

# Spinup-Evaluation

*A Physics-Based Benchmarking Platform for ML-Based Spinup of  
the NEMO Ocean Model*

Matt Archer<sup>1</sup> Isaac Akanho<sup>1</sup> Surbhi Goel<sup>1</sup>

Simon Sadler<sup>1</sup> Etienne Meunier<sup>2</sup> Guillaume Gachon<sup>2</sup>  
Julie Deshayes<sup>2</sup>

2026-02-24

<sup>1</sup>*Institute of Computing for Climate Science (ICCS), University of Cambridge*

<sup>2</sup>*Institut Pierre-Simon Laplace (IPSL)*



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Spinup-evaluation

## Accelerating Spinup



UNIVERSITY OF  
CAMBRIDGE

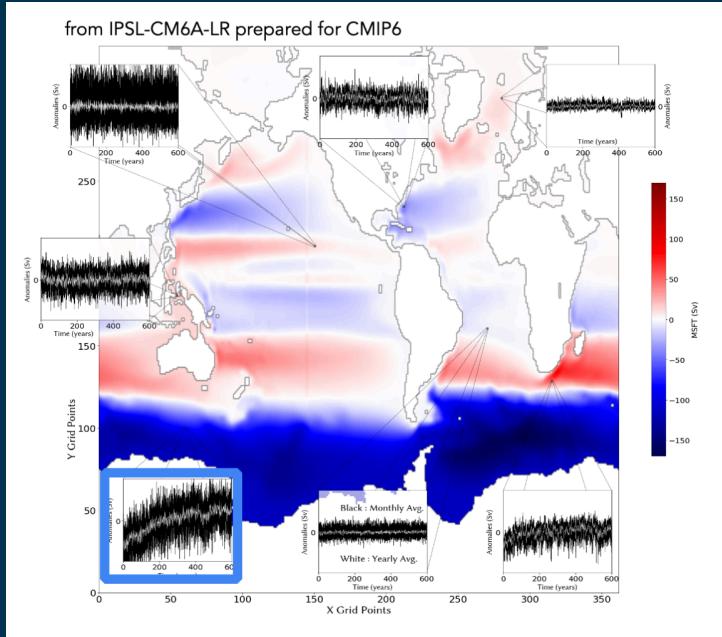


Institute of  
Computing for  
Climate Science

# Spinup-evaluation

## Accelerating Spinup

- Ocean spinup is **slow**. Upwards of a 1000 simulation years to achieve equilibrium state!



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Spinup-evaluation

## Accelerating Spinup

- Ocean spinup is **slow**. Upwards of a 1000 simulation years to achieve equilibrium state!
- Many machine learning emulators exist for ocean modelling that are well validated:
  - How do we know if they are correct *physically*?
  - Errors can easily propagate during rollout.



# Spinup-evaluation

## Accelerating Spinup

- Ocean spinup is **slow**. Upwards of a 1000 simulation years to achieve equilibrium state!
- Many machine learning emulators exist for ocean modelling that are well validated:
  - How do we know if they are correct *physically*?
  - Errors can easily propagate during rollout.
- Our focus is on the physical **evaluation** of ocean models:
  - Including **equation resolving**, pure **AI** emulators and **hybrid** approaches.



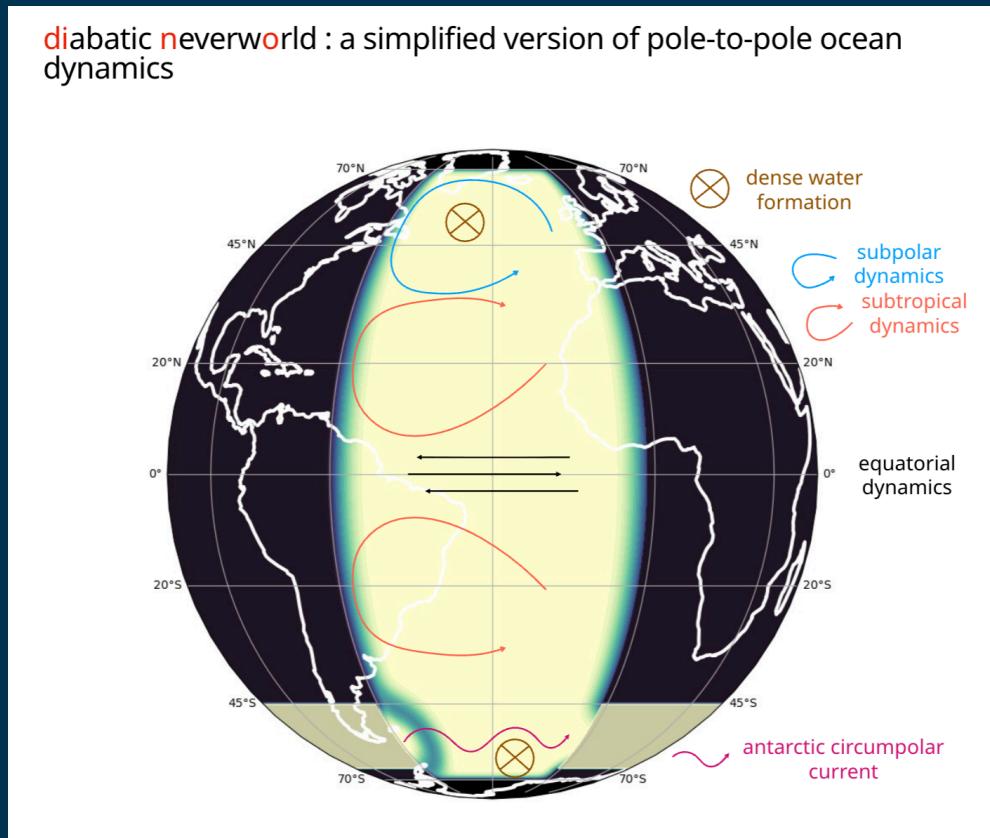
# Spinup-evaluation

## Accelerating Spinup

- Ocean spinup is **slow**. Upwards of a 1000 simulation years to achieve equilibrium state!
- Many machine learning emulators exist for ocean modelling that are well validated:
  - How do we know if they are correct *physically*?
  - Errors can easily propagate during rollout.
- Our focus is on the physical **evaluation** of ocean models:
  - Including **equation resolving**, pure **AI** emulators and **hybrid** approaches.
- Implement sustainable software that is easy to extend, robust and reusable.
  - Generalised evaluation benchmark for climate.



# NEMO/DINO ocean model



[/vopikamm/DINO](https://vopikamm/DINO)



We use the DINO configuration ([Kamm, Deshayes, and Madec 2025](#)) for development purposes.

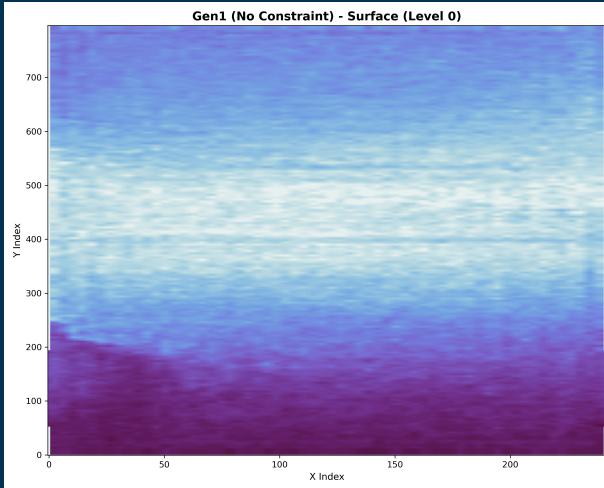


UNIVERSITY OF  
CAMBRIDGE

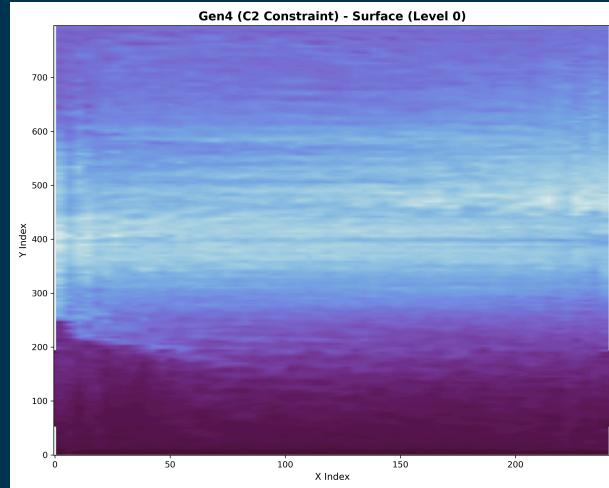


Institute of  
Computing for  
Climate Science

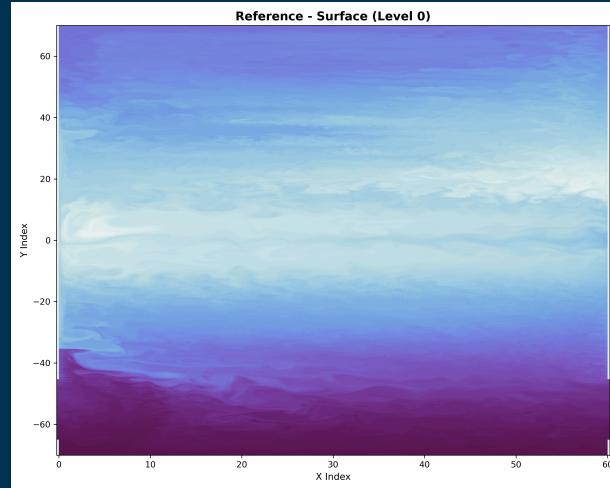
# Is RMSE enough?



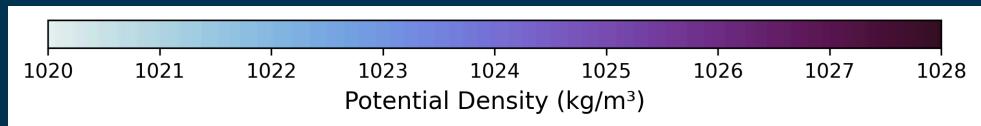
Unconstrained



✓ Constrained



Reference



*Unconstrained      Constrained      Reference*

	<i>Unconstrained</i>	<i>Constrained</i>	<i>Reference</i>
$T_{DW}$	2.9	2.6	2.6
$\rho$ error	26.8	0.4	0.4

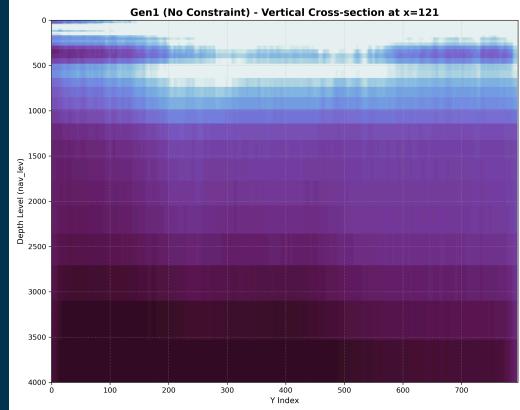
Evaluate even when resolution changes.



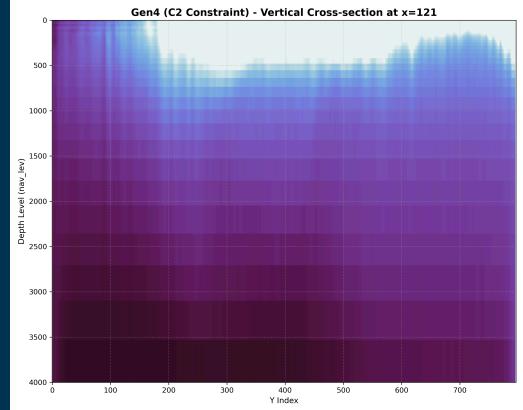
UNIVERSITY OF  
CAMBRIDGE



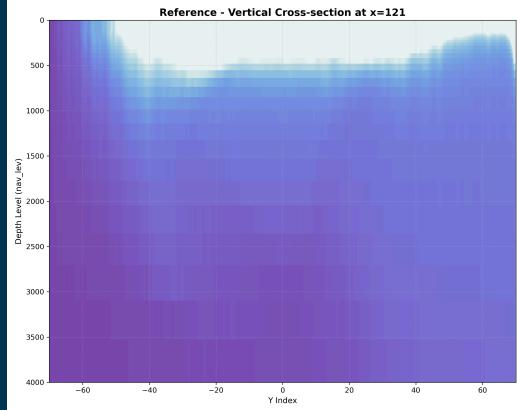
# Is RMSE enough?



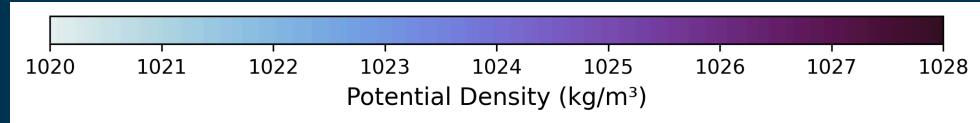
Unconstrained



Constrained



Reference



	<i>Unconstrained</i>	<i>Constrained</i>	<i>Reference</i>
--	----------------------	--------------------	------------------

$T$ DW	2.9	2.6	2.6
$\rho$ error	26.8	0.4	0.4

Evaluate even when resolution changes.



UNIVERSITY OF  
CAMBRIDGE



# Physical evaluation

## Metrics

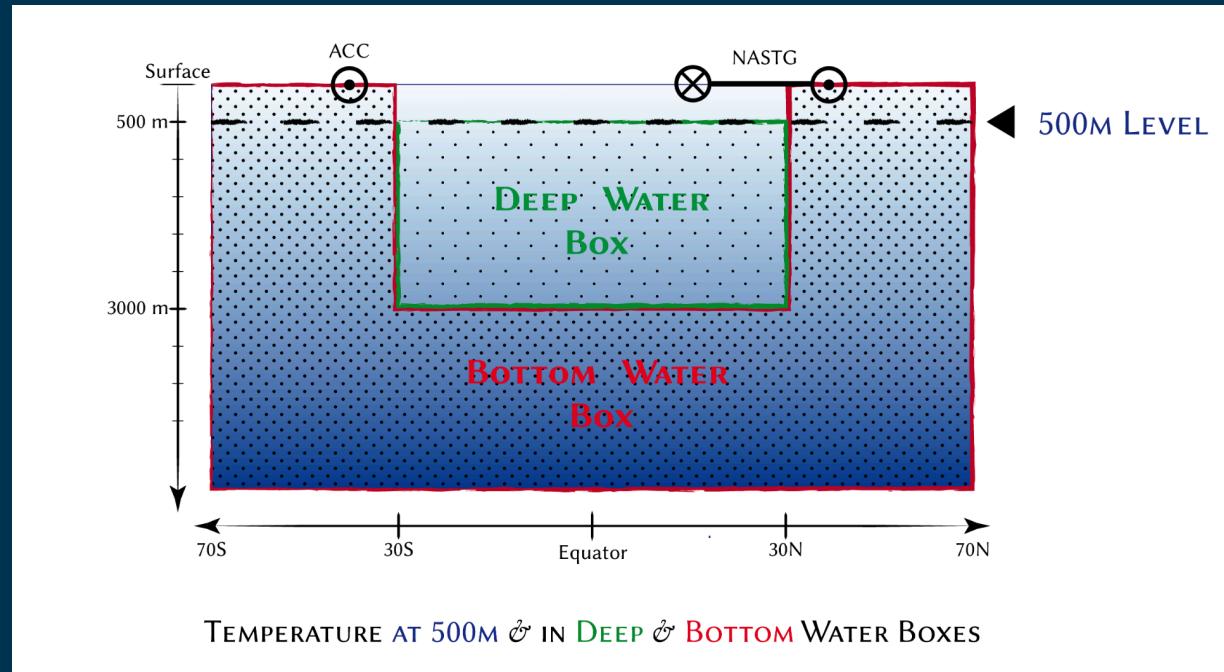


Figure from *Learning to generate physical ocean states*. Meunier et al. (2025)

Bring in physical knowledge of problem  
i.e. metrics specific to tracking **equilibrium** progress.

- *No ground truth*



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Evaluation-Benchmark



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Evaluation-Benchmark

Ocean  
Model



AI emulator or traditional time  
integrator of physical quantities.

AI Emulator  
/ Hybrid



[/m2lines/nemo-  
spinup-forecast](https://m2lines/nemo-spinup-forecast)



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Evaluation-Benchmark

Ocean  
Model



AI emulator or traditional time  
integrator of physical quantities.

AI Emulator  
/ Hybrid



[/m2lines/nemo-  
spinup-forecast](https://m2lines/nemo-spinup-forecast)

Physical Evaluation

- Physical evaluation of  
instantaneous and temporal  
fields [GridT, restart].nc



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Evaluation-Benchmark

Ocean  
Model



AI emulator or traditional time  
integrator of physical quantities.

AI Emulator  
/ Hybrid



[/m2lines/nemo-spinup-forecast](https://m2lines/nemo-spinup-forecast)

Translator  
Toolbox

- Create **restart** state from AI emulator
- Simple checks
- ‘Glue’ module for future time-stepping

Physical Evaluation

- Physical evaluation of instantaneous and temporal fields [**GridT, restart**].nc



# Evaluation-Benchmark

- How do we evaluate hybrid ocean emulators?

[DOI/m2lines/spinup-evaluation](https://doi.org/10.5281/m2lines/spinup-evaluation)



version 0.2

Pillars 1, 2 and 3 enabled with the use of two software components:



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Evaluation-Benchmark

- How do we evaluate hybrid ocean emulators?
- PILLAR 1:
  - Compute metrics at **end** time only.
  - Compute during **roll-out**

[DOI/m2lines/spinup-evaluation](https://doi.org/10.5281/m2lines/spinup-evaluation)



version 0.2

Pillars 1, 2 and 3 enabled with the use of two software components:



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Evaluation-Benchmark

- How do we evaluate hybrid ocean emulators?
- PILLAR 1:
  - Compute metrics at **end** time only.
  - Compute during **roll-out**
- PILLAR 2:
  - Feed emulated state back into ocean model to check **stability** and
    - evaluate metrics over time history

<https://github.com/m2lines/spinup-evaluation>



version 0.2

Pillars 1, 2 and 3 enabled with the use of two software components:



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Evaluation-Benchmark

- How do we evaluate hybrid ocean emulators?
- PILLAR 1:
  - Compute metrics at **end** time only.
  - Compute during **roll-out**
- PILLAR 2:
  - Feed emulated state back into ocean model to check **stability** and
    - evaluate metrics over time history
- PILLAR 3:
  - Sample from a **distribution** of possible emulated states

Pillars 1, 2 and 3 enabled with the use of two software components:

<https://github.com/m2lines/spinup-evaluation>



version 0.2



UNIVERSITY OF  
CAMBRIDGE

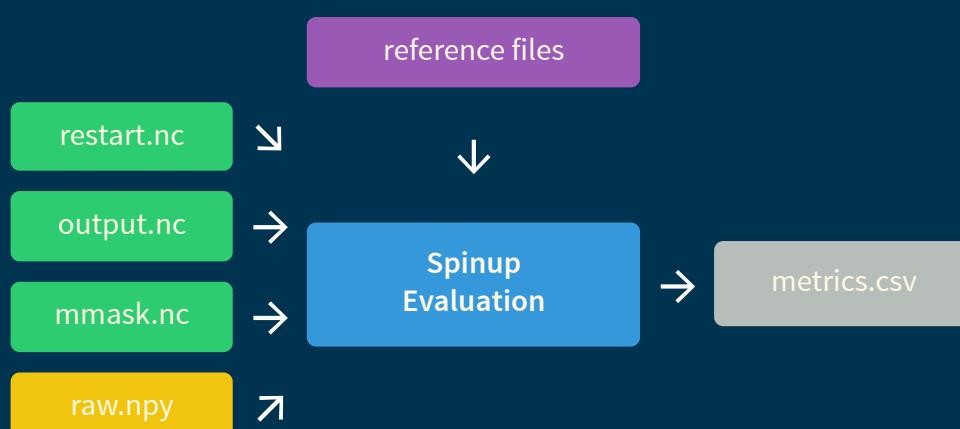


Institute of  
Computing for  
Climate Science

# Developer highlights

## Spinup-evaluation (1)

- ↳ Computes metrics on raw emulator outputs (`.npy`), instantaneous (`restart.nc`) and time averaged (`grid.nc`) output files & **statistical comparisons**
- ⌚ Strict mapping of ocean variables to `restart`, `grid` (output) and `mesh_data` names
- Provides a framework to easily add **new metrics**
- Implements temporal downsampling – all possible metrics computed even if data output at different cadences.



[Q/m2lines/spinup-evaluation](#)



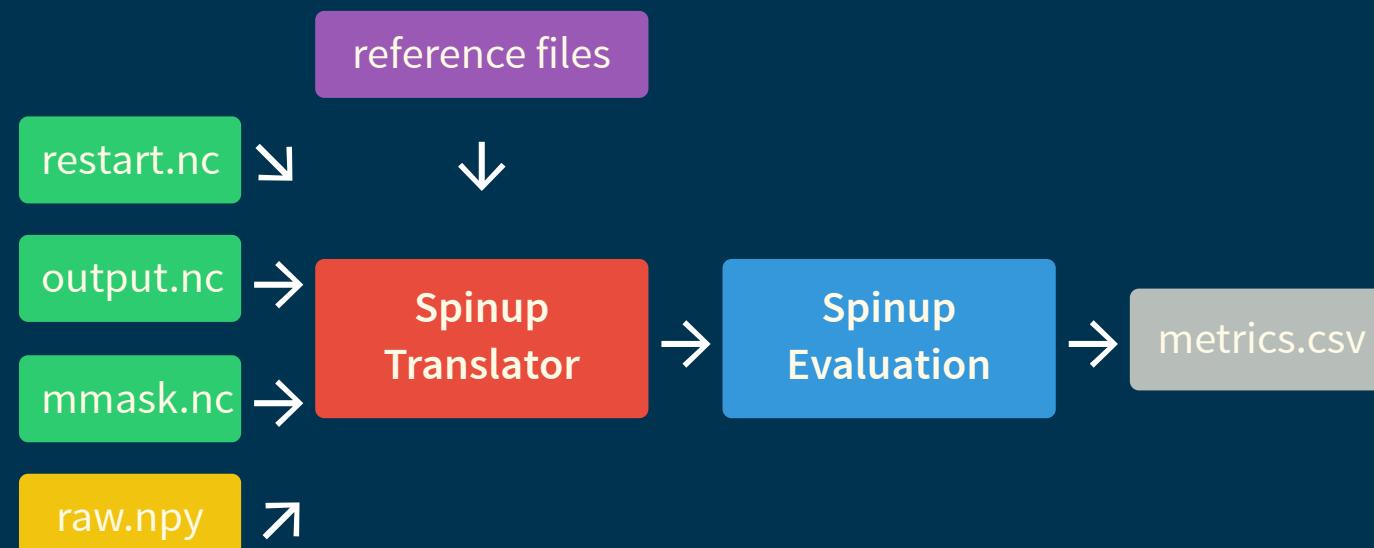
version 0.2



# Developer highlights

## Spinup-restart/translator (2)

- Take an **emulator** state and converts to NEMO compatible **restart.nc** file
- Basic validation gate; does **grid encoding** and **sanity checks** (existence of **NaNs** etc)
- *Regrid, upscale, downscale* resolution;
- Encode other **physical assumptions** (geostrophic velocities)



# Developer highlights

## Spinup-restart/translator (2)

- Take an **emulator** state and converts to NEMO compatible **restart.nc** file
- Basic validation gate; does **grid encoding** and **sanity checks** (existence of **NaNs** etc)
- *Regrid, upscale, downscale* resolution;
- Encode other **physical assumptions** (geostrophic velocities)



Translate between other ocean model restart files - potentially difficult due to grids (in progress..)



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Software engineering



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Software engineering

## Modular & Generalizable

- Reusable components across domains
- Extensible to new models (e.g., MOM6,  
*Oceananigans*)



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Software engineering

## Modular & Generalizable

- Reusable components across domains
- Extensible to new models (e.g., MOM6,  
*Oceananigans*)

## Quality & Standards

- Robust testing (Unit, Integration, Coverage) with CI
- Well documented end-to-end examples (hybrid)
- Data products adhere to FAIR (see [Wilkinson et al. 2016](#))
- **Collaborative Design:** Bridging Science & RSE



UNIVERSITY OF  
CAMBRIDGE

Institute of  
Computing for  
Climate Science

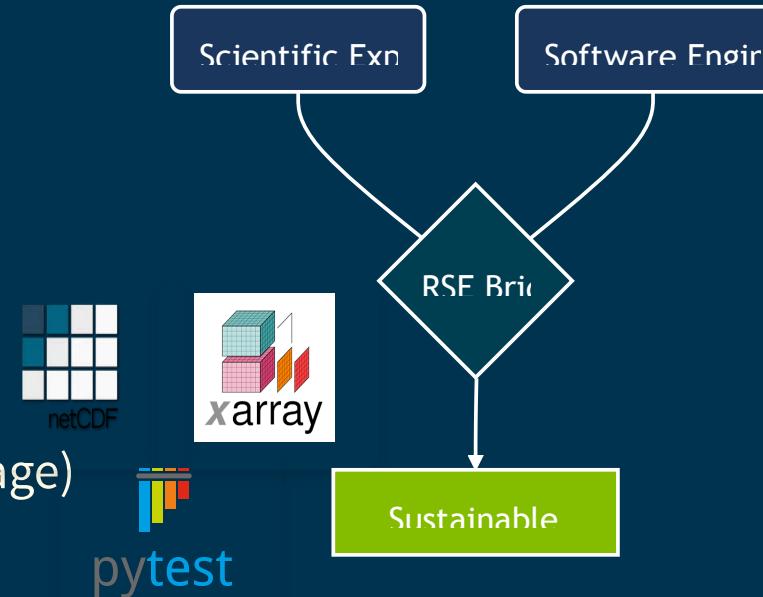
# Software engineering

## Modular & Generalizable

- Reusable components across domains
- Extensible to new models (e.g., MOM6, *Oceananigans*)

## Quality & Standards

- Robust testing (Unit, Integration, Coverage) with CI
- Well documented end-to-end examples (hybrid)
- Data products adhere to FAIR (see Wilkinson et al. 2016)
- Collaborative Design: Bridging Science & RSE



Eliminating silos through cross-disciplinary partnership.



# Conclusions

## Key Takeaways



UNIVERSITY OF  
CAMBRIDGE

Institute of  
Computing for  
Climate Science

# Conclusions

## Key Takeaways

### 1. Extend evaluation of ocean models:

- Evaluation of AI emulators must go beyond simple RMSE metrics.
- Enable evaluation in conjunction with equation resolving techniques.



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Conclusions

## Key Takeaways

### 1. Extend evaluation of ocean models:

- Evaluation of AI emulators must go beyond simple RMSE metrics.
- Enable evaluation in conjunction with equation resolving techniques.

### 2. Software enabled research:

- Research software engineering is a critical component of research bringing generalisability, extensibility and robustness to software products.



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# Conclusions

## Key Takeaways

### 1. Extend evaluation of ocean models:

- Evaluation of AI emulators must go beyond simple RMSE metrics.
- Enable evaluation in conjunction with equation resolving techniques.

### 2. Software enabled research:

- Research software engineering is a critical component of research bringing generalisability, extensibility and robustness to software products.

## Next Steps

- Extend to other ocean models and AI emulators
- Translation between different ocean checkpoint files



# Thanks for Listening

Get in touch:

 Matt Archer

 [ma595.github.io](https://ma595.github.io)

 [ma595@cam.ac.uk](mailto:ma595@cam.ac.uk)

 [iccs@cam.ac.uk](mailto:iccs@cam.ac.uk)

 [ma595](https://github.com/ma595)

 [/ma595/spineval-agu - Slides](https://github.com/ma595/spineval-agu)



 [/m2lines/spinup-evaluation](https://github.com/m2lines/spinup-evaluation)



The ICCS received support from Schmidt Sciences



UNIVERSITY OF  
CAMBRIDGE



Institute of  
Computing for  
Climate Science

# References

- Gorce, Blandine, Luther Ollier, David Kamm, and Etienne Meunier. 2025. “Physically Consistent Sampling for Ocean Model Initialization.” In *Advances in Neural Information Processing Systems*. Vol. 38. <https://neurips.cc/virtual/2025/loc/san-diego/poster/126936>.
- Kamm, D., J. Deshayes, and G. Madec. 2025. “DINO: A Diabatic Model of Pole-to-Pole Ocean Dynamics to Assess Subgrid Parameterizations Across Horizontal Scales.” *EGUsphere* 2025: 1–26. <https://doi.org/10.5194/egusphere-2025-1100>.
- Meunier, Etienne, David Kamm, Guillaume Gachon, Redouane Lguensat, and Julie Deshayes. 2025. “Learning to Generate Physical Ocean States: Towards Hybrid Climate Modeling.” arXiv. <https://doi.org/10.48550/arXiv.2502.02499>.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>.

