A community-driven cookbook of recipes for analysing global ocean-sea ice model output

by the **Consortium for Ocean and Sea Ice Modelling in Australia**

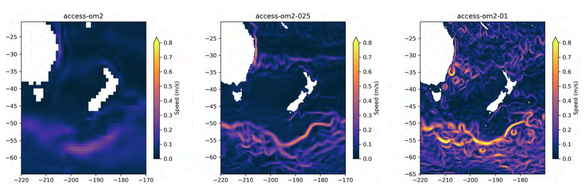
cosima.org.au

Motivation

Running climate models costs a lot! $$$$
How can we maximise the benefit of a model run?
How can we collaborate?

What is COSIMA?

- Consortium for Ocean and Sea Ice Modelling in Australia
- Formed in 2012 (∼10 people); now >150 people
- **Culture is key.** *Shared models, data and tools build community.*
- Developed **ACCESS-OM2** (MOM5-CICE5) global ocean–sea ice model suite at 1°, 0.25°, and 0.1° resolution
  - Adopted in climate models (ACCESS-CM2, ACCESS-CM2-025) and Australian Bureau of Meteorology ocean forecasts (Bluelink)
  - 550TB of shared output data, easily explored and analysed; 200 users
  - Underpinned >70 papers since 2019 (>1100 total citations)
  - >50 ongoing research projects



ACCESS-OM2 global ocean-sea ice model suite at 1°, 0.25°, and 0.1° lateral resolution

Collaborative ocean modelling within COSIMA

- Model runs are very expensive & output data is valuable for many research questions
- For example, a global simulation at 0.1° with biogeochemistry:
  - ∼15TB memory
  - >12,000 cores (>250 Cascade Lake nodes × 48 CPUs)
  - Up to 12hr (∼150,000 core hours) per simulated year
  - Runs for months on Gadi (Australia's HPC) to simulate 60 years
  - Costs $\mathcal{O}(10^7)$ core hours
  - Generates ∼40TB of output
- Few groups in Australia have resources to run these models, so in COSIMA we cooperate:
  - **Piggyback:** save outputs for multiple projects in each model run
  - **Share all data freely** on Gadi and via THREDDS: >550TB of shared data so far
  - How to make this useful to people?

The COSIMA community

COSIMA supports a diverse set of users:

- Most want to analyse output, not run models
  - so we **seek requests for diagnostics** to output from runs, and **make data discoverable, accessible and computable** via the **COSIMA Cookbook**
- Many want to run standard executables with configuration changes (perturbation experiments)
  - so we supply **12 standard configurations**, with input files and precompiled executables
- A few want to make code changes
  - easy! `git clone` downloads all model source code, with 1-line build ⟹ contribute source code improvements via pull requests with automated tests
- Core team of developers
  - model development, maintenance, performance, porting, bug fixes, PR reviews, etc

All of COSIMA's work is open source!

`github.com/COSIMA`

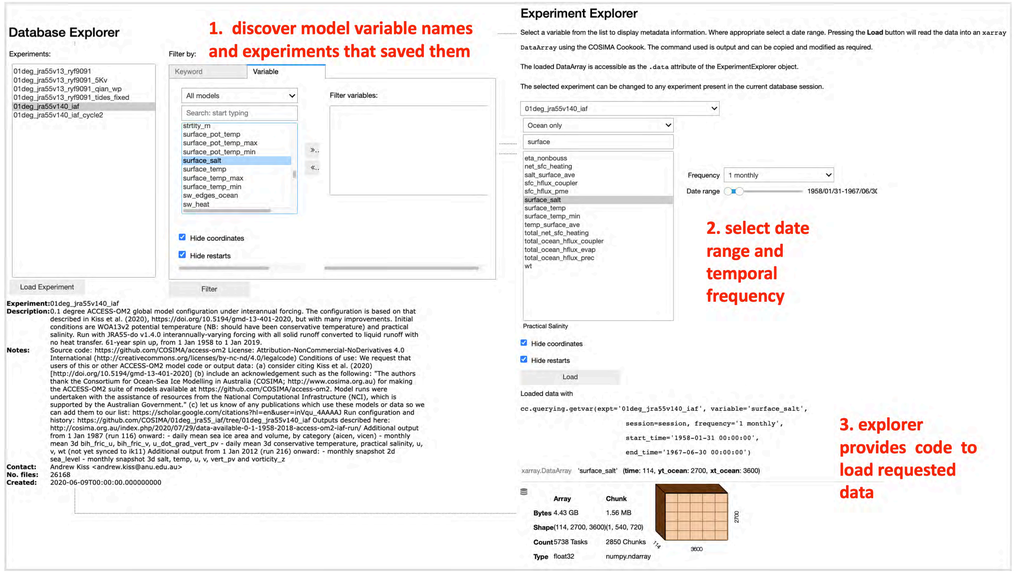
GitHub

COSIMA *is* a Community

COSIMA Annual Workshop 2022 in Hobart, Tasmania; only 2 out of >150 people are funded by COSIMA

Easy data discovery with Database Explorer

- Names of variables for the data you need
- Experiments that saved that variable
- Spatial & temporal resolution and time period of data
- Code needed to access the required data



Data: discoverable, accessible, computable

- **Discoverable**: allow everybody to find what they need without knowing technical details
- **Accessible**: easy to obtain and understand data, with provenance connecting it to all details of model configuration that created it
- **Computable**: systems to enable rapid calculation on huge datasets

Analyze data in-place

- Don't want to download 10s of TB
- Can't process on a local machine anyway; won't fit in memory
- Analyze data in-situ and in parallel using:
  - `xarray`: python package to use netCDF metadata to combine and subset individual files without loading data until it actually needed: lazy evaluation
  - `dask`: python package to process big computations in parallel without needing to fit it all in memory

What the COSIMA ocean data looks like

- 100s of experiments
- 20-100 output variables in each experiment
- u, v, w, temperature, salinity, sea surface height, 10 biogeochemical tracers, & many online-calculated diagnostics
- Large experiments may contain >100,000 files, tens of TB
- Each experiment may have 100s of run directories (due to queue limits)
- Each run directory may contain hundreds of output NetCDF files

In summary: quite confusing!

Abundance of output!

Climate models produce *a lot* of data! No need to know where each file is! `cosima-cookbook` remembers for you!

COSIMA Cookbook & Recipes

xarray dask

COSIMA Cookbook Python package
An experiment database that knows where all the files live

- Open-source Python package `github.com/COSIMA/cosima-cookbook`
- Framework for indexing and querying ocean-sea ice model output
- SQLite database of metadata from all experiments, updated nightly
- **Users don't need to know file names or directory structure**
- Calls `xarray` and returns a `dask` dataset with the requested output
- **GUI database explorer** to browse available experiments and variables

During a "Finding Nemo"-themed hackathon where we updated the COSIMA Recipes; Sep. 2023

COSIMA Recipes
Shared, community-contributed notebooks for model data analysis

- Excellent resource for new researchers/students to get them up to speed in 1-2 days!
- Community can learn from and teach one another
- `github.com/COSIMA/cosima-recipes`

Acknowledgments

- Australian Research Council grants LP160100073 & LP200100406
- Computing and storage resources supplied by Australia's National Computational Infrastructure (NCI)
- Supported by ACCESS-NRI, enabled by the Australian Government through the National Collaborative Research Infrastructure Strategy