Contribution ID: **73** Type: **Talk (15min + 5min)**

Reproducible Package Environments for Modeling Workflows

Wednesday, March 6, 2024 2:30 PM (20 minutes)

By default languages like R and Python load packages/modules from a system-wide package environment. Thus all projects use the same package versions which makes it hard to track which package versions were used for a specific project, and using different package versions for separate projects becomes impossible. Package updates which are required for one project may break other projects. These issues are especially problematic on large multi-user systems like an HPC (high performance computer cluster).
To solve this virtual package environments can be used. These isolate projects from each other and from the system-wide package environment, making it possible to use different package versions in separate projects. Also, this makes it clear which package versions were used for a project.
In this talk we present our package environment solution for developing scientific models on a large multi-user HPC. We describe how and why our approach is different from typical package environment setups, what the advantages and downsides are, and what lessons we have learned. In summary, at the cost of some additional complexity for users, our approach greatly improves reproducibility, robustness, and control over which package versions are used for each model run. Using a shared package cache we need 8.3 GB disk space for 3400 different package versions, and restoring entire package environments with over 200 packages takes a couple of seconds.

Slot length

Primary author: SAUER, Pascal (Potsdam Institute for Climate Impact Research)

Presenter: SAUER, Pascal (Potsdam Institute for Climate Impact Research)

Session Classification: Reproducible Packaging

Track Classification: Research Software: Reproducible Packaging