOUT THE AUTHOR(S)

- Name
- Bio

Tools and Applications, September 13, 2022, 10:50 AM - 12:30 PM

A Data Cube for Collaborative Analysis of Antarctic Geospatial Data

Ben Jolly
Manaaki Whenua – Landcare Research
jollyb@landcareresearch.co.nz

ABSTRACT / INTRODUCTION **(Up to 200 words)**

Large-scale geospatial data analysis is a conceptually simple process: define a small area of interest, test various analysis techniques, then scale up to a larger area. However, an inordinate amount of time is often spent on the scaling stage as researchers discover that even acquiring and pre-processing data is difficult at regional to continental scales, let alone running the actual analysis on it. As such, many studies are purposely constrained to more manageable areas despite there being value in larger-scale analysis. The Antarctic Science Platform (ASP) Projects 3 and 4 attempted to address this by the creation of a data cube on the NeSI High Performance Computer (HPC). This was designed to hold diverse geospatial data from different disciplines and research groups and act as a common environment for researchers to access and analyse these data, reducing some of the technical barriers and hopefully easing and encouraging inter-disciplinary collaboration. Despite delays and setbacks a proof-of-concept data cube is now up and running with further plans to expand both access and available datasets. This presentation covers the use of Snakemake and STAC to create the data cube, analyses the setbacks, and discusses future possibilities including implementation of the Open Data Cube.

ABOUT THE AUTHOR(S)

- Ben Jolly
- Ben is a Remote Sensing Researcher at Manaaki Whenua – Landcare Research with an Electronics and Computer Systems Engineering degree from Massey University and a PhD in Atmospheric Physics from the University of Canterbury. Current research interests range from sea-ice to vegetation mapping utilising both desktop and HPC resources.

Research Management and Digital Repository Solutions with Symplectic Elements and Figshare

Dr Anthony Dona
Digital Science
a.dona@digital-science.com

Demonstration

ABSTRACT / INTRODUCTION

Digital Science aims to deliver flexible research solutions that enable universities, research institutions or funding agencies to pursue their research goals and more seamlessly advance knowledge. **Symplectic Elements** is a research management system that supports the collection, collation, sharing and reporting of various outputs and professional activities that occur throughout the research lifecycle. It enables organisations to better capture research information, comply to assessment strategies and report internally on research outcomes. Furthermore, Symplectic Elements can be coupled with **Figshare**, a comprehensive off the shelf repository solution. Figshare enables researchers or research organisations to make traditional and non-traditional research outputs citable, shareable and more easily discoverable. Data, software, white papers, policy documents, publications and much more can be made available in a controlled and curated way through Figshare. All research outputs will be provided a persistent identifier (DOI) for citation metrics and to capture future usage and mentions. Join this demonstration to see how world leading research organisations are managing their research processes and making their research outputs open and available to others in the academic domain.

ABOUT THE AUTHOR

- Anthony Dona PhD
- Dr Anthony Dona completed his doctorate in Analytical Chemistry at the University of Sydney in 2010. From there, Anthony moved to London to help build and managed the National Phenome Centre at the Imperial College of London. He specialised in developing and implementing methodologies that marry clinical metadata with comprehensive metabolite measurements. Anthony collaborated with and helped implement these methodologies in labs across the globe. Anthony returned to Australia in 2015 and is now a Senior Director working with Government and Funders across Asia Pacific to enable them the better assess, fund and apply novel research. He continues to drive his research interests and mentor postgraduate students through an honorary position at the University of Sydney.

Data: Management, Visualisation and Manipulation, September 13, 2022, 1:30 PM - 2:40 PM

Show me the data! Large-scale interactive data visualisation with Xarray

Maxime Rio[1,2] and Amir Pirooz[1]
[1]NIWA, [2]NeSI
maxime.rio@nesi.org.nz, amir.pirooz@niwa.co.nz

ABSTRACT / INTRODUCTION

With the increase of computing power and storage capacity, scientific datasets, whether observational or of simulation origin, become larger every day. It is not uncommon to manipulate datasets that are an order of magnitude larger than the available memory on workstations. As an additional constraint, these large datasets tend to be stored on remote servers and transferring them locally is not an option. In this context, interactive exploration of the data can become very challenging.

In this talk, we will present how an ecosystem of Python packages interoperating with Xarray can greatly simplify the visualisation of large datasets stored on a remote server. In particular, we will highlight the role of Dask for handling data retrieval and Holoviz packages to build a simple web-based dashboard. As an illustration, we will show how these tools can help NIWA scientists visualise a large archive of high-resolution forecasts.

ABOUT THE AUTHOR(S)

Maxime Rio is a data science engineer and data scientist at NeSI and NIWA. He enjoys helping researchers to analyse their data, from visualisation to probabilistic modelling.

Amir Pirooz is a numerical weather prediction (NWP) / Computational Fluid Dynamics (CFD) modeller and analyst at NIWA. His work focuses on simulating wind flow, developing an NWP reanalysis model, and analysing observational and numerical data.

Bringing Order to the Chaos–Improving the Reliability of a Python Library for Data Ingestion **(Up to 15 words)**

Authors name(s) Lukas Trombach, Chris Seal
Organisation(s) University of Auckland
Authors Email(s) lukas.trombach@auckland.ac.nz, c.seal@auckland.ac.nz

ABSTRACT / INTRODUCTION **(Up to 200 words)**

Scientific instruments producing terabytes of data a day are becoming increasingly widespread. Although management of such large amounts of files and metadata can be challenging, it is even more important to do it efficiently than it is for smaller amounts.
Research funders, such as the Royal Society of New Zealand, have recognised this and now require comprehensive research data management plans to be submitted as part of their grant application process. Research data management is often simplified through the use of a centralised system, though this brings with it additional technical complications around ingestion of the data.
Here, we show how the data management tool MyTardis can be used to automate the ingestion of data by creating a python library which utilises the MyTardis API. While the ingestion project started out as a simple python script it has evolved into a more complex piece of software, which can be hard to maintain in pure python.
We share lessons learned and our experiences with improving the reliability, testability and maintainability of our python code base by using type hints in combination with tools and libraries like pylint, pytest, mypy and pydantic.

ABOUT THE AUTHOR(S)

- Lukas Trombach
- Lukas has a background in computational chemistry and software engineering. He finished his PhD in Chemistry at Massey University Auckland and then took on a role as a full-stack software developer at a medium sized telco company. He is currently working as an eResearch Engagement Specialist at the Centre for eResearch where he is mostly working on onboarding research instruments into MyTardis.
- Chris Seal
- Chris has a background in materials engineering, completing a PhD in Chemical and Materials Engineering at The University of Auckland. He worked in a range of roles in the materials production and energy generation sectors, before moving into a more computational space while completing a postdoctoral research project at the University of Manchester. Upon returning to New Zealand, he returned to The University of Auckland and is currently working as a Senior eResearch Solutions Specialist at the Centre for eResearch.

Data: Management, Visualisation and Manipulation, September 13, 2022, 1:30 PM - 2:40 PM

Introducing opendata.fit: a FAIR data analysis and publication platform

Varvara Efremova and James Wilmot
UNSW Sydney & Australian Characteristic Commons at Scale
varvara@echus.co and jameswilmot2000@gmail.com

ABSTRACT / INTRODUCTION **(Up to 200 words)**

This talk introduces the opendata.fit platform, developed as part of the Australian Characterisation Commons at Scale (ACCS) project. opendata.fit is a web-based data analysis and publication platform, focused on improving reproducibility in research and supporting FAIR (Findable, Accessible, Interoperable and Reusable) data principles. opendata.fit builds on the work of Bindfit, a web-based binding constant fitting platform developed at UNSW.
opendata.fit aims to provide users with a discipline-agnostic, common platform to manage scientific analysis workflows and datasets from development to publication. It provides tooling to upload instrument data, execute analysis algorithms in the cloud, visualise results, and publish complete workflows as citable entities. Users will be able to choose from a library of existing public workflows or develop their own. It currently supports execution of Python-based analysis algorithms with plans to extend support to other common scientific computing languages.
As part of this work, we have developed a novel approach to data portability and metadata management using and extending the open source Frictionless Data standard and Vega visualisation language.
This presentation will provide an overview of the opendata.fit platform, demonstrate a sample analysis workflow on small angle scattering data, and discuss the platform's future direction.

ABOUT THE AUTHOR(S)

Varvara is a research software engineer at the University of New South Wales with a background in astrophysics. She has extensive experience with the Python scientific stack, containerisation, web development, as well as cloud computing infrastructure in the context of scientific applications. She is interested in using modern web and containerisation technologies to develop tools that enable consistent, reproducible execution and publication of scientific workflows.

James is a research software engineer at UNSW Sydney with a background in physics. He has extensive experience in web application technologies, containerisation and python. When not coding he enjoys running, cycling and being outdoors.

Both Varvara and James help run and maintain a community makerspace based in Canberra.

-

Tools and Applications 2, September 13, 2022, 2:55 PM - 5:10 PM

Weather forecasting in the cloud - early experience with cloud HPC

Wolfgang Hayek, Paulo Almeida Rodenas, Kameron Christopher
NIWA
wolfgang.hayek@niwa.co.nz, paulo.almeidarodenas@niwa.co.nz, kameron.christopher@niwa.co.nz

ABSTRACT / INTRODUCTION

Numerical weather prediction (NWP) remains a challenging application for high-performance computing (HPC), combining the need for high computational performance, low-latency/high-bandwidth network communication, fast parallel IO, and high platform availability to ensure that forecasts are produced in a timely and reliable manner. In recent years, commercial cloud providers have started to offer suitable infrastructure in the public cloud, promising scalable and efficient execution of HPC workflows, and thus a potentially attractive alternative to on-premises systems for use cases such as disaster recovery for operational weather forecasting.
This talk will discuss hands-on experience with running a large weather forecasting model in the public cloud, including examples of available platforms and services, how to get started, and some comparisons with traditional on-premises HPC platforms from a user's perspective.

ABOUT THE AUTHOR(S)

Wolfgang Hayek is a HPC Research Software Engineer at NIWA, and group manager of NIWA's scientific programming group, with many years of experience in scientific computing and HPC.
Paulo Almeida Rodenas is a HPC Data Science Architect at NIWA. He gets to turn interesting research ideas into optimised software to run on NZ's supercomputers :) When he is not doing any of that, he writes technical deep dives for his blog: https://deepdives.medium.com
Kameron Christopher is NIWA's Chief Scientist for HPC and Data Science. He has a research and engineering background in AI methods for digital signal processing, computer graphics, and HCI.

A semi-automated HPC deployed system to create national flood maps

Authors name(s): Hisako Shiona, Emily Lane
Organisation(s): NIWA
Authors Email(s): hisako.shiona@niwa.co.nz

ABSTRACT / INTRODUCTION

Mā te haumaru ō te wai is a five-year MBIE-funded Endeavour programme to increase the flood resilience of Aotearoa New Zealand through providing accurate, nationally consistent flood hazard and risk information and working with government, iwi, and other stakeholders to use that information wisely. This information will help ensure robust adaptation decisions.
A critical task in understanding our national flood hazard is developing a semi-automated system to create national flood-maps. This allows us to scale up from modelling a single-catchment to the entire country consistently.
New Zealand was broken up into catchment-based domains based on a range of factors such as terrain, population, land use, domain-size, memory-limits, etc. For each domain, geofabrics datasets (DEM, roughness, river-network, etc), design storms developed from HIRDS (High Intensity Rainfall Design System), river flow data from hydrological models, TopNet and tide are gathered and fed into the two-dimensional hydrodynamic model, BG-Flood. Then, the results are combined back into national level.
We have chosen Cylc 8 (workflow engine for cycling systems) as a tool which allows to process each component in the correct order and cycle through catchments. I will describe the workflow, and discuss the obstacles we faced pulling the separate parts into one process.

ABOUT THE AUTHOR(S)

Name: Hisako Shiona
Bio: Hisako is an atmospheric technician at NIWA. Her work primarily focuses on data processing and analysis. Her research interests centre on software development on various disciplines.

ESTABLISHING AN AUTOMATED DEVICE VERIFICATION PROCESS FOR LUMI™ DRUG SCAN

Janet Stacey
Institute of Environmental Science and Research (ESR)
janet.stacey@esr.cri.nz

ABSTRACT / INTRODUCTION **(Up to 200 words)**

Lumi Drug Scan (Lumi™) is a low cost, rapid screening solution jointly developed by ESR and NZ Police. The solution assesses samples for the presence of illicit substances in the field. It consists of a near infrared (NIR) device attached to a mobile phone that carries out a scan of the substance and uses cloud-based machine learning algorithms to identify the presence of illicit drugs in the New Zealand market.

This presentation will describe the Lumi™ service and showcase the process that ESR have established to determine the working range of the Lumi™ devices and test individual devices against a set of chosen performance criteria using an automated process. The process can take scans and produces two reports: a complete report of results for an analyst, and a Certificate of Verification report for a customer in approximately 70 seconds per device. Automation of this process was considered essential as Lumi™ moves into an operational service.

A verification process gives reassurance that the device is fit for purpose and scans obtained are trustworthy and suitable for analysis. This process has been used for the initial verification of devices and will act as an annual check throughout the device's life.

ABOUT THE AUTHOR(S)

- Janet Stacey
- Janet Stacey is a Digital Sciences Engineer (Senior Scientist) at the Institute of Environmental Science and Research (ESR). She has two Master of Science degrees, one in Forensic Science and another in Bioinformatics. Janet worked in the Forensic Biology case work laboratory for 11 years processing DNA and RNA samples to assist with the resolution of criminal investigations before transitioning to a Data Science role. She has now been an active Research Software Engineer for four years. Her work includes furthering ESR's work with machine learning/AI and automated workflows, dashboard/visualizations and growing data science capability through mentoring and training. She is a member of RSE-aunz Steering Committee, and an representative on various advisory groups and working groups. She has a special interest in Forensic Intelligence, Responsible AI Development and Māori Data Sovereignty.

Extendable projection of social contact matrices

Nicholas Tierney(1), Aarathy Babu(1), Michael Lydeamore(2), Nick Golding(1,3)
1: Telethon Kids Institute; 2: Monash University; 3: Curtin University
nicholas.tierney@telethonkids.org.au

ABSTRACT / INTRODUCTION (Up to 200 words)

Contact matrices describe the degree of face-to-face contact between individuals of given age groups. They are commonly used to model how diseases such as COVID-19 spread in a population. Contact matrices are produced from empirical data resulting from a *contact survey*, which requires individuals to diary their daily amount and manner of contact with people. However, these surveys are expensive to run, so there are not many empirical datasets.

Existing statistical methodologies can project empirical contact matrices to new countries - e.g. those produce by Prem et al. Existing work provides these synthetic contact matrices for a fixed set of countries, at specific time points. Infectious disease modellers therefore face a problem when developing models for countries not considered in these analyses, or for subpopulations within those countries. We need software to implement these methods so they can be flexibly reused on different populations.

We have developed the {conmat} software, which contains methods for creating synthetic contact matrices using existing contact surveys, and can be used to create synthetic contact matrices for any new age population structure. In this talk, I demonstrate the use of {conmat} in epidemiological analyses, and discuss the software design process.

ABOUT THE AUTHOR(S)

- Nicholas Tierney
    - Dr. Nick Tierney completed his undergrad and honours in Psychological Science, then took an unconventional turn into a PhD in Statistics. He now works as a research software engineer with Dr. Nick Golding here at the Telethon Kids Institute. He is currently working on improving and maintaining the greta (https://greta-stats.org/) R package for statistical modelling. He is also interested in implementing workflows to automate data analysis.
    - Previously Dr. Tierney was at Monash University (2017-2020), working as a research fellow, then lecturer. He worked with Professor Di Cook, creating exploratory data

analysis techniques. He also taught ETC1010, introduction to Data Analysis (https://dmac.netlify.org/)

- Dr. Tierney's research interests are broad, but centred around improving data analysis. This includes exploratory data analysis, statistical modelling, diagnostics, and understanding how colour choice can impact decision making. Dr. Tierney is a strong believer in free and open source software, and has written several popular R packages to improve data analysis, which can be seen on his software page: http://njtierney.com/software.
- Dr. Tierney is a keen outdoors person, in particular hiking, rock climbing, and kayaking. He recently hiked 300K of the Australian Alpine Walking Track, a very rugged adventure, especially when hiking solo. Nick is also interested in coffee, music, and photography, especially analog film photography.

Tools and Applications 2, September 13, 2022, 2:55 PM - 5:10 PM

From Untitled1 to CRAN: The Why and How of Publishing My First R Package

Jie Kang[1], Chris Scott[2] and Albert Savary[2]
- Beef and Lamb New Zealand
1. New Zealand eScience Infrastructure (NeSI)

jie.kang@blnzgenetics.com
chris.scott@nesi.org.nz
albert.savary@nesi.org.nz

ABSTRACT / INTRODUCTION **(Up to 200 words)**

At the very beginning of my academic journey, I joined over two million R users across the globe to embrace my own research question as a fourth-year student. With many helps from my colleagues and mentors, I managed to carefully start punching R codes which were supposed to help me to conduct simulations, run statistical analyses and eventually address my research question. However, soon I realised that the new script I opened a few weeks ago (named “Untitled1.R”) gradually became a headache on its own: it contains more than one thousand lines of codes (excluding the random comments I made from time to time), yet a large chunk of them is repetitive and redundant. Needless to explain how confused I was, some suggested that I should be patient and consider modularising my script, and ideally, convert repetitive tasks into functions. To date, I still firmly believe that was one of the best programming advice I’ve ever had. Later, a seamless logic flow behind the codes started showing up and I can see this collection of R functions matured into an R package (so I know what I’m doing after the summer break). Then I talked to Chris and Albert to see if I could get some help with getting published on CRAN, and yes, they walked through the process with me and even helped me to test the package. Looking back, I would recommend this to anyone who just opened or is about to create their “Untitled1.R” scripts: it might pay if you code it like you’d publish it on CRAN.

ABOUT THE AUTHOR(S)

- Jie Kang

  Jie works at Beef and Lamb New Zealand as the genetic evaluation specialist, while finishing his PhD in Quantitative Genetics. Prior to this, Jie studied at the University of Otago, majoring in Mathematics and Statistics. Outside work, Jie enjoys practicing music with his band called ‘gingernuts’ and training mixed martial arts.
- Chris Scott

  Chris works as a Research Software Engineer for NeSI, helping researchers by developing, optimising and improving code through NeSI’s consultancy service. Outside work, Chris enjoys running, cycling, swimming and trying to train his dog to do agility.
- Albert Savary

  Albert works as a support specialist for NeSI. He supports the users of NeSI platforms. Outside of work, Albert enjoys hiking.

Scaling NWP workloads on AWS to achieve your research goals

Timothy Brown, Sean Smith
Amazon Web Services
tpbrown@amazon.com, seaam@amazon.com

ABSTRACT / INTRODUCTION

The use of cloud computing technologies within HPC has grown considerably over the last few years. With these advances there's more options on how to run Numerical Weather Prediction (NWP) than ever. In this talk we distill the options and show how researchers can get started in an environment they're familiar with. We'll discuss cluster orchestration with AWS ParallelCluster and Slurm, parallel filesystem with Amazon FSx for Lustre, high performance networking with Elastic Fabric Adapter (EFA), software management with Spack and then we'll present scaling and cost analysis of the Unified Forecast System on AWS HPC6a (AMD Milan).

ABOUT THE AUTHOR(S)

Timothy Brown is a Principal Solutions Architect for Compute & HPC at AWS. He has 15 years of HPC experience, spanning different roles related to compute, storage and network optimizations, with a focus on numerical weather prediction. Prior to AWS, Timothy was a Software Engineer at Spire Global. Timothy holds an MSc and BSc (Hons) from UWA, Australia.

Sean Smith is a Sr. Solution Architect for HPC at AWS. Prior to that, Sean worked as a Software Engineer on AWS Batch and CfnCluster, becoming the first engineer on the team that created AWS ParallelCluster. Sean holds a Masters and Bachelor's degree from Boston University in Computer Science.

Tools and Applications 2, September 13, 2022, 2:55 PM - 5:10 PM

Offloading work to a remote high performance computer using Globus and funcX

Chris Scott [1], Maxime Rio [1, 2]
(1) NeSI, (2) NIWA
chris.scott@nesi.org.nz, maxime.rio@nesi.org.nz

ABSTRACT / INTRODUCTION (Up to 200 words)

Some researchers may have a workflow that has one or a few steps requiring extra computational power but otherwise could run well on their local machine, where they are most comfortable working. Others could be generating data locally (e.g. from a piece of lab equipment) and want to push that data to NeSI for processing and then retrieve the results automatically. To answer these needs, the RemoteJobManager (RJM) tool has been developed to enable researchers to seamlessly offload work to the NeSI HPC platform. The benefit of using RJM in these situations is that the tool will do all the heavy lifting of transferring data to and from NeSI and interacting with the Slurm job scheduler on behalf of the user. In this talk we will share our experience of developing RJM using Globus and funcX and demonstrate how the tool works.

ABOUT THE AUTHOR(S)

- Chris Scott is a Research Software Engineer at NeSI
- Maxime Rio is a Data Science Engineer at NeSI and Data Scientist at NIWA