Next Generation Ocean Dynamical Core Roadmap Project: Final Report

Jason Holt, Adrian New, Hedong Liu, Andrew Coward (NOC)

Stephen Pickles, Mike Ashworth (STFC)

1. Introduction

This document should be read in conjunction with the "Next Generation Ocean Dynamical Core Roadmap Project: Summary and Recommendations" document, which provides a summary of the present "Final Report", and a Roadmap for ways forward for UK ocean modelling over the next 5-15 years. It describes two complementary ways forward: first is the incremental evolution of the NEMO model, and second a new modelling initiative drawing on the GungHo project.

1.1 Background

The National Oceanography Centre (NOC) was commissioned by NERC in December 2010 to undertake the Ocean Roadmap project as part of the Thematic Action Plan on "Next Generation Weather and Climate Prediction (NGWCP)". Consultation elsewhere within the NGWCP programme had revealed a compelling need for the UK to develop a new atmospheric dynamical core to provide scalable solutions for the massively parallel computer architectures expected on the 5-15 year timescale. This has now led to the initiation of the GungHo project (see below). However, the needs for the ocean were less clear, so the Ocean Roadmap project was commissioned to assess the development of ocean model dynamical cores required to meet UK operational and research needs on the timescale of 2015-2025. The project was required to undertake an assessment of the current and likely future trajectory of leading models, anticipated bottlenecks on computational performance, capabilities for interfacing to sub-grid scale parameterisations, data assimilation schemes and other component models of the earth system, extensive consultation with the UK and wider science and stakeholder communities, whether there exists a need for integrated modelling spanning the deep ocean to the coasts, and the preparation of a "roadmap" outlining potential development pathways. Alongside this document we also provide a 'Summary and recommendation' document that can be read as the Roadmap. However, this is not a static process and NOC will endeavour to keep that document up to date as part of its National Capability role, to reflect changes in scientific need and computational landscape. This will be done in on-going consultations with key stakeholders such as the Met Office and HEIs.

Following initiation of the project with a kick-off meeting in February 2011, and extensive consultation with users and stakeholders, an Interim Report was produced in September 2011, which can be read in conjunction with this Final Report, particularly the Review of Current Modelling Capabilities and the Technical Review will not be repeated here. The Interim Report identified a number of drivers for change in the present status of ocean modelling. These were, in particular, the need for better linkages between coastal, shelf and deep ocean models, the need to improve mixing in the models, and the need for the models to scale well on the massively parallel computer architectures. The Interim Report also identified a draft set of options for the future development of ocean models in the UK, for strategic oceanographic research, climate and operational activities.

Draft Options from the Interim Report:

1. **Separate structured and unstructured grid model development.** Here the development of NEMO (the structured model in widespread use already in the UK for global applications) continues along its current pathway, but with an enhanced emphasis on improving its ability to better represent coastal and shelf seas. The use of unstructured models was foreseen as continuing in the UK for specific coastal and shelf applications, and an assessment would be made of their global applicability in 5 years time (though the UK would not actively pursue their development for global applications through NC).

2. **Introducing an unstructured grid model into NEMO.** Here, either an existing or a new unstructured model was envisaged as being included within the NEMO framework. This would additionally require an interface to be developed to couple the unstructured model in shelf/ coastal areas, to the NEMO model in the open ocean. This would have the advantage of introducing specialized unstructured models to the large international NEMO community, and thereby moving to join up the two rather disparate communities (for global and shelf modelling).

3. **Transition to Fluidity-ICOM.** This would require further effort to develop the ICOM (Fluidity-ICOM) model towards being a fully-capable ocean general circulation model, with acceptably low numerical mixing. This would build on the substantial effort which has already been undertaken in the UK to develop the model so far, and the impressive support infrastructure that goes alongside it.

The Interim Report also considered but effectively ruled out the development of a completely new model from scratch, the incorporation of ICOM into NEMO, and the merger of ICOM and NEMO under a single framework.

Since the publication of the Interim Report, further consultation with the user community has been undertaken, (see section 1.2 below). This further period has led us to develop these initial options. In particular, option 2 is not now seen as worthy of serious consideration for further development because the data structures of the two models (NEMO and the unstructured model) would be so different that re-use of code would be limited, and much work would be needed to couple them together, and to interface to biogeochemical and sea-ice models. There may be some limited benefit in providing NEMO users with the same NEMO interface to enable them to run the unstructured model, but this would be outweighed by the issues above. Furthermore, we are unable to recommend option 3 for further explicit consideration since the ICOM model has not yet proven a convincing baroclinic global applicability. However, there is now a very considerable effort in the GungHo project aimed at developing a new dynamical atmosphere code that scales well on massively parallel computers, and we see much merit in developing closer links with GungHo to develop a comparable new ocean-shelf code. It should be noted that there is significant overlap between the planned development pathways for ICOM and GungHo, and they appear to be sharing many approaches. Hence, it makes sense to align strategically with the GungHo effort, and to draw on the expertise of the ICOM community through this process.

We therefore now recommend two principle options:

1. **Develop NEMO for global, shelf sea and 'global coastal ocean' applications.**
2. **Develop an ocean model within the GungHo framework.**

We envisage these would evolve concurrently, with effort divided appropriately given prevailing needs and state of the models. These options are detailed further below.

We envisage that near coastal and estuarine modelling (sub km scale) should continue using the current range of structured and unstructured models (see below). This area of modelling does not necessarily need a new strategic direction, but should benefit from Options (1) and (2) above as they develop

Global structured grid models are discussed in section 2 and coastal/shelf sea applications of NEMO (as regional/local models) in section 3. Issues relating the improvement of shelf seas in global models are considered in section 4 and the development of NEMO for the next generation computers in section 5. Section 6 describes current status in unstructured grid models and section 7 focusing on the prospects for developing an ocean model within the framework of the GungHo project. The remainder of this introduction is devoted to user perspectives and a forward look on High Performance Computing (HPC).

1.2 User perspectives and long-term needs

This work is aimed at three key users groups in the UK for physical ocean model development:

1. The Research Community: involved in fundamental and applied oceanographic research (including HEIs, Research Centres, the Met Office)
2. The Operational Community: involved in the production and dissemination of quality controlled, traceable model products (primarily the Met Office, but also Research Centres)
3. The assessment community: involved in providing evidence on environmental status and change to government and other stakeholders (include CEFAS, Marine Scotland, Met Office, Research Centres)

Noting that these roles overlap, and a particular process (such as climate projection) may involve all three. Other users include government departments/agencies and industry, but these are generally 'downstream' of the above mentioned three.

In the first phase of the project, leading to the Interim Report, key users and stakeholders were brought together for a kick-off meeting in February 2011. Extensive further individual visits to UK stakeholder institutions were then undertaken, together with a detailed literature investigation of technical aspects of relevant ocean and shelf models. Since then, in the second phase of the project, we have continued an active engagement with the user and stakeholder community. This has led to, in particular, a strong engagement with the GungHo community (including a focused workshop at the Issac Newton Institute (Cambridge) in 2012 between Ocean Roadmap, GungHo, and interested stakeholders, and Holt becoming an oceans representative on the GungHo Executive), further consultation with the international community (Stephen Griffies, Alistair Adcroft and Sergei Danilov in particular), detailed responses on the state of FVCOM modelling in the UK, and focused discussions with the Met Office. In addition, an early draft of this document was circulated for comment by all stakeholders and users ahead of the Final Workshop in February 2013.

From these user consultations, some principles have emerged:

1. The need for the best physics through process representation and resolution
2. The need for 'useful' models that can be readily applied to a wide range of research, operation and assessment questions, in a timely fashion.
3. The need for efficient models that can be run many times and for long periods.
4. That model data storage, transfer and analysis is a crucial part of the process and often more limiting than computation.

These must be tensioned against each other: while we may naturally strive for (1) it should not result in excluding (2), (3) and (4), and note that different communities may have very different needs.

As part of this process, we have identified three drivers for change (DFCs) in ocean modelling, which remain essentially as those described in the Interim Report. These are:

DFC1: Linking Coastal, Shelf and Open Ocean Models

- Downscaling to regions of interest
- Improving the representation of shelf seas in global models

DFC2: Improved Mixing

- Reduced spurious (numerical) mixing
- Improving the representation of sub-grid scale physics

DFC3: New Computer Architectures

- Maintain efficient use of new computer architectures as they evolve.

DFC1 will, for instance, enable better representation of fluxes and exchanges between the shelf seas and deep oceans and act to join up the shelf and global modelling communities, DFC2 will be of interest to all modellers: global modellers interested in decadal and longer timescales, and for improved representation of upper ocean and shelf sea physics. DFCs 1 and 2 will therefore have different priorities for different users, but are generally both important to take forward. DFC3 however, will affect all users needing to undertake internationally competitive simulations, be they fine resolution, multi-centenial, large ensembles, complex ecosystems, and/or involving sophisticated data assimilation. It should therefore be prioritized. One issue that has not been definitively resolved by this study is the demand (or not) for non-hydrostatic capability, so this should be part of the on-going consideration as these options evolve (noting that it is more readily accommodated in option 2).

This should be seen against the background of refining resolution and increasing computer power. We note that the current generation of high resolution global (~1/12^o) models 'cross the threshold' identified by Hecht and Smith (2008): "[for models ~0.1^o resolution] Ocean modelling of the North Atlantic has crossed over a threshold into a regime in which the variability of the circulation is comparable to that which is observed", and with this comes many benefits in the simulations. Shelf sea modelling is in a similar position as we move towards eddy permitting models on-shelf (~1km resolution) and coastal models down to ~50m that can (for example) differentiate between individual structures in a wind farm. Given that there are likely to be O(3) HPC computer refreshes over this period, e.g. starting with Archer for the UKRC community in 2014, we will move to the position where these are the workhorse models, and the question as to whether to go further with resolution or focus on other aspects lies primarily with the user community. To try and quantify this, Table 1 lists a series of possible future model configurations, and estimates of their cost with and without a timestep penalty. For these purposes a factor of 5 is assumed for models that use unstructured meshes and a factor of 8 for models run with biogeochemical models (an upper bound for their costs). To estimate when these could become routine models, an exponential fit to the growth of UKRC computers is used (see Figure 3 below).

Table 1: Possible model grids

Grid	S/US	Vertical	Size	Cost	Cost	When routine physics model	When routine ensemble/BGC model
			(k cells)	(time step)	(no time step)		
				cf ORCA025			
Global Scale							
¼	S	75 Z	904	1	1	2011	2014
1/12	S	75 ALE	8150	27	9	2016	2020
1/36	S	100 ALE	73350	973.7	108.2	2022	2026
1/4+1/12 multi-scale	US	75 ALE	2802	46.5	15.5	2017	2021
1/12+1/36 multiscale	US	100 ALE	8700	577.4	64.2	2021	2025
Basin Scale							
1/12 (NA)	S	75 s-Z	1080	3.6	1.2	2013	2016
1/12+1/60 NWS	US	75 ALE	2856	189.6	15.8	2020	2023
Shelf scale (NWS)							
1/12	S	50 s	111	0.2	0.1	2008	2012
1/60	S	50 s	1776	15.7	1.3	2015	2019
1/120	S	75 s	7104	125.7	5.2	2019	2022
Shelf-coast							
1/160+200m UK coast	US	50 s	3448	1346.2	12.7	2023	2026
1/160+200m NW Europe coast	US	50 s	5030	1963.8	18.5	2023	2027
1/160+200m UK + 50m Estuary x N	US	50 s	3448+N*400m	12358.4	27.5	2026	2030

S= structured, US = unstructured

There are many caveats to these estimates, not least the scientific development time needed to achieve the various stages, but they do serve as a reasonable guide to either encourage or constrain the aspirations of the ocean model user community. It also highlights the importance of trying to alleviate the timestep constraints in multi-scale models.

1.3 HPC forward look

High performance computing (HPC) has long been an underlying enabling technology for computer simulation in the ocean sciences. An ocean modelling roadmap should therefore be built on a roadmap for hardware and software HPC technologies. The overlying trend in HPC performance is captured in the data gathered for the TOP500[1] list which has been available since 1993. Figure 1 shows the performance trend over nearly two decades and an extrapolation to 2020. Note the logarithmic scale of the abscissa from which it is evident that performance has been increasing exponentially, doubling about every eighteen months. Also shown is the performance of the winners of the Gordon Bell prize, awarded for exceptional performance from real applications. There are two technology drivers for this. In the first decade, clock speeds were increasing so that each generation of processor ran faster than the previous one. There were also considerable architectural improvements allowing multiple instructions to be executed concurrently (Instruction Level Parallelism). In the second decade the increase has largely been driven by massive increases in parallelism, such that whereas the 1993 list still contained machines with only one processor, now the smallest system on the list has over 1000 cores, the largest has over 1.5 million cores and there are twenty-one systems with over 100,000 cores. The top systems are mostly in the US, but UK and Europe follows with a 2-3 year lag, and in 2012 the first Petascale system in the UK was made available for academic research with a peak performance of over 1 Petaflop/s (10^{15} operations per second) and with 98,304 cores available for a single job.

[1] http://www.top500.org/

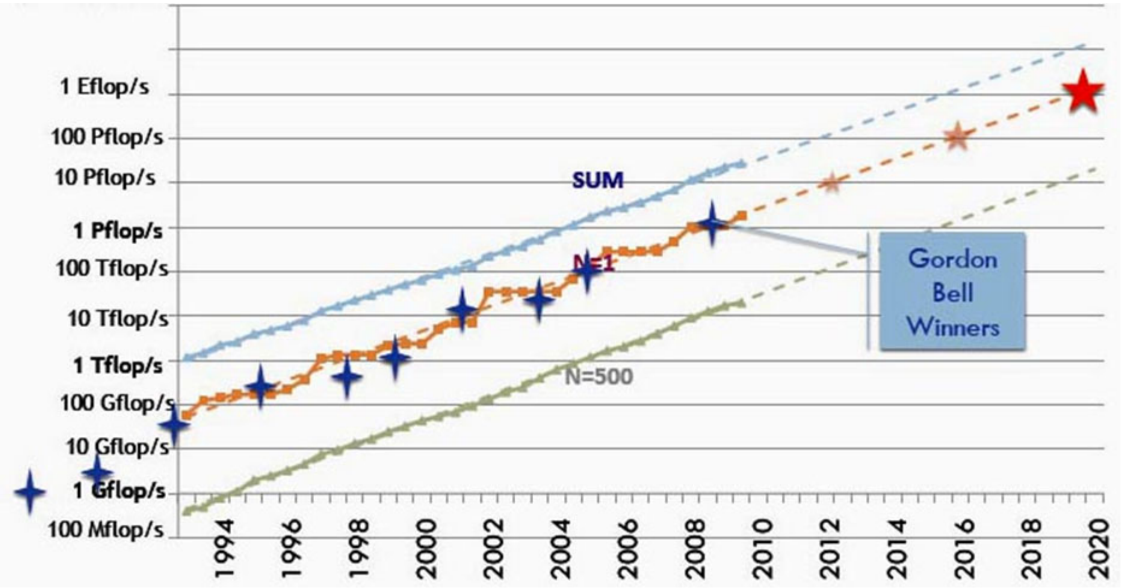


Figure 1: TOP500 data showing performance trends for the top machine in the list (N=1), the bottom machine (N=500) and the sum of all machines in the list (SUM). Also shown is the performance of the winners of the Gordon Bell prize (blue crosses) (credit~: Jack Dongarra, IESP).

Hardware technology is now limited by “The Power Wall” (Figure 2). This figure shows four hardware trends over the past four decades. The number of transistors per chip has been and still is increasing at an exponential rate – this is known as Moore’s Law after Gordon Moore one of the founders of Intel. Intel says that this increase is expected to continue until the end of the current decade, although physics dictates that it must end at some time. The clock speed of processors has peaked at around 2-3 GHz. The reason for this is that the power consumed by a chip increases at a high power of the clock frequency and power densities are limited at around 100W per chip (the “Power Wall”). Instruction level parallelism (ILP) has also peaked at around 4-8 instructions per clock cycle - memories are not fast enough to provide enough operands to justify greater ILP. Further performance increase will therefore be driven solely by increase in parallelism, through larger and larger number of processor cores. We envisage that this trend will continue for the foreseeable future.

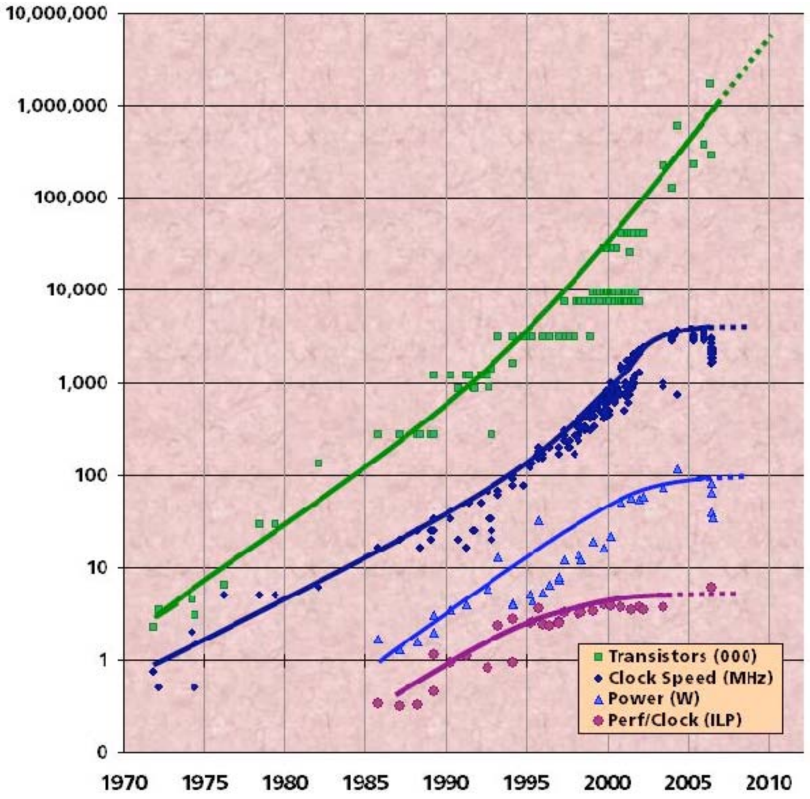


Figure 2: Trends for the number of transistors on a chip (green curve in thousands), the processor clock speed (dark blue curve), the processor power consumption (light blue curve) and the Instruction Level Parallelism (ILP) (purple curve). Credit: Sutter, Olukotun and Hammond)

Following the success of Petascale systems, the first of which appeared in 2008, there has naturally been much speculation about reaching the next thousand-fold increase in performance, the Exaflop or 10^{18} operations per second. Several reports have appeared describing the hardware and software technologies required to reach Exascale, notably the DARPA report "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems" by Kogge et al (2008). This report describes a strawman Exascale design achievable in 2017 (though most experts now believe 2020 or 2022 is more likely) with 166 million cores and consuming 68 MW. One assumption for this strawman is that the power budget for a supercomputing centre cannot grow beyond a reasonable limit. Current technology could be used to build an Exascale system, but that would draw around 1GW, costing $1B per annum, which is clearly unreasonable. Exascale within a reasonable power budget requires around a 100-fold increase in power efficiency over current technologies, and indeed it is this power efficiency constraint which is the limiting factor. It is also here where the major impact for ocean sciences lies. The advantage of Exascale for ocean modelling comes in two forms. There is the prospect of a single ocean model or an ocean model as a part of a coupled climate simulation running at Exascale performance levels on 100 million cores, which would enable huge increases in resolution and in the fidelity of representation of small-scale ocean processes. There will also be a knock-on effect down the pyramid of systems. With the power efficient technology developed for Exascale, Petascale systems will be available with 100,000 cores in a single rack consuming only around 100kw. Current technology will not be able to compete on price/performance.

To get the UK perspective, the past growth of the research council HPC facility provides a guide to future computing capability. There are other facilities available e.g. the Met Office computer, the joint Met Office/NERC computer, capability computing services such as HARTREE, and local clusters. Here we assume these will grow at a comparable rate to the UKRC service. Figure 3 shows the increase in UKRC HPC facility peak performance, from HPCx, through the four phases of Hector to the expected size to Archer. The peak performance of this facility has increased exponentially over the past ~8 years. Given this trend has somewhat flattened off since the rapid increase between HPCx and Hector Phase 2a, the conservative estimate is to extrapolate the trend from Phase 2a to Archer. This gives a peak performance of ~50 times HECTOR Phase 3 by 2019 (100Pflop/s) and ~1000 times Phase 3 by 2023 (5Eflop/s). This closely follows Moore's Law / TOP500 trends, and predicts the UK maintains a performance about a factor of ten lower than the US at any one time (or lags by 3-4 years). There are of course many unknowns in this prediction including the UK Government's continued commitment to HPC, and the share of the resource the UK marine community may receive, but there are presently no indications that these will falter. Figure 3 also shows how memory per core has steadily declined through HECTOR; a clearer picture on this trend will be apparent when the specifications of Archer are available, but all indications suggest it will continue to be downwards.

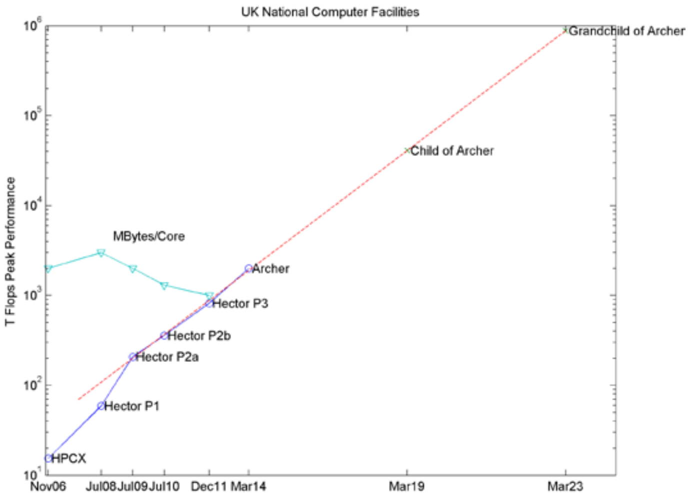


Figure 3 The UK research computer facility. Peak performance and memory per core.

What will this mean for design of next generation ocean codes? In order to effectively utilise large numbers of cores, codes will have to extract very high degrees of parallelism from the underlying numerical algorithms. This will also be multi-level; at least three-way nested parallelism with high-level coarse-grained parallelism at the node level probably using MPI as today, multi-threading on a node using OpenMP or OpenAcc, and fine-grained parallelism within a core, e.g. vectorisation at the loop level. Memory management is also becoming more and more important. The size of memory cannot increase to match the numbers of cores, for reasons of cost and power, so the amount of memory per core will reduce significantly (this has started already). Memory bandwidth per core and interconnect speed per core will also drop. Algorithm design must therefore focus on management and movement of data. Floating point operations are essentially free, so algorithms which repeat calculations in order to save on data movement are to be favoured.

All this means that in principle we will be able to achieve an ocean code running efficiently on hundreds of thousands or even millions of cores, exploiting Petascale or Exascale levels of performance. However, there will have to be major developments in the underlying algorithms in order to map them onto architectures with very high degrees of concurrency.

2. Global Structured grid models – Status and Plans

In order to form a useful component of a climate modelling system, probably the key issue facing global ocean models today is the need to preserve water mass structures in the adiabatic ocean interior over decadal to centennial timescales. This means that spurious or numerical mixing needs to be reduced to an absolute minimum, which is particularly an issue for geopotential (z-) coordinate and terrain following (s-) coordinate models. Several major modelling efforts are now moving towards, or actually employing, vertical coordinate systems that are adapted towards these ends. These are described below.

2.1 NEMO

NEMO (Nucleus for European Modelling of the Ocean, led by Gurvan Madec) is perhaps the foremost physical ocean model in Europe, and is developed through a coordinated effort between partners in France, the UK, and Italy, with scientific direction (in the global context) provided through the Drakkar project, which also involves partners in Germany and Canada. In particular, NEMO is in widespread use for global and shelf modelling applications in the UK. It is run operationally at the Met Office for ocean forecasting, seasonal forecasting and earth system modelling, and is developed jointly for these purposes between NOC and the Met Office. It is also in use in the HEI community for a wide range of PhD projects. It appears to be the model of choice for EC funding and the perception, at least, is that using this model improves funding prospects. The UK is, therefore, firmly committed to NEMO as a strategic model in the short to medium term at least, and NEMO provides an excellent platform for continued UK use into the future.

The most important issues that NEMO faces in terms of its suitability for decadal timescale ocean simulations, and earth system and climate modelling are those concerning ocean mixing, which affects water mass properties (and hence density gradients and circulation patterns) on decadal to centennial timescales.

Firstly, there is the general need to reduce spurious numerical mixing to the lowest possible levels. Such mixing can dominate the explicitly applied mixing in large regions of some ocean simulations (for example in the Southern Ocean) making the model simulations unrealistic. NEMO is already developing a new z-tilde vertical coordinate (Leclair and Madec, 2011) that is isopycnal to high frequency motions, and this will help reduce the spurious mixing from, for example, inertial and internal gravity waves. However, it is recommended that NEMO should move as quickly as possible to a fully generalized vertical coordinate using the ALE (Arbitrary Lagrangian Eulerian) framework. This will give NEMO additional functionality in reducing spurious numerical mixing, and will enable it to keep up with planned developments for other leading models such as MOM. In addition, effort should also be targeted at trials of new advection schemes: The Prather advection scheme (Prather, 1986) is a high order method (which advects second order moments) and has been shown to greatly improve the performance of the MITGCM in representing water masses in the Southern Ocean, again through reduced numerical mixing; the Piecewise Parabolic Method (Colella and Woodward, 1984), is widely in GFD applications, including POLCOMS (James, 2000) and more recently shelf sea applications of NEMO.

Secondly, NEMO should aim to incorporate realistic representations of the physical processes, which actually do mix the water column in nature. This would include for example, mixing by internal waves radiated from deep topography and localized shelf sources, inertial waves, Langmuir turbulence, shear spiking at the base of the mixed layer, and restratification by sub-mesoscale processes (building on the existing Fox-Kemper scheme).

Finally, NEMO should develop a better representation of the mixing of deep dense outflows such as those flowing southwards between Greenland and Scotland. This falls between the two areas outlined above. These waters typically mix too quickly (and spuriously) in NEMO, and in the case of the outflows between Greenland and Scotland, give rise to a North Atlantic Deep Water mass (critical for the overturning of the Atlantic circulation) that is too light and too high in the water column. This could be addressed by either the addition of a stream-tube or other bottom boundary layer scheme to NEMO, by a straightforward “plumbing” fix (to force the overflows down to their correct depths), or by using sigma- (terrain-following) coordinates. This latter method has been shown to help with the downslope flows of such water masses, but the resultant state may be (to some extent at least) an artifact of the depth to which the sigma-layers are taken. This aspect would also need a better understanding of the real-world processes which actually do mix these overflows, which in turn requires guidance from observations.

In addition to these global modelling efforts, NEMO also includes the AGRIF methodology as an important feature, allowing high-resolution regional models to be included as a two-way nest within a lower resolution outer model. This feature is proving extremely useful in the Drakkar programme, with multiple studies underway as PhD student projects. It is also useful for high-resolution studies in targeted areas of operational need. However, further development is needed to enable AGRIF to operate with cyclic wrap-around, in sea-ice regions, and to better distribute available processors if more than one AGRIF region is employed. It is unclear whether AGRIF is the best way forward, as another possibility is to develop a coupler, such as OASIS, for this purpose (essentially to allow it to interpolate and transfer variables in three-dimensions). An investigation is requires as to which would be the most efficient and flexible approach to meet the needs of 2-way nesting.

It is worth noting that the Grids work package in the GungHo project (see below) recommends that a C-grid using quadrilaterals (as used by NEMO and the other leading global models below) gives the best solution for the Shallow Water Equations on a sphere, hence this remains an attractive option.

2.2 Other Leading Models

GFDL/ Princeton. Three leading models are currently being maintained by GFDL/ Princeton. These are **MOM** (z-coordinates, led by Stephen Griffies), the **MITGCM** (z-coordinates, with a non-hydrostatic option, led by John Marshall), and **GOLD** (isopycnal coordinates, led by Bob Hallberg). Agreement has now been reached to merge the GOLD and MOM modelling efforts into a single model using the ALE (Arbitrary Lagrangian-Eulerian) approach for the vertical coordinate (led by Alistair Adcroft). In principle, the equations of motion can be written for an arbitrary vertical coordinate system, allowing those coordinate surfaces both to move in the vertical, and for mass exchange between the surfaces. Typically, horizontal (or lateral) operations (e.g. advection and mixing) are completed first, and then a re-mapping is undertaken in the vertical to restore the coordinate surfaces to their chosen values. The GFDL/ Princeton effort, which will appear as MOM6 in 2013/14, has chosen to adopt a C-grid in the horizontal and pressure-coordinates (z-coordinates) in the vertical, with restoration of pressure to the chosen values through the vertical re-mapping. This is similar to the “hybrid” vertical coordinate used by HyCOM (described below) save that that model employs largely isopycnal (constant-density) coordinate surfaces. The use of constant-pressure coordinates will enable an accurate representation of the equation of state for seawater, which is not so readily achievable in HyCOM.

HyCOM. The Hybrid-coordinate ocean model (HyCOM) is a leading structured model in the US (C-grid in the horizontal), and employs an arbitrary vertical coordinate system. It also forms the basis for the Norwegian operational model system for the Arctic Ocean. The scientific leaders are Rainer Bleck (NCAR and GISS) and Eric Chassignet (FSU), and the model is employed as the US Navy global short-range ocean forecasting system (managed through the Stennis Space Laboratory). Typically, the model uses isopycnal coordinates in the deep ocean (with very low or zero spurious numerical mixing) that transition to z-coordinates near the surface. Probably the outstanding issue with this model is an over-diffusive zone below the mixed layer where the coordinate surfaces transition from isopycnals to z-coordinates, requiring extensive re-mapping in the vertical. An earlier problem with non-conservation of heat (leading to a net heat flux imbalance across the sea-surface of order 0.5 W/m^2) has now been largely reduced (to about 0.1 W/m^2) due to adoption of appropriate time-smoothing of the layer interfaces, together with ensuring that different aspects of the sea-surface height are consistently coded.

3. Development of NEMO for coastal and shelf seas

Here the evolution of NEMO as a regional model for coastal and shelf seas, “NEMO-shelf”, is considered. NEMO-shelf currently includes tides, terrain following and hybrid s- z- coordinates and an appropriate horizontal pressure gradient scheme, and advanced vertical mixing and advection schemes. An option for “wetting and drying” of grid cells is currently under development. The model is now being effectively used as a shelf sea model with similar capability to ROMS, POLCOMS and POM. It is the model of choice in several NERC Research Programmes (RPs, Ocean Acidification, Ocean-Shelf Exchange, Shelf Seas Biogeochemistry) and forms the North West Shelf (NWS) (O’Dea et al., 2012) and IBIROOS (Maraldi et al., 2012) operational capability in the GMES Marine Core Services. The UK is a leading player in the development of NEMO for shelf seas, noting that the NWS application is the only MYOCEAN domain to utilise the full suite of shelf sea features (the IBIROOS application uses z-coordinates). The effort directed towards NEMO-shelf in the UK is modest and benefits from having amalgamated shelf and ocean modelling groups at both the Met Office and NOC. We note, however, there is no analogue to the global ocean Drakkar group for shelf seas. It may be desirable for the UK to convene such a group, particularly as the current cycle of operational oceanography focused EC projects (MYOCEAN II) draws to a close in 2015. NEMO-shelf has only limited use in the HEI sector in the UK. It still lacks certain capabilities needed for local/coastal (sub km) scale modelling and so we treat shelf scale and coastal modelling separately here.

3.1 Shelf Scale

At shelf scale NEMO is routinely run at ~7km resolution for the North West European Continental shelf (the AMM7 Atlantic Margin Model, which is run operationally at the Met Office) and is increasingly used in larger area domains at this resolution (Northern North Atlantic and Arctic). These take advantage of its pedigree as an open ocean model (e.g. using hybrid z-, s- coordinates). NEMO AMM7 scales well to 320 processors on HECToR (Phase3) taking typically 3 hours per model year. Calls on increased computer resources above these modest levels arise from ensemble simulations (data assimilation and uncertainty analysis), ecosystem modelling (particularly those aimed at resolving ecological rather than biogeochemical processes), long transient climate impact simulations and refined resolution. Here we focus on the latter.

Currently there are no finer resolution versions of NEMO-shelf such as the ~1.8km resolution models developed with POLCOMS across the shelf (Holt and Proctor, 2008), in the Irish Sea (Holt and Proctor, 2003) and in the southern North Sea (unpublished). Models at this resolution have the advantage of capturing motions at the 1st Baroclinc Rossby radius across much of the shelf, and so permit eddies and internal tides. They also better resolve the tidal excursion, and so allow tidal straining processes to be resolved (strain induced stratification and convection), and give better geographic representation. As part of the NERC FASTNet RP and in the Joint Coastal Ocean Modelling Programme (for joint work between NOC and the Met Office on NEMO-shelf) a 1/60° Atlantic margin model is under development. This has ~ 1.6 x 10^6 sea points and so is ~2 times larger than the ORCA025 NEMO configuration. Alongside the science objectives, this configuration is aimed at replacing AMM7 as the primary operational model, although use of AMM7 in climate downscaling work will endure. Over the next ~10 years a further refinement to 1/120° would be desirable. This would give a model that comfortably resolves eddies on the shelf and permits them in near coastal currents. It would also match the resolution of available observations (e.g. scanfish and gliders). Refinement below this level at shelf scale would be of limited benefit in a hydrostatic model. Scaling to achieve these refinements should not be a particular issue assuming the computational resource is available as the 1/120° model would still be ~0.5 the size of the current ORCA0083 model (a 1/12° global NEMO model currently running on HECToR). The challenge lies in the appropriate representation of the subgrid scale processes. The role of eddies in shelf seas is less well established than in the open ocean, and little work has been done on the appropriate parameterisation of sub-mesoscale processes in this context (aside from simply maintaining stability). The goal is to find scale-selective parameterisations that lead to a realistic level of mesoscale activity without excessively damping the finer scale motions permitted by the resolution. The need to constrain numerical mixing in a shelf sea context is just as prevalent as in the open ocean, and there has been some success here through the use of Adaptive Vertical Coordinates (Hofmeister et al., 2010), hence a move toward an ALE approach for global ocean models would also benefit the shelf sea models.

Shelf sea modelling is characterised by a demand for many different configurations to meet multiple science and user needs, much more so than in the global case. The process of setting up these model domains and validating them is time consuming with little novel scientific benefit. The capability to rapidly configure shelf sea models would benefit all aspects of their use from research, operational oceanography and regional assessments. It also has benefits for shelf sea modelling in the global context (described below). *Ad hoc* capabilities for this already exist, but these need to be formalised within the NEMO framework.

Regional modelling has generally operated without feedback to other components of the Earth System, despite benefits of interactive ocean-shelf (e.g. with AGRIF) and sea-atmosphere (Schrum et al., 2003) coupling having long been identified. Moreover, land-sea coupling (land-surface – hydrological model coupling), although generally 1-way, is often neglected. Hence, it is timely to consider developing a regional analogue to global earth systems models for example through JWCRP WorkStrand II (Integrated Environmental Modelling, IEM) and the Met Office effort on UK Environmental Prediction.

3.2 Coastal Scale

NEMO has yet to be deployed at coastal and local scales (e.g. in estuaries). This work is currently undertaken using POLCOMS and a variety of unstructured mesh models – both 3D and 2D. The NERC research centres (NOC, SAMS, PML) have largely adopted FVCOM as the model of choice for this, and there is a general tendency to move work from POLCOMS to FVCOM. NEMO currently lacks a wetting/drying capability or coupling to a wave model, both crucial aspects of near shore modelling. Development of the former capability is underway in MYOCEAN II and there are plans to couple WaveWatch III to NEMO, through an NCOF working group. The near coastal zone is an arena where the unstructured approach has distinct advantages since accurately representing flow around complex coastlines and between regions of restricted exchange is crucial, and there is generally a need to focus on a region of interest, while still including far field effects. None-the-less there is still a distinct advantage for pursuing NEMO as a near coastal model. A particular challenge for sub-km scale modelling is the extremely short time steps needed, so coastal simulations are often limited to comparatively short (e.g. seasonal) periods. Hence, the structured approach, being more computationally efficient, becomes attractive for the longer term simulations (multiannual to decadal) needed for (e.g.) climate downscaling and sediment transport. Curvilinear grids offer some multi-scale capability and can be used to refine resolution in (e.g.) estuaries and through straits, but the issues with wave propagation along staircase coastlines (see below) remain.

Therefore, it makes sense to complete the transition of near-coastal capabilities from POLCOMS to NEMO, to complement the capability of FVCOM. NEMO has a curvilinear grid capability unlike POLCOMS, which now has an essentially static dynamical code base, and being a B-grid model is less suited to this work in the first place. Such a transition would allow efficient working and sharing of ideas between shelf seas and near coastal modellers.

4. Improving the representation of shelf seas in global NEMO models

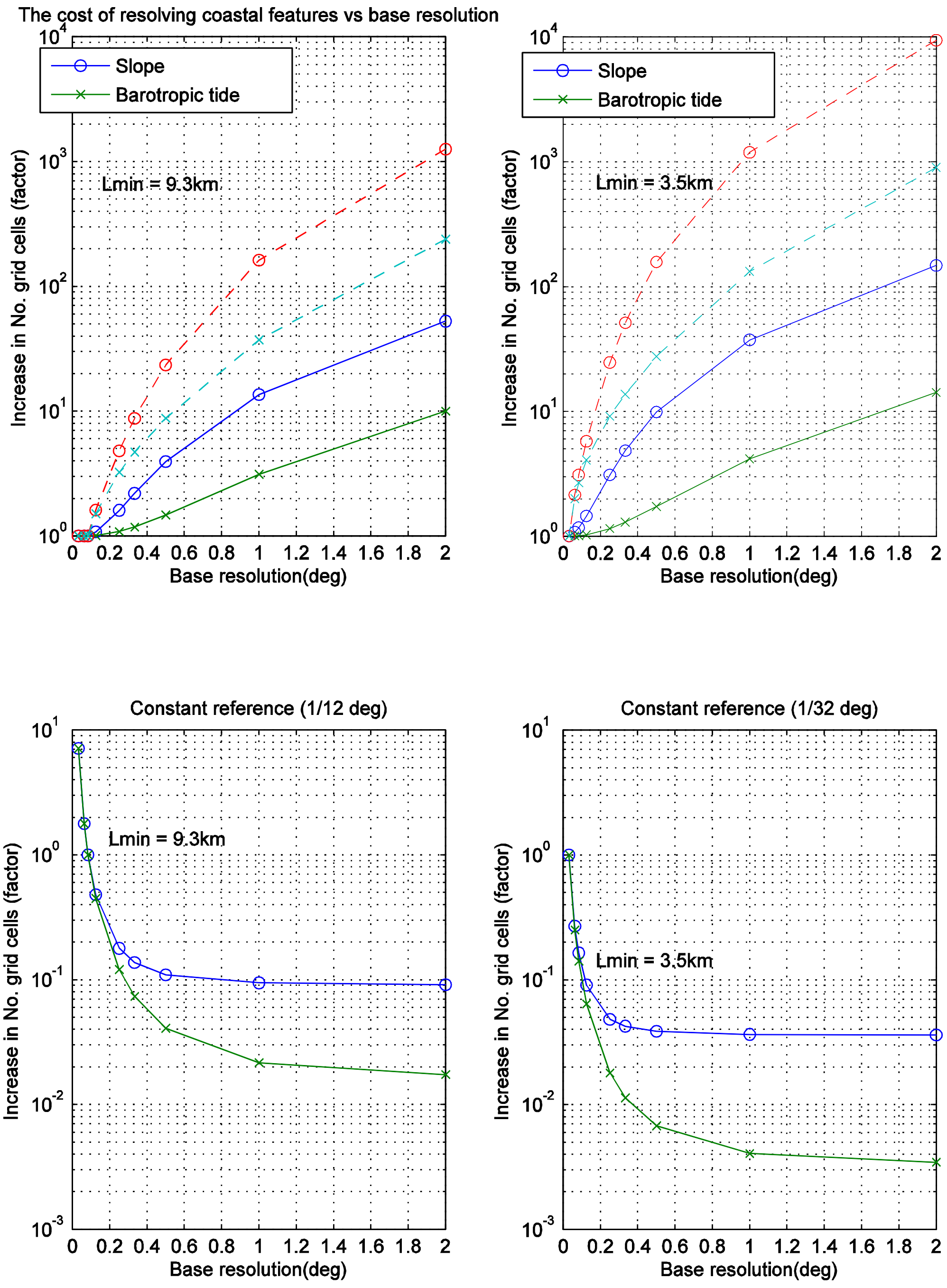
Improving the representation of coastal and shelf seas in global models represents one of the grand challenges in ocean modelling and indeed Earth System science, and so requires some detailed consideration. There are several motivations for this including dynamics, biogeochemistry and ecosystem, global scale impacts and regional scale impacts. Specific examples include: the effects of tides on dense water production in Arctic shelves (Postlethwaite et al., 2011), the fate of carbon sequestered by highly biologically productive shelf seas, global food security implications of climate change effects on fish meal production (Merino et al., 2012) and the need to understand the structure and functioning of shelf sea ecosystems and so aid policy objectives such as the requirement to reach and maintain Good Environmental Status in the Marine Strategy Framework Directive.

4.1 The Challenges

There are two facets to the specific challenge of improving shelf seas in global scale models: the processes and the scales. The first of these is comparatively straightforward. There are specific processes that are important in shelf seas and often neglected in global models. To improve shelf sea representation these should be introduced into global models, drawing on the approaches developed in regional shelf sea models. The challenge is to ensure this does not compromise the accuracy of the open-ocean solution or substantially increase the computational cost. For example, global 1/4^{o} global models can be run with tides, but potentially at the risk of enhanced spurious mixing. The bigger challenge relates to the small scales needed to represent the processes and geography (coastline, bathymetry, straits) of shelf and coastal seas; it is issues of scale that have perpetuated this overall challenge. To illustrate this issue it is worth considering how many extra grid cells a global model would need to resolve particular features of shelf seas globally. Two simple examples are: barotropic Coastal Trapped Waves (CTWs, e.g. tides) with length scale: $L_{\mathrm{bt}}=(gH)^{0.5}/f$; and topographic depth relative to slope: $L_{\mathrm{sl}}=(\mathrm{H.}|\mathrm{dH/dx+dH/dy}|^{-1})$. The latter is derived from the depth integrated continuity equations and is the scale below which barotropic motions would be expected to be topographically controlled (Greenberg et al., 2007). Hence the number of cells needed globally can readily be calculated from a fine resolution bathymetry and mask (here we use that of the 1/12^{o} NEMO model), given the number of cells needed to resolve the feature, a coarsest expectable resolution (the base resolution), and a minimum length scale (needed because both length scales tend to zero with water depth, H). Results can then be compared with an unrefined model (at the base resolution) or a ‘high resolution everywhere’ option at the minimum length scale.

Examples results are shown in Figure 4 depending on the base resolution for two minimum length scales: 9.3km (~1/12^{o}) and 3.5km (~1/32^{o}). The former is typical of fine resolution global models (e.g. ORCA0083 but not yet used for coupled climate modelling) and coastal ocean models coupled to ecosystems or used for climate downscaling. The latter has been proposed for global scale modelling, but not commonly realisaed, and is used in some regional applications (e.g. Maraldi et al., 2012). So for example, a 1^{o} global model refined where necessary to resolve CTWs down to 9.3 km has 3 times more grid cells or is ~13 times more expensive if some account of timestep is made. These factors increase to 4 and 150 if a minimum resolution of 3.5 km is required, and resolution of the bathymetry is substantially more expensive. Hence, in this case resolving these features quickly dominates the rest of the model. At a 1/4^{o} base resolution these figures are substantially reduced, e.g. resolving the slope at 9.3 km needs a factor of 4 at a 1/4° base resolution compared with ~180 at 1° base resolution. So the 1/4^{o} model with refinement to resolve slopes is ~256 times more expensive than the 1° without such refinements, but only ~1.4 times the 1° with refinements. Hence, in a global refinement exercise the starting point is likely to be a reasonably fine resolution global model, and these need to comfortably fit into the computational resource before attempting the refinement. A particular caveat of these results is that no account is made of the overhead in achieving this refinement, which would be expected to be considerable.

Figure 4 also compares the refined models with two global fine resolution options. The global 1/12^{o} is ~5 times more expensive than a 1/4^{o} model refined to resolve slopes, and a 1/32^{o} is 20 times more expensive. These figures need to be seen alongside the needs of the open ocean model. For example, Griffies et al (2009) note (in the context of mesoscale eddies): “There is no obvious place where grid resolution is unimportant”. Hence, unless very efficient methods of multi-scale modelling are developed the added benefits of the higher resolution global model (e.g. improved Gulf Stream separation) are likely to outweigh the marginal improvements in efficiency achievable by a multi-scale method if only modest coastal-ocean representation is required. If fine resolution coastal ocean representation is desirable or if the processes of interest are less widespread than those considered here (e.g. eastern boundary upwelling or cascading in the Arctic) then the scaling begins to favour multi-scale modelling.



The cost of resolving coastal features vs base resolution

Constant reference (1/12 deg)

Constant reference (1/32 deg)

Figure 4. Number of grid cells to achieve process representation in shelf seas. Top panels are relative to base resolution i.e. fractional cost to refine the global grid to resolve barotropic tides and slope, dotted lines adds in a factor for explicit time stepping. Bottom panels are relative to a fine resolution global model set to match the minimum length scale of ~9.3km (1/12 deg; left) and ~3.5km (1/132 deg; right), i.e. representing the saving that might be achieved by a multi-scale approach.

Temporal resolution is less of a challenge and the short timescale processes of interest in shelf seas (at ~ diurnal frequency) are increasingly already considered in global models, particularly when explicit methods are used to treat fast barotropic waves. The model timestep is crucial when considering multi-scale modelling approaches.

Care is needed when using implicit methods that accuracy (of the representation of these waves) as well as stability remains a prime consideration. A particular consideration is, however, that it is difficult to accommodate reduced timesteps by increased parallelism in the same way that increased horizontal resolution can be.

4.2 The Options

For our purposes, there are four approaches to improving the representation of coastal and shelf seas in the global context: process representation, parameterisation, structured grid approaches, and unstructured grid approaches (considered below in sections 6 and 7). Structured grid approaches naturally divide into 'high resolution everywhere' approaches, curvilinear coordinate approaches and nested approaches.

4.2.1 Process representation in fine resolution global models

As noted above the representation of coastal processes in open ocean models is straightforward at least in concept, particularly now the NEMO model has the capability of simulating both open ocean and shelf sea cases; this essentially allows these processes to be included as the global model resolution is refined. It is worth noting that the present generations of global NEMO ocean models (ORCA025 – 1/4° resolution, and ORCA0083 – 1/12 ° resolution) are at a comparable resolution to many shelf sea modelling efforts, historic and on-going. Specific physical processes include tides, multiple boundary layer mixing (e.g. with Generic Length Scale (GLS) approaches), riverine fluxes, coastal up- and down-welling, and dense water cascades. The need to resolve benthic boundary layers means these processes are facilitated by fine vertical resolution. This can readily be achieved using arbitrary terrain following coordinates, but comes at the cost of the need for more advanced treatments of the horizontal pressure gradient and diffusion terms, and often the need to smooth the bathymetry to reduce the slope of the coordinate surfaces. Hybrid z-s approaches can alleviate this (O'Dea et al., 2012). The open question is whether features pertinent to the coastal-ocean can be introduced into global models without degrading the solution of these models in the open ocean or significantly increasing their computational cost. This requires a systematic exercise in model development and assessment by both global and coastal-ocean modelling communities.

It should be noted that the resolution used for global (and indeed regional) coupled hydrodynamic-ecosystem simulations has increased very slowly over the past decade, and certainly lags behind the increase in computer power over this period. Most still use ~1^{o}, despite clearly identified benefits of improved physics (e.g. Hecht and Smith, 2008) and the importance of this physics on the ecosystem (e.g. Sinha et al., 2010). Hence, noting that many of the motivations behind this driver for change lie in biogeochemistry and ecosystems, this option is closely coupled to the need to improve computational efficiency of the NEMO model and its ability to exploit current/future computer systems.

A concerted research effort is required to explore the differences in shelf seas between global and shelf sea NEMO model versions, so that these differences can be specifically justified on scientific and/or practical grounds, and where appropriate harmonised. Both NOC and the Met Office now have open ocean and shelf sea models in the same groups and the coordinating programmes, JOMP and JCOMP should facilitate a cross comparison exercise.

4.2.2 Parameterisation

For cases where a significant increase in computer resource is not a practical option, parameterisation is the primary solution. This essentially involves using theoretical knowledge and empirical information to develop a representation of subgrid scale processes and geographic features. Examples include tidal parameterisations (Koch-Larrouy et al., 2008), immersed boundary methods (Tseng and Ferziger, 2003) and porous barriers (Adcroft, 2013). Central issues of this approach are 'testability' and 'uniqueness' i.e. achieving an improvement in the model for scientifically sound reasons that could not be achieved if another absent process were considered instead. Solutions to these issues lie in having a sound theoretical basis for the parameterisation (e.g. its functional or dynamical form), using multiple data sources (particularly of flux rather than state variables) and assessing as many components of the system as possible to avoid error compensation. While, these shortcomings must be born in mind, this is the only approach available for some classes of model, and so must be given serious consideration if any progress is to be made in the representation of shelf seas in these. The most obvious example is an Earth System Model used to simulate biogeochemical cycles on centennial to millennial timescales. There has yet to be a comprehensive UK effort in parameterising shelf seas in global models; there was some consideration of this in the QUEST programme (Allen et al., 2010), but this was not fully realised. Beyond the simple statistical relations common in parameterisations are embedded dynamical models or emulators of these (termed "super-parameterisation"). Examples include a surface-slope driven 1D tidal model which was embedded in a 2-degree NEMO model in MARQUEST and deep convection which has been parameterised using embedded 2D non-hydrostatic models.

The development of subgrid-scale parameterisation is a key output of much process oriented research and so can be followed by doctoral training, responsive mode, and research programme funding lines. However, the step from an empirical or theoretical relation to a formulation that is well tested in a 3D General Circulation Model (GCM) is often elusive, and very difficult to achieve within the scope of a PhD studentship or standard grant. Hence, approaches for the pull through of new process knowledge to GCMs needs to be better established. This requires model test platforms that are readily accessible and usable in the wider community – including a degree of support and training. A working group on this topic could be convened to sit within the JWCRP UKESM framework – with the role of horizon scanning and facilitating pull through.

4.2.3 Structured grid multi-scale models: curvilinear approaches

Structured grid models have some scope for multi-scale capability by distorting their horizontal coordinates. This is generally limited to a single region of interest and requires considerable investment in model configuration (a new global model must be configured and tested); so this area must be likely to persist as a focus of interest. An obvious example to facilitate regional impact studies would be to focus resolution on European Seas, through the use of (e.g.) a rotated polar grid (e.g. Gröger et al., 2012). Given its lack of flexibility we do not recommend pursuing this avenue.

4.2.4 Structured grid multi-scale models: nesting

Nesting is the most common approach to multi-scale modelling. In its simplest form boundary conditions for a fine resolution regional model are taken from a previous run of larger area oceanic and atmospheric models. The underlying assumption to such one-way nesting is that feedbacks between the regional and global simulations are small, at least on the timescales of interest. Such feedbacks can occur both in the ocean and atmosphere leading to some effort in two-way ocean model nesting and regional coupled models. However, when the regional model becomes large (e.g. basin scale), then there is an issue of dynamic balance as the forcing atmospheric heat flux arises from a different underlying ocean temperature.

The general downside to nesting is the accuracy to which information can be exchanged between the two domains and the degradation of the solution at the boundary; it is usual to linearise the boundary conditions and to only exchange a limited subset of information at lower frequency than the model timesteps of the region in question. That said, there has been extensive work on regional model boundary conditions and by using a careful combination of active and passive approaches good solutions can be obtained. One way nesting can be straightforwardly extended to a global scale using multiple regional nests. The NERC eScience GCOMS project undertook this by developing a system to automatically configure and run simultaneously an arbitrary number shelf sea regions, that could cover all the world's shelf seas (Holt et al., 2009). Its practical application in the NERC QUEST_FISH project (Barange et al., 2011) was, however, limited to 21 of the 42 regions needed for global coverage, although these covered ~66% of the global fish catch.

A natural extension of the nesting approach is to move to two-way global scale nesting, (e.g. using AGIF or OASIS; but noting both would need substantial development for this). The main advantages of the nesting approach is that the system is less expensive (as it needs fewer grid cells) than a fine resolution global model, and can be run at a different timesteps and with different processes represented in global and shelf sea regions. In the one-way context it gives the flexibility of running the shelf sea system without re-running the coupled global model(s). The challenge for global scale nesting is to maintain these advantages to a significant level that makes the exercise worthwhile compared with a global high resolution model. Ocean-shelf coupling is best achieved by placing the open boundaries in the open-ocean, so the shelf-slope region is included in the finer resolution model. Hence, the (irregularly shaped) definition of shelf sea regions in GCOMS covers about 30% of the global ocean area, i.e. substantially more than the shelf sea area itself. The minimum fractional increase in cost of such a nested system relative to an unrefined ("outer") model is $1+rf^3$, where r is the fraction covered by nests and f the increase in resolution, i.e. this rapidly reaches the position of the nest using all the resource. Relative to a global increase in resolution this is: $1/f^3+r$, i.e. quickly asymptoting to r. This excludes the costs of coupling, which can also be considerable (both in computation and scientist time). Hence this approach is practical so long as the fraction of the global ocean to be considered is small, so lends itself to downscaling studies (such as QUEST_FISH) rather than upscaling (such as effects on basin scale dynamics or global carbon cycles).

Nesting remains the most practical approach to meet the regional impacts and to some extent global impacts motivations, these can be met with the improved relocatability capability (described above) and this can be developed to cover multiple regions using the GCOMS approach. For cases where fast two-way interaction is needed (e.g. dynamics) a ‘high resolution everywhere’ and parameterization approach is favoured. A two-way nested system for global biogeochemical cycles is conceivable, but given the likely technical effort needed this would better be spent on developing appropriate process representation and scalability in global models.

5. Developing NEMO for next generation computers

5.1 Introduction

Given the robustness of the quadrilateral C-grid approach and the strong position of NEMO nationally and internationally, it is important to develop it for the next generation of computer architectures. Such a process would need to be carefully coordinated with the NEMO consortium, which needs to ‘own’ the process. It potentially would need to accommodate a wider range of computer architectures than an approach tailored to a specific system. While this may end up not being optimal for any partner it does make the system more portable.

NEMO was originally designed as a static memory code optimised for vector computers, and when adopted by the UK both NERC group engaged in its used (NOCS and POL) both suffered a reduction in computational performance compared with their ‘in-house’ models (OCCAM and POLCOMS), which were well configured for parallel machines (e.g. Ashworth et al., 2004). Initial tests showed POLCOMS being 4-6 times faster than NEMO V2: I/O issues, computation on land and cache reuse (array ordering) were cited as explanations at the time. Since then the performance of NEMO has steadily improved and the current dynamic memory version has slightly better performance than POLCOMS with simple partitioning (Figure 5; but noting this comparison is weighted in favour of NEMO, being the larger grid). The NEMO AMM7 scales well to ~800 cores, but no further.

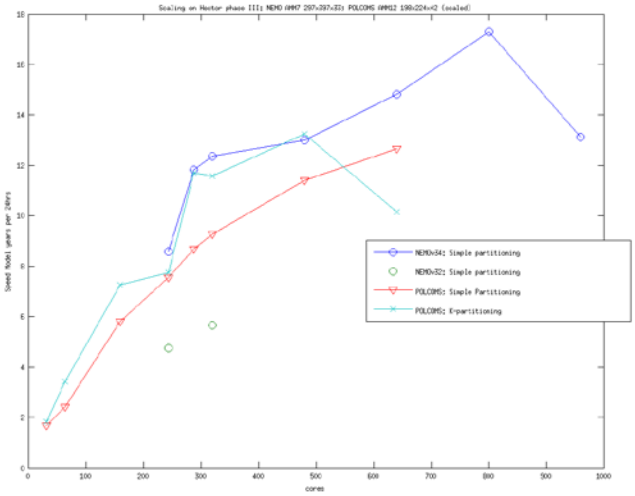


Figure 5. NEMO AMM7 scaling on HECToR Phase III, compared with POLCOMS AMM12 (scaled).

The 'Flagship' model of the Drakkar consortium (1/12^{o} Global ORCA0083) scales to at least 4684 cores on HECToR Phase 3 (Figure 5); the speedup between 1784 and 4684 cores is ~40% for a ~60% increase in cost. Also, the CNRS group has been able to achieve reasonable scaling up to 10,000 cores (S. Masson).

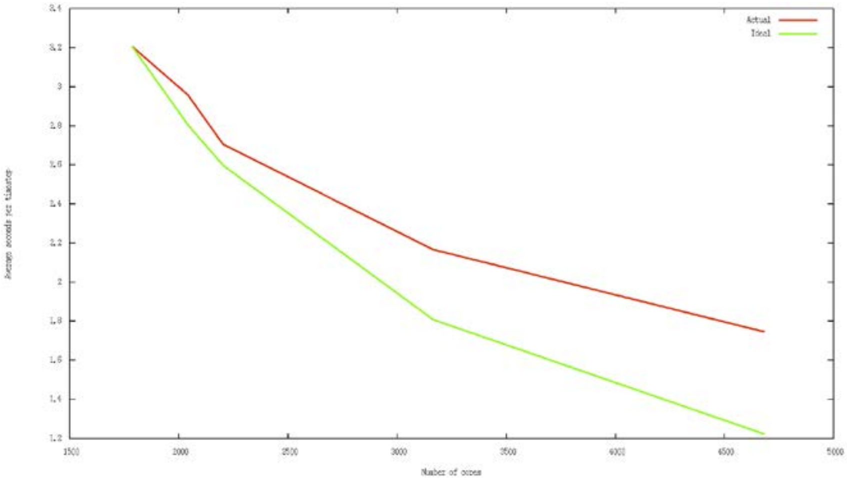


Figure 6. Scaling of ORCA0083 on HECToR Phase 3.

Critical in achieving this level of scalability is the introduction of a new I/O server (XIO, which defines a set of cores dedicated to I/O). Figure 7 shows that there is not a systematic increase in I/O costs with increasing core count and with some partition arrangements (e.g. 2040 cores on 110x25) the asynchronous I/O effectively smoothes out the impact on run-time. For example, in the simulation with 4684 cores (light blue curve), the processors running the model pass their data to 120 I/O processors as a message. These I/O processors then collate all the incoming data and write to disk, while the model is still running, and complete the task before the next I/O step.

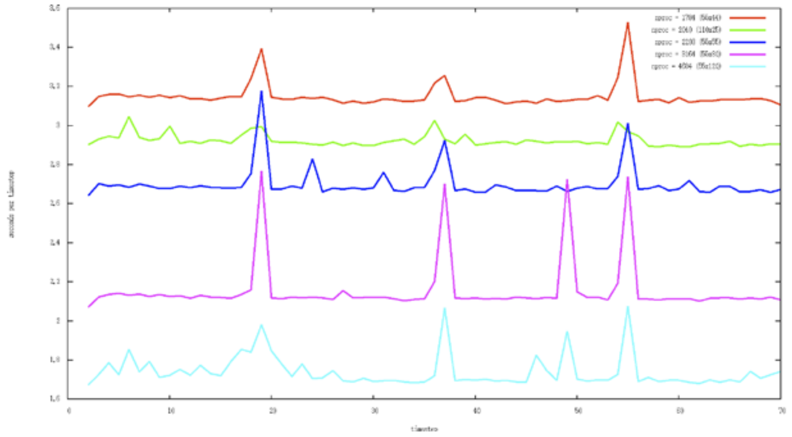


Figure 7. The number of seconds taken for each timestep during a NEMO ORCA0083 simulation on HECToR. History files are written at timesteps 18, 36,54 and 72.

5.2 Current issues with NEMO Optimisation

The current release of NEMO is written in Fortran90 and parallelised using MPI alone. A domain is decomposed geographically into equally sized rectangular sub-domains (c.f. the K-partitioning used in POLCOMS, which recursively splits the domain into rectangles of different size to achieve a balanced load). Fields are represented as 3-dimensional arrays with the level (vertical) index last, a choice which is well suited to vector architectures. NEMO performs redundant computations on land and beneath the seabed, the results of which are masked out by multiplication with 3-d mask arrays. Like many climate modelling codes, NEMO's profile is fairly flat; in many configurations, no single routine accounts for more than 10% of run-time. Consequently, attempts to improve the performance of NEMO by expedients such as single-processor optimisation, introducing a second level of parallelism, or enlisting accelerators such as GPUs, must touch a large number of subroutines. By and large, the computational intensity of NEMO is low, and performance is limited more by memory bandwidth and latency than by floating-point speed.

The main factors affecting the performance of NEMO include:

1. Effective load imbalance due to redundant computations.
2. Load imbalance in the presence of sea-ice.
3. I/O performance.
4. The large number of MPI halo exchanges needed when using the time split bartropic solution method

The effective load imbalance due to redundant computations on land can easily account for 50% or more of run-time in coastal applications; additional redundant computations below the sea-bed (in z-coordinate applications) only makes matters worse. However, NEMO does have the ability to eliminate land-only sub-domains; this alleviates the effective load imbalance somewhat, and has the fortunate property of tending to eliminate more land as the number of processors is increased keeping the problem size fixed.

Load imbalance due to sea-ice is difficult to treat. Re-distributing fields before a sea-ice calculation (which do not occur at ever time-step) is in principle possible, but would involve trading off additional expensive phases of communication against improvements to load balance. Weighting a k-partitioning approach to the presence/absence of ice would be a compromise option. Transitioning to a hybrid MPI+OpenMP approach might open more opportunities for exploiting functional parallelism, provided that a numerically acceptable method can be found for working with out-of-date sea-ice results.

Historically, the scalability of input and output performance has been limited by having separate files per MPI process (which tends to overload the file system at higher process counts), and the lack of asynchronous I/O, which causes computations to block until the I/O completes. As demonstrated above, the introduction of an I/O server model within NEMO alleviates both of these inhibitors. This is heading in the right direction, but these results demonstrate that careful case-by-case tuning is required, which limits flexibility.

5.3 Recent optimization work for NEMO

Here we describe recent and ongoing development work on the NEMO code that is relevant to making the software better suited to future computer architectures.

The project “Developing NEMO for Large, Multi-Core Scalar Systems” (Pickles and Porter., 2012) was a serious attempt to reduce the effective load-imbalance due to redundant computations on land. Their approach featured: (1) decomposing the domain into variably-sized rectangular sub-domains of approximately equal computational load (i.e. number of land-points) using technology proven in POLCOMS (Pickles, 2010), (2) changing the definition of all 3-d arrays so that the level index varies fastest, (3) permuting loop nest orderings to match the new index layout, and (4) changing loop bounds to avoid land points. Key to their approach is having the level index first in array definitions. Without this, loop-level avoidance of the dry points that remain in a sub-domain after partitioning cannot be done efficiently. They found that the level-first approach improves the performance of halo exchanges (and hence overall scalability) noticeably, and is slightly favourable or at least neutral for single-processor performance on most subroutines on cache-based systems. A further advantage of this approach is that in the strong scaling limit where the sub-domain size becomes small, the size of the vertical dimension over which the compiler vectorises is unaffected. However, the vertical implicit solvers (used for mixing and turbulence parameterization) involve the solution of a tri-diagonal system and do not readily vectorise in the level-first ordering, due to loop-carried dependencies between layers. As they ran out of time to implement loop-level avoidance of dry points (step (4)) their conclusions at the end of the project were ambiguous. It is clear that this kind of transformation is both invasive and pervasive, would take considerable effort to apply to whole NEMO code base, and could be disruptive to the NEMO development cycle.

A follow-up project FINISS (commenced January 2013) aims to implement sufficient loop-level avoidance of dry points so that a properly informed recommendation concerning the competitiveness of this approach can be made. FINISS will also attempt to eliminate avoidance of redundant computation beneath the seabed in both the level-first and level-last orderings, although in the latter case the avoidance will be limited to the points beneath the deepest layer in a sub-domain.

We are aware of several investigations into the introduction of a second layer of parallelism in NEMO using OpenMP threads:

a) Italo Epicoco and Silvia Mocavero of the Euro-Mediterranean Center on Climate Change (CMCC) have been working on an MPI+OpenMP hybrid version of NEMO. They have already produced a complete hybrid version for the limited-area Mediterranean configuration in use at CMCC. They introduced OpenMP at the level of the outer loop over threads (in the existing level-last ordering) and thus model levels are shared amongst threads. They calculated that some 86% of the code can be parallelised readily using this approach.

b) As part of the PRACE IIiP project, Andrew Porter of STFC investigated introducing OpenMP to some of the most costly NEMO subroutines. He found that performance is improved by further sub-dividing the MPI sub-domains into 'tiles' and parallelising the loop over these tiles with OpenMP. This reduces the number of thread synchronisation points and encourages cleaner, better-structured code. In addition, the tile size can be tuned independently of the number of OpenMP threads to improve cache reuse.

c) Currently, STFC and CMCC are collaborating in further investigations into the use of tiles with OpenMP in NEMO. While STFC has experimented with 2-dimensional tiles in different array-index orderings, CMCC are investigating the use of 3-dimensional cuboids in the level-last ordering. As part of their commitment to the 2013 NEMO work packages, CMCC are continuing the implementation of a hybrid parallel approach based on the 3D domain decomposition as well as functional parallelism for the tracers and momentum equations.

We are also aware of two investigations into extending NEMO to make use of general-purpose Graphics Processing Units (GPUs):

a) In the gNEMO project, Andrew Porter of STFC compared two different directives-based approaches for programming GPUs, examined the impact on the code base of incorporating GPU functionality and finally, and compared GPU performance with that obtained from an OpenMP version of the same code on multi-core CPUs (Porter et al., 2012). He ported five NEMO subroutines to GPUs in this work, and also parallelised two of them using OpenMP. He obtained speedups for four of the five subroutines on NVIDIA Fermi GPUs (sea-ice being the exception as it is modelled in 2D and offers limited parallelism). However, the pure OpenMP versions running on a full socket of an Intel Westmere CPU yielded 60-70% of the GPU performance. He found that programming for heterogeneous memory spaces is difficult, code bloat can be a problem, and that great care must be taken to minimise the effect of the bottleneck resulting from the PCIe bus between CPU and GPU.

b) Recent work by NVIDIA is a little more encouraging. Using the OpenACC directives-based standard (which was not available for the gNEMO project), NVIDIA ported the core of NEMO version 3.4 to GPUs. Their port is able to run the GYRE configuration, with I/O disabled. Since data transfer between CPU and GPU is so slow, the whole model runs on the GPU(s) and only halo data is transferred to the CPU for the calls to the MPI library. To improve the efficiency of the halo swaps, NVIDIA have added functionality so that halo swaps for multiple arrays can be performed in one operation. For GYRE_50 they found 5x speed-up when using four Tesla M2090's compared to 24 Intel Xeon X5670 cores (2.93 GHz).

5.4 Strategies for optimising NEMO

To remain competitive, it is expected that simulation codes must be able to exploit ever increasing numbers of core. If single core performance declines this may be required just to deal with current workloads on future systems, let alone achieve the next order-of-magnitude resolution improvements demanded by scientific ambitions.

The arguments (and evidence) in favour of introducing an additional thread-based layer of parallelism into an MPI code are more compelling than ever. Exploiting the shared memory within a cluster node helps to reduce the cost of communications thereby improving scalability, and to alleviate the impact of decreasing memory per core by reducing the need to store redundant copies in memory. On many architectures, it is also possible to hide much of the effect of memory latency by oversubscribing execution units with threads.

For NEMO, being written in Fortran, the easiest route towards an additional layer of thread-based parallelism is OpenMP. However, there would be profound consequences for how a hybrid MPI+OpenMP code is managed (and developed, tested, and maintained), that go far beyond dealing with the possibility of small changes in the numerical results. The changes required to the current NEMO code base to get to a hybrid version are both invasive and pervasive. It is very easy for a change to a code fragment to break the correctness and/or performance of the OpenMP version, and the compiler cannot do very much to help. Given the significant number of different options in NEMO, careful code configuration management and quality control would be required, placing an additional burden on the Systems Team.

Although in principle the transition to a hybrid version of NEMO can be done in an evolutionary, incremental way, in practice, it will feel much more revolutionary. This is for two main reasons: (1) there are at any time many parallel (science) development strands, which would overlap with the hybridisation work, and (2) in NEMO (as in most codes), there is very little separation of the scientific/numerical concerns from the concerns of performance and parallelism. For example, in NEMO, halo exchanges are sprinkled throughout what many would call "user code". An approach that requires OpenMP directives to be mixed into almost every piece of user code demands a much increased, and potentially unreasonable, level of sophistication from scientific contributors.

Despite the growing interest in the use of accelerators such as GPUs, we are not convinced that they offer a viable, cost-effective and attractive near-term solution for NEMO. We would want to see further convergence on programming models, and more compelling evidence for their cost effectiveness in codes from kindred fields. In the near term, the dependence on the PCIe bus for communications between host and accelerator is a proven bottleneck and difficult to deal with. Many vendors' roadmaps anticipate its eventual elimination, with the many cores of the accelerator moving onto the same piece of silicon as the host. In that world of many heterogeneous cores on a single chip, it is not unreasonable to hope that a hybrid MPI+OpenMP code might be in no worse a position than a code that has already been optimised for current-generation accelerators. Although this is a space that needs close monitoring, we do not currently recommend further major investments in porting NEMO to GPUs.

We consider instead three main ways forward for improving the performance and scalability of NEMO:

Approach 1. The first approach is the most incremental and evolutionary of the three. Persist with the current MPI-only parallelism, and the current array index ordering. Concentrate on functionality. Make incremental improvements to performance and scalability where necessary and possible. Continue to develop the I/O server (which is on the right track).

Approach 2. The second approach is more revolutionary, involving invasive and pervasive changes to the current code base. Conclude investigations into the ordering of array indices, and decide once and for all whether to stick with the level last ordering or to transition to a level first ordering, bearing in mind that the latter will require considerable effort and the former will doom NEMO to at least some degree of redundant computation. Having decided on the array index ordering, choose a tiling strategy and begin the transition to a hybrid MPI+OpenMP programming model.

Approach 3. The third approach is to start afresh. The main argument for a fresh start from a software engineering point of view is the prospect of achieving cleaner separation of the "user code", which deals with the scientific and numerical concerns, from the infrastructure, where the concerns of performance and parallelism are addressed. If this can be achieved, it should result in a code which is more future-proof against changes in machine architecture. But this is not yet proven. Whether the returns would justify the major investment of a complete re-write is clearly moot. Moreover, careful consideration would be needed before embarking on a 'science neutral' rewrite as potential benefits of novel approaches (such as multi-scale capability) would not be realized. However, this could be seen as a 'low risk' next generation option (and could possibly be conducted in the GungHo framework, see below, as the science approach is at least well tried and tested).

The incremental approach (1) will result in a NEMO whose performance steadily diminishes in competitiveness over time. Eventually, perhaps in five years' time, the scalability shortcomings will be noticeable and users will look for alternatives.

The more revolutionary approach (2) might buy NEMO an additional five years or so of good service. We estimate that at least ten (an educated guess) FTE-years effort would be required to make the transition. The ability to ingest new functionality during the transition will be diminished, and the NEMO release schedule might be temporarily impacted. After the transition, we expect that new functionality will take a little longer to incorporate and validate than it did before, owing to the increased complexity of the programming model. Nonetheless, this is our recommendation. The viability of option (3) may depend on how the computational science layer of GungHo evolves and, given that no new functionality would be gained it is not currently a preferred option.

We further recommend that:

- If the decision is taken to adopt the level first ordering, then some effort should be invested in finding a way to get the tri-diagonal solves in the vertical to vectorise. It would seem desirable to try to put as much of this as possible into a common subroutine, since there are currently several routines that follow the same pattern and would need modifying in similar ways.
- Once the exploratory work is out of the way, it will be important to minimize the time in which the NEMO trunk is in an inconsistent state. Thus we are recommending a “freeze and sprint” strategy.
- The NEMO core development team would own the transition. They will need more resources to do it effectively. We recommend that the core team in Paris be augmented by at least two dedicated developers with OpenMP skills for the transition period, and (say) one thereafter.
- Invest in training developers throughout the NEMO consortium, updating the NEMO style guide, and improving testing and validation procedures.

6. Unstructured models – Status and plans

The most relevant currently-available unstructured models, which could potentially be considered as the basis for a UK unstructured model are now considered. These are either Finite Volume (FV) or Finite Element (FE) codes, as opposed to the global models described earlier which are Finite Difference (FD). Sergey Danilov (AWI, Bremerhaven) is currently preparing a review article describing the methods used by these and other related models in more detail.

MPAS-Ocean. This development is being led by Tod Ringler at Los Alamos National Laboratory (LANL) in the US. MPAS is an FV-FD model using Voronoi tessellations in the horizontal (mostly hexagons for practical applications) and the ALE coordinate system in the vertical. Importantly, this has been run in a global configuration at 15 km resolution (Ringler et al., 2013), and has achieved a performance only order 2-3 times slower than an equivalent configuration of the POP (Parallel Ocean Program) model (an optimized, structured, FD code). It is also being coupled as part of an Earth System Modelling framework (CESM) and as such is likely to form the ocean component of the next NCAR IPCC climate modelling effort, and to overtake POP as the leading ocean model at LANL. The model uses the TRiSK solution method, however, which gives an inconsistent reconstruction of the Coriolis term if irregular hexagons are used in the horizontal (see GungHo grids workpackage report). Because it uses an orthogonal hexagon horizontal grid, the model representation of complex coastlines will be sub-optimal, compared with an arbitrary triangular mesh for example. Momentum conservation maybe another issue with this approach.

FVCOM. This is an FV model developed through a joint effort between UMASSD and WHOI in the US (led by Changsheng Chen), employing a triangular mesh in the horizontal, sigma-coordinates in the vertical, and including the GOTM vertical turbulence scheme. It has largely been applied to coastal and estuarine contexts so far, although there are global and Arctic Ocean applications. It is in use in several research centres in the UK (NOC, PML and SAMS) for near coastal applications. Its triangular mesh is ideally suited to representing complex coastlines, but may lead to problems with the representation of waves for global applications (a dispersion relation analysis shows that such grids may give rise to spurious inertia-gravity wave modes, which are easily excited at small Rossby deformation radius, Danilov, 2010). In addition, the sigma-coordinate system is not optimal for deep open ocean vertical mixing and FVCOM may be overly-diffusive (owing to low-order tracer advection schema), which may affect its ability to represent global water mass distributions over decadal time-scales. In regimes dominated by eddies the model may suffer from having too many velocity degrees of freedom (see below).

FESOM. This is an FE model development led by Sergey Danilov (AWI, Bremerhaven), which couples an EVP sea-ice model to FEOM (the Finite Element Ocean circulation Model). FEOM supports the KPP and MY mixing schemes, and any combination of z- and sigma- coordinates in the vertical. The horizontal surface mesh is triangular, and employs the P1-P1 discretisation (so that model quantities are linear interpolations of nodal values). While most of the applications of FESOM are for regional studies, the model has been run globally and coupled to the ECHAM6 atmosphere. It is fully MPI parallelized.

FV-AWI. This development (again led by Sergey Danilov at AWI) will result in an FV equivalent to the FESOM system, and will provide a similar functionality to that in MPAS. At present, an FV model has been constructed, using a triangular horizontal mesh (with the same variable placement as FVCOM) and prisms in the vertical. It is coupled to a sea-ice model and is running in a global configuration. Being an FV code it is ~2-3 times faster than a similar configuration of FEOM. Again, since it uses a quasi-B triangular grid, the comments made above for FVCOM on its too large velocity space also apply here. However, the noise in the velocity field can be controlled with a special algorithm to compute momentum advection (Danilov, 2012).

Fluidity-ICOM. This is an FE development led by Chris Pain at Imperial College. It provides a non-hydrostatic code and can be configured to be unstructured in all three dimensions, with a grid which can adapt to optimally resolve the flow structures via dynamic mesh adaptivity, thereby placing enhanced grid resolution in regions where fine-scale features develop. Again, a triangular grid is used in the horizontal and has been tested with z and sigma layers as well as fully unstructured meshing in the vertical. The model uses higher order interpolation schemes than FESOM, so it may be relatively more accurate at the cost of computational expense. Significant effort has gone into the parallel performance of the model with hybrid OpenMP+MPI approaches recently allowing for scaling on over 10k cores. Whilst originally developed for engineering computational fluid dynamics (CFD) applications, it has now been successfully applied to a number of coastal and regional ocean applications (e.g. barotropic tides and tsunamis), as well as high-resolution buoyancy driven process studies (overflows and convection), and with a niche area being its ability to couple CFD with marine systems for the study of fluid structure interactions in coastal engineering, for example marine renewables and flood defence. However, it has so far not demonstrated a convincing global baroclinic capability, in spite of significant NERC support for such a development. One of the main issues appears to be the accurate representation of the higher-order elements on the sphere, and the need to parallelise these elements. One possibility would be to simplify ICOM by employing prisms in the vertical (as used by other unstructured models) rather than the current tetrahedral elements. This could inform the development of a new ocean core within the GungHo framework (see below), and ICOM’s alignment with GungHo could potentially facilitate relatively straightforward coupling to estuarine and sub-km scale modelling.

7. Developing an ocean model within the GungHo framework

7.1 Introduction

GungHo is a joint NERC and Met Office funded project to develop a new atmospheric dynamical core suitable for operational NWP on massively parallel computers. It is a collaboration between the Met Office, Bath, Exeter, Leeds, Manchester, Reading and Warwick Universities, Imperial College and STFC. Total effort is ~50 years FTE. Phase I (Feb 2011 - Jan 2013) addressed basic science questions including quasi-uniform grids, transport schemes, time schemes (implicit, explicit, etc), test cases and computational science. Phase II (Feb 2013 - Jan 2016) is aimed at developing and implementing a three-dimensional scheme, working on vertical aspects, code design and development and testing.

Conclusions from Phase I are focusing on 3 Finite Element approaches on triangular, hexagonal and quadrilateral grids with various arrangements. Conclusion on timestepping are still awaited, but are expected to be implicit in vertical and some mixture of explicit/implicit in the horizontal.

One of the main objectives of the Gung-Ho project is to tackle the poor scalability issue caused by the “meridian clustering around poles” in the traditional longitude-latitude grid atmosphere models. This grid “clustering” inhibits the large scale model parallelisation and becomes the bottleneck in large scale scalability. This specific issue can be readily circumvented in ocean models by using a tripolar grid (Murray, 1996).

A Key aspect of the development in GungHo is the computational science area. This proposes a clear division between the science layer and the computer science layer, in this way the basic numerical functions can be developed in such as way as to scale efficiently without over complicating the scientific code. This may proceed by either manual or automatic code generation.

The GungHo project represents an immense opportunity for ocean modelling in the UK, and given the on-going investment in this, it would be perverse to consider a new ‘strategic’ ocean model development exercise that did not align strongly with GungHo. This is particularly the case given the challenging objective of achieving efficient usage of future computer architectures. However, the science requirements of ocean models differ significantly from the atmospheric case, as does the nature of the modelling community in the UK. These aspects are considered below in more detail, but the principle is that the ocean model needs flexibility within its science layer to develop in the direction that best suites the needs of all aspects of marine modelling, and that this needs to be achievable within a comparatively small community.

7.2 Properties of an ocean model

While the governing equations are largely very similar, the differing nature of the oceans and the atmosphere implies that significantly different modelling approaches may be required for the two cases. The presence of the lateral boundaries of the ocean (they only cover ~71% of the surface with ~356k km of coastline) and O(1) variations of depth are particular examples. The former gives rise to distinct circulation features (such as western boundary intensification) and also alleviates the North Pole issue, and this strong driver towards alternative grid approaches to the more traditional latitude-longitude grids. The large variations in water depth means that dominant scales (e.g. Rossby Radii) vary by several orders of magnitude in the global ocean, especially between shelf seas and open oceans (the multi-scale issue). Moreover, pinch points and regions of restricted exchange (e.g. sills, overflows, straits) are geographic features not so relevant to the global atmosphere, but key to ocean circulation (e.g. inter-basin deep water mass transport). Another issue peculiar to the geometry of the oceans is the coastline fitting problem. Traditional quadrilateral grid represents the natural coastline as a “stair-case”. This limits the geographic accuracy that can be achieved and often leads to dynamical problems such as the retarding of coastal trapped waves (Adcroft and Marshall, 1998; Greenberg et al., 2007; Madec et al., 1991). It should be noted that, in contrast to the North Pole Problem, these issues are obstacles to accuracy rather than necessarily effecting scalability.

Models of the coastal-ocean and the open-ocean have had somewhat different evolutions, owing to different physical characteristics of these regions. They, are however, tightly coupled components of the Earth system; the movitation for improving the interaction between open and coastal-ocean are summarised above, along with some structured grid based options. However, only an unstructured grid approach offers truly flexible multi-scale modelling that can seamlessly simulate the oceans, shelf seas and potentially coastal regions as well. Aligning with the developments in GungHo potentially provides an important multi-scale and unstructured grid option with the potential to meet many of these challenges.

It is helpful to try and identify desirable properties of an ideal ocean model from previous experience, and a list of these is:

1. Good discrete dissipation properties: i.e. minimizes numerical diffusion
    a. Permits features commensurate to resolution
    b. Deep water mass properties are maintained
    c. Can accommodate theoretically/empirically sound subgrid scale models
2. Conservation of mass, tracer, energy, momentum, PV, and perhaps enstrophy
3. Good discrete dispersion properties
    a. Numerical modes are controllable without breaking other requirements (particularly 1)
4. Computationally efficient
    a. For large, high resolution models and smaller models needed for ensemble and long simulations
    b. Increase in computational cost for similar accuracy is much less than increase in computer power over the development time

5. Fits coastline at least as well as quadrilaterals, ideally as well as triangles
6. Flexible vertical coordinates and methods (including ALE)
7. Multi-scale horizontal mesh capability and so geometric flexibility
8. Accurate in realistic and idealized test cases
9. Portable, adaptable and easy to implement
    a. Can be effectively used by a small community and small groups (e.g. 1PI, 1 post doc, 1 phd student)
    b. Has a well supported infrastructure
    c. Can accommodate future developments

Against these properties the range of possible solution approaches (FE, FV, FD) and grid arrangement need to be tabulated to provide a 'Properties-Approaches' matrix. It is expected that no combination of solution approach and grid arrangement will meet all of these criteria, and moreover it is not practical to objectively assess all of these in realistic tests (indeed in several cases an approach for testing is not well established). Hence we must fall back on expert judgment, past lessons and compromise. The conclusion will also be dependent on the user demand; we are unlikely to be able to construct a model framework that is the best choice for all conceivable applications, but it should span a reasonable spectrum of applications, e.g. of these properties multi-scale horizontal mesh capability and good coastline fitting are crucial for meeting the science driver of a seamless shelf to ocean model (DFC1), and we would hope to be able to accommodate these in a single model configuration.

7.3 Approaches for an ocean model within the GungHo framework

Unstructured grids are an important option that permit coastal boundary fitting (e.g. solving the stair-case problem with triangular elements) and efficient local grid refinement for resolving specific areas of interest, lending much more flexibility than a curvilinear approach. They also facilitate multi-scale ocean modelling, and so allow significantly finer resolution than the current generation of fine resolution structured grid models. Hence, they should be seen as a truly next generation approach that demands state of the art numerical and computational science. However, it is very hard to construct a numerical scheme, using unstructured meshes that has comparable properties to the quadrilateral grids and is free of numerical modes (Staniforth and Thuburn, 2012). Moreover, issues of the complexity of numerical methods and indirect addressing on unstructured grids mean these approaches are inherently less efficient than structured grids. Testing and quantifying this in real world like-for-like cases is problematic, but anecdotal experience suggests unstructured grid models are between 3-10 times less efficient than comparable structured grid models. Hence, it is very difficult to beat the traditional quadrilateral grids in the open ocean area. This is also the case for the global coastal ocean modelling (discussed above), until a fine resolution global model has been achieved.

Three FE grid schemes have been recommended in GungHo phase I, which have good numerical properties based on a theoretical analysis of the shallow water equations (Cotter and Shipton, 2012). As pointed out in GungHo grid work package phase I report, these recommended FE grids/methods still need further analysis and case testing. Here we anticipate, maybe optimistically, that the GungHo framework will be flexible enough to accommodate FE, FV or even FD approaches, but accept the closer the ocean model is to the atmospheric (FE) approach then the less effort it would take to develop.

There are concerns about the FE approach in the global ocean modelling context both in terms of its relatively high computational cost compared with FV and FD methods and also the limit success in its application as a global ocean model over past years. Despite much effort internationally on FE modelling (our interim report lists 7 efforts), only the FEOM model has reached the published literature as a global model for realistic applications, and we note this model is transitioning to a FV approach. We should, therefore, consider very carefully before adopting a framework that could not readily accommodate an FV approach. It may be the case that the significant effort directed at the GungHo project and the considerable experience within (e.g.) Imperial College, means that this is the model that 'breaks the mould' for FE global ocean modelling. However, FE within GungHo should be seen as a relatively high risk option, given the historical track record and the unfamiliarity with this approach in the ocean modelling community in the UK. Hence a thorough investigation of the various approaches available and how they relate to the properties listed above is needed.

The FV approach is a lower risk option as far as global ocean modelling is concerned. The choice of element shape and grid arrangement is still problematic, and all options have some drawbacks. A mixed quadrilateral, triangular element mesh (e.g. Figure 8) may provide a very good option to meet the needs of seamless coast to ocean modelling, which could be readily achieved with a general grid FV model. It could potentially benefit from the multi-scale and coastline fitting capability of triangles, but retain the good numerical properties of quadrilaterals over most of the ocean. Stabilization of numerical modes would only be needed in the regions gridded by triangles, but how to build a grid such as this with an FE approach is not clear. Other element shapes such as hexagons may have good numerical properties (Ringler et al., 2010; although this approach has been drawn into question in the GungHo project), but lack the multi-scale capability and coastline fitting of a triangular mesh. C-grid structured quadrilaterals have been identified as an excellent global ocean option. Whether a NEMO-like model could be accommodated within GungHo could be further explored, noting that this would meet the computational objectives, but without bringing the science benefits of the unstructured approach.

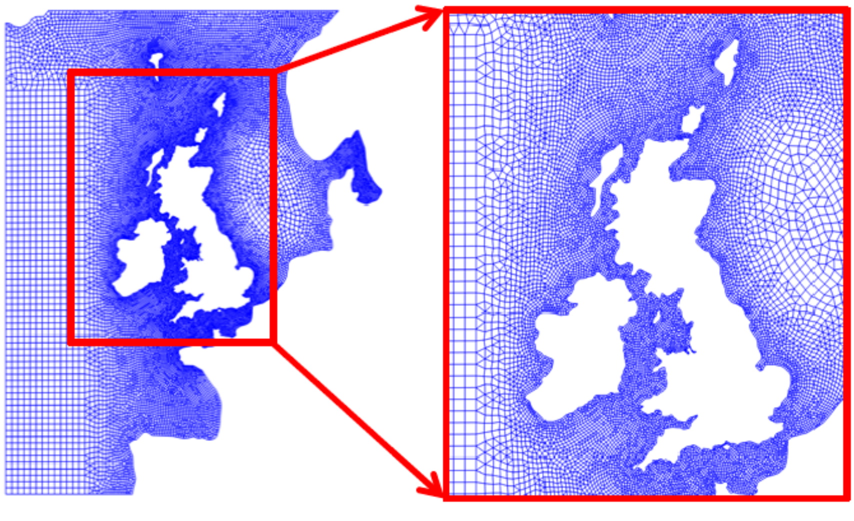


Figure 8. Grids for the North-west European shelf based on mixed triangular and quadrilateral elements.

The choice of grid arrangement remains problematic for FV models (just as with FE models). For the variable arrangement, corresponding to the Arakawa A, B, C grids for the traditional quadrilateral grid, there are four options in the framework of FV methods (Figure 9).

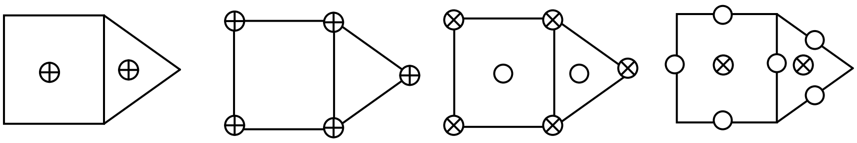


Figure 9 Variable arrangement in a mixed triangle quadrilateral grid. From left to the right, we name them A, A1, B, C grid. ⊕ *Represents all variables,*⊗ *represents pressure/sea surface elevation ,*○ *represents full velocity in B-Grid and velocity components normal to the edge of polygons in C-Grid.*

MPAS-ocean uses a hexagon C-grid, UnTRIM and SUNTANs use a triangular C-grid. Because these models use FD methods to calculate the pressure gradient term their grids must be orthogonal, which is a strong restriction on their flexibility. FVCOM and FV-AWI use a triangular B-grid and so do not have this limitation. The commercial, (DHI) Mike series of models (unpublished) uses an A grid and one of the approaches described by Danilov (2012) uses an A1 grid. Generally, the Arakawa C-grid has been considered the ideal candidate for the quadrilateral grid because of its good discrete dispersion relations and conservation properties. However, these properties cannot be retained in the unstructured cases (Danilov, 2010). This appears to arise because the ratio of degrees of freedom (DOF) of velocity variables to pressure variables is no longer 2:1 (in 2-D). This ratio is important since in this case the discrete shallow water equations (SWE) have two branches of inertia-gravity waves modes and one set of geostrophic wave modes, and so match the original continuous SWEs. It can be proved that all staggered grids (B or C) with non-quadrilateral cells cannot retain this DOF ratio (e.g. the triangular C grid has the ratio of 3:2 and the isotropic triangular B grid has the ratio of 4:1), and so permit spurious numerical modes. Collocated grids (A- and A1- grid) have the required DOF ratio. These, however, produce the so-called ‘check-board’ mode, because its pressure gradient operator has some null spaces. Hence all options have issues with numerical modes and the question remains as to whether these can be constrained without adversely affecting the solution (e.g. by making the model excessively dissipative).

At this stage there is not sufficient evidence to definitively choose a particular grid arrangement and solution approach, and this is also supported by the diversity of approaches currently in use, and the lack of consensus this reflects. To some extent this may arise from a mismatch between theoretical expectations of what makes a good approach and practical experience. For example, the triangular B-grid of FVCOM has a range of theoretical problems (as noted above), yet is probably the most widely used baroclinic unstructured grid model in the coastal modelling community. Similarly the GungHo project identified the TRiSK approach as having an inconsistent reconstruction of the Coriolis term (with non-orthogonal grids), yet MPAS (which uses TRiSK) is set to become a world leading ocean model, through its adoption by NCAR. This is not to say that theoretical considerations should not be taken into account in model design, but they must be complemented by practical 'real world' tests at the earliest opportunity. Hence a new phase of detailed assessment and prototyping is needed, going beyond the desk study that we have conducted here to consider theoretical analysis, and both idealized and realistic test cases for an ocean model in this framework. We must also accept that the GungHo framework may not be as flexible as we would want and all options may not be open to us.

7.4 How an ocean model might fit into the GungHo framework.

While it is proposed to align an ocean modelling effort with GungHo, our suggestion is to develop an ocean model with a distinct identity (e.g. gungho-ocean or g-ocean), which utilises aspects from the GungHo development, but is not 'the same model'. An alternative approach would be to have the ocean model as a set of 'options' within the GungHo science layer. This could make code management very difficult and potentially limit the flexibility of development of the ocean model. That said cross fertilisation of ideas and movement of code between the ocean and atmosphere science layers would be vital and so it would be important to establish an IP model and a set of Working Groups that facilitates this. Hence we propose the following structure:

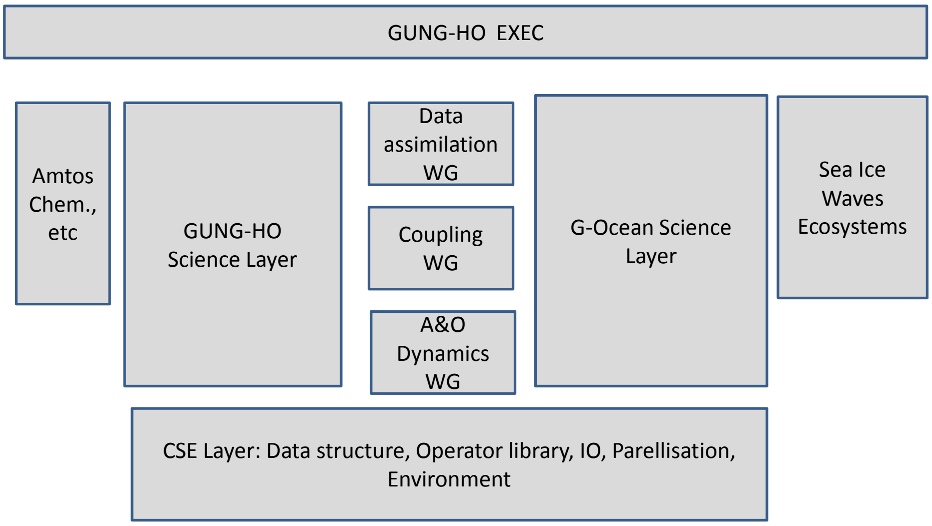


Figure 10. How an ocean model might fit into the GungHo framework.

The G-Ocean science layer would have an independent 'development committee' drawn from HEIs, NERC centres, the Met Office and (ideally) international partners which would work within the coding rules and data structures defined by the CSE layer. Opportunities to develop this as a European collaboration should be explored. Previous experience suggests that co-ownership within the community is an important aspect. The ocean community would also have representation on the GungHo exec and within the CSE layer, but it may be fair to assume that the needs of the ocean model would not be driving the priorities within these.

We propose a three stage approach in engaging with the GungHo project:

Stage 1: Planning, prototyping and consortium building, 2013-2014

Activity A: (Planning) A G-Ocean steering committee will be convened in 2013 with the objective of exploring the Properties-Approaches matrix. This will be a group of O(10) PIs and experts largely drawn from outside the GungHo community, but with some cross-over representation .

Activity: B: (Prototyping) a group of modelling practitioners will be gathered to explore test cases and help to inform (A). Hopefully this will be able to use the same tools as GungHo and will be able to give modellers early sight/experience of how the code is developing.

Activity C: (International consortium building). We will open discussions with the partners in the NEMO consortium as to the viability of developing G-ocean as the follow on to OPA as the core code in NEMO, i.e to become NEMO2. Mechanisms include: Presenting Ocean Road Map documents to the NEMO steering committee as the UK's vision for the evolution of the NEMO consortium; lobbying to get appropriate call texts into Horizons 2020; informal discussion with key players.

Gateway 1: Will this approach perform better than NEMO (OPA) given its current trajectory? Is funding in place for Stage 2? 2014

Stage 2: Construction of the basic dynamic model.

A consortium (UK and EU if possible) would be assembled to construct a global and regional 3D model. 2014-2017

Gateway 2: Commitment for this to become the next strategic UK/ European ocean model; formal engagement with the NEMO apparatus begins here. 2017

Stage 3: Development of a fully-fledged ocean model. Expand the consortium to include sea ice, waves, biogeochemistry, ecosystems and data assimilation. 2017-2021

References

Adcroft, A., 2013. Representation of topography by porous barriers and objective interpolation of topographic data. Ocean Modelling 67, 13-27.

Adcroft, A., Marshall, D., 1998. How slippery are piecewise-constant coastlines in numerical ocean models? Tellus 50A, 95-108.

Allen, J.I., Aiken, J., Anderson, T.R., Buitenhuis, E., Cornell, S., Geider, R., Haines, K., Hirata, T., Holt, J., Le Quéré, C., Hardman-Mountford, N., Ross, O.N., Sinha, B., While, J., 2010. Marine ecosystem models for earth systems applications: The MarQUEST experience. Journal of Marine Systems 81, 19-33.

Ashworth, M., Holt, J.T., Proctor, R., 2004. Optimization of the POLCOMS Hydrodynamic Code for Terascale High-Performance Computers, Proceedings of the 18th International Parallel & Distributed Processing Symposium, 26th-30th April 2004, Santa Fe, New Mexico.

Barange, M., Allen, I., Allison, E., Badject, M.-C., Blanchard, J., Drakeford, B., Dulvy, N.K., Harle, J., Holmes, R., Holt, J., Jennings, S., Lowe, J., Merino, G., Mullon, C., Pilling, G., Rodwell, L., Tompkins, E., Werner, F., 2011. predicting the

Impacts and socio-economic consequences of climate change on global marine ecosystems and fisheries: The QUEST_Fish Framework, in: Ommer, R.E., Perry, R.I., Cochrane, K., Cury, P. (Eds.), World Fisheries: a social-ecological analysis. Wiley-Blackwell, p. 417.

Bergman, K., Borkar, S., Campbell, D., Carlson, W., Dally, W., Denneau, M., Franzon, P., Harrod, W., Hiller, J., Karp, S., Keckler, S., Klein, D., Lucas, R., Richards, M., Scarpelli, A., Scott, S., Snavely, A., Sterling, T., Williams, S., Yelick, K., Kogge, P., 2008. ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems Peter Kogge, Editor & Study Lead.

Colella, P., Woodward, P.R., 1984. The Piecewise Parabolic Method (PPM) for Gas-Dynamical Simulations. Journal of Computational Physics 54, 174-201.

Cotter, C.J., Shipton, J., 2012. Mixed finite elements for numerical weather prediction. Journal of Computational Physics 231, 7076-7091.

Danilov, S., 2010. On utility of triangular C-grid type discretization for numerical modeling of large-scale ocean flows. Ocean Dynamics 60 (6), 1361-1369.

Danilov, S., 2012. Two finite-volume unstructured mesh models for large-scale ocean modeling. Ocean Modelling 47, 14-25.

Greenberg, D.A., Dupont, F., Lyard, F.H., Lynch, D.R., Werner, F.E., 2007. Resolution issues in numerical models of oceanic and coastal circulation. Continental Shelf Research 27, 1317-1343.

Griffies, S.M., Adcroft, A., Banks, H., al, e., 2009. Problems and prospects in large-scale ocean circulation models. Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, doi: 10.5270/OceanObs5209.cwp.5238

Gröger, M., Maier-Reimer, E., Mikolajewicz, U., Moll, A., Sein, D., 2012. NW European shelf under climate warming: implications for open ocean – shelf exchange, primary production, and carbon absorption. Biogeosciences Discuss. 9, 16625-16662, doi:16610.15194/bgd-16629-16625-12012.

Hecht, M.W., Smith, R.D., 2008. Toward a physical understanding of the North Atlantic: A review of model studies in an eddying regime, in: M. W. Hecht, M.W., Hasumi, H. (Eds.), Ocean Modeling in an Eddying Regime, Geophys. Monogr. Ser. AGU.

Hofmeister, R., Burchard, H., Beckers, J.M., 2010. Non-uniform adaptive vertical grids for 3D numerical ocean models. Ocean Modelling 33, 70-86.

Holt, J., Harle, J., Proctor, R., Michel, S., Ashworth, M., Batstone, C., Allen, J.I., Holmes, R., Smyth, T., Haines, K., Bretherton, D., Smith, G., 2009. Modelling the global coastal-ocean Phi. Trans Roy. Soc. Lon. A doi:10.1098/rsta.2008.0210.

Holt, J.T., Proctor, R., 2003. The role of advection in determining the temperature structure of the Irish Sea. Journal of Physical Oceanography 33, 2288-2306.

Holt, J.T., Proctor, R., 2008. The seasonal circulation and volume transport on the northwest European continental shelf: a fine-resolution model study. Journal of Geophysical Research 113, C06021, doi:06010.01029/02006JC004034.

James, I.D., 2000. A high-performance explicit vertical advection scheme for ocean models: how PPM can beat the CFL condition. Applied Mathematical Modelling 24, 1-9.

Koch-Larrouy, A., Madec, G., Iudicone, D., Atmadipoera, A., Molcard, R., 2008. Physical processes contributing to the water mass transformation of the Indonesian Throughflow. Ocean Dynamics 58, 275-288.

Leclair, M., Madec, G., 2011. (z)over-tilde-Coordinate, an Arbitrary Lagrangian-Eulerian coordinate separating high and low frequency motions. Ocean Modelling 37, 139-152.

Madec, G., Chartier, M., Delecluse, P., Crepon, M., 1991. A three-dimensional numerical study of deep-water formation in the northwestern Mediterranean Sea. . Journal of Physical Oceanography 21, 1349-1371.

Maraldi, C., Chanut, J., Levier, B., Reffray, G., Ayoub, N., De Mey, P., Lyard, F., Cailleau, S., Drévillon, M., A. Fanjul, E.A., Sotillo, M.G., Marsaleix, P., Team, t.M., 2012. NEMO on the shelf: assessment of the Iberia–Biscay–Ireland configuration. Ocean Sci. Discuss. 9, 499-583.

Merino, G., Barange, M., Blanchard, J.L., Harle, J., Holmes, R., Allen, I., Allison, E.H., Badjeck, M.C., Dulvy, N.K., Holt, J., Jennings, S., Mullon, C., Rodwell, L.D., 2012. Can marine fisheries and aquaculture meet fish demand from a growing human population in a changing climate? Global Environmental Change 22, 795-806.

Murray, R.J., 1996. Explicit generation of orthogonal grids for ocean models. Journal of Computational Physics 126, 251-273.

O'Dea, E.J., Arnold, A., Edwards, K.P., Furner, R., Hyder, P., Martin, M.J., Siddorn, J.R., Storkey, D., While, J., Holt, J.T., Liu, H., 2012. An operational ocean forecast system incorporating NEMO and SST data assimilation for the tidally driven European North-West shelf. Journal of Operational Oceanography 5(1), 3-17.

Pickles, S.M., 2010. Multi-Core Aware Performance Optimization of Halo Exchanges in Ocean Simulations, Cray User Group Proceedings.

Pickles, S.M., Porter., A.R., 2012. Developing NEMO for Large Multi-core Scalar Systems: Final Report of the dCSE NEMO projec, STFC Technical Report, , pp. http://epubs.cclrc.ac.uk/bitstream/8209/DLTR-2012-8202.pdf.

Porter, A.R., Pickles, S.M., Ashworth, M., 2012. Final report for the gNEMO project: Porting the oceanographic model NEMO to run on many-core devices, STFC Technical Report, pp. http://epubs.cclrc.ac.uk/bitstream/7538/DLTR-20120-20101.pdf.

Postlethwaite, C.F., Maqueda, M.A.M., Fouest, V.l., Tattersall, G.R., Holt, J., Willmott, A.J., 2011. The effect of tides on dense water formation in Arctic shelf seas. Ocean Sci., 7, 203–217.

Prather, M.J., 1986. Numerical advection by conservation of 2nd-order moments. J. Geophys. Res.-Atmos. 91, 6671-6681.

Ringler, T., Petersen, M., Higdon, R.L., Jacobsen, D., Jones, P.W., Maltrud, M., 2013. A multi-resolution approach to global ocean modeling. Ocean Modelling, http://dx.doi.org/10.1016/j.ocemod.2013.1004.1010.

Ringler, T.D., Thuburn, J., Klemp, J.B., Skamarock, W.C., 2010. A unified approach to energy conservation and potential vorticity dynamics for arbitrarily-structured C-grids. Journal of Computational Physics 229, 3065-3090.

Schrum, C., Hubner, U., Jacob, D., Podzun, R., 2003. A coupled atmosphere/ice/ocean model for the North Sea and the Baltic Sea. Climate Dynamics 21, 131-151.

Sinha, B., Buitenhuis, E.T., Quere, C.L., Anderson, T.R., 2010. Comparison of the emergent behavior of a complex ecosystem model in two ocean general circulation models. Progress in Oceanography 84, 204-224.

Staniforth, A., Thuburn, J., 2012. Horizontal grids for global weather and climate prediction models: a review. Quarterly Journal of the Royal Meteorological Society 138, 1-26.

Tseng, Y.H., Ferziger, J.H., 2003. A ghost-cell immersed boundary method for flow in complex geometry. J. Comput. Phys. 192, 593-623.