DEMOCRATIZE THE DATA

A NEW WAY TO ANALYZE AND DESIGN OCEAN MODELS

By Thomas W.N. Haine

ABSTRACT. Ocean circulation models running on the latest supercomputers can cover the globe with resolutions of a few kilometers. These virtual ocean datasets are increasingly realistic and provide insight into processes at scales that are inaccessible with conventional observations. Because these datasets are far too massive for individual researchers to download and analyze, new cloud-based, open-source, cyberinfrastructure resources are being developed. These tools provide a new analysis paradigm that is scalable, accessible, and inclusive, and that democratizes access to ocean circulation model output. They also accelerate the pace of analysis of ocean models and thereby increase the pace of discovery in oceanography. Another challenge concerns the priorities for next-generation ocean circulation models. In particular, to improve circulation model simulations, how should increased supercomputer power be spent? Input on this question from the oceanographic community is sought.

INTRODUCTION

Simulation of ocean currents by numerical models has been revolutionized by information technology advances in the last 50 years. New discoveries have resulted from improved observing technologies, such as the global Argo network of autonomous profiling floats (Riser et al., 2016; Argo, 2020) and satellite observations of sea level (Lee et al., 2010; Vinogradova et al., 2025). Improved ocean circulation models have also resulted in new discoveries (Fox-Kemper et al., 2019; Haine et al., 2021), particularly those based on better model grid resolution. The growth in ocean circulation model fidelity brings challenges, however. One challenge concerns the difficulty of providing access to the very large volumes of data ocean circulation models produce, and another concerns the priorities for future cutting-edge ocean circulation model simulations.

This commentary introduces and explains these topics and outlines some possible ways ahead. Developments in cloud storage and cloud computing are providing open cyberinfrastructure platforms that lower the barrier to data access. Open discussion on future circulation model priorities is also beginning. These services for, and engagement with, the oceanographic community aim to make cutting-edge ocean current simulations as widely accessible and as useful as possible.

GRID CELL AND DATA GROWTH

Global ocean general circulation models (OGCMs) show exponential growth in grid cell resolution. This remarkable expansion ultimately derives from Moore's law, which states that the density of microelectronic devices doubles every two years (Moore, 1975). To illustrate, Figure 1 shows the number of horizontal grid cells used to discretize the global ocean in five cutting-edge OGCMs since 1980 (with black dots). The number of horizontal grid cells doubles every 2.5 yr, keeping up with Moore's law (some of the increase in computer power is used to refine OGCM vertical resolution). Nowadays, cutting-edge OGCMs have horizontal resolutions of around 1 km, with hundreds of millions of grid cells covering the surface of the global ocean.

Coupled Earth system models of the kind used to project global climate change by the Intergovernmental Panel on Climate Change (IPCC) also show exponential refinement of the horizontal grid resolution in their ocean models (Figure 1, colored dots). For these models, the doubling time is 3.7 yr, somewhat slower than for OGCMs because other components of the Earth system model compete for the computer speedup.

Observations of the global ocean have been revolutionized by information technology advances too. Figure 1 shows, for example, the number of annual deep stations with high-quality temperature measurements (CTD stations deeper than 1,000 m). In the early 2000s, the rate of such observations increased by a factor of 10 as the global Argo network came online. Today, about 100,000 deep temperature stations are reported each year.

Consider next the relative rates of growth of OGCM resolution and deep temperature measurements. Figure 1 shows that OGCMs outstrip the observations, so there are now around 1,000 horizontal grid cells for every deep temperature station. Put another way, the average spacing between Argo CTD profiles is 300 km,

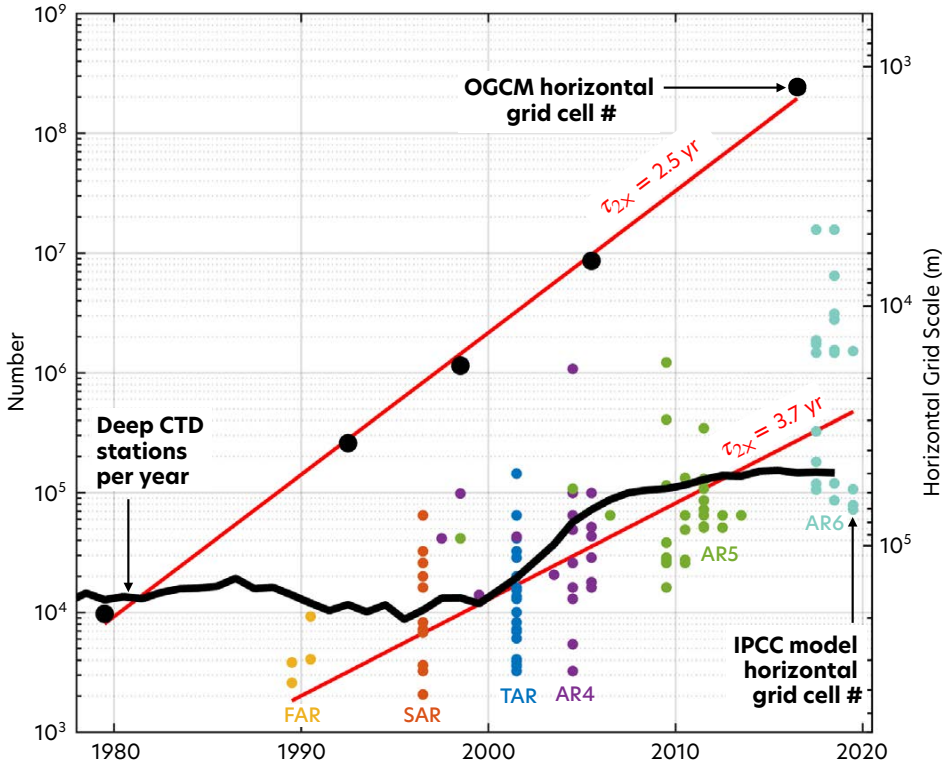

FIGURE 1. Growth over time of the number of horizontal grid cells in global ocean general circulation models (OGCMs, see the black dots), the number of horizontal grid cells in the global coupled climate model from the Intergovernmental Panel on Climate Change (IPCC, see the colored dots), and the number per year of deep (greater than 1,000 m depth) CTD stations. Note that the *y*-axis is logarithmic and the straight red lines indicate exponential growth (the doubling times, τ_{2x} are shown). The black dot in 2016 is for the LLC4320 OGCM (see text and Figures 2 and 3). The three-letter abbreviations in color refer to the IPCC assessment reports. *Modified from Figure 2 in Haine et al. (2021)*

whereas the average spacing between cutting-edge OGCM grid cells is 1 km. In this sense, cutting-edge OGCMs are becoming unconstrained by data because the data are sparse compared to the OGCM degrees of freedom (and notice that this is not true for the ocean components of cutting-edge IPCC models). The unequal growth of OGCM resolution and data density reflects the so-called maturation of computational oceanography (Haine et al., 2021). Cutting-edge OGCMs are thus becoming more and more valuable as a resource in oceanography.

OGCM SOLUTIONS AND DATA ACCESS

LLC4320

For example, the 2016 black dot in Figure 1 is a model solution called LLC4320 (the name refers to the latitude-longitude-cap horizontal grid with 4320 × 4320 grid cells in each of 13 faces that tile the global ocean; Rocha et al., 2016; Arbic et al., 2018). The LLC4320 simulation provides hourly output for one year in 2011–2012 using the Massachusetts Institute of Technology OGCM code. A few similar solutions exist using other circulation models and different configurations. Collectively, such solutions are called "nature runs" or "digital twins" of the ocean currents (Boyes and Watson, 2022; Chen et al., 2023; NASEM, 2024; Vance et al., 2024). They are useful for many purposes that include understanding ocean dynamics, designing observing systems, and machine learning.

Indeed, the oceanographic community is eagerly adopting these cutting-edge OGCM solutions. To illustrate, the red dots in Figure 2 show the number of papers each year that utilize the LLC4320 solution. As in Figure 1, the *y*-axis of Figure 2 is logarithmic, and straight lines indicate exponential growth. Thus, Figure 2 shows that the number of LLC4320 papers per year has grown roughly as an exponential with a doubling time of around 3 yr; dozens of papers now employ the LLC4320 simulation per year.

Despite this growing popularity, the data from LLC4320-type cutting edge simulations are very challenging to use. The main problem is the massive size of the datasets, which means that access to these data is difficult and time-consuming. For LLC4320, the total uncompressed data volume is four petabytes (one petabyte is 10^{15} bytes), and it takes many months to obtain accounts on the NASA supercomputers where the LLC4320 simulation was run. Moreover, the datasets are far too massive for individual researchers to download and analyze personal copies.

POSEIDON PROJECT

Making the LLC4320 (and similar) simulation data easy to use is therefore an important priority. Evidence from a neighboring field in fluid mechanics shows the benefits of opening massive simulation datasets to easy community access. Specifically, the blue dots in Figure 2 show the number of papers each year that utilize the Johns Hopkins Turbulence Database (JHTDB; Li et al., 2008). The JHTDB is an open numerical turbulence laboratory that provides free access to benchmark numerical solutions for various canonical turbulence problems. Figure 2 shows that the number of JHTDB papers per year has also grown exponentially, with a doubling time of 3.0 yr. In total, more than 6×10^{14} individual model grid cells have been queried using the JHTDB. A recent paper states that

"since its publication, the JHTDB had become a gold standard and an hypothesis testing tool in the turbulence community" (Shnapp et al., 2023). This opening up of cutting-edge benchmark simulations has been termed "democratizing the data." In addition, such databases significantly reduce carbon emissions by reusing extant data rather than recomputing them (Yang et al., 2024).

Inspired by the JHTDB, an initiative called the Poseidon Project has been democratizing the LLC4320 (and similar) OGCM data. Figure 3 illustrates some key features of the Poseidon Project and the modular workflows it supports. The left panel of Figure 3 is a screenshot from the public Poseidon Viewer showing surface relative vorticity in the LLC4320 North Atlantic Ocean. The first Poseidon Project design goal is for users to access the data with very low latency (time delay). The Poseidon Viewer achieves this goal by visualizing the LLC4320 simulation data interactively, including on mobile devices in a few seconds (try the Poseidon Viewer interactive LLC4320 visualization tool).

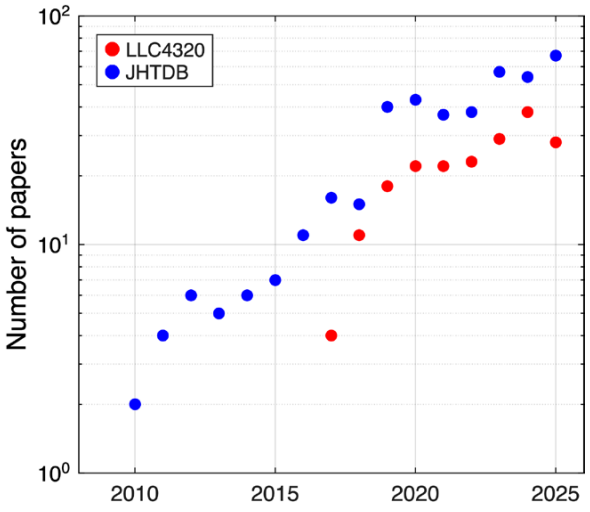

FIGURE 2. Growth over time of the number of papers per year citing the LLC4320 global OGCM and the Johns Hopkins Turbulence Database (JHTDB). Note that the *y*-axis is logarithmic (the $\tau_{2\times}$ doubling time for the annual JHTDB citations is 3.0 yr). The data are taken from the LLC4320 and JHTDB websites as of March 2025.

The second Poseidon Project design goal is to provide a simple software interface for accessing the data. The Poseidon Project (like the JHTDB) is hosted on SciServer, which is a collaborative cloud environment for analysis of extremely large datasets (Medvedev et al., 2016). The SciServer supports Jupyter notebooks for data analysis. The middle panel of Figure 3 shows a screenshot of a SciServer Jupyter notebook using the OceanSpy Python software to analyze LLC4320 data (Almansi et al., 2019). In this example, a synthetic hydrographic section is being plotted. The OceanSpy software is an interface to scalable, open-source tools from the Pangeo community (which can be used directly in SciServer, for example, by using xarray without the OceanSpy interface). The right panel of Figure 3 shows trajectories of drifting particles in the LLC4320 surface currents. The trajectories were computed in a SciServer Jupyter notebook using the Seaduck Python software (Jiang et al., 2023).

The third Poseidon Project design goal is to focus on final computation and rendering of high-quality figures. SciServer achieves these goals by performing data-proximate, lazy calculations (no data downloads are necessary, although they are possible) and providing a robust, stable, fully functional programming environment in the cloud. Thus, anyone with internet access can interact with the LLC4320 data, make calculations, and produce publication-ready figures. This is another sense in which the simulation data are being "democratized" (made open to everyone).

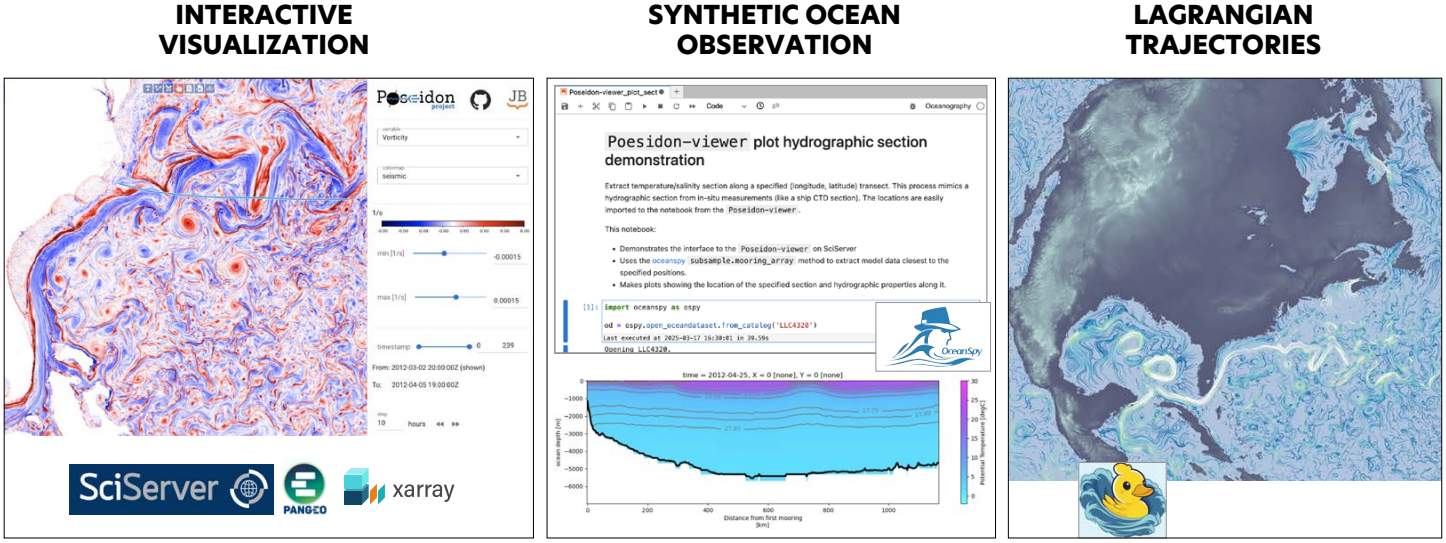

FIGURE 3. The Poseidon Project makes high-resolution OGCM solutions publicly available, such as the global LLC4320 simulation. Users can interact with the data using a mobile-friendly, interactive visualization tool and Python application programming interface software such as OceanSpy (Almansi et al., 2019), which samples the OGCM data using synthetic oceanographic instruments, along with Seaduck (Jiang et al., 2023), which computes Lagrangian trajectories. The data can also be accessed using Pangeo tools such as xarray. Run the Poseidon Viewer interactive LLC4320 visualization tool.

FUTURE OGCM PRIORITIES

Returning to Figure 1, notice that the LLC4320 simulation is already a decade old. Moore's law has continued in the years since NASA computed LLC4320, and the time is ripe to make a new benchmark cutting-edge calculation. Extrapolating the OGCM red line in Figure 1 suggests that such a new simulation could have 3×10^9 horizontal grid cells, which corresponds to a horizontal grid scale of 350 m. This resolution captures part of the unexplored regime of submesoscale dynamics in which rotational, inertial, and buoyancy effects are all of similar importance (Taylor and Thompson, 2023), and which is very hard to observe with current oceanographic instruments.

Alternatively, the extra computational power could be spent on other priorities. For example, the simulation could be run for longer than one year at the same resolution as LLC4320. Or the initial condition could be improved to avoid transient adjustments during the simulation. The question is, what are the most important priorities and, in particular, how should the extra computational power be spent?

This question was asked during a town hall meeting at the 2024 Ocean Sciences Meeting. Participants in the town hall responded to an online survey that asked them to rank 11 different priorities for designing the next cutting-edge global benchmark OGCM simulation. Participants could also write in their own priorities. Figure 4 shows the results of the survey, summarizing the opinions of 44 respondents (the survey is still open—take the survey).

The survey results show no consensus for future benchmark OGCM solutions because all the priorities were ranked as important by some respondents and as unimportant by others. Nevertheless, preferences are clear overall. The most highly ranked priorities include longer run time and better horizontal and vertical resolution. These priorities are relatively easy to implement because they require little OGCM code development and little pre-computation before the main OGCM code is run. Better model spin-up/initial conditions and better air-sea forcing are also highly ranked. These priorities are harder to implement because they involve improvements (which need to be precisely defined) to input data from other large, complex modeling systems. The four middle-ranked priorities are: better constraints to observations, better model parametrizations, better model topography, and better mean circulation and stratification. These are desirable scientific goals that are easy to state but hard to achieve. One reason is that they involve detailed tuning of OGCM parameters and input data, or improvements to OGCM software. Another reason is that these priorities are interrelated because, for example, improving the mean circulation probably requires better parametrizations and topography, which will inevitably improve agreement with observations. Two priorities were ranked as unimportant overall, namely an ensemble of LLC4320 runs (easy to implement) and better diversity in model code (relatively easy to implement using existing OGCM systems).

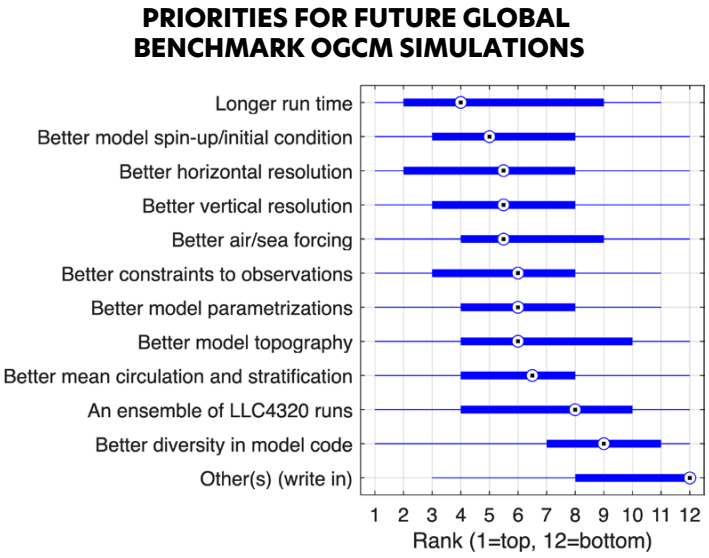

FIGURE 4. Results from a 2024 Ocean Sciences Meeting survey on priorities for the next benchmark global OGCM simulation. Forty-four respondents ranked the priorities on the *y*-axis on a scale of 1 to 12 (1 is the top priority). The median value is shown with the dotted circle, the 25th and 75th percentiles are shown with the thick bar, and the thin bars indicate maximum and minimum values. "Other(s) (write in)" priorities included adding biogeochemistry, better documentation and tutorials, and better evaluation with observations. Take the survey.

Other priorities listed by a few respondents included adding biogeochemistry, better documentation, and better comparison with observations.

OUTLOOK

Given the ongoing advances in computational hardware, software, and infrastructure, the time is ripe for a new cutting-edge OGCM solution (or more than one) to be computed. Efforts like LLC4320 and the Poseidon Project require significant resources and therefore need broad support from academia, industry, funding agencies, and non-professional oceanographers. To date, these efforts have been supported by government agencies and private foundations with standalone projects every few years. The need to sustain open shared cyberinfrastructure like SciServer and digital twins like LLC4320 is widely recognized (Barker et al., 2019; Grossman, 2023; Le Moigne et al., 2023; NASEM, 2024). The future sources of support and the pathway for migrating from research project funding to community infrastructure funding are uncertain, however.

One notable example of a stable, long-term, cloud-based data analysis environment for ocean sciences is the Mercator Ocean International and Copernicus Marine Service resource, funded by the European Commission. It provides real-time global ocean hindcasts, analyses, and forecasts using ocean circulation models, in situ and remote observations, and data assimilation (although not presently at the LLC4320 horizontal resolution). Their focus is on operational oceanography and the state of the ocean for diverse stakeholders (von Schuckmann et al., 2024). Apart from academic users, people have applied the Copernicus Marine Service

to oil spill modeling, shipping route optimization, and maritime tourism, to name a few. The value of such resources for catalyzing research and expanding the community of users engaged with ocean currents is tremendous.

As this commentary outlines, the track record of ocean model advancements is remarkable, with no obvious end in sight. The knowledge and tools for disseminating and analyzing massive ocean current simulations currently exist. Decisions on future priorities with broad community input and engagement are now required. The prospects for future ocean model improvements and refinement are very bright, and many are straightforward to implement.

REFERENCES

Almansi, M., R. Gelderloos, T. Haine, A. Saberi, and A. Siddiqui. 2019. OceanSpy: A Python package to facilitate ocean model data analysis and visualization. *Journal of Open Source Software* 4(39):1506, https://doi.org/10.21105/joss.01506.

Arbic, B.K., M.H. Alford, J.K. Ansong, M.C. Buijsman, R.B. Ciotti, J.T. Farrar, R.W. Hallberg, C.E. Henze, C.H. Hill, C.A. Luecke, and others. 2018. A primer on global internal tide and internal gravity wave continuum modeling in HYCOM and MITgcm. Pp. 307–392 in *New Frontiers in Operational Oceanography*. E. Chassignet, A. Pascual, J. Tintoré, and J. Verron, eds, GODAE OceanView, https://doi.org/10.17125/gov2018.ch13.

Argo. 2020. Argo float data and metadata from Global Data Assembly Centre (Argo GDAC), https://doi.org/10.17882/42182.

Barker, M., S.D. Olabarriaga, N. Wilkins-Diehr, S. Gesing, D.S. Katz, S. Shahand, S. Henwood, T. Glatard, K. Jeffrey, B. Corrie, and others. 2019. The global impact of science gateways, virtual research environments and virtual laboratories. *Future Generation Computer Systems* 95:240–248, https://doi.org/10.1016/j.future.2018.12.026.

Boyes, H., and T. Watson. 2022. Digital twins: An analysis framework and open issues. *Computers in Industry* 143:103763, https://doi.org/10.1016/j.compind.2022.103763.

Chen, G., J. Yang, B. Huang, C. Ma, F. Tian, L. Ge, L. Xia, and J. Li. 2023. Toward digital twin of the ocean: From digitalization to cloning. *Intelligent Marine Technology and Systems* 1(3), https://doi.org/10.1007/s44295-023-00003-2.

Fox-Kemper, B., A. Adcroft, C.W. Böning, E.P. Chassignet, E. Curchitser, G. Danabasoglu, C. Eden, M.H. England, R. Gerdes, R.J. Greatbatch, and others. 2019. Challenges and prospects in ocean circulation models. *Frontiers in Marine Science* 6:65, https://doi.org/10.3389/fmars.2019.00065.

Grossman, R.L. 2023. Ten lessons for data sharing with a data commons. *Scientific Data* 10:120, https://doi.org/10.1038/s41597-023-02029-x.

Haine, T.W.N., R. Gelderloos, MA. Jiminez-Urias, A.H. Siddiqui, G. Lemson, D. Medvedev, A. Szalay, R.P. Abernathy, M. Almansi, and C.N. Hill. 2021. Is computational oceanography coming of age? *Bulletin of the American Meteorological Society* 102(8):E1481–E1493, https://doi.org/10.1175/BAMS-D-20-0258.1.

Jiang, W., T.W.N. Haine, and M. Almansi. 2023. Seaduck: A Python package for Eulerian and Lagrangian interpolation on ocean datasets. *Journal of Open Source Software* 8(92):5967, https://doi.org/10.21105/joss.05967.

Le Moigne, J., M.M. Little, R.A. Morris, N.C. Oza, K.J. Ranson, H. Riris, L.J. Rogers, and B.D. Smith. 2023. *Earth System Digital Twin (ESDT) Architecture Framework*. NASA Technical Report, Earth Science Technology Office, NASA, 12 pp.

Lee, T., S. Hakkinen, K. Kelly, B. Qiu, H. Bonekamp, and E. Lindstrom. 2010. Satellite observations of ocean circulation changes associated with climate variability. *Oceanography* 23(4):70–81, https://doi.org/10.5670/oceanog.2010.06.

Li, Y., E. Perlman, M. Wan, Y. Yang, C. Meneveau, R. Burns, S. Chen, A. Szalay, and G. Eyink. 2008. A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence. *Journal of Turbulence* 9:N31, https://doi.org/10.1080/14685240802376389.

Medvedev, D., G. Lemson, and M. Rippin. 2016. SciServer compute: Bringing analysis close to the data. Pp. 1–4 in *Proceedings of the 28th International Conference on Scientific and Statistical Database Management - SSDBM '16*.ACM Press, https://doi.org/10.1145/2949689.2949700.

Moore, G.E. 1975. Progress in digital integrated electronics. Pp. 11–13 in International Electron Devices Meeting. IEEE, http://www.eng.auburn.edu/~agrawvd/COURSE/E7770_Spr07/READ/Gordon_Moore_1975_Speech.pdf.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2024. *Foundational Research Gaps and Future Directions for Digital Twins*. The National Academies Press, Washington, DC, 202 pp., https://doi.org/10.17226/26894.

Riser, S.C., H.J. Freeland, D. Roemmich, S. Wiffjels, A. Troisi, M. Belbéoch, D. Gilbert, J. Xu, S. Pouliquen, A. Thresher, and others. 2016. Fifteen years of ocean observations with the global Argo array. *Nature Climate Change* 6(2):145–153, https://doi.org/10.1038/nclimate2872.

Rocha, C.B., T.K. Chereskin, S.T. Gille, and D. Menemenlis. 2016. Mesoscale to submesoscale wavenumber spectra in Drake Passage. *Journal of Physical Oceanography* 46(2):601–620, https://doi.org/10.1175/JPO-D-15-0087.1.

Shnapp, R., S. Brizzolara, M.M. Neamtu-Halic, A. Gambino, and M. Holzner. 2023. Universal alignment in turbulent pair dispersion. *Nature Communications* 14:4195, https://doi.org/10.1038/s41467-023-39903-6.

Taylor, J.R., and A.F. Thompson. 2023. Submesoscale dynamics in the upper ocean. *Annual Review of Fluid Mechanics* 55:103–127, https://doi.org/10.1146/annurev-fluid-031422-095147.

Vance, T.C., T. Huang, and K.A. Butler. 2024. Big data in Earth science: Emerging practice and promise. *Science* 383(6688), https://doi.org/10.1126/science.adh9607.

Vinogradova, N.T., T.M. Pavelsky, J.T. Farrar, F. Hossain, and L.-L. Fu. 2025. A new look at Earth's water and energy with SWOT. *Nature Water* 3:27–37, https://doi.org/10.1038/s44221-024-00372-w.

von Schuckmann, K., L. Moreira, M. Grégoire, M. Marcos, J. Staneva, P. Brasseur, G. Garric, P. Lionello, J. Karstensen, and G. Neukermans, eds. 2024. *8th edition of the Copernicus Ocean State Report (OSR8)*. Copernicus Publications, State Planet, 4-osr8, https://doi.org/10.5194/sp-4-osr8.

Yang, X., W. Zhang, M. Abkar, and W. Anderson. 2024. Computational fluid dynamics: Its carbon footprint and role in carbon reduction. *Journal of Renewable and Sustainable Energy* 16:055906, https://doi.org/10.1063/5.0217320.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under grants 1835640 and 2103874, by the Institute for Data Intensive Engineering and Science at Johns Hopkins University, and by the Alfred P. Sloan Foundation.

AUTHOR

Thomas W.N. Haine (thomas.haine@jhu.edu), Earth & Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA.

ARTICLE CITATION

Haine, T.W.N. 2025. Democratize the data: A new way to analyze and design ocean models. *Oceanography* 38(3), https://doi.org/10.5670/oceanog.2025.e303.

COPYRIGHT & USAGE