# Projected Computing Resource Requirements of the Canadian Ocean Modelling Community

Katja Fennel[1*], Susan Allen[2], Entcho Demirov[3], Brad DeYoung[3], Dany Dumont[4], Eric Galbraith[5], Kevin Lamb[6], Paul Myers[7], Richard Peltier[8], Francis Poulin[6], Andrea Scott[9], Jinyu Sheng[1], Jennifer Shore[10], Marek Stastna[6], Keith Thompson[1], Bruno Tremblay[5], Kirsten Zickfeld[11]

[1] Oceanography Department, Dalhousie University, Halifax
[2] Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver
[3] Department of Physics and Physical Oceanography, Memorial University of Newfoundland, St. John's
[4] Institut des sciences de la mer de Rimouski, Université du Québec à Rimouski, Rimouski
[5] Department of Atmospheric and Oceanic Sciences, McGill University, Montreal
[6] Department of Applied Mathematics, University of Waterloo, Waterloo
[7] Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton
[8] Department of Physics, University of Toronto, Toronto
[9] Department of Systems Design Engineering, University of Waterloo, Waterloo
[10] Royal Military College of Canada, Kingston
[11] Department of Geography, Simon Fraser University, Burnaby

[*] katja.fennel@dal.ca

## Scope

This whitepaper describes the computing resource needs of the Canadian ocean modelling community. Ocean modelling includes modelling of the physics of ocean circulation and sea ice dynamics, coupled physical-biological and biogeochemical modelling and data assimilation. Ocean modellers from 10 universities have been consulted and are represented among the 17 authors. The limit in scope to ocean modelling was necessary in order for our community to provide a specific list of computing resource requirements by the SPARC deadline. Several of us are also involved in lake modelling for which computational resource requirements are similar. Requirements by atmospheric modellers are likely similar as well, but due to time constraints we did not consult with them.

## Science Description

The ocean offers Canadians great economic opportunities that include resource extraction, transportation, renewable energy and aquaculture, but is arguably also the most difficult environment on Earth for human use. The ocean is a key player in the global climate system, presents a range of challenges and threats to human activity and is, at the same time, under human pressure. Man-made additions of fossil-fuel carbon dioxide to the atmosphere are leading to ocean warming, ocean acidification and loss of dissolved oxygen threatening marine ecosystems and likely increasing the frequency at which extreme events such as flooding and tropical storms will occur in the future. The Canadian Arctic is undergoing dramatic changes that come with economic opportunities, but also risks and dramatic impacts on local communities. A large proportion of the Canadian population lives in the coastal zone (and on or near the Great Lakes) and depends on ocean-driven economies. This reality motivates academic research to improve our basic understanding of oceanographic processes and our ability to simulate, predict and project physical, biological and chemical ocean characteristics on time scales from days, weeks and seasons to centuries. Numerical models are essential to this endeavour. Research foci of Canadian ocean modellers include the development of numerically more accurate, stable, realistic and sophisticated model codes and modelling methods, the study of fundamental ocean processes and of gradual long-term changes such as sea level rise, the implementation of prediction systems for early warning of ocean-related extreme events (e.g. flooding, ice cover) and improved response to emergencies (e.g. search and rescue, oil spill trajectories). The ocean modelling community also simulates ocean ecosystem processes (from harmful algae blooms to fish catches) and projects how the occurrence of extreme events and the health of ecosystems may change with global warming and due to other man-made pressures. Furthermore, we contribute to the development of coupled atmosphere-ocean models and thus the generation of more accurate forecasts and long-term projections of the state of this coupled system. Our research questions are particularly relevant to Canada surrounded by three oceans, with a long coastline and economic dependence on the ocean. Our work in the academic sector feeds significantly into supporting modelling done by the federal government (at Environment Canada, the Department of Fisheries and Oceans Canada, and National Defence and the Canadian Armed Forces), and thus our science directly contributes to the wellbeing of Canadians and addresses various policy questions.

## Current Use of Advanced Research Computing

All of the authors of this whitepaper make regular use of Compute Canada systems (e.g., ACEnet, Calcul Quebec, SHARCNET, SCINET, WestGRID, HPCVL) and would not be able to meet their research goals without access to these systems. Some authors have and have had special resource allocations.

| Name | Affiliation | CCDB ID (recent RAC if applicable) |
|---|---|---|
| Susan Allen | UBC | sbf-885 |
| Entcho Demirov | Memorial | anh-433 (2014 RAC: 80 cores, 4TB storage on Mahone, ACEnet) |
| Brad deYoung | Memorial | mbq-845 |
| Dany Dumont | Rimouski | xcu-601 |
| Katja Fennel | Dalhousie | qqh-593 (2014 RAC: 151 core years, 15 TB on guillimin/Calcul Quebec) |
| Eric Galbraith | McGill | ayu-503 (2014 RAC: 330 core years) |
| Kevin Lamb | Waterloo | gdt-055 (2013 RAC: 342 core years, 20 TB on orca, SHARCNET) |

| | | |
|---|---|---|
| Paul Myers | Alberta | wgq-844-01 (2014 RAC: 40 core years [has already used 100 core years in 2014], 30 TB storage on Jasper/Silo/Westgrid) |
| Richard Peltier | Toronto | cif-100-01 (2014 RAC: 800 core years on gpc_ib/SciNet, 1950 core years on tcs/SciNet, 50 TB on tcs, 350 TB on HPSS) |
| Francis Poulin | Waterloo | sxs-140 |
| Andrea Scott | Waterloo | iwu-994-02 |
| Jennifer Shore | Royal Military College of Canada | hpc1513 |
| Marek Stastna | Waterloo | pim-260 |
| Keith Thompson | Dalhousie | qyw-621 |
| Bruno Tremblay | McGill University | rqf-100-aa,ab,ac |

Many of us also make use of other computing resources, primarily systems housed at our home universities. In some cases, individual researchers have purchased and are purchasing additional cores to supplement Compute Canada resources with the understanding that these researchers have priority access to the donated cores (e.g. Marek Stastna, Kevin Lamb, Francis Poulin, Dany Dumont). However, these systems tend not to keep pace with our growing computing demands. A nationally shared resource through Compute Canada is the best long-term model if our needs can be met (without long queue wait times, etc.).

Currently we experience wait times, sometimes of up to days, which restrict our progress significantly. At present, there are not enough Compute Canada resources available and more competitions for dedicated resources per year would be helpful. One concern is that free access to Compute Canada resources comes with the risk of encouraging inefficient use; perhaps the competitions for special allocations are a way around this problem. Another presently limiting factor is the inability to schedule runs. Real-time ocean prediction is an emerging and expanding research area. In order to test and develop predictive ability, researchers need to run regular predictions. In order to be useful, such predictions need to be done in a timely manner. Thus our community would benefit from the ability to apply for regular, scheduled access to a set number of nodes for a set run time.

Many of us have also purchased large storage arrays/NAS systems/back up hard disks to archive and analyze the voluminous output produced by our models (e.g. Paul Myers has 60-70 TB of disk space on various machines/disks beyond his Compute Canada allocations; Eric Galbraith has 200 TB of storage space; Kevin Lamb, Francis Poulin and Marek Stastna have approximately 500 TB of storage space).

Each of us uses, on average, on the order of a 100 to 1000 cores and needs to store data on the order of 10 to 100 TB. Services like cloud, portals/gateways are less relevant for our community, but accelerators may become useful for us.

Standard software requirements include FORTRAN, C, C++, NetCDF, FFTW, PETSc, Matlab and Python.

## Future Growth

The ocean modelling community is currently limited by computing resources. Our research would immediately benefit from more and faster cores, and faster interconnections between cores. Ocean models are typically solved in parallel with parallelization accomplished through domain decomposition (or tiling). This necessitates continuous communication between cores and results in generally mediocre scaling. At present the largest number of cores for a single job is somewhere between 100 and 1000 cores (computations are continuous without peaks and valleys). Fast interconnections between cores are key to improving the scaling of ocean models. It is becoming increasingly necessary to run model ensembles (e.g. in ensemble-based data assimilation), which increase the numbers of cores needed by one or two orders of magnitude, but have the advantage of scaling very well.

The presently available computational infrastructure limits the spatial and temporal model resolution we can afford, the time periods over which model simulations can be run, and numbers of biological and chemical species and of ensemble members that can be included. Since insufficient model resolution is a major barrier to improving model realism, we are constantly striving to run at higher resolution. Due to numerical constraints, doubling the resolution of an ocean model increases the computational demand by about an order of magnitude (without considering scaling).

Groups that develop novel models, especially those for process studies, offer the potential for a new generation of ocean models that are higher order, and scalable by their construction. These efforts require a stable, and predictable infrastructure as well as support for scientific libraries (e.g. PETSc, Metis, Proj).

A frontier of oceanographic research is moving rapidly towards the small spatial scales at which most of the ocean's energy is focused: on the order of 1 to 10s of km. These scales must be resolved for the entire globe in order to simulate climatically relevant timescales, which results a very large computational burden. Nonetheless, successes based on recent increases in computing resources suggest it is feasible. A high-level example would be the ocean models used in the Intergovernmental Panel on Climate Change (IPCC) assessments. Ocean models represented in the IPCC's Assessment Report 5 (2014) have roughly doubled in spatial resolution compared to those in Assessment Report 4 (2007), which implies that computational effort has increased by an order of magnitude during the past 7 years. We anticipate a further increase in computational needs by at least an order of magnitude over the next five years as the international community is moving aggressively to ultra-high resolution models. A resolution of 1/12 of a degree is now becoming common for basin/global experiments, and we expect a push to using 1/36 degree within the next few years. Also, there is an increasing need/interest in coupling ecosystem and biogeochemical dynamics with the physical models and carry additional variables at higher resolutions (e.g. spatial, temporal, spectral) than in the past.

With respect to research into fundamental processes much work has been done with two-dimensional models in the past. Cutting-edge research is now shifting to three-dimensional simulations, which immediately increase needs by a 2-3 orders of magnitude.

**If Compute Canada cannot meet this growing need for computational resources the Canadian ocean modelling community will be at a distinct competitive disadvantage internationally.**

We anticipate that storage needs will increase commensurate with increases in number and speed of cores. With respect to process studies and high-resolution ocean models, storage is already a serious issue. Fully three-dimensional runs with ecology and biogeochemistry and more complex phenomena to be explored (e.g. flow separation during tidal flow over topography) will require large, fast storage, and the tools to make comparisons between suites of simulations. Fast connections between the computer and the storage system are less of a factor than fast connections between our workstations and Compute Canada storage. We anticipate needing hundreds of TB for storage per research group (including long-term storage) by 2016 and onwards.

Memory requirements vary by application. For some of us 2 to 4 GB per node is sufficient. For a three-dimensional simulation of shear instabilities in an internal solitary wave, memory requirements on 960 cores were 1020 GB; however, an order of magnitude more would have been useful if the cores and memory were available.

Accelerator systems may benefit our model development, for example, by handling Lagrangian particle tracking, but at present these are not a standard tool in ocean modelling.

We are producing large quantities of model data, and require the incorporation of ever larger observational data sets, hence innovative ways of analyzing these (e.g. pattern recognition and extraction) may become useful. They are not standard tools for us at present.

We do not foresee any dramatic changes in our software/middleware needs and do not have a clearly identified software development need that could be articulated to a Compute Canada development team. It is possible that groups developing high-order ocean or process study models may reach the stage where involvement with a software development team (including documentation) would lead to a product that is of utility to the broader community.

Strong network connectivity is important for us because it would enable visualization of data on remote servers and because we do need to move model data to users for analysis and interpretation. While the current network connectivity is not the limiting factor for our work and we do not expect a dramatic increase in our connectivity needs (those are secondary to our needs for more and faster cores, faster interconnects and commensurate storage needs), shared access to model output would facilitate collaborative work where researchers from different universities collaborate on analyses and where analyses must be done remotely. If model output will be stored at researchers' institutions after being produced on a Compute Canada cluster, this could result in relatively large amounts of data transfer (order of 1 TB per week for a continuously running 300 cpu job).