

# Oceananigans.jl: A model that achieves breakthrough resolution, memory and energy efficiency in global ocean simulations

Simone Silvestri<sup>1</sup>, Gregory Wagner<sup>1</sup>, Christopher Hill<sup>1</sup>, Matin Raayai Ardakani<sup>2</sup>, Johannes Blaschke<sup>3</sup>, Jean-Michel Campin<sup>1</sup>, Valentin Churavy<sup>1</sup>, Navid Constantinou<sup>4</sup>, Alan Edelman<sup>1</sup>, John Marshall<sup>1</sup>, Ali Ramadhan<sup>1</sup>, Andre Souza<sup>1</sup>, and Raffaele Ferrari<sup>1</sup>

<sup>1</sup>Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>2</sup>Northeastern University, Boston, MA, USA

<sup>3</sup>Lawrence Berkeley National Laboratory, Berkeley, CA, USA

<sup>4</sup>Australian National University, Canberra, ACT, Australia

March 2023

## Abstract

Climate models must simulate hundreds of future scenarios for hundreds of years at coarse resolutions, and a handful of high resolution decadal simulations to resolve localized extreme events. Using Oceananigans.jl, written from scratch in Julia, we report several achievements: First, a global ocean simulation with breakthrough *horizontal resolution* — 488m — reaching 15 simulated days per day (0.04 simulated years per day; SYPD). Second, Oceananigans simulates the global ocean at 488m with breakthrough *memory efficiency* on just 768 Nvidia A100 GPUs, a fraction of the resources available on current and upcoming exascale supercomputers. Third, and arguably most significant for climate modeling, Oceananigans achieves breakthrough *energy efficiency* reaching 0.95 SYPD at 1.7 km on 576 A100s and 9.9 SYPD at 10 km on 68 A100s — the latter representing the highest horizontal resolutions employed by current IPCC-class ocean models. Routine climate simulations with 10 km ocean components are within reach.

## 1 Justification

Oceananigans.jl — a new ocean model written from scratch in Julia — achieves ocean simulations with breakthrough resolution, memory and energy efficiency, realizing 0.041 simulated years per day (SYPD) at 488 m on 768 Nvidia A100s, 0.95 SYPD at 1 km on 576 A100s, and 9.9 SYPD at 10 km on 68 A100s.

## 2 Performance Attributes

Categories	Scalability, time-to-solution, energy-to-solution.
Type of method	Fully explicit with sub-cycling.
Results basis	Whole application excluding I/O.
Numerical precision	Both 64- and 32-bit cases measured.
System scale	Results measured on full-scale systems.
Measurement mechanism	Timers, memory used and energy used.

### 3 Overview of the Problem

Climate models are essential for predicting where, when, and how climate change threatens Earth’s ecosystems and human civilization. But current climate models, which capture only the broadest aspects of global warming, fall far short of providing the needed accuracy and granularity required to design and implement costly adaptation and mitigation strategies [14]. Significant reduction of the uncertainty of climate predictions is potentially worth trillions of dollars [20].

Climate models simulate the three-dimensional fluid dynamics, thermodynamics, chemistry, and biology of the atmosphere, ocean, and land to predict the hydrological cycle, carbon cycle and the net energy imbalance of the Earth system. While typical climate models use coarse resolutions of 25-100 km to simulate the numerous climate scenarios required by the Intergovernmental Panel on Climate Change (IPCC) [25], a handful of state-of-the-art climate simulations have been performed at higher resolutions of O(10 km) at astronomical expense. At either resolution there are many processes, such as clouds and ocean turbulence, that cannot be explicitly simulated and are instead approximated by empirical formulae called *parameterizations*. Biases due to inadequate parameterizations dominate the uncertainty of climate predictions over the next few decades [34, 23].

The prevailing strategy to reduce climate model uncertainty is to refine model resolution as much as possible [34]. For example, at horizontal resolutions of 1 km a substantial fraction of atmospheric convection and ocean turbulence are explicitly modeled by Newton’s laws of motion, greatly reducing the impact of parameterizations [34]. High-resolution climate modeling is further required to make predictions for specific regions, providing information for local decision makers on adaptation and mitigation [14].

Yet the “resolution strategy” is fundamentally limited: even at 1 km resolution many climate-relevant physical processes remain unresolved [49]. Worse, processes such as sea ice dynamics, biology, or cloud-aerosol interaction will never be resolved because accurate macroscopic laws do not exist. Absent theoretical breakthroughs, such “irreducible” uncertainties can be addressed only by leveraging Earth system observations through advances in data assimilation and machine learning [41]. Data-driven optimization of climate models requires *ensembles* of climate predictions, rather than single predictions at the highest affordable resolution. Ensembles of simulations are also required to explore emission scenarios and to estimate the impact of initial condition uncertainty and internal variability.

Consequently, reducing the uncertainty of climate predictions demands not *just* higher resolution, but *more efficient resource utilization* to enable hundreds to thousands of relatively high-resolution simulations. As an example, we consider the computational requirements to enable 100-simulation ensembles using all 37,888 AMD MI250 GPUs of the Frontier exascale supercomputer: completing an ensemble of 300-year simulations (200 years of spin-up + 100 years of prediction) within one month of wall clock time requires a climate model that can achieve 10 simulated years per day (SYPD) using 378 GPUs, or 1/100th of Frontier’s resources. Disruptive progress on climate modeling requires not just scalable performance for a single, high-resolution simulation, but advances in *efficiency* to meet this ensemble-based “10 per 100th” benchmark [40].

Our submission uses the ocean component of a new climate model being developed by the Climate Modeling Alliance [44]. The ocean contributes key uncertainty to climate predictions due to its prominent role in the Earth system’s heat and carbon cycles. At 10 km resolutions, where ocean model uncertainties are significantly reduced, the ocean often is the most expensive climate model component [19]. This calls for a step change in ocean model performance.

## 4 Current State of the Art

We are aware of only three global ocean simulations that have achieved resolutions finer than 5 kilometers — all at tremendous computational expense. In 2014, MITgcm [29] was used to perform the one year, tidal-forced ice-ocean simulation “LLC4320” [45], which exhibits 2.2 km horizontal resolution with 90 vertical levels. LLC4320 achieved 0.047 simulated years per day (SYPD) using 70,000 cores of the NASA Pleiades system.

FIO-COM32 [50] ran at  $\sim$ 2.5 km ( $1/32^{\text{nd}}$  degree) horizontal resolution with 90 vertical levels for 3.5 years. [48] ported LICOM3 to GPUs to realize 0.51 SYPD at  $1/20^{\text{th}}$  horizontal resolution with 60 vertical levels using 384 MI50 AMD GPUs, and further managed to scale to 26200 MI50s with strong scaling efficiency of 8%.

The largest ocean simulations used in current IPCC-class climate models, which typically require faster-time-to-solution to support longer simulations, have horizontal resolutions of roughly 10 km. [8] describes output from four 60-year ocean simulations following the OMIP-2 protocol with 8 km ( $1/12^{\text{th}}$  degree), 10 km, and two with 11 km ( $1/10^{\text{th}}$  degree). [11] report a 110-year simulation at 10 km ( $1/10^{\text{th}}$  degree) horizontal resolution, the longest high resolution OMIP-2-style run. Some of the highest resolution climate models are the iHESP CESM-based model with 25km-10km atmosphere-ocean resolution [51], achieving 3.4 SYPD, and the 50km-10km HadGEM3-GC3.1 submission to HighResMIP [18, 37], achieving 0.4 SYPD.

At 3.4 SYPD, the iHESP CESM achieves sufficient time-to-solution for hundreds to thousands of years of simulated climate. But such a simulation is purchased for a high price, requiring the 40% of the Sunway TaihuLight supercomputer [51] and 4 million cores consuming 6 MW for hundreds of days of wall clock time. Enabling the large ensembles of high-resolution simulations needed to improve climate prediction requires both performance at scale as well as efficient *resource utilization*.

Figure 1 plots simulated years per *mega-watt-hour* (SYPMWh) against resolution for state-of-the-art ocean models. The SYPMWh metric encodes the efficiency requirement needed to make progress on climate uncertainty with next-generation climate models: in particular, we require both higher-resolution models (moving rightwards in figure 1) as well as more efficient models (moving upwards in figure 1). For completeness, we report SYPMWh also for two GPU-based models: Veros [22], an ocean model, and COSMO [15], an atmospheric model. The present nomination is shown with stars from whence we see significant performance gains compared to the existing state-of-the art.

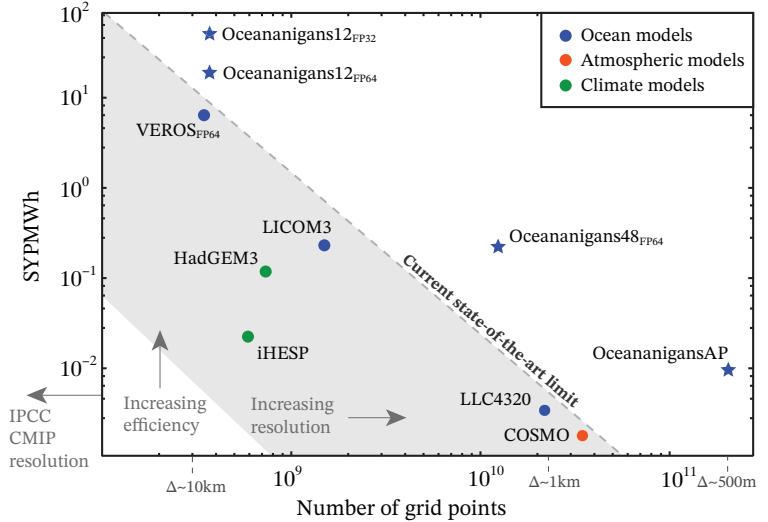


Figure 1: Simulated years computed by a megawatt-hour of energy (SYPMWh) versus number of grid points for state-of-the-art atmosphere and ocean models. Stars show the performance of our ocean model in a realistic and “aqua planet” (AP) setup.

## 5 Innovations

Our achievement is three-fold: first, using new software written in the Julia programming language called Oceananigans.jl [35], we report a near-global ocean simulation with highest-ever horizontal resolution (488 m) reaching 15 simulated days per day (0.04 SYPD). Second, Oceananigans performs this simulation with breakthrough *memory efficiency* on just 768 NVidia A100 GPUs, and thus a fraction of the available resources on current and upcoming exascale supercomputers. Third, and arguably most important, Oceananigans achieves breakthrough *energy efficiency*, simulating the global ocean at 0.95 SYPD with 1.7 km resolution on 576 A100s, and at 10 km — the highest horizontal resolution employed by an IPCC-class ocean model — achieving 9.9 SYPD on 68 Nvidia A100s. This final milestone proves the feasibility of *routine* climate simulations with 10 km ocean components, a crucial resolution threshold at which ocean macroturbulence (the most energetic ocean motions with scales between 10–100 km) is fully resolved.

We attribute these achievements first and foremost to a high-risk, high-reward strategy to develop a new ocean model from scratch in Julia with a specific focus on GPU performance and memory efficiency. Additional crucial ingredients include advances in numerical methods for finite volume fluid dynamics on the sphere and a novel optimization for simulating ocean free surface dynamics that achieves unprecedented GPU scalability.

### 5.1 Starting from scratch with Julia

Oceananigans.jl is an open-source library for ocean-flavored fluid dynamics written from scratch in Julia [7]. Julia is a dynamic high-level programming language that leverages Just-In-Time (JIT) compilation and LLVM [24] to achieve performance competitive with traditional HPC languages like C or Fortran. Julia has gathered interest as potential language for HPC [17, 10, 16, 21, 27] and provides easy integration with MPI [47, 38]. Most of Oceananigans.jl software is hardware-agnostic through the Julia package KernelAbstractions.jl [10], which enables performance portability targeting CPUs and different GPU vendors using the JuliaGPU [6, 5] software stack, similar to the capabilities provided by Kokkos [9], OCCA [30], and HIP [1].

To our knowledge, Oceananigans is the first ocean model written from scratch for GPUs, rather than ported from existing CPU code. Starting from scratch and using the Julia programming language allowed us to rethink the typical patterns used in ocean and atmosphere dynamical cores. In particular, we developed a system of composable atomic operators that leverages Julia’s functional programming paradigm and effective inlining capabilities to recursively construct large expression trees for calculus on staggered finite volume grids. Using this composable operator system, we fuse the entire tendency computation for each prognostic variable into a single compute-heavy kernel, each of which depends on only two intermediate diagnostic variables representing hydrostatic pressure and vertical diffusivity (which is treated implicitly using a predictor-corrector method).

Such a high degree of abstraction yields a number of innovations: first, kernel fusion maximizes efficiency on GPUs. Second, almost all intermediate quantities are computed on-the-fly, so that Oceananigans is extremely memory efficient and can perform global ocean simulations at resolutions up to 1/4<sup>th</sup> degree on a single Nvidia V100. Finally, because all compute-heavy kernels rely on a single “tendency kernel function” applied at each grid index  $i, j, k$ , we can easily optimize performance by rapidly prototyping techniques to overlap computation and communication. The sparsity of kernels per time-step and small number of temporary variables mean that Oceananigans’ algorithmic structure is markedly different from current ocean models, which typically allocate 10 to 100 *times*

the minimum necessary memory [4] and distribute computations across many small kernels [51]. We argue these algorithmic differences are a major factor in Oceananigans' energy-efficiency and time-to-solution on GPU systems.

## 5.2 New numerical methods for finite volume fluid dynamics on the sphere

Our results use `Oceananigans.HydrostaticFreeSurfaceModel`, which solves the hydrostatic Boussinesq equations in a finite volume framework on staggered C-grids [3]. Oceananigans' hydrostatic model employs an implicit-explicit second-order Adams-Bashforth time stepping scheme. Vertically-implicit diffusion is implemented with a backward Euler time-discretization and tridiagonal solver.

A major innovation is a new adaptive-order scheme based on weighted essentially non-oscillatory (WENO) reconstructions [42] for advecting momentum and tracers on curvilinear finite-volume grids [43]. This new scheme automatically adapts to changing spatial resolution and permits stable, high-fidelity simulations of ocean turbulence without explicit dissipation or hyper-dissipation. This innovation reduces setup time when changing or increasing resolution while guaranteeing high-fidelity solutions that exhibit the minimum necessary dissipation of sharp, near-grid scale features.

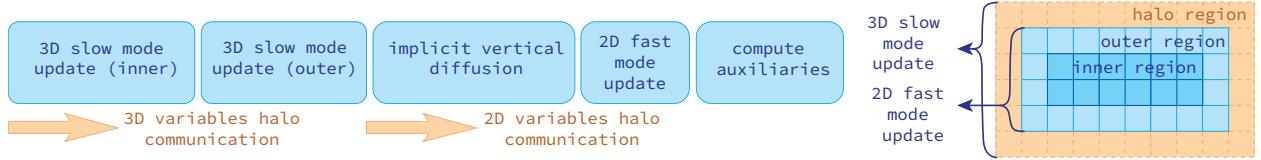


Figure 2: Left: time-stepping sequence. Right: different domains over which 2D fast and 3D slow mode updates take place (here assuming 1 barotropic substep per baroclinic step – halo region of size 1 – and second-order methods – outer region of size 1)

## 5.3 Optimization of ocean free surface dynamics for unprecedented GPU scalability

In hydrostatic ocean models with a free surface, the vertically-averaged, two-dimensional “barotropic mode” has dynamics orders of magnitude faster than the three-dimensional “baroclinic” component, and must be treated by a special “barotropic solver”. Due to communication overhead, barotropic solvers in current ocean models — whether implicit or explicit — are a major bottleneck that accounts for between 40% [22] to 60% [48, 36] of the cost of a typical IPCC-class ocean simulations.

Oceananigans' excellent scalability is enabled by an innovative optimization of the parallel barotropic solver. An increase in computation is traded in for decreased communication latency by leveraging the two-dimensionality of the barotropic problem. Our new barotropic solver is based on explicit subcycling of the barotropic mode. Increasing the width of the barotropic halo to equal the number of explicit subcycles (typically between 10–30) greatly decreases the frequency of communication. As a result, communication is required once per time-step rather than every subcycle, reducing the frequency of communication by a factor of 10 to 30. The cost of the barotropic solver is therefore less than 10% of the total cost of a time step. Due to the sparsity of communication enabled by our novel barotropic solver, all communication operations can be overlapped with computational workloads as sketched in figure 2.

## 6 How performance was measured

The Oceananigans model performance is estimated for two near-global ocean simulations with different domains: a realistic (R) domain and an aqua planet (AP) domain. Both domains span the entire longitudinal extent of the sphere and cover a latitude range of 75°S to 75°N.

The Realistic domain has realistic bathymetry and is forced by realistic surface momentum, heat, and salinity fluxes derived from the ECCO2 state estimate[31] at three resolutions:

- **OceananigansR12** with 1/12<sup>th</sup> degree horizontal resolution ( $\sim 7$  km) and 48 vertical levels
- **OceananigansR24** with 1/24<sup>th</sup> degree horizontal resolution ( $\sim 3.4$  km) and 100 vertical levels
- **OceananigansR48** with 1/48<sup>th</sup> degree horizontal resolution ( $\sim 1.7$  km) and 100 vertical levels

Figure 3 shows surface vertical vorticity after one year integration of **OceananigansR12** and **OceananigansR48** over the global ocean and also for selected regions to show further detail. Both **OceananigansR48** and **OceananigansR12** exhibit macroscale turbulent ocean features that are currently unresolved by most IPCC-class models. The **OceananigansR48** solution exhibits fronts, filaments, and other “submesoscale” vorticity features realized only a handful of times in global simulations.

The idealized **OceananigansAP** suite of simulations [13], which has idealized bathymetry and surface forcing that does not require interpolation to different resolutions, is used for weak scaling experiments. All **OceananigansAP** experiments have 100 vertical levels and two latitudinal ridges that divide the world ocean into two basins. We vary the horizontal resolution of **OceananigansAP** from 1/6<sup>th</sup> of a degree ( $\sim 14$  km) to 1/196<sup>th</sup> of a degree ( $\sim 488$  m).

None of our simulations require explicit horizontal diffusion of momentum or tracers owing to the adaptive WENO advection scheme described in section 5.2. All simulations use a Richardson-number-based parameterization for vertical mixing due to unresolved shear and convective turbulence at 1–100 m scales.

To assess the time-to-solution for each experiment in simulated years per day (SYPD), we measure average wall clock time per time-step. Wall clock time is sampled through NVIDIA’s Nsight System and recorded by NVIDIA Tool Extension Library via the NVTX.jl Julia package.

To assess the efficiency of each solution in simulated years per mega-watt-hour (SYPMWh), we combine SYPD with an estimate of the mean power draw over the duration of an experiment. On MIT Satori [2], which has 256 Nvidia V100s, we have access to precise, billing-grade power metering. For all simulations with Nvidia A100s we estimate power consumption  $P$  with

$$P = 250D + 300N \text{ Watts}, \quad (1)$$

where  $D$  is the number of A100s and  $N$  is the number of nodes.

We further note that power estimates are provided by LICOM3 and COSMO, but not for LLC4320 or Veros. To estimate the power consumption of LLC4320, we assume that each of the 1000 dual CPU nodes draws 500W. We estimate the power consumption of iHESP CESM [51] and HadGCM3 [37] as a percentage of the peak power consumption of their respective clusters. We use equation (1) to estimate Veros’ power consumption on 1 node with 16 A100s.

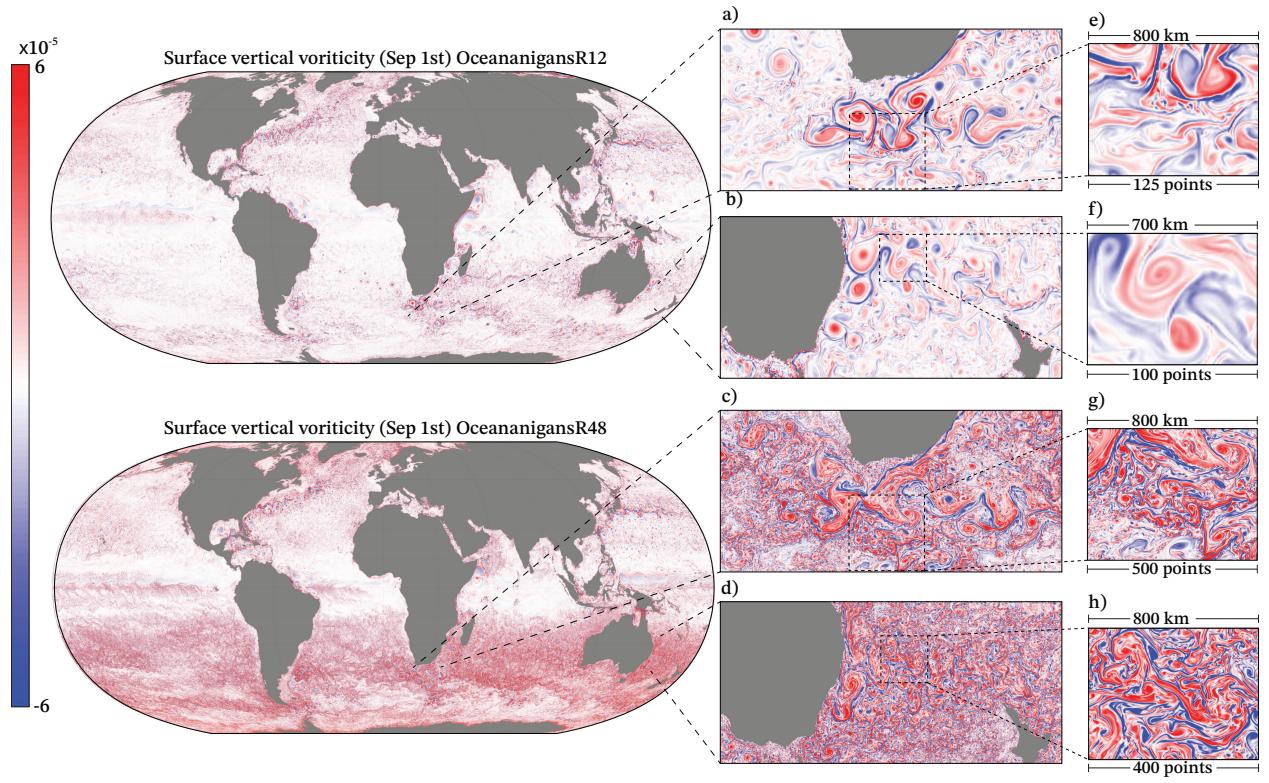


Figure 3: Vertical vorticity as simulated by **Oceananigans12** (top left) and **Oceananigans48** (bottom left) after a one year integration on September 1st. To the right, insets zoom on particularly energetic current systems: the Aghulas and the East Australian Currents. While major ocean currents with widths of 10-100 km are resolved in both simulations, the sharp density fronts and associated currents that develop at the ocean surface in winter at scales between 1-10 km (the ocean weather) are only resolved by **Oceananigans48**. On September 1 — spring in the southern hemisphere, fall in the northern hemisphere — such sharp frontal features populate the southern ocean but are suppressed in the north.

## 7 Performance Results

We report both scaling results via time-to-solution in SYPD and efficiency results via energy-to-solution in SYPMWh.

### 7.1 Scaling Results

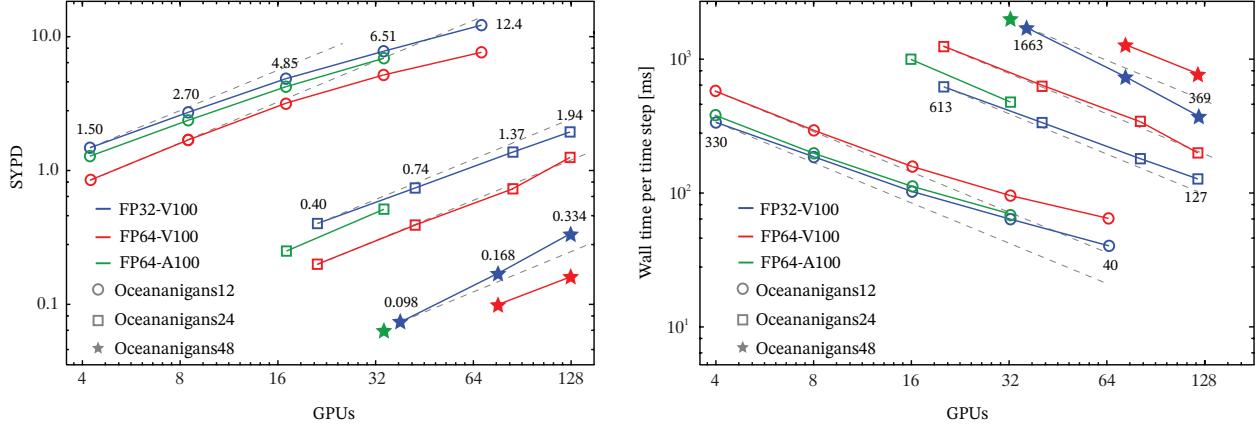


Figure 4: Strong scaling tests for the realistic setups **OceananigansR12** ( $1/12^\circ$ ), **OceananigansR24** ( $1/24^\circ$ ), and **OceananigansR48** ( $1/48^\circ$ ). The left plot reports simulated years per wall clock day (SYPD) while the right plot wall clock milliseconds per time steps. All results are averaged over 1500 time steps.

**Realistic ocean simulations (Satori and Engaging clusters).** We report strong scaling tests using the realistic global setup shown in figure 3 on two clusters: (i) the MIT Satori cluster [2], a high-performance Power 9 system composed of 64 Power 9 nodes hosting four Nvidia V100 GPUs with 32GBs memory each, and (ii) the Engaging MIT cluster, using 8 nodes that host 4 NVlinked A100s with 80GBs memory each. The resulting wall clock time per time step, averaged over 1500 time steps, is presented in Figure 4 for both single precision (FP32) and double precision (FP64) computations. On a single node, **OceananigansR12** attains 0.9 SYPD in double precision and 1.4 SYPD in single precision, with a wall clock time per time step ranging from 330 to 550 ms. When increasing the number of nodes up to 16 (64 GPUs), the communication overhead increases, resulting in 12.4 SYPD in single precision and 7.75 SYPD in double precision. We measure a strong scaling efficiency of 52% in single precision and 55% in double precision over 64 GPUs, because the computational workload (40 ms wall clock time per time-step) eventually becomes too short to completely mask the communication overhead.

For higher-resolution ocean weather-permitting simulations, the scaling is almost ideal across the range we investigate. For **OceananigansR24** (FP64-V100) and **OceananigansR48** (FP32-V100), we measure larger than ideal scaling. This counter-intuitive result is a product of a load balance improvement as the number of GPUs increases. In summary, we attain 1.94 SYPD on 120 V100 GPUs with a kilometer-scale resolution (**OceananigansR24**) and 0.33 SYPD with an ocean weather-resolving simulation (**OceananigansR48**). Finally, we have tested the **OceananigansR48** setup on 144 Perlmutter nodes (576 A100 GPUs), reaching the 0.95 SYPD. This is the *first instance* of a kilometer-scale ocean achieving  $\sim 1$  SYPD. We have also tested the **OceananigansR12** setup on 17 nodes obtaining 9.9 SYPD (see fig. 5).

**Aqua-planet simulation (Perlmutter cluster).** We report weak scaling tests on the NERSC supercomputer (Perlmutter). Perlmutter is a HPE (Hewlett Packard Enterprise) Cray EX super-

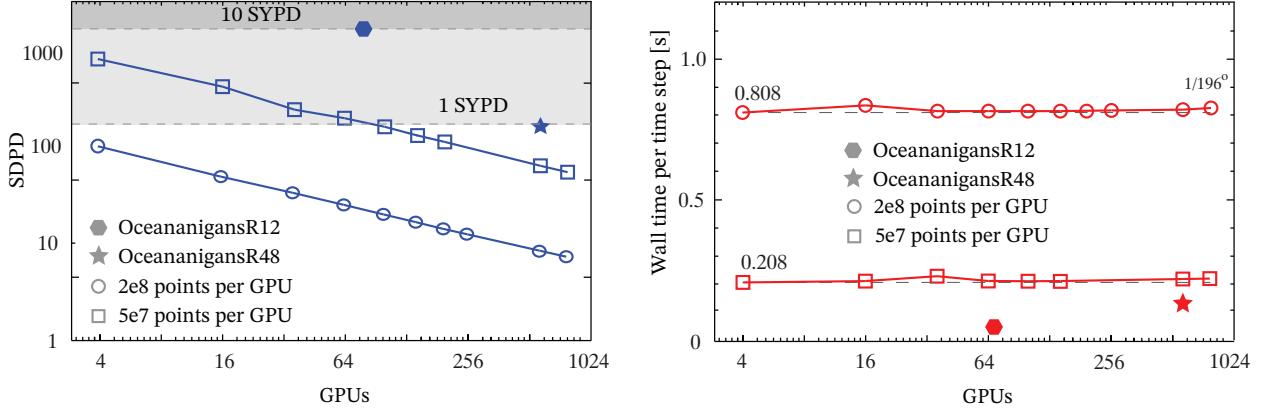


Figure 5: Weak scaling tests performed in double precision with the **OceananigansAP** setup. Each GPU has a grid equivalent to a global  $1/6^\circ$  and 100 vertical layers. The weak scaling is performed up to a horizontal resolution of  $1/168^{\text{th}}$  of a degree ( $\sim 488$  m resolution) where we achieve 15 simulated days per wall clock day (1 year in roughly 25 days). The star marks the performance of **OceananigansR48** (figure 3) on 144 Perlmutter GPU nodes. All results are averaged over 500 time steps.

computer that hosts four A100 GPUs with 40GB per node, linked through a NVLink3 interconnect. All weak scaling tests are performed using the **OceananigansAP** setup on double precision. We allocate two different horizontal resolutions ( $1/12$  and  $1/6$  of a degree), progressively increasing them with the number of GPUs while maintaining 100 vertical levels. As shown in figure 5, we obtain 100% weak scaling efficiency for the whole investigated range (1 to 196 nodes – 4 to 768 A100s).

## 7.2 Energy efficiency

In table 1 we summarize the energy metrics for our computations as well as the other investigated models. Figure 1 is derived from the data outlined in this table. HadGEM3 and iHRES entries are estimated by including the whole coupled climate model (atmosphere and ocean). Unavailable data is marked with -. Our Oceananigans simulations are the highest in each of their columns. This reflects our attention to memory and energy efficiency.

Model	Time step	Grid size	CPU/GPU	SYPD	wtime/tstep	Power est.	In fig 1
HadGEM3 <sub>FP64</sub> (Climate) [37]	-	$\sim 7.22 \times 10^8$	9396 (Cray XC40)	0.4	-	141KW	✓
iHRES <sub>FP64</sub> (Climate) [51]	-	$\sim 6 \times 10^8$	Sunway TaihuLight	3.7	-	6500KW	✓
LLC4320 <sub>FP64</sub> (Ocean)	25 s	$8.7 \times 10^{10}$	2000 (Intel)	0.041	1.6	500KW	✓
Veros <sub>FP64</sub> (Ocean) [22]	180 s	$3.5 \times 10^8$	16 (A100)	0.8	0.62	5.2KW	✓
Veros <sub>FP32</sub> (Ocean) [22]	180 s	$3.5 \times 10^8$	16 (A100)	1.3	0.38	5.2KW	
LICOM3 <sub>FP64</sub> (Ocean) [48]	60 s	$1.5 \times 10^9$	384 (MI50)	0.51	0.32	92KW	✓
LICOM3 <sub>FP64</sub> (Ocean) [48]	60 s	$1.5 \times 10^9$	26200 (MI50)	2.72	0.06	6300KW	
COSMO <sub>FP64</sub> (Atmos) [15]	6 s	$3.46 \times 10^{10}$	4888 (P100)	0.043	0.4	1000KW	✓
Oceananigans <sub>R12</sub> <sub>FP32</sub> (Ocean)	180 s	$3.7 \times 10^8$	4 (V100)	1.5	0.33	1.2KW	✓
Oceananigans <sub>R12</sub> <sub>FP32</sub> (Ocean)	180 s	$3.7 \times 10^8$	64 (V100)	12.4	0.04	18KW	
Oceananigans <sub>R12</sub> <sub>FP64</sub> (Ocean)	180 s	$3.7 \times 10^8$	68 (A100)	9.9	0.05	22	✓
Oceananigans <sub>R48</sub> <sub>FP32</sub> (Ocean)	45 s	$1.24 \times 10^{10}$	120 (V100)	0.33	0.37	36KW	
Oceananigans <sub>R48</sub> <sub>FP64</sub> (Ocean)	45 s	$1.24 \times 10^{10}$	576 (A100)	1.0	0.13	187KW	✓
Oceananigans <sub>R48</sub> <sub>FP64</sub> (Ocean)	45 s	$1.24 \times 10^{10}$	32 (A100)	0.063	1.9	10.4KW	
Oceananigans <sub>AP</sub> <sub>FP64</sub> (Ocean)	11 s	$2.1 \times 10^{11}$	768 (A100)	0.063	0.81	252KW	✓

Table 1: Performance details of state-of-the art climate, ocean, and atmosphere models. Larger grid sizes correspond to finer spatial resolution. Computations belonging to this submission are shown in bold.

## 8 Implications

By developing a new model from scratch specifically for GPUs, and wielding a handful of key ocean-model-specific innovations, Oceananigans achieves 9.9 SYPD at 10 km resolution using less than 1% of the resources of current state of the art supercomputers. This achievement means that most climate model runs submitted to IPCC will be able use 10 km ocean models — *precipitating a step change in the accuracy of climate prediction.*

At scales between 10–100 km, macroscale ocean turbulence exerts a key control on ocean carbon and heat uptake. However, attempts to accurately parameterize this key process in coarse resolution models have frustrated generations of oceanographers. The inadequacies of macroscale parameterizations are associated with major biases and uncertainty in climate predictions [33, 39]. At resolutions of 10 km, the need for macroscale turbulence parameterization is eliminated, and ocean simulations capture key ocean features such as sharp sea surface temperature gradients supporting the formation of marine stratus clouds above narrow eastern boundary currents like the California and Benguela Current [28], and changes in the meridional overturning circulation due to the effect of Antarctic meltwater on deep convection in austral winter [26].

Additionally, by achieving 0.95 SYPD at 1.7 km resolution, we pave the way for decadal ocean simulations of the ocean “submesoscale” — the ocean analogue to atmospheric weather — which exhibits hourly fluctuations, high spatial and seasonal variability, and which exerts a strong control on ocean air-sea fluxes, biological productivity and fish stocks [46]. The granularity and accuracy provided by 1.7 km resolution is further required to plan local mitigation strategies and predict local extreme events.

Third, the unparalleled speed of execution and memory efficiency of Oceananigans allows global computations at never-before-seen sub-kilometer resolutions. The capacity for ultra-high-resolution simulations aligns with current advancements in resolution of ocean sampling platforms from satellites [32, 12] to fleets of floats and drones. While this wealth of data is likely to provide new insights and scientific knowledge about the nature of small scale processes, global high-resolution ocean simulations will be needed to explore their impact on global climate scales.

Finally, our results pave the way for marked increase in energy efficiency of climate simulations. The very reason to develop climate models, as stated by the Coupled Model Intercomparison Project (CMIP), for example, is to provide the necessary information to effectively reduce emissions and mitigate the effects of global warming — while, counterproductively, the carbon footprint of climate simulations that contribute to CMIP increases rapidly. Oceananigans’ achievements represent a milestone towards decreased energy consumption by climate modeling efforts.

## 9 Acknowledgments

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC award DDR-ERCAP0025591. This work is partly supported by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program and by NSF grant AGS-1835576. N.C.C. is supported by the Australian Research Council DECRA Fellowship DE210100749.

## References

- [1] HIP: C++ Heterogeneous-Compute Interface for Portability. <https://github.com/ROCm-Developer-Tools/HIP>.
- [2] MIT Satori User Documentation. <https://mit-satori.github.io/index.html>.
- [3] A. Arakawa and V.R. Lamb. Computational design of the basic dynamical processes of the UCLA general circulation model. *General Circulation Models of the Atmosphere*, 17, 1977.
- [4] V. Balaji, E. Maisonnave, N. Zadeh, B.N. Lawrence, J. Biercamp, U. Fladrich, G. Aloisio, R. Benson, A. Caubel, J. Durachta, et al. CPMIP: measurements of real computational performance of Earth system models in CMIP6. *Geoscientific Model Development*, 10(1):19–34, 2017.
- [5] T. Besard, V. Churavy, A. Edelman, and B. De Sutter. Rapid software prototyping for heterogeneous and distributed platforms. *Advances in engineering software*, 132:29–46, June 2019.
- [6] T. Besard, C. Foket, and B. De Sutter. Effective Extensible Programming: Unleashing Julia on GPUs. *IEEE Trans. Parallel Distrib. Syst.*, (4):827–841, December 2017.
- [7] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, January 2017.
- [8] E.P. Chassignet, S.G. Yeager, B. Fox-Kemper, A. Bozec, F. Castruccio, G. Danabasoglu, C. Horvat, W.M. Kim, N. Koldunov, Y. Li, et al. Impact of horizontal resolution on global ocean–sea ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2). *Geoscientific Model Development*, 13(9):4595–4637, 2020.
- [9] T.R. Christian, D. Lebrun-Grandie, D. Arndt, J. Ciesko, V. Dang, N. Ellingwood, R. Gayatri, E. Harvey, D.S. Hollman, D. Ibanez, et al. Kokkos 3: Programming model extensions for the exascale era. *IEEE Transactions on Parallel and Distributed Systems*, 33(4):805–817, 2021.
- [10] V. Churavy, D. Aluthge, J. Samaroo, A. Smirnov, J. Schloss, L. C Wilcox, S. Byrne, M. Waruszewski, A. Ramadhan, S. Schaub, N. C Constantinou, J. Bolewski, M. Ng, T. Besard, B. Arthur, C. Kawczynski, C. Hill, C. Rackauckas, J. Cook, J. Liu, M. Schanen, O. Schulz, P. Haraldsson, T. Arakaki, and T. Chor. JuliaGPU/KernelAbstractions.jl: v0.9.1, March 2023.
- [11] M. Ding, H. Liu, P. Lin, Y. Meng, W. Zheng, B. An, Y. Luan, Y. Yu, Z. Yu, Y. Li, J. Ma, J. Chen, and K. Chen. A century-long eddy-resolving simulation of global oceanic large- and mesoscale state. *Sci. Data*, 9:661, 2022.
- [12] C. Donlon, B. Berruti, S. Mecklenberg, J. Nieke, H. Rebhan, U. Klein, A. Buongiorno, C. Mavrocordatos, J. Frerick, and B. Seitz. The sentinel-3 mission: Overview and status. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 1711–1714. IEEE, 2012.
- [13] D. Ferreira, J. Marshall, and J.-M. Campin. Localization of deep water formation: Role of atmospheric moisture transport and geometrical constraints on ocean circulation. *J. Climate*, 23:1456–1476, 2010.

- [14] T. Fiedler, A.J. Pitman, K. Mackenzie, N. Wood, C. Jakob, and S.E. Perkins-Kirkpatrick. Business risk and the emergence of climate analytics. *Nature Climate Change*, 11:87–94, 2021.
- [15] O. Fuhrer, T. Chadha, T. Hoefer, G. Kwasniewski, X. Lapillon, D. Leutwyler, D. Lüth, C. Osuna, C. Schär, T.C. Schulthess, and H. Vogt. Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0. *Geosci. Model Dev.*, 14:2781–2799, 2021.
- [16] M. Giordano, M. Klöwer, and V. Churavy. Productivity meets performance: Julia on A64FX. In *2022 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 549–555, September 2022.
- [17] W. F. Godoy, P. Valero-Lara, E. T. Dettling, C. Trefftz, I. Jorquera, T. Sheehy, R. G. Miller, M. Gonzalez-Tallada, J. S Vetter, and V. Churavy. Evaluating performance and portability of high-level programming models: Julia, Python/Numba, and kokkos on exascale nodes. March 2023.
- [18] J M. Gutiérrez and A.-M. Tréguier. IPCC, 2021: Annex II: Models. *Climate Change 2021: The Physical Science Basis*, page 2087–2138, 2021.
- [19] R. Haarsma, M. Acosta R., Bakhshi, P.-A. Bretonnière, L.-P. Caron, M. Castrillo, S. Corti, P. Davini, E. Exarchou, F. Fabiano, U. Fladrich, R. Fuentes Franco, J. García-Serrano, J. von Hardenberg, T. Koenig, X. Levine, V.L. Meccia, T. van Noije, G. van den Oord, F.M. Palmeiro, M. Rodrigo, Y. Ruprich-Robert, P. Le Sager, E. Tourigny, S. Wang, M. van Weele, and K. Wyser. HighResMIP versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR – description, model computational performance and basic validation. *Geoscientific Model Development*, 13(8):3507–3527, 2020.
- [20] C. Hope. The \$10 trillion value of better information about the transient climate response. *Phil. Trans. R. Soc. A*, 3:20140429, 2015.
- [21] S. Hunold and S. Steiner. Benchmarking Julia’s communication performance: Is Julia HPC ready or full HPC? In *2020 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*, pages 20–25. IEEE, November 2020.
- [22] D. Häfner, R. Nuterman, and M. Jochum. Fast, Cheap, and Turbulent—Global Ocean Modeling With GPU Acceleration in Python. *JAMES*, 13(12):e2021MS002717, 2021.
- [23] E.J. Kendo, N.M. Roberts, H.J. Fowler, M.J. Roberts, S.C. Chan, and C.A. Senior. Heavier summer downpours with climate change revealed by weather forecast resolution model. *Nature Climate Change*, 4(7):570–576, 2014.
- [24] C Lattner and V Adve. LLVM: A compilation framework for lifelong program analysis & transformation. In *International Symposium on Code Generation and Optimization, 2004. CGO 2004*. IEEE, 2004.
- [25] H. Lee and J. Romero (eds.). IPCC, 2023: Climate Change 2023: Synthesis Report. A Report of the Intergovernmental Panel on Climate Change. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 2023.

- [26] Q. Li, M.H. England, A. McC Hogg, S.R. Rintoul, and A.K. Morrison. Abyssal ocean overturning slowdown and warming driven by Antarctic meltwater. *Nature*, 615(7954):841–847, 2023.
- [27] W. C. Lin and S. McIntosh-Smith. Comparing julia to performance portable parallel programming models for HPC. In *2021 International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*, pages 94–105. ieeexplore.ieee.org, November 2021.
- [28] J. Ma, S. Xu, and B. Wang. Warm bias of sea surface temperature in eastern boundary current regions—a study of effects of horizontal resolution in CESM. *Ocean Dynamics*, 69:939–954, 2019.
- [29] J. Marshall, A. Adcroft, C. Hill, L. Perelman, and C. Heisey. A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. *Journal of Geophysical Research: Oceans*, 102(C3):5753–5766, 1997.
- [30] D.S. Medina, A. St-Cyr, and T. Warburton. OCCA: A unified approach to multi-threading languages. *arXiv preprint arXiv:1403.0968*, 2014.
- [31] Dimitris Menemenlis, Jean-Michel Campin, Patrick Heimbach, Chris Hill, Tong Lee, An Nguyen, Michael Schodlok, and Hong Zhang. Ecco2: High resolution global ocean and sea ice data synthesis. *Mercator Ocean Quarterly Newsletter*, 31(October):13–21, 2008.
- [32] R. Morrow, L-L. Fu, F. Ardhuin, M. Benkiran, B. Chapron, E. Cosme, F. d’Ovidio, J.T. Farrar, S.T. Gille, G. Lapeyre, P.-Y. Le Traon, A. Pascual, A. Ponte, B. Qiu, Rasclle N, C. Ubelmann, J. Wang, and E.D. Zaron. Global Observations of Fine-Scale Ocean Surface Topography with the Surface Water and Ocean topography (SWOT) Mission. 2019.
- [33] D.R. Munday, H.L. Johnson, and D.P. Marshall. Impacts and effects of mesoscale ocean eddies on ocean carbon storage and atmospheric pCO<sub>2</sub>. *Global Biogeochemical Cycles*, 28(8):877–896, 2014.
- [34] T. Palmer. Build high-resolution global climate models. *Nature*, 515:338–339, 2014.
- [35] A. Ramadhan, G. Wagner, C. Hill, J.-M. Campin, V. Churavy, T. Besard, A. Souza, A. Edelman, R. Ferrari, and J. Marshall. Oceananigans.jl: Fast and friendly geophysical fluid dynamics on GPUs. *Journal of Open Source Software*, 5(53), 2020.
- [36] T. Ringler, M. Petersen, R.L. Higdon, D. Jacobsen, P.W. Jones, and M. Maltrud. A multi-resolution approach to global ocean modeling. *Ocean Modelling*, 69:211–232, 2013.
- [37] M.J. Roberts, A. Baker, E.W. Blockley, D. Calvert, A. Coward, H.T. Hewitt, L.C. Jackson, T. Kuhlbrodt, P. Mathiot, C.D. Roberts, R. Schiemann, J. Seddon, B. Vannière, and P.L. Vidale. Description of the resolution hierarchy of the global coupled hadgem3-gc3.1 model as used in cmip6 highresmip experiments. *Geosci. Model Dev.*, 12(12):4999–5028, 2019.
- [38] V. Churavy S. Byrne, L.C. Wilcox. MPI.jl: Julia bindings for the message passing interface. *Proceedings of the JuliaCon Conferences*, 2021.
- [39] O.A. Saenko, D. Yang, and J.M. Gregory. Impact of mesoscale eddy transfer on heat uptake in an eddy-parameterizing ocean model. *Journal of Climate*, 31(20):8589 – 8606, 2018.

- [40] T. Schneider, S. Behera, G. Boccaletti, C. Deser, K. Emanuel, R. Ferrari, L.R. Leung, N. Lin, T. Müller, A. Navarra, O. Ndiaye, A. Stuart, J. Tribbia, and T. Yamagata. Harnessing AI, Data, and Computing to Advance Climate Modeling and Prediction. *Nature Climate Change*, submitted.
- [41] T. Schneider, S. Lan, A. Stuart, and J. Teixeira. Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations. *Geophysical Research Letters*, 2017.
- [42] C. Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. Icase report no. 97-65, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, 1997.
- [43] S. Silvestri, G.L. Wagner, A. Souza, N. Constantinou, C. Hill, J.-M. Campin, and R. Ferrari. A WENO-based Vector Invariant advection scheme for Implicit Ocean LES. *In preparation*.
- [44] D. Struck. Meet the team shaking up climate models. *The Christian Science Monitor*, 2021.
- [45] Z. Su, J. Wang, P. Klein, A. Thompson, and D. Menemenlis. Ocean submesoscales as a key component of the global heat budget. *Nat. Commun.*, 9:775, 2018.
- [46] J.R. Taylor and A.F. Thompson. Submesoscale Dynamics in the Upper Ocean. *Annual Review of Fluid Mechanics*, 55(1):103–127, 2023.
- [47] MPI Team. MPI: A Message-Passing interface standard. Technical report, USA, 1994.
- [48] P. Wang, J. Jiang, P. Lin, M. Ding, J. Wei, F. Zhang, L. Zhao, Y. Li, Z. Yu, W. Zheng, Y. Yu, X. Chi, and H. Liu. The GPU version of LASG/IAP climate system ocean model version 3 (LICOM3) under the heterogeneous-compute interface for portability (HIP) framework and its large-scale application. *Geosci. Model Dev.*, 14(5):2781–2799, 2021.
- [49] N.P. Wedi, I. Polichtchouk, P. Dueben, V.G. Anantharaj, P. Bauer, S. Boussetta, P. Browne, W. Deconinck, W. Gaudin, I. Hadade, et al. A baseline for global weather and climate simulations at 1 km resolution. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002192, 2020.
- [50] B. Xiao, F. Qiao, Q. Shu, X. Yin, G. Wang, and S. Wang. The development and validation of a global  $1/32^\circ$  surface-wave-tide-circulation coupled ocean model: FIO-COM32. *Geosci. Model Dev.*, 2022.
- [51] S. Zhang, H. Fu, L. Wu, Y. Li, H. Wang, Y. Zeng, X. Duan, W. Wan, L. Wang, Y. Zhuang, et al. Optimizing high-resolution community earth system model on a heterogeneous many-core supercomputing platform. *Geoscientific Model Development*, 13(10):4809–4829, 2020.